НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2025)

МАТЕРИАЛЫ
XXIV Международной конференции
имени А. Ф. Терпугова
20–25 октября 2025 г.

ТОМСК
Издательство Томского
государственного университета
2025

УДК 519
ББК 22.17
И74
Информационные технологии и математическое моделирование (ИТММ-2025) : материалы XXIV Международной конференции имени А. Ф. Терпугова (20–25 октября 2025 г.). — Томск : Издательство Томского государственного университета, 2025. — 444 с.

ISBN 978-5-908040-36-5

Сборник содержит избранные материалы XXIV Международной конференции имени А. Ф. Терпугова по следующим направлениям: теория массового обслуживания и ее приложения, прикладной вероятностный анализ, интеллектуальные системы, информационные технологии и программная инженерия, математическое и компьютерное моделирование.

Для специалистов в области информационных технологий и математического моделирования.

УДК 519
ББК 22.17

Редколлегия:
С. П. Моисеева, доктор физико-математических наук, профессор
Д. В. Семенова, кандидат физико-математических наук, доцент
Е. А. Фёдорова, кандидат физико-математических наук
А. В. Зорин, доктор физико-математических наук, доцент
О. Д. Лизюра

ISBN 978-5-908040-36-5

© Авторы (текст), 2025
© Томский государственный университет (оформление, дизайн), 2025

NATIONAL RESEARCH TOMSK STATE UNIVERSITY
SIBERIAN FEDERAL UNIVERSITY

INFORMATIONAL TECHNOLOGIES AND MATHEMATICAL MODELLING (ITMM-2025)

PROCEEDINGS
of the 24th International Conference
named after A. F. Terpugov
2025 October, 20–25

TOMSK
Tomsk State
University Publishing
2025

UDC 519
LBC 22.17
I60
Informational technologies and mathematical modelling (ITMM-2025): Proceedings of the 24th International Conference named after A. F. Terpugov (2025 October, 20–25). — Tomsk: Tomsk State University Publishing, 2025. — 444 p.

ISBN 978–5–908040–36–5

This volume presents selected papers from the 24th International Conference named after A. F. Terpugov. The papers are devoted to new results in the following areas: queuing theory and its applications, applied probabilistic analysis, intelligent systems, information technology and software engineering, mathematical and computer modeling.

UDC 519
LBC 22.17

E d i t o r s:
S. P. Moiseeva, Doctor of Physical and Mathematical Sciences, Professor
D. V. Semenova, Candidate of Physical and Mathematical Sciences, Associate Professor
E. A. Fedorova, Candidate of Physical and Mathematical Sciences
A. V. Zorine, Doctor of Physical and Mathematical Sciences, Associate Professor
O. D. Lizyura

ISBN 978-5-908040-36-5

© Authors (Text), 2025
© Tomsk State University (Publishing, Design), 2025

Прикладной вероятностный и статистический анализ

НОВЫЕ ЭКОНОМИЧНЫЕ АЛГОРИТМЫ С УРАВНИВАНИЕМ ВЕРОЯТНОСТЕЙ ДЛЯ МОДЕЛИРОВАНИЯ ГАУССОВСКОГО РАСПРЕДЕЛЕНИЯ

А. В. Войтишек[1,2] Ю. А. Иванова[2], У. П. Сейтмуратов[2]

[1] *Институт вычислительной математики и математической геофизики СО РАН, г. Новосибирск, Россия*

[2] *Новосибирский государственный университет, г. Новосибирск, Россия*

В работе построены новые алгоритмы с уравниванием вероятностей для компьютерного моделирования стандартного гауссовского распределения. Показано, что эти алгоритмы являются значительно экономичнее (двусторонний алгоритм метода исключения с кусочно-постоянными мажорантой и минорантой – в 1,26 раза, а модифицированный зиккурат-метод – в 2,11 раз) широко используемых в расчетах по методу Монте-Карло формул Бокса–Мюллера. **Ключевые слова:** *компьютерное моделирование случайных величин, гауссовское распределение, формулы Бокса – Мюллера, алгоритмы с уравниванием вероятностей, двусторонний алгоритм метода исключения с кусочно-постоянными мажорантой и минорантой, модифицированный зиккурат-метод*

1. Введение: применения и моделирования гауссовского распределения

Функция плотности гауссовского распределения задается формулой:

$$f_{\xi}^{m,\sigma^2}(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-m)^2}{2\sigma^2}}, \quad -\infty < u < +\infty, \tag{1}$$

с параметрами $m = \mathbf{E}\xi^{(m,\sigma^2)}$, $\sigma^2 = \mathbf{D}\xi^{(m,\sigma^2)}$.

Для получения значений $\xi_i^{(m,\sigma^2)}$, распределенных согласно плотности (1), достаточно моделировать значения $\xi_i^{(0,1)}$ стандартной гауссовской случайной величины $\xi^{(0,1)}$, имеющей плотность распределения

$$f_{\xi}^{(0,1)}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \quad -\infty < u < +\infty, \tag{2}$$

Исследования выполнены в рамках государственного задания ИВМиМГ СО РАН FWNM-2025-0002.

а затем использовать преобразование $\xi_i^{(m,\sigma^2)} = m + \sigma\xi_i^{(0,1)}$.

Случайные величины $\xi^{(m,\sigma^2)}$ применяются при решении современных задач, связанных с моделированием и анализом атмосферных процессов и климата, океанологических и гидрологических процессов и т. д. (см., например, [1]).

Справедливо следующее утверждение.

УТВЕРЖДЕНИЕ (раздел 1.10.2 книги [2]). *Если* $\boldsymbol{\xi}$ *– d-мерный изотропный случайный вектор, квадрат длины которого имеет* χ^2*-распределение с d степенями свободы, то его компоненты* $\xi^{(1)}, \ldots, \xi^{(d)}$ *независимы и нормальны с параметрами* $(0, 1)$ *(т. е. распределены согласно плотности* (2)*).*

В разделе 13.1 книги [3] показано, что для экономичного компьютерного моделирования стандартного гауссовского распределения (2) следует использовать сформулированное утверждение для случая $d = 2$, который дает широко известные формулы Бокса – Мюллера вида:

$$\xi_1 = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2, \qquad \xi_2 = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2 \tag{3}$$

где $\alpha_1, \alpha_2 \in U(0,1)$ – стандартные случайные числа, т. е. независимые выборочные значения случайной величины $\alpha \in U(0,1)$, равномерно распределенной в интервале $(0,1)$.

В работе [4] было установлено, что компьютерные вычисления функций $\ln x$, $\sin x$ и $\cos x$, используемых в формулах Бокса-Мюллера (3), требуют значительных вычислительных ресурсов. Этот вывод основан на вычислительных экспериментах, проведенных на различных компьютерах с использованием системы NMPUD (Numerical Modelling of Probabilistic Univariate Distributions). В ходе исследований было показано, что затраты на вычисление этих функций сопоставимы с затратами на одно обращение к генератору стандартных случайных чисел, что является существенным фактором при оценке эффективности метода.

Альтернативный подход представлен в наших работах [5, 6], где демонстрируется, что для широкого класса распределений с монотонными плотностями наиболее экономичным оказывается универсальный двусторонний метод исключения. Этот метод использует кусочно-постоянные мажоранту и миноранту с уравниванием вероятностей. Аналогичные преимущества имеет существенно модифицированный (по сравнению с работой [7]) зиккурат-метод (см. далее раздел 3).

Однако применение упомянутых алгоритмов с уравниванием вероятностей к нормальному распределению (2) имеет ряд особенностей по сравнению с работами [5, 6], одной из которых является наличие «хво-

стов» распределения, в то время как в работах [5, 6] рассматривались только случайные величины на конечных интервалах распределения.

2. Двусторонний алгоритм метода исключения

В соответствии с подходом, представленным в работах [5, 6], следует построить мажоранту и миноранту для функции $f(u) = e^{-\frac{u^2}{2}}$, при этом наиболее универсальным является применение кусочно-постоянных приближений, которые обеспечивают уравнивание вероятностей на неравномерной сетке $\{-x_M < ... - x_2 < x_1 = 0 < x_2 < ... < x_M\}$ [6]. Особенностью рассматриваемого случая является то, что плотность (2) определена на всем интервале $u \in (-\infty; +\infty)$. Поэтому для «хвостов» распределения в качестве мажоранты предлагается использовать функцию $f^{(1)}(u) = e^{-\frac{x_M u}{2}}$.

АЛГОРИТМ 1. *1. Для выбора участка монотонности моделируется случайная величина* $\alpha_1 \in U(0,1)$. *Если* $\alpha_1 \le \frac{1}{2}$, *выбирается интервал* $(-\infty, 0]$, *в противном случае – интервал* $[0, +\infty)$.

2. С целью избежать повторного обращения к генератору случайных величин для выбора номера отрезка вычисляется следующее выражение:

$$\beta(\alpha_1) = \begin{cases} 2\alpha_1, & \text{если } 0 < \alpha_1 \le 1/2; \\ 2\alpha_1 - 1, & \text{если } 1/2 < \alpha_1 < 1. \end{cases} \tag{4}$$

3. Далее вычисляется значение $m_0 = [M\beta(\alpha_1)] + 1$ *([A] здесь обозначает целую часть числа A); таким образом,* m_0 *будет обозначать номер выбранного отрезка* $[x_{m_0}, x_{m_0+1}]$.

СЛУЧАЙ 1 *(если* $m_0 < M$*):*

4. Моделируется значение $\xi_0^{(1)}$ *по формуле* $\xi_0^{(1)} = x_{m_0} + (x_{m_0+1} - x_{m_0})\gamma(\alpha_1)$; *здесь* $\gamma(\alpha_1) = M\beta(\alpha_1) - m_0 + 1$.

5. Моделируется величина $\eta_0 = \alpha_2 f(x_{m_0}); \alpha_2 \in U(0,1)$.

6. Проводится проверка условия $\eta_0 \le f(x_{m_0+1})$. *В случае выполнения данного условия принимается значение* $\xi_0 = \xi_0^{(1)}$, *если в пункте* 1 *выбран интервал* $[0, +\infty)$, *или* $\xi_0 = -\xi_0^{(1)}$, *если выбран интервал* $(-\infty, 0]$ *(при этом алгоритм завершает свою работу), в противном случае осуществляется проверка* $\eta_0 < f(\xi_0^{(1)})$. *Если последнее условие выполняется, алгоритм завершается, и принимается значение* $\xi_0 = \xi_0^{(1)}$ *(или* $\xi_0 = -\xi_0^{(1)}$*). В противном случае происходит возврат к пункту* 1.

СЛУЧАЙ 2 *(если* $m_0 = M$*):*

4. Моделируется значение $\xi_0^{(1)}$ *по формуле* $\xi_0^{(1)} = x_M - 2\gamma(\alpha_1)/x_M$ *(или* $\xi_0^{(1)} = -x_M + 2\gamma(\alpha_1)/x_M$*).*

5. Моделируется величина $\eta_0 = \alpha_2 e^{-\frac{x_M \xi_0^{(1)}}{2}}$; $\alpha_2 \in U(0,1)$.

6. Осуществляется проверка $\eta_0 < f(\xi_0^{(1)})$. *Если данное условие выполняется, алгоритм завершается, и принимается значение* $\xi_0 = \xi_0^{(1)}$ *(или* $\xi_0 = -\xi_0^{(1)}$*); иначе происходит возврат к пункту* 1.

3. Модифицированный зиккурат-метод

Здесь точки $\{-x_M < ... - x_2 < x_1 = 0 < x_2 < ... < x_M\}$ выбираются так, что равны площади прямоугольников с вершинами $(0, f(x_m)); (0, f(x_{m-1})); (x_m, f(x_{m-1})); (x_m, f(x_m))$ (справа от оси ординат) и площадь нижнего элемента.

АЛГОРИТМ 2. *1. Алгоритм начинается с выбора участка монотонности функции плотности. Для этого генерируется случайная величина* α_1, *равномерно распределенная на интервале* $(0,1)$. *В зависимости от ее значения определяется рабочий интервал: при* $\alpha_1 \le \frac{1}{2}$ *выбирается отрицательная полуось* $(-\infty, 0]$ *в противном случае – положительная* $[0, +\infty)$.

2. С помощью этой функции определяется номер отрезка разбиения $m_0 = [M\beta(\alpha_1)] + 1$, *где* $\beta(\alpha_1)$ *вычисляется по формуле* (4).

СЛУЧАЙ 1 *(если* $m_0 < M$*):*

3. На этом этапе генерируется значение $\tilde{\xi}_0$ *по формуле* $\tilde{\xi}_0 = x_{m_0}\gamma(\alpha_1)$.

4. Если полученное значение $\tilde{\xi}_0$ *не превышает* x_{m_0-1}, *оно принимается как результат работы алгоритма:* $\xi_0 = \tilde{\xi}_0$, *если в пункте* 1 *выбран интервал* $[0, +\infty)$, *или* $\xi_0 = -\tilde{\xi}_0$, *если выбран интервал* $(-\infty, 0]$. *В противном случае генерируется дополнительная случайная величина* $\eta_0 = f(x_{m_0}) + \alpha_2[f(x_{m_0-1}) - f(x_{m_0})]$; $\alpha_2 \in U(0,1)$, *и проверяется условие* $\eta_0 < f(\tilde{\xi}_0)$. *При выполнении этого условия значение* $\xi_0 = \tilde{\xi}_0$ *(или* $\xi_0 = -\tilde{\xi}_0$*) принимается как результат, иначе алгоритм перезапускается с первого шага.*

СЛУЧАЙ 2 *(если* $m_0 = M$*):*

3. Вычислим величину $\tilde{\xi}_0^{(1)} = \frac{x_2[f(x_1)-f(x_2)]}{f(x_M)}\gamma(\alpha_1)$.

4. Далее проверяется условие $\tilde{\xi}_0^{(1)} \le x_M$. *При его выполнении получаем* $\xi_0 = \tilde{\xi}_0^{(1)}$ *(или* $\xi_0 = -\tilde{\xi}_0^{(1)}$*). В противном случае генерируются величины* $\xi_0^{(2)} = x_M - 2\delta(\alpha_1)/x_M$ *и* $\eta_0 = \alpha_2 e^{-x_M \tilde{\xi}_0^{(2)}/2}$ *здесь* $\alpha_2 \in U(0,1); \delta(\alpha_1) = \frac{\gamma(\alpha_1)-P_M}{1-P_M}$; $P_M = \frac{x_M f(x_M)}{x_2[f(x_1)-f(x_2)]}$. *Далее проводится проверка условия* $\eta_0 < f(\tilde{\xi}_0^{(2)})$. *Если условие выполняется, то* $\xi_0 = \tilde{\xi}_0^{(2)}$ *(или* $\xi_0 = -\tilde{\xi}_0^{(2)}$*) становится итоговым результатом моделирования, в противном случае алгоритм возвращается к пункту* 1.

4. Результаты вычислений

Для сравнительного анализа эффективности трех методов генерации нормально распределенных случайных величин были проведены вычислительные эксперименты. Тестирование проводилось на выборках различного объема: от $N = 10^5$ до $N = 10^8$. Вычисления проводились на персональном компьютере MacBook Air M1, а код был написан на языке Python. Для всех трех методов наблюдается устойчивая линейная зависимость времени выполнения от объема выборки.

Было проведено исследование зависимости времени выполнения от количества прямоугольников разбиения M в диапазоне от 10 до 300.

Таблица 1

Время выполнения алгоритмов (в секундах) при различном количестве разбиений M (10^6 генерируемых значений).

M	формулы Б–М	Дв. мет. иск.	Зиккурат–метод
150	3,035	2,343	1,370
240	3,035	2,302	1,345
300	3,035	2,309	1,346

Экспериментальные данные показывают существование оптимального диапазона количества разбиений ($150 \leqslant M \leqslant 300$), в котором оба метода с уравниванием вероятностей демонстрируют стабильную производительность. Заметим, что увеличение количества разбиений свыше $M = 150$ не приводит к значительному улучшению производительности, что позволяет рекомендовать это значение как минимальное достаточное для практического применения. Зиккурат-метод демонстрирует существенное преимущество в производительности (около 70 %) по сравнению с двусторонним методом исключения во всем исследованном диапазоне значений M.

Заключение

В работе проведен сравнительный анализ методов генерации нормально распределенных случайных величин, который показал, что классические формулы Бокса–Мюллера (3), несмотря на свою распространенность, являются достаточно трудоемкими из-за затрат на вычисление функций $\ln x$, $\sin x$ и $\cos x$. Предлагаемый двусторонний метод исключения, основанный на использовании кусочно-постоянных мажорант и минорант с уравниванием вероятностей, продемонстрировал высокую эффективность и экономичность (затраты – по сравнению с формулами Бокса–Мюллера – снизились более чем на четверть) – несмотря на некоторые сложности его реализации, связанные с наличием «хво-

стов» у нормального распределения (2). Также была предложена новая модификация зиккурат–метода, которая оказалась более чем в два раза экономичнее классических формул Бокса–Мюллера (3).

СПИСОК ЛИТЕРАТУРЫ

1. *Пригарин С. М.* Модели случайных процессов и полей в методах Монте-Карло. Palmarium Academic Publishing, 2014. 160 с.
2. *Михайлов Г. А., Войтишек А. В.* Статистическое моделирование. Методы Монте-Карло. М.: Изд-во «Юрайт», 2025. 323 с.
3. *Войтишек А. В.* Лекции по численным методам Монте-Карло. Новосибирск: НГУ, 2018. 314 с.
4. *Войтишек А. В., Гаджиахмедов М. Г., Рыжов И. А., Трофимов И. А.* Сравнение вычислительных затрат на основные математические операции с помощью системы NMPUD // Информационные технологии и математическое моделирование (ИТММ-2021): Материалы XX Международной конференции имени А. Ф. Терпугова (1–5 декабря 2021 года). Томск: Издательство Томского государственного университета, 2022. С. 329–334.
5. *Брызгалов В. Л., Войтишек А. В.* Анализ трудоемкости формул метода обратной функции распределения для случайных величин с конечным интервалом распределения // Информационные технологии и математической моделирование (ИТММ-2023): материалы XXII Международной конференции имени А. Ф. Терпугова (4–9 декабря 2023 года). Томск: Издательство Томского государственного университета, 2023. Т. 1. С. 314–321.
6. *Брызгалов В. Л., Войтишек А. В.* Конструктивный алгоритм уравнивания вероятностей // Информационные технологии и математической моделирование (ИТММ-2024): материалы XXIII Международной конференции имени А. Ф. Терпугова (20–26 октября 2024 года). Томск: Издательство Томского государственного университета, 2024. С. 363–368.
7. *Marsaglia G., Tsang W. W.* The ziggurat method for generating random variables // Journal of Statistical Software. 2000. Vol. 5. № 8. P. 1–7.

Войтишек Антон Вацлавович — доктор физико-математических наук, профессор, главный научный сотрудник, лаборатория стохастических задач ИВМиМГ СО РАН; профессор, кафедра вычислительной математики НГУ. E-mail: *vav@osmf.sscc.ru*

Иванова Юлия Андреевна — студентка магистратуры механико-математического факультета НГУ. E-mail: *y.ivanova@g.nsu.ru*

Сейтмуратов Убайдулла Пахтамурат улы — студент магистратуры механико-математического факультета НГУ. E-mail: *u.seitmuratov@g.nsu.ru*

СПЕЦИАЛЬНЫЕ ПОКАЗАТЕЛИ КАЧЕСТВА ПРИБЛИЖЕНИЯ ПРИ ВЫБОРЕ ЯДЕРНОЙ ФУНКЦИИ В КОМПЬЮТЕРНОМ ФУНКЦИОНАЛЬНОМ АЛГОРИТМЕ ПРИБЛИЖЕНИЯ ВЕРОЯТНОСТНОЙ ПЛОТНОСТИ ПО ЗАДАННОЙ ВЫБОРКЕ

А. В. Войтишек[1,2], У. П. Сейтмуратов[2], Н. Х. Шлымбетов[2]

[1] *Институт вычислительной математики и математической геофизики СО РАН, г. Новосибирск, Россия*

[2] *Новосибирский государственный университет, г. Новосибирск, Россия*

В данной работе рассматриваются вопросы выбора ядерной функции при построении специальных компьютерных функциональных ядерных алгоритмов приближения вероятностной плотности по заданной выборке. **Ключевые слова:** *вычислительный функциональный ядерный алгоритм, выбор ядерной функции, среднеквадратическая погрешность*

1. Введение: экономичная версия вычислительного функционального ядерного алгоритма

В наших работах [1–4] отмечено, что при обработке больших данных (например, при применении технологий машинного обучения) актуальным оказывается использование экономичных алгоритмов компьютерного приближения неизвестной плотности $f_{\boldsymbol{\xi}}(\mathbf{x})$ по заданной выборке $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n\}$ с заданным уровнем погрешности на всей компактной области распределения $X \subset \mathbb{R}^d$. Подробное изучение возможных вычислительных схем (здесь главным образом сравнивались различные версии т. н. вычислительных функциональных проекционных и ядерных алгоритов) привело нас к выводу о том, что наилучшими свойствами (экономичность в сочетании с точностью) обрадает следующая частная версия вычислительного функционального ядерного алгоритма.

Не ограничивая общности, предположим, что конечные границы $a^{(k)}, b^{(k)}$, $k = 1, ...d$ области X можно ввести таким образом, чтобы в X можно было бы ввести равномерную сетку с шагом h вида

$$\mathbf{y}_i = \left(a^{(1)} + j_i^{(1)} h, ..., a^{(d)} + j_i^{(d)} h\right); \quad i = 1, ..., M; \tag{1}$$

Исследования выполнены в рамках государственного задания ИВМиМГ СО РАН FWNM-2025-0002.

$$j_i^{(m)} = 0, 1, ..., M_m;\ m = 1, ..., d;\ \ M = (M_1 + 1) \times ... \times (M_d + 1);$$

с шагом $h = (b^{(m)} - a^{(m)})/M_m;\ m = 1, ..., d.$

АЛГОРИТМ. *Выберем ядерную функцию* $\varkappa^{(\mathbf{x})}(\mathbf{z})$ *(см., например,* [5]*) такую, что* $f_{\boldsymbol{\xi}}(\mathbf{x}) \approx \int_X f_{\boldsymbol{\xi}}(\mathbf{z})\varkappa^{(\mathbf{x})}(\mathbf{z})\, d\mathbf{z} = \mathbf{E}\varkappa^{(\mathbf{x})}(\boldsymbol{\xi})$ *для всех* $\mathbf{x} \in X$. *Реализуем на компьютере соответствующие приближения* $f_{\boldsymbol{\xi}}(\mathbf{y}_i) \approx \tilde{f}_{\boldsymbol{\xi}}^{(\mathbf{y}_i)}(n) = \frac{1}{n}\sum_{j=1}^{n} \varkappa^{(\mathbf{y}_i)}\left(\boldsymbol{\xi}_j\right);\ i = 1, ..., M$ *(здесь* $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n\}$ *– заданная выборка или ее подмножество) и строим окончательную классическую компьютерную аппроксимацию плотности* $f_{\boldsymbol{\xi}}(\mathbf{x})$

$$f_{\boldsymbol{\xi}}(\mathbf{x}) \approx L^{(M,n)}\tilde{f}_{\boldsymbol{\xi}}(\mathbf{x}) = \sum_{i=1}^{M} \tilde{f}_{\boldsymbol{\xi}}^{(\mathbf{y}_i)}(n)\chi^{(i)}(\mathbf{x}), \tag{2}$$

здесь $\boldsymbol{\chi}^{(M)} = \left\{\chi^{(1)}(\mathbf{x}), ..., \chi^{(M)}(\mathbf{x})\right\}$ – базис т. н. «мультилинейного восполнения»:

$$\chi^{(i)}(\mathbf{x}) = B\left[\frac{x^{(1)} - a^{(1)}}{h} - j_i^{(1)}\right] \times ... \times B\left[\frac{x^{(d)} - a^{(d)}}{h} - j_i^{(d)}\right]; \tag{3}$$

$$B(u) = \begin{cases} u + 1 \text{ при } -1 \le u \le 0; \\ -u + 1 \text{ при } 0 \le u \le 1; \\ 0 \quad \text{при} \quad |u| > 1; \end{cases} \qquad \mathbf{x} = \left(x^{(1)}, ..., x^{(d)}\right).$$

2. О выборе ядерной функции $\varkappa^{(\mathbf{x})}(\mathbf{z})$

По аналогии с работой [5] будем выбирать ядерную функцию в виде

$$\varkappa^{(\mathbf{x})}(\mathbf{z}) = \prod_{s=1}^{d} \frac{1}{\hat{h}} K\left(\frac{x^{(s)} - z^{(s)}}{\hat{h}}\right); \mathbf{z} = \left(z^{(1)}, ...z^{(d)}\right), \tag{4}$$

где положительное число $\hat{h}$ определяют носитель («коэффициент размытости») ядерной функции $\varkappa^{(\mathbf{x})}(\mathbf{z})$, а ограниченная неотрицательная функция $K(y)$ является симметричной относительно нуля (четной) вероятностной плотностью: $K(y) = K(-y)$ и $\int_{-\infty}^{+\infty} K(y)\, dy = 1$.

Обратим внимание на то, что конструкция ядерной функции (4) аналогична мультипликативной конструкции

$$\chi^{(i)}(\mathbf{x}) = \chi\left[t_1^{(\mathbf{y}_i)}(x^{(1)})\right] \times ... \times \chi\left[t_d^{(\mathbf{y}_i)}(x^{(d)})\right];\ \mathbf{x} = \left(x^{(1)}, ..., x^{(d)}\right), \tag{5}$$

использованной в нашей работе [3] для построения аппроксимационных базисов; здесь $\chi(w)$ – одинаковая для всех компонент т. н. «образующая базис» функция одного переменного, а $t_j^{(\mathbf{y}_i)}(w)$; $j = 1, ..., d$ – единообразные (как правило, линейные) функции, связанные с соответствующими компонентами узла $\mathbf{y}_i$. В частности, базисные функции мультилинейного восполнения (3) является практически значимым примером конструкции (5).

Конструкции вида (3)–(5) позволяют перенести свойства одномерных ядерных и базисных функций, определяемых теми или иными образующими функциями $K(y)$ и/или $\chi(w)$, на многомерный случай, реализуя соответствующую индукцию по размерности d; поэтому далее – для простоты изложения – рассмотрим одномерный случай функции (4) вида

$$\varkappa^{(x)}(z) = \frac{1}{\tilde{h}} K\left(\frac{x-z}{\tilde{h}}\right).$$

Далее отметим, что оптимальная (с точки зрения минимальной трудоемкости) компьютерная реализация алгоритма 1 получается в случае, когда «коэффициент размытости» $\tilde{h}$ равен шагу сетки h (см. формулу (1)), и, кроме того, функция $K(y)$ должна быть положительна только на интервале $|y| < 1/2$:

$$\varkappa^{(x)}(z) = \frac{1}{h} K\left(\frac{x-z}{h}\right);\ K(y) > 0 \ \text{ при } \ |y| < 1/2.$$

В этом случае затраты сформулированного выше ядерного алгоритма не зависят от числа узлов сетки (1).

3. Использование $\mathbb{L}_2$–подхода

По аналогии с работой [5] рассмотрим следующий вариант представленного нами в работе [4] $\mathbb{L}_2$–подхода для «точечной» ядерной оценки $\tilde{f}_\xi^{(x)}(n) = \frac{1}{n}\sum_{j=1}^n \varkappa^{(x)}(\xi_j)$ (здесь $\{\xi_1, ..., \xi_n\}$ – заданная одномерная выборка) – для нее сеточное приближение (2) не рассматривается.

Применяя неравенство Коши–Буняковского и теорему Фубини, можно получить представление $\mathbb{L}_2$–погрешности в виде суммы компоненты смещения и стохастической компоненты [4]:

$$\left[\mathbf{E}\left\|f_\xi(x) - \tilde{f}_\xi^{(x)}(n)\right\|_{\mathbb{L}_2(\mathbb{R})}\right]^2 \le \int_{-\infty}^{+\infty} \mathbf{E}\left[f_\xi(x) - \tilde{f}_\xi^{(x)}(n)\right]^2 dx =$$

$$= \int_{-\infty}^{+\infty} \left(f_\xi(x) - \mathbf{E}\tilde{f}_\xi^{(x)}(n)\right)^2 dx + \int_{-\infty}^{+\infty} \mathbf{D}\tilde{f}_\xi^{(x)}(n)\, dx.$$

Наконец, используя замену $v = (x-u)/h$ и разлагая функцию $f_\xi(x-hv)$ в ряд Тейлора по переменной $s = hv$ при $h \downarrow 0$:

$$f_\xi(x-hv) = f_\xi(x) - hvf'_\xi(x) + \frac{1}{2}(hv)^2 f''_\xi(x) + o(h^2),$$

получаем

$$\int_{-\infty}^{+\infty} \mathbf{E}\left[f_\xi(x) - \tilde{f}_\xi^{(x)}(n)\right]^2 dx \approx C_1(h)\left[\int_{-1/2}^{+1/2} v^2 K(v)\,dv\right]^2 + \\ + C_2(n,h)\int_{-1/2}^{+1/2} K^2(v)\,dv, \qquad (6)$$

где $C_1(h) = \dfrac{h^4 \int_{-\infty}^{+\infty}\left[f''_\xi(x)\right]^2 dx}{4}$ и $C_2(n,h) = \dfrac{1}{nh}$.

4. Выбор показателя качества приближения

Из соотношения (6) следует, что при выборе функции $K(v)$ взятие в качестве показателя малости погрешности приближения $\tilde{f}_\xi^{(x)}(n)$ плотности $f_\xi(x)$ величины $E_1^{(K)} = \int_{-1/2}^{+1/2} K^2(v)\,dv$ – как это предложено в работе [5] – является не вполне правомерным: нужно еще каким-то образом учитывать величину $E_2^{(K)} = \int_{-1/2}^{+1/2} v^2 K(v)\,dv$ (в работе [5], на наш взгляд, неверно полагается, что $E_2^{(K)} = 1$).

Следуя рекомендациям работы [4], можно предложить приравнять (при фиксированных h и n) слагаемые суммы в правой части неравенства (6), соответствующие компоненте смещения и стохастической компоненте; при этом возникает пропорция вида

$$\int_{-1/2}^{+1/2} v^2 K(v)\,dv = C_3 \times \sqrt{\int_{-1/2}^{+1/2} K^2(v)\,dv} \quad \text{или} \quad E_2^{(K)} = C_3\sqrt{E_1^{(K)}}, \quad (7)$$

здесь $C_3 = \sqrt{\frac{C_2(n,h)}{C_1(h)}}$.

В связи с соотношением (7) неудачным видится также предложение из статьи [6], где при выборе функции $K(v)$ взятие в качестве показателя малости погрешности приближения $\tilde{f}_\xi^{(x)}(n)$ плотности $f_\xi(x)$ предложено брать величину $E_3^{(K)} = E_1^{(K)} \times \sqrt{E_2^{(K)}}$.

Для рассматриваемого в этой статье ядерного алгоритма мы должны рассматривать ядра $K(v)$, определенные на интервале $v \in$

$(-1/2, 1/2)$. Кроме того, с учетом соотношений (6), (7) мы предлагаем использовать при выборе функции $K(v)$ показатели малости погрешности приближения $\tilde{f}_\xi^{(x)}(n)$ плотности $f_\xi(x)$ вида $E_4^{(K)} = \sqrt{E_1^{(K)}} \times E_2^{(K)}$ и/или даже $E_5^{(K)} = \sqrt{E_1^{(K)}} + E_2^{(K)}$.

В таблице показаны наиболее употребимые постоянное, линейное и квадратичное ядра, взятые на интервале $(-1/2, 1/2)$. В первом столбце таблицы показаны различные типы функции $K^*(v)$, определенной на полуинтервале $[0, 1/2)$ и такой, что $K(v) = K^*(v)$ при $v \geq 0$ и $K(v) = K^*(-v)$ при $v \leq 0$. В остальных столбцах приведены рассмотренные выше показатели малости погрешности приближения $\tilde{f}_\xi^{(x)}(n)$ плотности $f_\xi(x)$.

Таблица 1

Сравнение показателей малости погрешности приближения $\tilde{f}_\xi^{(x)}(n)$ плотности $f_\xi(x)$ для разных ядер

$K^*(v)$	$E_1^{(K)}$	$E_2^{(K)}$	$E_3^{(K)}$	$E_4^{(K)}$	$E_5^{(K)}$
1	1,000	0,083	0,289	0,083	1,083
$2 - 4v$	1,333	0,042	0,273	0,048	1,197
$3/2 - 6v^2$	0,300	0,050	0,067	0,027	0,598

С одной стороны, табл. 1 показывает, что квадратичное ядро $K^*(v) = 3/2 - 6v^2$ дает наилучшие – малые – показатели качества приближения плотности (т. е. здесь получилась определенная аналогия с результатами работы [5]). С другой стороны, проведенные нами расчеты показали, что функция $K^*(v) = 1$ дает при использовании ядерного алгоритма приемлемую точность приближения за гораздо меньшее время вычислений, чем квадратичное ядро.

Заключение

В работе проведен анализ различных (в том числе, предложенных авторами статьи) показателей качества приближения неизвестной плотности $f_\xi(x)$ по заданной выборке $\{\xi_1, ..., \xi_n\}$ при выборе формы ядерной функции $K(v)$, $|v| < 1/2$ для использования ее в ядерном вычислительном алгоритме. Эти показатели выглядят как различные комбинации интеграла от квадрата $E_1^{(K)} = \int_{-1/2}^{+1/2} K^2(v)\, dv$ и «дисперсии» $E_2^{(K)} = \int_{-1/2}^{+1/2} v^2 K(v)\, dv$ ядерной функции. Расчеты показали, что, несмотря на «средние» показатели качества приближения, кусочно-постоянная ядерная функция (соответствующая многомерному аналогу полигона частот [1–4]) дает приемлемую точность приближения плотности за го-

раздо меньшее время, чем кусочно-линейная и кусочно-квадратичная ядерные функции.

СПИСОК ЛИТЕРАТУРЫ

1. *Войтишек А. В., Шлымбетов Н. Х.* Экономичные алгоритмы приближения вероятностной плотности по заданной выборке // Информационные технологии и математической моделирование (ИТММ–2022): материалы XXI Международной конференции имени А. Ф. Терпугова (25–29 октября 2022). Томск: Издательство Томского государственного университета, 2023. С. 319–324.
2. *Войтишек А. В., Шлымбетов Н. Х.* Требования к аппроксимационным базисам, используемым в вычислительных алгоритмах приближения вероятностной плотности по заданной выборке // Информационные технологии и математической моделирование (ИТММ–2023): материалы XXII Международной конференции имени А. Ф. Терпугова (4–9 декабря 2023). Томск: Издательство Томского государственного университета, 2023. Часть 1. С. 345–350.
3. *Войтишек А. В., Шлымбетов Н. Х.* Выбор аппроксимационных базисов, используемых в компьютерных функциональных алгоритмах приближения вероятностных плотностей по заданной выборке // Сибирский журнал вычислительной математики. 2024. Т. 27, № 2. С. 147–164.
4. *Войтишек А. В., Шлымбетов Н. Х.* Компоненты погрешности функциональных вычислительных алгоритмов приближения вероятностной плотности по заданной выборке: $\mathbb{L}_2$-подход // Информационные технологии и математическое моделирование (ИТММ–2024): материалы XXIII Международной конференции имени А. Ф. Терпугова (20–26 октября 2024). Томск: Издательство Томского государственного университета, 2024. С. 475–480.
5. *Епанечников В. А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. Т. 14, вып. 1. С. 156–161.
6. `https://ru.wikipedia.org/?curid=7083878\&oldid=146697015` — Википедия. Ядро (статистика). 16.07.2025.

Войтишек Антон Вацлавович — доктор физико-математических наук, профессор, главный научный сотрудник, лаборатория стохастических задач ИВМиМГ СО РАН; профессор, кафедра вычислительной математики НГУ. E-mail: *vav@osmf.sscc.ru*

Сейтмуратов Убайдулла Пахтамурат улы — студент магистратуры механико-математического факультета НГУ. E-mail: *u.seitmuratov@g.nsu.ru*

Шлымбетов Нурлыбай Хамдуллаевич — аспирант механико-математического факультета НГУ. E-mail: *n.shlymbetov@g.nsu.ru*

РЕГРЕССИЯ В ПРОСТРАНСТВЕ РАСПРЕДЕЛЕНИЙ, ОСНОВАННАЯ НА ЭРМИТОВЫХ СГЛАЖИВАЮЩИХ СПЛАЙНАХ

Б. С. Добронец[1], А. Р. Кемельбаева[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

Распределительная регрессия представляет собой современный подход к моделированию зависимости между переменными, позволяющий учитывать не только среднее значение целевой переменной, но и распределение в целом. В статье рассмотрен вычислительный вероятностный анализ (ВВА) – новое направление в области вычислительной статистики, ориентированное на обработку данных с неопределенностями и моделирование сложных систем. Основная идея ВВА заключается в представлении эмпирических данных в виде функций распределений, что позволяет учитывать вероятностную природу входных и выходных параметров. Также в статье указаны ключевые аспекты ВВА, предложен метод повышения точности аппроксимации плотности ошибок, основанный на использовании эрмитовых сплайнов пятой степени и метода наименьших квадратов. Результаты демонстрируют эффективность предложенного метода при наличии асимметрии, тяжелых хвостов или мультимодальности в распределении ошибок. **Ключевые слова:** *распределительная регрессия, вычислительный вероятностный анализ, сглаживающие эрмитовы сплайны, метод наименьших квадратов*

Введение

Традиционные методы регрессионного анализа, такие как линейная и логистическая регрессия, сосредоточены на оценке условного математического ожидания целевой переменной Y при заданных значениях предикторов X. Однако в реальных задачах важно учитывать не только среднее значение, но и всю совокупность параметров распределения $Y \mid X$, включая дисперсию, асимметрию и хвостовые характеристики. Распределительная регрессия позволяет моделировать параметры распределения $Y \mid X$ как функции от X, обеспечивая гибкость и адаптивность к сложным структурам данных.

Одним из подходов, который работает с распределениями, является символьный анализ данных (Symbolic Data analysis). В символьном ана-

лизе данных уже реализована регрессия в пространстве эмпирических распределений, но она работает только с данными представленными гистограммами [1].

GAMLSS – это современный подход, который также строит распределения, выделяет квантили этих распределений, но ещё и сохраняет местоположение, масштаб и форму [3]. Данный подход работает с наборами моделей, представленными различными аналитическими функциями, у которых есть свои параметры, но не работает с эмпирическими распределениями.

Функциональный анализ данных (Functional Data Analysis) рассматривает случайные функции и их распределения. Основной данного подхода является использование функционального метода главных компонент (FPCA) [5]. Если в качестве данных используются функции плотности вероятности, то интеграл от них должен быть равен единице и они должны быть положительны, что нарушается при использовании функционального метода главных компонент.

BBA предлагает принципиально иной подход, основанный на представлении данных в виде функций распределений и использовании кусочно-полиномиальных моделей для их обработки [4].

В работе предлагается метод распределительной регрессии, который учитывает произвольное распределение ошибок без параметрических ограничений. Особое внимание уделено практической реализации, включающей использование эрмитовых сплайнов пятой степени для аппроксимации плотности вероятностей и минимизации функционала метода наименьших квадратов.

1. Вычислительный вероятностный анализ

Вычислительный вероятностный анализ (BBA) – это методология численного моделирования, ориентированная на работу с данными, содержащими неопределенности. Основная цель BBA – снижение уровня неопределенности на всех этапах информационного процесса за счет использования вероятностных моделей. В отличие от классических подходов, BBA оперирует не точечными значениями, а функциями распределения, которые описывают вероятностную природу входных и выходных параметров [2].

Ключевыми элементами BBA являются:

- численные арифметики для плотностей вероятности;
- вероятностные расширения;
- анализ неопределенностей.

Эти элементы позволяют работать с неопределенностями в данных, что важно для построения регрессионных моделей в пространствах распределений. Регрессия в пространствах распределений предполагает, что входные X и выходные Y переменные представлены распределениями. Модель строится следующим образом:

$$Y = f(X) + \varepsilon, \tag{1}$$

где f – функция, которая отображает распределения X в распределения Y, а ε – ошибка модели.

BBA работает над кусочно-полиномиальными функциями, которые определяются сеткой ω

$$\omega = \{-\infty, x_0, x_1, \ldots, x_n, \infty\}, \tag{2}$$

где x – это полиномиальное выражение, которое может быть представлено: кусочно-постоянной функцией (гистограммы); кусочно-линейной функцией (частотный полигон); сплайном; обобщенной кусочно-полиномиальной функцией. Для построения регрессии от их параметров, необходимо приведение к некоторому стандартному шаблону, затем там строить регрессии. Кусочно-полиномиальные функции x_i и сетки ω_i преобразуем с помощью специального отображения $S(t)$ а затем, с помощью обратного отображения переходим в исходное состояние.

Для аппроксимации используются три базисные функции Эрмита:

- $\varphi_0(x)$ – значение функции в узле,
- $\varphi_1(x)$ – первая производная,
- $\varphi_2(x)$ – вторая производная.

Аппроксимирующая функция $S(t)$ строится как линейная комбинация базисных функций:

$$S(t) = \sum_{j=1}^{m} a_j \cdot \varphi_j(t), \tag{3}$$

где a_j – коэффициенты, определяемые методом наименьших квадратов, $\varphi_j(t)$ – базисные функции Эрмита пятой степени.

Задача МНК формулируется как минимизация функционала ошибки:

$$E = \sum_{i=1}^{n} \left(y_i - S(x_i)\right)^2 \to \min, \tag{4}$$

где (x_i, y_i) – исходные данные, n – количество точек. Это приводит к системе линейных уравнений:

$$\mathbf{Aa} = \mathbf{b}, \tag{5}$$

где $\mathbf{A}$ – матрица скалярных произведений базисных функций:

$$A_{ij} = \sum_{k=1}^{n} \varphi_i(x_k) \cdot \varphi_j(x_k), \tag{6}$$

$\mathbf{b}$ – вектор скалярных произведений базисных функций и данных:

$$b_i = \sum_{k=1}^{n} y_k \cdot \varphi_i(x_k). \tag{7}$$

2. Эксперимент

Была построена аппроксимация температур за 1960 год в г. Красноярске. Были использованы три базисные функции Эрмита: $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x)$ на компактныом носителе ($x \in [-1, 1]$):

$$\varphi_0(x) = \begin{cases} 1 - 10x^3 + 15x^4 - 6x^5, & x > 0, \\ 1 + 10x^3 + 15x^4 + 6x^5, & x \le 0. \end{cases} \tag{8}$$

$$\varphi_1(x) = \begin{cases} x(1 - 6x^2 + 8x^3 - 3x^4), & x > 0, \\ x(1 - 6x^2 - 8x^3 - 3x^4), & x \le 0. \end{cases} \tag{9}$$

$$\varphi_2(x) = \begin{cases} x^2 - 3x^3 + 3x^4 - x^5, & x > 0, \\ x^2 + 3x^3 + 3x^4 + x^5, & x \le 0. \end{cases} \tag{10}$$

Для построения сплайновой аппроксимации функции плотности вероятности f на первом этапе использовались ядерные оценки с параметром $h = 1$. Для разбиения данных на интервалы были заданы узлы $\omega = \{-22, -2, 5, 25, 39\}$. Интервалы h_i между узлами:

$$h_1 = 20, \quad h_2 = 7, \quad h_3 = 20, \quad h_4 = 14. \tag{11}$$

Каждая функция $\varphi_j(t)$ из формулы (3) адаптирована к конкретным интервалам h_i:

- $\varphi_0(t) = \varphi_0(x)$,
- $\varphi_1(t) = h_i \cdot \varphi_1(x)$,
- $\varphi_3(t) = h_i^2 \cdot \varphi_2(x)$.

В результате применения метода наименьших квадратов с базисными функциями Эрмита пятой степени была получена гладкая кривая, аппроксимирующая исходные данные. На рис. 1 представлены исходные точки (синие маркеры) и результат аппроксимации (красная линия).

Кривая демонстрирует плавное изменение значений, сохраняя непрерывность как самой функции, так и её первых и вторых производных на границах интервалов. Это подтверждается отсутствием видимых изломов или разрывов на графике, что особенно заметно в областях с высокой плотностью точек данных.

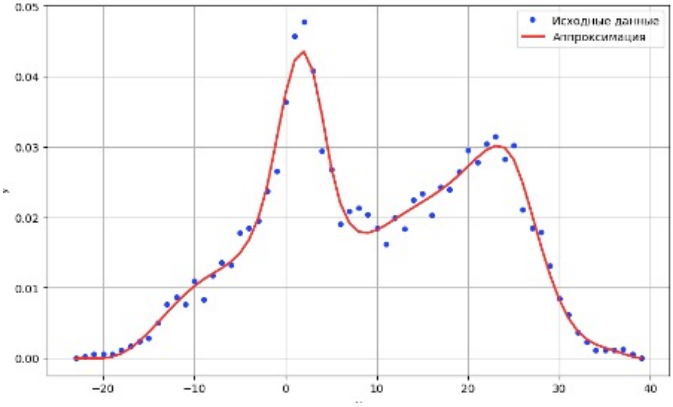


Рис. 1. Аппроксимация данных методом наименьших квадратов

Точность модели была оценена по среднеквадратичной ошибке (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - S(x_i))^2}. \tag{12}$$

Расчёт показал $\mathrm{RMSE} = 0,12 \cdot 10^{-2}$, что свидетельствует о высокой степени соответствия модели исходным данным. Для получения регрессии в пространсвах распределений необходимо повторить этот эксперемент для температурных данных в г. Красноярске за другие периоды.

Заключение

Распределительная регрессия открывает новые возможности для анализа сложных данных, позволяя моделировать не только среднее значение, но и рассматривать распределение в целом. ВВА представляет собой мощный инструмент для анализа данных с неопределенностями, сочетающий методы теории вероятностей, численных вычислений и визуализации. Подход позволяет строить точные модели даже при ограниченном объеме данных, что делает его актуальным для задач

цифровой экономики, оценки надежности и климатического прогнозирования.

Предложенный подход на основе базисных функций Эрмита и МНК демонстрирует эффективность для задач аппроксимации данных с требованием гладкости. Преимущества метода включают гибкость моделирования, устойчивость к выбросам и возможность интерпретации через базисные функции. Перспективы исследования включают адаптацию метода для многомерных данных и интеграцию с алгоритмами машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. *Billard L., Diday E.* Symbolic Data Analysis: Conceptual Statistics and Data Mining. Chichester: John Wiley & Sons, 2006. 336 p.
2. *Добронец Б. С., Попова О. А.* Вычислительный вероятностный анализ: модели и методы. Красноряск: СФУ, 2020. 237 с.
3. *Rigby R. A., Stasinopoulos D. M.* Generalized additive models for location,scale and shape // Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005. № 3. С. 507–554.
4. *Добронец Б. С., Попова О. А., Кемельбаева А. Р.* Новыеподходыврегрессионном анализе данных // Информатизация и связь. 2022. № 5. С. 78–81.
5. *Yoshiyuki A.* A Functional Linear Regression Model in the Space of Probability Density Functions // Research Institute of Economy, Trade and Industry. 2017.
6. *Fahrmeir L., Kneib T., Lang S., D. Marx B.* Regression: Models, Methods and Applications. Berlin: Springer Berlin Heidelberg, 2022. 746 p.

Добронец Борис Станиславович — доктор физико-математических наук, профессор кафедры систем искусственного интеллекта, институт космических и информационных технологий. E-mail: *bdobronets@yandex.ru*

Кемельбаева Айжамал Рысбековна — ассистент кафедры информационных систем, институт космических и информационных технологий. E-mail: *k_ aichka_ r@mail.ru*

РАСПРЕДЕЛЕНИЕ СУММЫ ЭКСПОНЕНЦИАЛЬНЫХ ВЕЛИЧИН СО СТРУКТУРОЙ ЗАВИСИМОСТИ, ОПРЕДЕЛЯЕМОЙ *FGM*-КОПУЛОЙ

А. Р. Ерилова[1], Д. В. Семенова[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

В статье представлено доказательство утверждений о виде функции распределения и плотности распределения взвешенной суммы двух зависимых случайных величин, имеющих маргинальные экспоненциальные распределения с различными параметрами, и со структурой зависимости, определяемой FGM-копулой. **Ключевые слова:** *распределение суммы зависимых случайных величин, экспоненциальное распределение, FGM-копула*

Введение

Согласно теореме Скляра [1] для произвольного абсолютно непрерывного распределения с функцией распределения $H(x_1, x_2)$ и с маргинальными распределениями $F_1(x_1)$, $F_2(x_2)$ существует единственная копула $C(u_1, u_2)$ такая, что для всех $(x_1, x_2) \in \mathbb{R}^2$ выполнятся равенство $H(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. Такое представление совместного распределения "делит" особенности исходного распределения на маргинальные — совокупность распределений $F_1(x_1)$, $F_2(x_2)$, и коррелятивные — копулу $C(u_1, u_2)$, в которой упакованы все связи между компонентами исходного распределения, т.е. в ней сосредоточены все модели зависимостей между случайными величинами.

Копула Фарли-Гумбеля-Моргенштерна (FGM) относится к числу наиболее известных параметрических семейств копул [1, 2]. Впервые она была введена Моргенштерном в 1956 для маргинальных распределений Коши. Позднее Гумбель в 1960 применил аналогичную формулу к экспоненциальному распределению, получив таким образом копулу для двумерного экспоненциального закона. Дальнейшее развитие было предложено Фарли в 1960, расширившим область применения формулы на произвольные распределения. Многомерный случай детально изучался Джонсоном и Котцем в 1975–1977 гг., именно они ввели название «распределение FGM». Семейство копул FGM имеет ограниченный диапазон зависимости, который она допускает меж-

ду случайными величинами и в двумерном случае задается формулой $C_\theta(u,v) = uv + \theta uv(1-u)(1-v)$. Параметр зависимости θ копулы строго ограничен $\theta \in [-1,1]$, а меры ассоциации, такие как τ-Кендалла и ρ-Спирмена, задаются соответственно как $\tau = 2\theta/9$ и $\rho = \theta/3$. Таким образом, копула FGM позволяет моделировать слабую зависимость между случайными величинами [1, 3]. Использование FGM копулы в данной работе обосновано тем, что она привлекательна своей простотой и формой, допускающей явные вычисления многих величин в замкнутой форме и получения точных результатов.

В работе [4] был предложен метод нахождения распределения, плотности и квантилей взвешенной суммы зависимых случайных величин, где структура зависимости полностью определяется абсолютно непрерывной копулой. Согласно этому методу в [5] сформулированы утверждения для двумерного случая функции плотности вероятности и кумулятивной функции распределения взвешенной суммы двух зависимых экспоненциальных величин, структура зависимостей которых определяется однопараметрической FGM-копулой. В данной работе приводится доказательство утверждений.

1. Основной результат

Теорема 1. Пусть случайные величины X_1 и X_2 распределены на одном и том же вероятностном пространстве и имеют экспоненциальное распределение с параметрами λ_1 и λ_2 соответственно. И пусть их совместное распределение H определяется FGM-копулой с параметром $\theta \in [-1,1]$. Тогда функция распределения случайной величины $Z = w_1X_1 + w_2X_2$, $w_1, w_2 > 0$, имеет вид

$$F_Z(z) = 1 - e^{-\frac{\lambda_1}{w_1}z} - B \cdot \frac{1 - e^{-\frac{\lambda_1}{w_1}(1-A)z}}{1-A} + \theta B\left(I_{2a} - BI_{2b}\right), \qquad (1)$$

где $A = \dfrac{\lambda_2 w_1}{\lambda_1 w_2}, B = e^{-\frac{\lambda_2 z}{w_2}}$,

$$I_{2a} = -\frac{1 - e^{-\frac{\lambda_1}{w_1}(1-A)z}}{1-A} + \frac{2\left(1 - e^{-\frac{\lambda_1}{w_1}(2-A)z}\right)}{2-A},$$

$$I_{2b} = -\frac{1 - e^{-\frac{\lambda_1}{w_1}(1-2A)z}}{1-2A} + \frac{\left(1 - e^{-\frac{\lambda_1}{w_1}(2-2A)z}\right)}{1-A}.$$

Доказательство. Пусть $F_i = 1 - e^{-\lambda_i x}$, $\lambda_i > 0$, $x \geq 0$ – функция распределения случайной величины X_i, $i = 1,2$. Поскольку F_i — строго

возрастающая на $x \geq 0$, то в соответствии с преобразованием Смирнова можем записать $X_i = F_i^{-1}(U_i)$, где $F_i^{-1}(u) = -\dfrac{\ln(1-u)}{\lambda_i}$ — квантильная функция распределения, а случайная величина $U_i = F_i(x)$ имеет непрерывное равномерное распределение на $(0,1)$.

В результате, можем представить выражение для случайной величины Z в виде $Z = w_1 F_1^{-1}(U_1) + w_2 F_2^{-1}(U_2)$ с функцией распределения $F_Z(z) = \mathbb{P}\left(w_1 F_1^{-1}(U_1) + w_2 F_2^{-1}(U_2) \leq z\right)$.

Обозначим $A_z = \left\{(x_1, x_2) \in \mathbb{R}^2 : w_1 x_1 + w_2 x_2 \leq z\right\}$ — полуплоскость в $\mathbb{R}^2$, отсекающую значения, где взвешенная сумма $w_1 x_1 + w_2 x_2$ не превышает z. Пусть образ допустимой области A_z в пространстве копул есть область $B_z = \left\{(u_1, u_2) \in [0,1]^2 : u_1 \in (0,1), 0 < u_2 \leq \tau_z(u_1)\right\}$. Здесь $\tau_z(u_1)$ — функция границы, которая при фиксированных значениях $u_1 \in (0,1)$ и $z \geq 0$ принимает вид

$$\tau_z(u_1) = F_2\left(\frac{1}{w_2}\left(z - w_1 F_1^{-1}(u_1)\right)\right) = 1 - (1-u_1)^{-\frac{\lambda_2 w_1}{\lambda_1 w_2}} \cdot e^{-\frac{\lambda_2 z}{w_2}}$$

и удовлетворяет условию $w_1 X_1 + w_2 X_2 = z$. Поскольку носители X_1 и X_2 положительны, то полагаем $\tau_z(u) = 0$ при

$$\frac{z}{w_2} - \frac{w_1}{w_2} F_1^{-1}(u_1) < 0 \Longrightarrow \frac{z}{w_2} - \frac{w_1}{w_2}\left(-\frac{\ln(1-u_1)}{\lambda_1}\right) < 0 \Longrightarrow u_1 > 1 - e^{-\frac{\lambda_1}{w_1} z}.$$

Согласно теореме 2 из [4] имеем следующее выражение для $F_Z(z)$

$$F_Z(z) = \int_0^1 \frac{\partial C(u_1, \tau_z(u_1))}{\partial u_1} du_1. \tag{2}$$

Отсюда получим вид подынтегральной функции для FGM-копулы

$$\frac{\partial C(u_1, \tau_z(u_1))}{\partial u_1} = \tau_z(u_1) + \theta \tau_z(u_1)(1 - \tau_z(u_1))(1 - 2u_1). \tag{3}$$

Учитывая условие $u_1 > 1 - e^{-\frac{\lambda_1}{w_1} z}$, подставим (3) в интеграл (2) и в результате получим следующий интеграл

$$F_Z(z) = \int_0^{1 - e^{-\frac{\lambda_1}{w_1} z}} \left[\tau_z(u_1) + \theta \tau_z(u_1)(1 - \tau_z(u_1))(1 - 2u_1)\right] du_1. \tag{4}$$

Разделим интеграл (4) по свойству линейности интегралов на два:

$$\underbrace{\int\limits_0^{1-e^{-\frac{\lambda_1}{w_1}z}} \tau_z(u_1)du_1}_{I_1} + \theta \underbrace{\int\limits_0^{1-e^{-\frac{\lambda_1}{w_1}z}} \tau_z(u_1)\left(1-\tau_z(u_1)\right)(1-2u_1)du_1}_{I_2}. \qquad (5)$$

Обозначим $A = \dfrac{\lambda_2 w_1}{\lambda_1 w_2}, B = e^{-\frac{\lambda_2 z}{w_2}}$. Тогда интегралы I_1 и I_2 примут вид

$$I_1 = \int\limits_0^{1-e^{-\frac{\lambda_1}{w_1}z}} \tau_z(u_1)du_1 = \left(1-e^{-\frac{\lambda_1}{w_1}z}\right) - B\cdot\frac{1-e^{-\frac{\lambda_1}{w_1}(1-A)z}}{1-A}, \qquad (6)$$

$$I_2 = \theta B\left(I_{2a} - BI_{2b}\right). \qquad (7)$$

Применяя методы математического анализа, получим следующие выражения для интегралов I_{2a} и I_{2b} из (7):

$$I_{2a} = \int\limits_0^{1-e^{-\frac{\lambda_1}{w_1}z}} \frac{(1-2u)}{(1-u)^A}du = -\frac{1-e^{-\frac{\lambda_1}{w_1}(1-A)z}}{1-A} + \frac{2\left(1-e^{-\frac{\lambda_1}{w_1}(2-A)z}\right)}{2-A},$$

$$I_{2b} = \int\limits_0^{1-e^{-\frac{\lambda_1}{w_1}z}} \frac{(1-2u)}{(1-u)^{2A}}du = -\frac{1-e^{-\frac{\lambda_1}{w_1}(1-2A)z}}{1-2A} + \frac{\left(1-e^{-\frac{\lambda_1}{w_1}(2-2A)z}\right)}{1-A}.$$

После подстановки (6) и (7) в (5) и упрощения, получим функцию распределения $F_Z(z)$ в виде (1). ■

Заметим, что функция (1) распределения содержит выражения, которые имеют неопределенности вида $\dfrac{0}{0}$ при следующих соотношениях параметров $w_1, w_2, \lambda_1, \lambda_2$: $A = \dfrac{\lambda_2 w_1}{w_2 \lambda_1} \in \left\{1, \dfrac{1}{2}, 2\right\}$. В этих ситуациях требуется дополнительный анализ с использованием теории пределов.

Следствие 1. Пусть случайная величина $Z = w_1X_1 + w_2X_2$ имеет закон распределения, заданный формулами (1), (8). Тогда при значениях параметра $A \in \left\{1, \dfrac{1}{2}, 2\right\}$ справедливы следующие утверждения:

1) при $A = 1$: $F_Z(z) = F_\Gamma(z;2,\lambda) + \theta e^{-\lambda z}\left(3 - \lambda z - e^{-\lambda z}(2\lambda z + 3)\right)$, где $\frac{\lambda_2}{w_2} = \frac{\lambda_1}{w_1} = \lambda$, а $F_\Gamma(z;2,\lambda)$ – кумулятивная функция Гамма-распределения;
2) при $A = \frac{1}{2}$: $F_Z(z) = 1 - e^{-\frac{\lambda_1}{w_1}z} - 2B\left(1 - e^{-\frac{\lambda_1}{2w_1}z}\right) + \theta B\left(-2\left(1 - e^{\frac{-\lambda_1 z}{2w_1}}\right)\right) + \theta B\left(\frac{4}{3}\left(1 - e^{\frac{-3\lambda_1 z}{2w_1}}\right) - B\left(\frac{-\lambda_1 z}{w_1} + 2\left(1 - e^{\frac{-\lambda_1 z}{w_1}}\right)\right)\right)$;
3) при $A = 2$: $F_Z(z) = 1 - e^{-\frac{\lambda_1}{w_1}z} + B\left(1 - e^{\frac{\lambda_1}{w_1}z}\right) + \theta B\left(1 - e^{\frac{\lambda_1 z}{w_1} + \frac{2\lambda_1 z}{w_1}} - B\left(\frac{1 - e^{\frac{3\lambda_1 z}{w_1}}}{3} + e^{\frac{2\lambda_1 z}{w_1} - 1}\right)\right)$.

Выражение для плотности случайной величины Z можно получить путём дифференцирования функции (1).

Следствие 2. В условиях теоремы 1 функция плотности случайной величины $Z = w_1X_1 + w_2X_2$, $w_1, w_2 > 0$, имеет вид

$$
\begin{aligned}
f_Z(z) = {} & \frac{\lambda_2}{w_2} e^{-\frac{\lambda_2 z}{w_2}} \left(\frac{1 - e^{-\frac{\lambda_1(1-A)z}{w_1}}}{1 - A}\right) + \\
& + \theta e^{-2\frac{\lambda_2 z}{w_2}} \left[\frac{2\lambda_2}{w_2}\left(\frac{1 - e^{-\frac{\lambda_1(2-2A)z}{w_1}}}{1 - A} - \frac{1 - e^{-\frac{\lambda_1(1-2A)z}{w_1}}}{1 - 2A}\right)\right] - \\
& - \theta e^{-\frac{\lambda_2 z}{w_2}} \left[\frac{2\lambda_2}{w_2}\frac{1 - e^{-\frac{\lambda_1(2-A)z}{w_1}}}{2 - A} + \frac{\lambda_2}{w_2}\frac{1 - e^{-\frac{\lambda_1(1-A)z}{w_1}}}{1 - A} + \right. \\
& \left. + \frac{\lambda_1}{w_1}\left(2e^{-\frac{\lambda_1(2-A)z}{w_1}} - e^{-\frac{\lambda_1(1-A)z}{w_1}}\right) - \right. \\
& \left. - \frac{\lambda_1}{w_1} e^{-\frac{\lambda_2 z}{w_2}} \left(2e^{-\frac{\lambda_1(2-2A)z}{w_1}} - e^{-\frac{\lambda_1(1-2A)z}{w_1}}\right)\right].
\end{aligned} \tag{8}
$$

Заключение

В работе представлены аналитические выражения функций распределения и плотности вероятности взвешенной суммы двух зависимых экспоненциальных величин, связанных однопараметрической FGM-копулой. В настоящее время FGM-копулы находят практическое применение в различных сферах, таких как управление капиталом страховых компаний, для описания различных корреляционных отношений на финансовых рынках, выбор оптимальных моделей медицинского страхования, анализ финансовых временных рядов и построение моделей надежности [6, 7]. Перспективы дальнейшего исследования заклю-

чаются в обобщении полученных результатов для суммы $n > 2$ случайных величин, что потенциально открывает новые возможности для решения практических задач, таких как оценка риска банкротства, проведение стресс-тестирования и оценка риска инвестиционных портфелей.

СПИСОК ЛИТЕРАТУРЫ

1. *Nelsen R. B.* Introduction to Copulas. New York: Springer Science+Business Media, 2007. 276 p.
2. *Joe H.* Dependence Modeling with Copulas (1st ed.) // Chapman and Hall/CRC., 2014. https://doi.org/10.1201/b17116
3. *Durante F.* A new class of symmetric bivariate copulas. // Journal of Nonparametric Statistics, 2006. Vol. 18, № 7-8. P. 499–510. https://doi.org/10.1080/10485250701262242
4. *Gijbels I., Hermann K.* On the distribution of the sum of random variables with copula-induced dependence // Insurance: Mathematics and Economics, 2014. V. 59. P. 27–44.
5. *Ерилова А. Р., Семенова Д. В.* О распределении взвешенной суммы двух экспоненциальных величин со структурой зависимости, определяемой FGM-копулой // ИТ. НАУКА. КРЕАТИВ. Системы управления, информационные технологии и математическое моделирование (Омск, 13–15 мая 2025 г.). Омск: Издательство ОмГТУ, 2025.
6. *Blier-Wong C., Cossette H., Marceau E.* Risk aggregation with FGM copulas [Electronic resource] // Insurance Mathematics and Economics, 2022. Vol. 111, no. 2 https://www.researchgate.net/publication/362385607_Risk_aggregation_with_FGM_copulas.
7. *Cossette H., Côté M.-P., Marceau E., Moutanabbir K.* Multivariate distribution defined with Farlie–Gumbel–Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation // *Insurance: Mathematics and Economics*, 2013. Vol. 52, № 3. P. 560-572. https://doi.org/10.1016/j.insmatheco.2013.03.006.

Семенова Дарья Владиславовна — канд. физ.-мат. наук, доцент, доцент кафедры высшей и прикладной математики СФУ. E-mail: *DVSemenova@sfukras.ru*

Ерилова Анастасия Романовна — студент Института математики и фундаментальной информатики СФУ. E-mail: *erilova.an2004@mail.ru*

ОБ АДАПТИВНОЙ КАЛМАНОВСКОЙ ФИЛЬТРАЦИИ СИСТЕМ С МАЛЫМ ШУМОМ

Ю. А. Кутоянц [1,2]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*
[2] *Le Mans University, Le Mans, France*

Рассматривается частично наблюдаемая линейная система калмановской фильтрации, коэффициенты которой зависят от неизвестного конечно-мерного параметра. Предлагается алгоритм построения адаптивного фильтра реализация которого осуществляется в несколько шагов. Сначала на малом интервале наблюдений строится предварительная оценка, затем эта оценка позволяет построить асимптотически эффективную рекуррентную оценку неизвестного параметра. Подстановка её в уравнения Калмана и позволяет построить адаптивный фильтр. Показана его асимптотическая эффективность.

Ключевые слова: *Калмановская фильтрация, рекуррентная оценка параметров, адаптивная фильтрация*

Введение

Предлагается обзор недавно полученных результатов по построению адаптивных фильтров для нескольких частично наблюдаемых марковских процессов зависящих от неизвестных параметров. Предлагаемый метод построения таких фильтров проиллюстрируем символически на следующей модели. Рассмотрим частично наблюдаемую систему

$$\begin{aligned} \mathrm{d}X_t &= f(\vartheta,t)\,Y_t\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \qquad X_0, \quad 0 \le t \le T, \\ \mathrm{d}Y_t &= a(\vartheta,t)\,Y_t\mathrm{d}t + b(\vartheta,t)\,\mathrm{d}V_t, \qquad Y_0 = y_0, \quad 0 \le t \le T, \end{aligned}$$

где $W_t, 0 \le t \le T$ и $V_t, 0 \le t \le T$, – независимые винеровские процессы, $f(\vartheta,t), \sigma(t), a(\vartheta,t), b(\vartheta,t), t \in [0,T]$, – известные функции, зависящие от неизвестного параметра $\vartheta \in \Theta \subset \mathcal{R}^d$. Коэффициент диффузии $\sigma(t)^2 \ge c > 0$ и начальное значение y_0 известны.Наблюдаемый процесс – $X^T = (X_t, 0 \le t \le T)$. Основная задача: по наблюдениям X^T построить оценку условного ожидания $m(\vartheta,t) = \mathbf{E}_\vartheta\,(Y_t | X_s, 0 \le s \le t)$.

Работа поддержана грантом РНФ 24-11-00191.

Напомним, что, если ϑ известно, то $m(\vartheta,t)$ задано уравнениями фильтрации Калмана-Бьюси

$$\mathrm{d}m(\vartheta,t)=\left[a(\vartheta,t)-\frac{\gamma(\vartheta,t)\,f(\vartheta,t)^2}{\sigma(t)^2}\right]m(\vartheta,t)\,\mathrm{d}t+\frac{\gamma(\vartheta,t)\,f(\vartheta,t)}{\sigma(t)^2}\mathrm{d}X_t,$$

$$\frac{\partial\gamma(\vartheta,t)}{\partial t}=2a(\vartheta,t)\,\gamma(\vartheta,t)-\frac{\gamma(\vartheta,t)^2\,f(\vartheta,t)^2}{\sigma(t)^2}+b(\vartheta,t)^2,$$

с начальными условиями $m_0=\mathbf{E}_\vartheta(Y_0|X_0)$ и $\gamma_0=\mathbf{E}_\vartheta(Y_0-m(\vartheta,0))^2$. **Основная проблема:** *Получить хорошее рекуррентное приближение* $m_t^\star, 0<t\le T$ *процесса* $m(\vartheta,t), 0<t\le T$.

Поэтому нам нужна *хорошая оценка* неизвестного параметра. «Хорошо» в этой ситуации может означать следующее:

A. *легко вычисляется для каждого* $t\in(0,T]$;
B. *в момент* t *зависит от наблюдений* $X_s, 0\le s\le t$;
C. *ошибка оценки* ϑ *мала.*

Эта оценка будет использоваться для построения $m_t^\star, 0<t\le T$.

Первыми кандидатами, конечно, являются оценка максимального правдоподобия (MLE) $\hat\vartheta_t, 0<t\le T$ и BE $\tilde\vartheta_t, 0<t\le T$. Напомним, что функция отношения правдоподобия для этой модели:

$$L\left(\vartheta,X^t\right)=\exp\left\{\int_0^t\frac{f(\vartheta,s)\,m(\vartheta,s)}{\sigma(s)^2}\,\mathrm{d}X_s-\int_0^t\frac{f(\vartheta,s)^2\,m(\vartheta,s)^2}{2\sigma(s)^2}\,\mathrm{d}s\right\}.$$

MLE и BE определяются стандартными соотношениями. Эти оценки асимптотически эффективны (в регулярном случае). Следовательно, условия **B** и **C** выполняются, а условие **A** — нет. Это потому, что для построения этих оценок нам нужно вычислить решения $m(\vartheta,s),\gamma(\vartheta,s), 0\le s\le t$ уравнений фильтрации для всех или многих $\vartheta\in\Theta$ и всех $t\in(0,T]$. Конечно, эти вычисления трудно реализовать.

Для аппроксимации $m(\vartheta,\cdot)$ мы предлагаем следующую программу:

1) *построим предварительную оценку* $\bar\vartheta_\tau$ *на относительно небольшом интервале наблюдений* $[0,\tau]$ *(интервал обучения) таким образом, что* $\tau/T\to0$;
2) *используя* $\bar\vartheta_\tau$, *реализуем одношаговый MLE-процесс* $\vartheta_t^\star, \tau<t\le T$:

$$\vartheta_t^\star=\bar\vartheta_\tau+\mathbf{I}_\tau^t\left(\bar\vartheta_\tau\right)^{-1}\int_\tau^t\frac{\dot{\mathrm{M}}\left(\bar\vartheta_\tau,s\right)}{\sigma(s)^2}\left[\mathrm{d}X_s-M\left(\bar\vartheta_\tau,s\right)\mathrm{d}s\right],$$

где $\mathbf{I}_\tau^t(\vartheta), \tau\le t\le T$ *— это семейство информационных матриц Фишера,* $M(\vartheta,s)=f(\vartheta,s)\,m(\vartheta,s)$ *и* $\dot{\mathrm{M}}(\vartheta,s)=\partial M(\vartheta,s)/\partial\vartheta$;

3) в качестве приближения $m(\vartheta,t), \tau<t\le T$ мы предлагаем $m_t^\star, \tau<t\le T$, полученное с помощью уравнений фильтрации, где ϑ заменяется на $\vartheta_t^\star, \tau<t\le T$:

$$\mathrm{d}m_t^\star=\left[a\left(\vartheta_t^\star,t\right)-\frac{\gamma_t^\star f\left(\vartheta_t^\star,t\right)^2}{\sigma\left(t\right)^2}\right]m_t^\star\mathrm{d}t+\frac{\gamma_t^\star f\left(\vartheta_t^\star,t\right)}{\sigma\left(t\right)^2}\mathrm{d}X_t,\quad \tau<t\le T,$$

$$\frac{\partial\gamma_t^\star}{\partial t}=2a\left(\vartheta_t^\star,t\right)\gamma_t^\star-\frac{\gamma_t^{\star 2} f\left(\vartheta_t^\star,t\right)^2}{\sigma\left(t\right)^2}+b\left(\vartheta_t^\star,t\right)^2;$$

4) последний шаг — описать ошибку $m_t^\star-m(\vartheta,t)$ и обсудить асимптотическую оптимальность $\mathbf{E}_\vartheta\left(m_t^\star-m(\vartheta,t)\right)^2$. Мы предлагаем следующую минимаксную нижнюю границу для всех оценок $\bar m_t$ процесса $m(\vartheta,t)$: существует функция $\mathcal{S}_t^\star(\vartheta)^2, \vartheta\in\Theta$ такая, что

$$\lim_{\nu\to 0}\lim_{\varphi\to 0}\sup_{\|\vartheta-\vartheta_0\|\le\nu}\varphi^{-2}\mathbf{E}_\vartheta\left(\bar m_t-m(\vartheta,t)\right)^2\ge\mathcal{S}_t^\star\left(\vartheta_0\right)^2.$$

Здесь $\varphi\to 0$ — нормализующая скорость. Адаптивный фильтр $m_t^\star, \tau\le t\le T$ называется асимптотически эффективным, если для всех $\vartheta_0\in\Theta$ и $t\in(\tau,T]$ имеет место равенство.

Одношаговый MLE-процесс также можно записать в рекурсивной форме. Следовательно, оценка-процесс $\vartheta_t^\star, \tau\le t\le T$ и адаптивный фильтр $m_t^\star, \tau\le t\le T$ имеют рекурсивную структуру.

1. Основной результат

Мы применяем эту конструкцию и предлагаем адаптивные фильтры для различных моделей наблюдений в [1]:

1) малый шум в обоих уравнениях;
2) малый шум только в наблюдениях;
3) условно гауссовские процессы;
4) EKF (extended Kalman filter);
5) скрытый процесс Орнштейна-Угленбека;
6) скрытый телеграфный сигнал;
7) скрытый процесс авторегрессии (AR);
8) несколько приложений (локализация неподвижного и движущегося источника на плоскости, идентификация процесса цены ценной бумаги, проблемы с разладками в наблюдаемом и скрытом процессах, приближение решения BSDE).

Ниже мы рассмотрим одну реализацию предлагаемого алгоритма в ситуации, когда имеется слабый шум в обоих уравнениях.

2. Малый шум

Рассмотрим модель

$$\begin{aligned} \mathrm{d}X_t &= f(\vartheta,t)\,Y_t\mathrm{d}t + \varepsilon\sigma(t)\,\mathrm{d}W_t, \qquad X_0, \quad 0\le t\le T,\\ \mathrm{d}Y_t &= a(\vartheta,t)\,Y_t\mathrm{d}t + \psi_\varepsilon b(\vartheta,t)\,\mathrm{d}V_t, \qquad Y_0=y_0, \quad 0\le t\le T, \end{aligned}$$

где $\varepsilon\in(0,1)$, $\varepsilon\to 0$ и $\psi_\varepsilon=\varepsilon^\delta$, $\delta\in[0,1]$.

В условиях регулярности (гладкости функций) ОМП и БО асимптотически нормальны

$$\varphi_\varepsilon^{-1}\left(\hat\vartheta_\varepsilon-\vartheta_0\right)\Longrightarrow\zeta(\vartheta_0)\sim\mathcal{N}\left(0,\mathbf{I}(\vartheta_0)^{-1}\right), \qquad \varphi_\varepsilon^{-1}\left(\tilde\vartheta_\varepsilon-\vartheta_0\right)\Longrightarrow\zeta(\vartheta_0).$$

Какова связь между ψ_ε и φ_ε.

Мы имеем 5 разных ситуаций:

1. $\psi_\varepsilon=\varepsilon$, $(\delta=1)$, $\varphi_\varepsilon=\varepsilon$,
2. $\psi_\varepsilon=1$, $(\delta=0)$, $\varphi\varepsilon=\sqrt{\varepsilon}$,
3. $\psi_\varepsilon=\varepsilon^\delta$, where $1/3<\delta<1$, $\varphi_\varepsilon=\varepsilon^\delta$,
4. $\psi_\varepsilon=\varepsilon^{1/3}$, $\varphi_\varepsilon=\varepsilon^{\frac{1}{3}}$,
5. $\psi_\varepsilon=\varepsilon^\delta$, where $0<\delta<1/3$, $\varphi_\varepsilon=\varepsilon^{\frac{1-\delta}{2}}$.

Во всех случаях предлагается предварительные оценки минимальных расстояний или метода моментов, затем One-step MLE-process и адаптивный фильтр.

3. Оценка мимимальных расстояний ($\psi_\varepsilon=\varepsilon$)

Предельная система ($\varepsilon=0$, ϑ_0 истинное значение) имеет вид

$$x_t(\vartheta_0)=\int_0^t f(\vartheta_0,s)\,y_s(\vartheta_0)\,\mathrm{d}s, \quad y_t(\vartheta_0)=y_0\exp\left(\int_0^t a(\vartheta_0,s)\,\mathrm{d}s\right).$$

Введём обучающий интервал $[0,\tau_\varepsilon]$, где $\tau_\varepsilon=\varepsilon^{2/3}\to 0$) и по наблюдениям $X^{\tau_\varepsilon}=(X_t,0\le t\le\tau_\varepsilon)$ строится оценка минимальных расстояний (ОМР)

$$\check\vartheta_{\tau_\varepsilon}=\arg\inf_{\vartheta\in\Theta}\int_0^{\tau_\varepsilon}\left[X_t-x_t(\vartheta)\right]^2\mathrm{d}t.$$

Предположим, что выполнены условия регулярности (гладкости функций модели), тогда ОМР состоятелмна и асимптотически нормальна

$$\frac{\tau_\varepsilon^{1/2}}{\varepsilon}\left(\check\vartheta_{\tau_\varepsilon}-\vartheta_0\right)\Longrightarrow\mathcal{N}(0,D(\vartheta_0)), \qquad D(\vartheta_0)=\frac{6}{5}\frac{\sigma(0)^2}{y_0\dot f(\vartheta_0,0)^2}.$$

Скорость сходимости $\tau_\varepsilon^{1/2}\varepsilon^{-1}=\varepsilon^{-2/3}$.

4. Фильтр Калмана

Обозначим $\gamma(\vartheta,t)=\varepsilon^2\gamma_*(\vartheta,t)$. Тогда

$$\mathrm{d}m(\vartheta,t)=\left[a(\vartheta,t)-\frac{\gamma_*(\vartheta,t)\,f(\vartheta,t)^2}{\sigma(t)^2}\right]m(\vartheta,t)\,\mathrm{d}t+\frac{\gamma_*(\vartheta,t)\,f(\vartheta,t)}{\sigma(t)^2}\mathrm{d}X_t,$$

$$\frac{\partial\gamma_*(\vartheta,t)}{\partial t}=2a(\vartheta,t)\,\gamma_*(\vartheta,t)-\frac{\gamma_*(\vartheta,t)^2\,f(\vartheta,t)^2}{\sigma(t)^2}+b(\vartheta,t)^2.$$

Обозначим $y_t(\vartheta,\vartheta_0)=m(\vartheta,t)\,|_{\varepsilon=0}$м $S_*(\vartheta,\vartheta_0,t)=f(\vartheta,t)\,y_t(\vartheta,\vartheta_0)$,

$$\mathbf{I}_\tau^t(\vartheta_0)=\int_\tau^t\frac{\dot S_*(\vartheta_0,\vartheta_0,s)^2}{\sigma(s)^2}\mathrm{d}s.$$

Здесь $\mathbf{I}_\tau^t(\vartheta_0),\vartheta_0\in\Theta$ информация Фишера.

5. One-step MLE-process

$\vartheta^\star_{t,\varepsilon},\tau_\varepsilon<t\le T$

$$\vartheta^\star_{t,\varepsilon}=\check\vartheta_{\tau_\varepsilon}+\mathbf{I}^t_{\tau_\varepsilon}(\check\vartheta_{\tau_\varepsilon})^{-1}\int_{\tau_\varepsilon}^t\frac{\dot S_*(\check\vartheta_{\tau_\varepsilon},\check\vartheta_{\tau_\varepsilon},s)}{\sigma(s)^2}\left[\mathrm{d}X_s-M(\check\vartheta_{\tau_\varepsilon},s)\mathrm{d}s\right].$$

Здесь $M(\vartheta,s)=f(\vartheta,s)\,m(\vartheta,s)$, $\dot S_*(\vartheta,\vartheta,s)=\dot S_*(\vartheta,z,s)\,|_{z=\vartheta}$ и мы имеем уравнение К-Б для $M(\check\vartheta_{\tau_\varepsilon},s),\tau_\varepsilon\le s\le T$,

$$\mathbf{I}_\tau^t(\vartheta_0)=\int_\tau^t\frac{\dot S_*(\vartheta_0,\vartheta_0,s)^2}{\sigma(s)^2}\mathrm{d}s>0.$$

Обозначим

$$\zeta_{t,\varepsilon}=\frac{\vartheta^\star_{t,\varepsilon}-\vartheta_0}{\varepsilon},\qquad\zeta_t(\vartheta_0)=\mathbf{I}_0^t(\vartheta_0)^{-1}\int_0^t\frac{\dot S_*(\vartheta_0,\vartheta_0,s)}{\sigma(s)}\mathrm{d}w(s).$$

Доказано, что для любого $\tau_0\in(0,T)$ случайный процесс $\zeta_{t,\varepsilon},\tau_0\le t\le T$ сходится по распределению в $(\mathcal{C}[\tau_0,T],\mathcal{B})$ к гауссовскому процессу $\zeta_t(\vartheta_0),\tau_0\le t\le T$. Рекуррентное представление

$$\begin{aligned}\mathrm{d}\vartheta^\star_{t,\varepsilon}=\mathbf{I}_0^t(\check\vartheta_{\tau_\varepsilon})^{-1}\frac{\dot S_*(\check\vartheta_{\tau_\varepsilon},\vartheta_{\tau_\varepsilon},s)^2}{\sigma(t)^2}\left[\check\vartheta_{\tau_\varepsilon}-\vartheta^\star_{t,\varepsilon}\right]\mathrm{d}t+\\+\mathbf{I}_0^t(\check\vartheta_{\tau_\varepsilon})^{-1}\frac{\dot M(\check\vartheta_{\tau_\varepsilon},t)}{\sigma(t)^2}\left[\mathrm{d}X_t-M(\check\vartheta_{\tau_\varepsilon},t)\mathrm{d}t\right],\end{aligned}$$

позволяет относительно просто построить эту оценку “online”.

6. Адаптивный фильтр

Предположим, что уже есть

1) Интервал обучения $[0, \tau_\varepsilon]$, $\tau_\varepsilon = \varepsilon^{2/3}$. 2) Предварительный MDE $\check{\vartheta}_{\tau_\varepsilon}$. 3) Уравнения К-Б для $M(\check{\vartheta}_{\tau_\varepsilon}, s)$ 4) Одношаговый MLE-процесс $\vartheta^\star_{t,\varepsilon}, \tau_\varepsilon < t \leq T$:

$$\vartheta^\star_{t,\varepsilon} = \check{\vartheta}_{\tau_\varepsilon} + \mathbf{I}^t_{\tau_\varepsilon}(\check{\vartheta}_{\tau_\varepsilon})^{-1} \int_{\tau_\varepsilon}^t \frac{\dot{S}_*(\check{\vartheta}_{\tau_\varepsilon}, \check{\vartheta}_{\tau_\varepsilon}, s)}{\sigma\,(s)^2} \left[\mathrm{d}X_s - M(\check{\vartheta}_{\tau_\varepsilon}, s)\mathrm{d}s\right].$$

Адаптивный фильтр

$$\mathrm{d}m^\star_{t,\varepsilon} = \left[a\left(\vartheta^\star_{t,\varepsilon}, t\right) - \frac{\gamma_{t,\varepsilon}\, f\left(\vartheta^\star_{t,\varepsilon}, t\right)^2}{\sigma\,(t)^2}\right] m^\star_{t,\varepsilon}\mathrm{d}t + \frac{\gamma_{t,\varepsilon}\, f\left(\vartheta^\star_{t,\varepsilon}, t\right)}{\sigma\,(t)^2}\mathrm{d}X_t,$$

$$\frac{\partial\gamma_{t,\varepsilon}}{\partial t} = 2a\left(\vartheta^\star_{t,\varepsilon}, t\right)\gamma_{t,\varepsilon} - \frac{\gamma^2_{t,\varepsilon}\, f\left(\vartheta^\star_{t,\varepsilon}, t\right)^2}{\sigma\,(t)^2} + b\left(\vartheta^\star_{t,\varepsilon}, t\right)^2, \qquad \gamma_{0,\varepsilon} = 0.$$

Начальное значение $m^\star_{\tau_\varepsilon,\varepsilon} = m(\check{\vartheta}_{\tau_\varepsilon}, \tau_\varepsilon)$. Заметим, что этот фильтр не оптимален, но может быть модифицированным до оптимальмного, но не рекуррентного. При других значениях ψ_ε адаптивные фильтры могут быть и эффективными.

Заключение

В работе предложен алгоритм построения адаптивного фильтра в ситуации, когда частично наблюдаемая система зависит от неизвестного конечно-мерного параметра. Оценка параметра и фильтр имеют рекуррентный вид и допускают относительно простую численную реализацию.

СПИСОК ЛИТЕРАТУРЫ

1. *Kutoyants Y. A.* Hidden Markov Processes and Adaptive Filtering. Cham: Springer, 2025. 576 с.

Кутоянц Юрий Артёмович — д.ф-м.н., профессор эмеритус, Лаборатория Математики Мансо, Ле Ман Университет, Франция. E-mail: *yury.kutoyants@univ-lemans.fr*

КРИТЕРИИ СОГЛАСИЯ ДЛЯ СМЕСИ РАСПРЕДЕЛЕНИЙ ВЕЙБУЛЛА

Г. В. Мартынов[1]

[1] *Институт проблем передачи информации им. А. А. Харкевича, г. Москва, Россия*

Предельные распределения статистик критериев согласия, основанных на эмпирических процессах, и при параметрических гипотезах зависят как от вида гипотетического параметрического семейства, так и чаще всего от неизвестных параметров распределений, входящих в такое семейство. Настоящая работа имеет дело с пятипараметрическим семейством распределений, образованным смесью двух различных распределений Вейбулла. Такое семейство распределений возникает в одномерном кластерном анализе. Обычно на практике предположения о распределениях наблюдаемых случайных величин считается известным априори. Однако, такая уверенность не всегда оправдана. В предлагаемой статье находятся предельные ковариационные функции эмпирического процесса при различных значениях параметров. Обсуждаются методы применения рассматриваемых критериев, когда таблица распределения статистики не может быть получена ввиду её обширности. В этом случае критические значения критерия должны вычисляться или точными методами или путём моделирования при полученных оценках параметров. **Ключевые слова:** *критерии согласия, распределение Вейбулла, смеси распределений*

Введение

В работе рассматривается применение критериев согласия, основанных на эмпирических процессах, для проверки гипотезы том, что функция распределения наблюдаемой случайной величины является смесью двух функций распределения Вейбулла с неизвестными различными в общем случае параметрами. Альтернативное распределение может быть любым распределением, не входящим в гипотетическое семейство распределений. Такие множества распределений рассматривают в одномерном кластерном анализе. В общей теории непараметрических критериев для проверки гипотез для параметрических семейств распределения наблюдается случайная величина X и её наблюдения пусть будут $X^n = \{X_1, X_2, ..., X_n\}$. Распредение X неизвестно, и его мы обозначим

как $F(x),\ x \in R_1$. Будем проверять гипотезу

$$H_0 : \quad F(x) \in \mathcal{G} = \{G(x,\theta),\ \ \theta = (\theta_1, \theta_2, ...\theta_k)^\top \in \Theta \subset R_k\},$$

где вектор параметров θ неизвестен. Распределение G задано. Обозначим истинное значение параметра как θ_0. Для проверки этой гипотезы здесь будут использоваться статистика Крамера–Мизеса–Смирнова

$$\omega_n^2(\hat{\theta}_n) = n\int_{-\infty}^{\infty}(F_n(x) - G(x,\hat{\theta}_n))^2\, dG(x,\hat{\theta}_n) = n\int_0^1 (\hat{F}_n(t) - t)^2 dt,$$

и статистику Колмогорова–Смирнова

$$D_n(\hat{\theta}_n) = \sqrt{n}\sup_{x\in R_1}|F_n(x) - G(x,\hat{\theta}_n)| = \sqrt{n}\sup_{t\in[0,1]}|\hat{F}_n(t) - t|.$$

Здесь $\hat{\theta}_n$ является оценкой максисмального правдоподобия вектора параметров θ, $F_n(x)$ есть эмпирическая функция распределения, построенная по выборке X^n, $\hat{F}_n(t)$ является оцененной эмпирической функцией распределения, построенной по преобразованной выборке $T^n = (t_1, \ldots, t_n)$, где $t_i = G(X_n, \hat{\theta}_n),\ i = 1, ..., n$. Обе статистики, преобразованные к интервалу $[0, 1]$, являются функциями от эмпирического процесса

$$\xi_n(t, \hat{\theta}_n) = \sqrt{n}\psi(t)(\hat{F}_n(t) - t).$$

Теорема 1. При некоторых условиях регулярности (см. [2, 5, 7, 8]) ковариационная функция процесса $K(t, \tau, \hat{\theta}_n) = \xi_n(t, \hat{\theta}_n)$ сходится к ковариационной функции

$$K(t, \tau, \theta_0) = E(\xi(t, \theta_0)\xi(\tau, \theta_0)) = (K_0(t, \tau) - q^\top(t, \theta_0)I^{-1}(\theta_0)q(\tau, \theta_0)), \quad (1)$$

где $K_0(t, \tau) = \min(t, \tau) - t\tau,\ \ t, \tau \in (0,\ 1)$,

$$q(t, \theta_0) = (\partial G(x, \theta)/\partial\theta_1, ..., \partial G(x, \theta)/\partial\theta_k)_{t=G(x,\theta_0),\theta=\theta_0},$$

$I(\theta_0)$ – информационная матрица Фишера,

$$I(\theta_0) = (E((\partial/\partial\theta_i)\log g(X, \theta)(\partial/\partial\theta_j)\log g(X, \theta)))_{\theta=\theta_0,\ 1\leqslant i,j\leqslant k},$$

$g(x, \theta_0) = (\partial G(x, \theta)/\partial x)_{\theta=\theta_0}$.

При нулевой гипотезе математическое ожидание процесса $\xi_n(t, \hat{\theta}_n)$ равно нулю.

В предположении слабой сходимости эмпирического процесса к предельному процессу $\xi(t,\theta_0)$ в соответствующих метрических пространствах распределение статистик $\omega_n^2(\hat{\theta}_n)$ и $D_n(\hat{\theta}_n)$ сходятся к распределениям функционалов

$$\omega^2(\theta_0)=\int_0^1 \psi^2(t)\xi^2(t,\theta_0)dt \ \text{ и } \ D(\theta_0)=\sup_{t\in[0,1]}\psi(t)|\xi(t,\theta_0)|.$$

Для статистики типа Крамера–Мизеса–Смирнова отметим условие

$$\int_0^1 K(t,t,\theta_0)dt<\infty.$$

Из формулы (1) следует, что предельное распределении статистики $\omega_n^2(\hat{\theta}_n)$ в общем случае зависит при гипотезе от неизвестного вектора параметров θ_0, а также и от функции распределения G. Поэтому, для нахождения критических уровней используют вместо неизвестных параметров их оцененные значения, которые получают путем интерполяции по таблице, или как в предлагаемой работе они вычисляются каждый раз заново. При большом числе неизвестных параметров составление таблиц невозможно.

Отметим также, что известно преобразование Хмаладзе (см., например, [9]) эмпирического процесса такое, что предельное распределение не зависит ни от неизвестного вектора параметров,ни от функции распределения G.

1. Семейство смесей двух распределений Вейбулла

Мы будем применять Теорему 1 для вывода формулы для предельной ковариационной функции эмпирического процесса для пятипараметрического семейства смеси двух различных распределений Вейбулла.

$$F(x,\lambda_1,k_1,\lambda_2,k_2,\rho)=\rho W(x,\lambda_1,k_1)+(1-\rho)W(x,\lambda_2,k_2),$$

где

$$W(x,\lambda,k)=1-e^{-\left(\frac{x}{\lambda}\right)^k},$$

ρ является параметром смеси. Соответствующая функция плотности есть

$$f(x,\lambda_1,k_1,\lambda_2,k_2,\rho)=$$
$$=\rho\tfrac{k_1}{\lambda_1}\left(\tfrac{x}{\lambda_1}\right)^{k_1-1}e^{-\left(\frac{x}{\lambda_1}\right)^{k_1}}+(1-\rho)\tfrac{k_2}{\lambda_2}\left(\tfrac{x}{\lambda_2}\right)^{k_2-1}e^{-\left(\frac{x}{\lambda_2}\right)^{k_2}}.$$

Будем обозначать вектор параметров $(\lambda_1, k_1, \lambda_2, k_2, \rho)$ символом θ. Таким образом, $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \equiv (\lambda_1, k_1, \lambda_2, k_2, \rho)$. Далее индексы элементов соответствующих вектора будут обозначать как цифрами, так и их первоначальными представлениями.

2. Вектор Q

Далее будем поэтапно находить составляющие ковариационной функции предельного эмпирического процесса. Нетрудно получить следующее выражение для вектора Q из Теоремы 1:

$$Q(x,\theta) = (\partial_{\lambda_1}, \partial_{k_1}, \partial_{\lambda_2}, \partial_{k_2}, \partial_\rho) F(x,\theta) =$$

$$= \left(\frac{-e^{-\left(\frac{x}{\lambda_1}\right)^{k_1}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho}{\lambda_1}, e^{-\left(\frac{x}{\lambda_1}\right)^{k_1}} \left(\frac{x}{\lambda_1}\right)^{k_1} \rho \ln\left(\frac{x}{\lambda_1}\right), \frac{e^{-\left(\frac{x}{\lambda_2}\right)^{k_2}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (\rho-1)}{\lambda_2}, \right.$$

$$\left. e^{-\left(\frac{x}{\lambda_2}\right)^{k_2}} \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) \ln\left(\frac{x}{\lambda_2}\right), -e^{-\left(\frac{x}{\lambda_1}\right)^{k_1}} + e^{-\left(\frac{x}{\lambda_2}\right)^{k_2}} \right).$$

3. Производные логарифма функции плотности распределения Вейбулла по параметрам

Здесь опущены промежуточные преобразования и приводится только окончательный результат.

$$f_{\lambda_1}(x,\theta) = \frac{\partial \ln f(x,\theta)}{\partial \lambda_1} = \frac{e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1^2 \left(1-\left(\frac{x}{\lambda_1}\right)^{k_1}\right) \left(\frac{x}{\lambda_1}\right)^{k_1} \rho}{\lambda_1 \left(e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho\right)},$$

$$f_{k_1}(x,\theta) = \frac{\partial \ln f(x,\theta)}{\partial k_1} = \frac{e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} \left(\frac{x}{\lambda_1}\right)^{k_1} \rho \left(1 + k_1 \left(1-\left(\frac{x}{\lambda_1}\right)^{k_1}\right) \ln\left(\frac{x}{\lambda_1}\right)\right)}{e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho},$$

$$f_{\lambda_2}(x,\theta) = \frac{\partial \ln f(x,\theta)}{\partial \lambda_2} = -\frac{e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2^2 \left(1-\left(\frac{x}{\lambda_2}\right)^{k_2}\right) \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho)}{\lambda_2 \left(e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho\right)},$$

$$f_{k_2}(x,\theta) = \frac{\partial \ln f(x,\theta)}{\partial k_2} = \frac{e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) \left(1 + k_2 \left(-1 + \left(\frac{x}{\lambda_2}\right)^{k_2}\right) \ln\left(\frac{x}{\lambda_2}\right)\right)}{e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho},$$

$$f_\rho(x,\theta) = \frac{\partial \ln f(x,\theta)}{\partial \rho} = \frac{e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} - e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2}}{e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho}.$$

4. Элементы матрицы Фишера

Пятнадцать различных элементов информационной матрицы Фишера теперь могут быть записаны следующим образом

$$S_{ij}(\theta) = \int_0^\infty f_i(x,\theta) f_j(x,\theta) f(x,\theta)\,dx \equiv \int_0^\infty s_{ij}(x,\theta)\,dx, \quad 1 \leqslant i \leqslant j \leqslant 5.$$

Опустив преобразования, приведём две типичные формулы для $s_{ij}(x,\theta)$:

$$s_{\lambda_1\lambda_1}(x,\theta) = f^2_{\lambda_1}(x,\theta) f(x,\theta) =$$
$$= -\frac{e^{-\left(\frac{x}{\lambda_1}\right)^{k_1}} k_1^4 \left(1-\left(\frac{x}{\lambda_1}\right)^{k_1}\right)^2 \left(\frac{x}{\lambda_1}\right)^{-1+2k_1} \rho^2}{\lambda_1^3 \left(e^{\left(\frac{x}{\lambda_1}\right)^{k_1} - \left(\frac{x}{\lambda_2}\right)^{k_2}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho\right)},$$

$$s_{\lambda_1 k_2}(x,\theta) = f_{\lambda_1}(x,\theta) f_{k_2}(x,\theta) f(x,\theta) =$$
$$= \frac{k_1^2 \left(1-\left(\frac{x}{\lambda_1}\right)^{k_1}\right) \left(\frac{x}{\lambda_1}\right)^{-1+k_1} \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) \rho \left(1 + k_2 \left(1-\left(\frac{x}{\lambda_2}\right)^{k_2}\right) \ln\left(\frac{x}{\lambda_2}\right)\right)}{\lambda_1^2 \left(e^{\left(\frac{x}{\lambda_1}\right)^{k_1}} k_2 \left(\frac{x}{\lambda_2}\right)^{k_2} (1-\rho) + e^{\left(\frac{x}{\lambda_2}\right)^{k_2}} k_1 \left(\frac{x}{\lambda_1}\right)^{k_1} \rho\right)}.$$

Все интегралы оказываются сходящимися.

5. Ковариационная функция

Используя полученные ранее формулы мы получаем следующее выражение для предельной ковариационной матрицы эмпирического процесса:

$$C(t,\tau,\theta) = \min(t,\tau) - t\tau - Q^\top(W^{-1}(t,\theta),\theta) \times$$
$$\times \left(\begin{pmatrix} S_{\mu_1\mu_1} & S_{\mu_1\sigma_1} & S_{\mu_1\mu_2} & S_{\mu_1\sigma_2} & S_{\mu_1\rho} \\ & S_{\sigma_1\sigma_1} & S_{\sigma_1\mu_2} & S_{\sigma_1\sigma_2} & S_{\sigma_2\rho} \\ & & S_{\mu_2\mu_2} & S_{\mu_2\sigma_2} & S_{\mu_2\rho} \\ & & & S_{\sigma_2\sigma_2} & S_{\sigma_2\rho} \\ & & & & S_{\rho\rho} \end{pmatrix}(\theta) \right)^{-1} Q(W^{-1}(\tau,\theta),\theta).$$

Эта функция оказывается зависящей от неизвестного параметра θ. При оцененном параметре могут быть найдены собственные значения корреляционного оператора, которые скорее всего окажутся однократными. Эти значения используются в формуле Смирнова для вычисления функций распределения квадратичных форм от нормальных случайных величин. Подробности можно найти в работах [1] и [2].

6. Замечание об оценках параметров

Оценивание параметров в таких моделях является трудоёмкой работой. Некоторые исследования в этом направлении можно найти в работах [4] и [6].

Заключение

Из полученных в работе результатов следует, что применение критериев согласия к смесям распределений является вполне выполнимой, хотя и трудоёмкой работой. Необходимо продолжить исследования для получения надёжных методов оценивания в таких моделях. То же самое заключение можно сделать и из результатов работы [3].

СПИСОК ЛИТЕРАТУРЫ

1. *Мартынов Г. В.* Вычисление функций распределения квадратичных форм от нормально распределённых квадратичных форм // Теория вероятностей и её применения. 1976. Т. 20, № 4. С. 782–793.
2. *Мартынов Г. В.* Критерий омега-квадрат. Москва: Наука, 1978. 79 с.
3. *Martynov G. V.* Cramer-von Mises test for Gaussian distribution mixtures // Journal of Mathematical Sciences. Springer. 2025. Vol. 289, № 3. P. 359–363
4. *Mezbahur Rahman, Sabiha Mahzaben.* Maximum likelihood parameter estimation for the mixtures of normal distributions // Advances and applications in statistics. 2018. Vol. 53, № 5. P. 501–518.
5. *Neuhaus G.* Asymptotic properties of the Cramér-von-Mises statisticswhen parameters are estimated // Proc. Prague Symp. Asymptotic Stat. Prgue Charles Univ., 1973. 1974. Vol. 2. P. 257–297.
6. *Tianyu Yan, Kai-Tai Fang, Hong Yin.* A novel approach for parameter estimation of mixture of two Weibull distributions in failure data modeling // Statistics and computing, Springer, 2024. V. 34. № 221. P. 1–16
7. *Durbin J.* Weak Convergence of the Sample Distribution Function when Parameters are Estimated // Ann. Statist. 1973. V. 1. № 2. P. 279–290.
8. *Kac M. Kiefer J., Wolfowitz J.* On tests of normality and other tests of goodness-of-fit based on distance methods // Ann. Math. Statist. 1955. V. 30. P. 420–447.
9. *Khoul H. L., Swordson E.* Khmaladze Transformation // International Encyclopedia of Statistical Science. Berlin Heidelberg: Springer. 2011. P. 715–718.

Мартынов Геннадий Владимирович — канд. физ.-мат. наук, снс, внс, ИППИ РАН. E-mail: *magevl@gmail.com*

АНАЛИЗ ВЛИЯНИЯ НАЧАЛЬНОГО ШУМА НА ПРОЦЕСС ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ ДИФФУЗИОННЫХ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

М. А. Мирза[1]

[1] *Сибирский Федеральный Университет, г. Красноярск, Россия*

Диффузионные вероятностные модели в последние годы произвели революцию в области генерации изображений благодаря своей способности масштабироваться за счёт увеличения вычислительных мощностей и количества данных, используемых в процессе обучения. Они способны генерировать разнообразные изображения по фиксированному текстовому запросу, используя различные значения начального шума. Однако влияние выбора начального шума на качество изображений не изучено до конца. Эта работа анализирует, как различные значения начального шума влияют на качество изображений, создаваемых по набору текстовых описаний. Результаты исследования показывают, что выбор начального шума значительно влияет на конечный результат, подчёркивая важность его оптимизации для улучшения качества генерации изображений. **Ключевые слова:** *генерация изображений, диффузионные вероятностные модели, генеративные модели*

Введение

Диффузионные вероятностные модели [1–3] стали настоящим прорывом в области генерации изображений. Их успех объясняется способностью эффективно обучаться при увеличении объема данных и размера моделей. Если ранее внимание исследователей было направлено на изучение влияния текстовых промптов [4] и совершенствование диффузионного процесса [5], то в последние годы всё больше работ концентрируются на оценке влияния начального шума на процесс генерации [6–9].

Настоящая работа направлена на изучение влияния выбора начального шума на процесс генерации изображений с помощью диффузионной вероятностной модели, что может позволить улучшить генерацию без увеличения размера модели. Вдохновляясь работами [7–9], где реализован подбор оптимального шума для улучшения генерации через предварительно обученные модели, была проведена оценка того,

как выбор начального шума влияет на качество изображений по метрикам CLIP, ImageReward и BRISQUE [10–12]. Используя модель Stable Diffusion v2.0 [3], был сгенерирован набор из 80 000 изображений на основе 100 фиксированных случайных шумов и 800 текстовых запросов из набора данных PartyPrompts [13].

Анализ оценок сгенерированных изображений показал значительное влияние начального шума на качество генерируемых изображений по всем используемым метрикам качества [10–12], а также существенные статистические различия между распределениями оценок для разных вариантов начальных шумов.

1. Генерация данных

В рамках исследования были проведены вычислительные эксперименты, в ходе которых сгенерировали 80 000 изображений с использованием модели Stable Diffusion v2.0 [3] и планировщика PNDM. Для генерации применили 800 текстовых запросов из набора PartyPrompts [13], включающих различные промпты на английском языке с аннотированными категориями и уровнями сложности. Это позволяет анализировать их влияние на качество изображений. Аннотации, составляющие менее 10% выборки, были объединены в категорию «Other». Распределение аннотаций представлено на рис. 1.

Было использовано 100 фиксированных значений начального шума, выбранных фиксированием значения начального состояния генератора случайных чисел в диапазоне от 0 до 99, что способствовало созданию уникальных изображений. Такой подход позволил изучить влияние начального шума и текстовых подсказок на генерацию изображений, обеспечивая разнообразие и детальный анализ взаимодействия семантики и стилистики в процессе генерации.

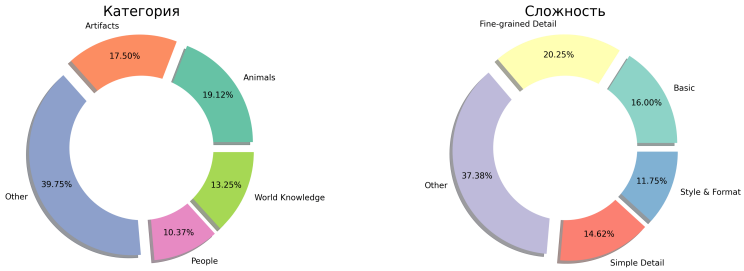


Рис. 1. Распределение категорий и сложности запросов в наборе данных PartyPrompts

2. Результаты эксперимента, анализ групповых оценок

В ходе работы для каждого изображения рассчитали значения всех метрик. Полученные оценки сгруппировали по используемому начальному шуму. Затем все варианты шума ранжировали по средним значениям полученных оценок. Результаты ранжирования показаны на рис. 2 и демонстрируют, что начальный шум существенно влияет на качество изображений. Значения наилучших и наихудших средних оценок равны: 0,3250 против 0,3154 для CLIPScore, 0,5766 против 0,1732 для ImageReward и 10,7742 против 25,2483 для BRISQUE.

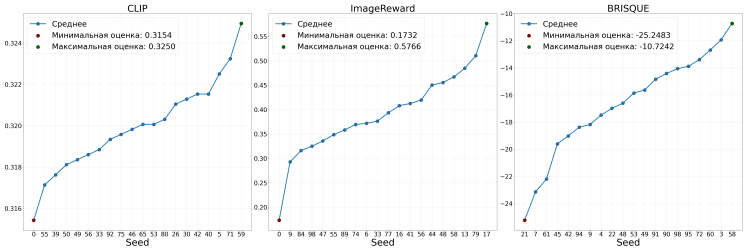


Рис. 2. График распределения средних оценок для начальных состояний отсортированных по возрастанию



Рис. 3. Примеры генерации для пяти худших (слева) и пяти лучших (справа) значений начальных состояний

Значения начального шума с более высокими оценками чаще генерируют более визуально приятные и понятные для человека изображения,

в то время как худшие варианты обладают скудной цветовой палитрой и более низким качеством, что видно на рис. 3.

3. Анализ распределения групповых оценок изображений

В ходе оценки изображений, сгенерированных из различных значений начального шума, были выявлены значительные различия в средних оценках. Непараметрический критерий Крускала–Уоллиса подтвердил их неслучайный характер. Результаты, представленные в табл. 1, показывают, что выбор начального шума оказывает существенное влияние на оценки изображений.

Таблица 1

Статистический тест гипотезы о едином распределении оценок

Критерий Крускала-Уоллиса	CLIP	ImageReward	BRISQUE
F-Статистика	224,0451	373,0522	4222,2286
p-value	0,0001	0,0001	0,0001

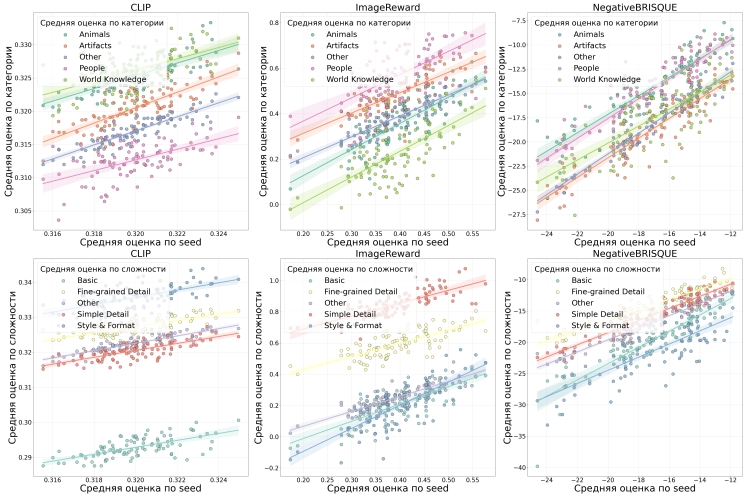


Рис. 4. Корреляция средних и групповых оценок для начальных шумов отсортированных по средней оценке

В ходе регрессионного анализа оценок изображений были вычислены средние оценки для групп, определенных типом или сложностью

запроса и фиксированным значением шума. Исследование корреляции между средними для групп и средними оценками для различных вариантов начального шума (пример на рис. 4) позволило оценить отклонение внутригрупповых оценок от общего среднего для каждого начального шума. Результаты показывают сильную линейную связь между этими параметрами.

Для изучения степени влияния шума на формирование оценки изображения был использован тест ANCOVA, оценивающий вклад различных факторов в предсказании оценки изображения. Результаты, приведенные в табл. 2 и табл. 3, демонстрируют, что для метрик CLIPScore и ImageReward шум менее значим, чем текстовый запрос, но всё же оказывает статистически значимое влияние на значения оценок. Влияние шума на метрику BRISQUE сравнимо с влиянием текстового запроса. Таким образом, при генерации изображений выбор значения шума оказывает значительное влияние на формирование оценки.

Таблица 2

F-статистика для факторов, влияющих на оценку изображения (начальный шум, категория, сложность)

Предикторы	CLIP	ImageReward	BRISQUE
Категория	331,6917	115,2502	197,622832
Сложность	2677,2224	1029,3812	417,4126
Начальный шум	2,1151	4,2083	34,1107

Таблица 3

F-статистика для факторов, влияющих на оценку изображения (начальный шум, текстовый запрос)

Предикторы	CLIP	ImageReward	BRISQUE
Текстовый запрос	226,0866	130,6875	53,9409
Начальный шум	6,8374	9,1216	50,4453

Заключение

Исследование показало, что начальный шум существенно влияет на стиль и оценки изображений, созданных диффузионной вероятностной моделью. Независимо от текстового промпта, настройки начального шу-

ма остаются значимыми для конечного результата. Это подчеркивает важность внимательной настройки шума для улучшения качества генерации и открывает возможности для дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Sohl-Dickstein, et al.* Deep Unsupervised Learning using Nonequilibrium Thermodynamics // arXiv preprint arXiv:1503.03585. 2015.
2. *Ho J., et al.* Denoising Diffusion Probabilistic Models // arXiv preprint arXiv:2006.11239, 2020.
3. *Rombach, R., et al.* High-resolution image synthesis with latent diffusion models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 10684–10695.
4. *Kim, G., Kwon, T., Chul Ye, J.* DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation // arXiv preprint arXiv:2110.02711. 2021.
5. *Kawar, B., Zada, S., et al.* Text-based real image editing with diffusion models // arXiv preprint arXiv:2210.09276. 2022.
6. *Ahn, D., Kang, J., et al.* A Noise is Worth Diffusion Guidance // arXiv preprint arXiv:2412.03895. 2024.
7. *Samuel, D., et al.* Generating images of rare concepts using pre-trained diffusion models // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. P. 4695–4703.
8. *Ma, N., Tong, S., Jia, H., et al.* Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps // arXiv preprint arXiv:2501.09732. 2025.
9. *Xu, K., Zhang, L., Shi, J.* Good Seed Makes a Good Crop: Discovering Secret Seeds in Text-to-Image Diffusion Models // arXiv preprint arXiv:2405.14828. 2024.
10. PartiPrompts. HuggingFace datasets. *Radford, A., et al.* Learning Transferable Visual Models From Natural Language Supervision // Proceedings of the International Conference on Machine Learning (ICML). 2021. P. 8748–8763.
11. *Xu, J., et al.* Learning and Evaluating Human Preferences for Text-to-Image Generation // arXiv preprint arXiv:2304.05977. 2023.
12. *Mittal, A., Moorthy, A. K., Bovik, A. C.* No-reference image quality assessment in the spatial domain // IEEE Transactions on Image Processing. 2012. P. 4695–4708.
13. `https://huggingface.co/datasets/nateraw/parti-prompts` — PartiPrompts. HuggingFace datasets.

Мирза Михаил Алексеевич — аспирант Сибирского федерального университета. E-mail: *mihailmirza98@gmail.com*

РОБАСТНОЕ ЭФФЕКТИВНОЕ НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ В АДДИТИВНОЙ РЕГРЕССИИ

Н. И. Никифоров [1], С. М. Пергаменщиков [1,2], Е. А. Пчелинцев [1]

[1] *Томский государственный университет, г. Томск, Россия*
[2] *Руанский университет, г. Руан, Франция*

В статье рассматривается задача робастного оценивания неизвестной функции в аддитивной регрессионной модели в дискретном времени. Предложена новая процедура оценивания, для которой скорость сходимости робастного риска равна параметрической (с точностью до логарифмически возрастающего множителя). **Ключевые слова:** *Непараметрическое оценивание, аддитивные регрессионные модели, робастный риск, эффективность*

1. Введение

Пусть наблюдения $(y_j)_{j\geq 1}$ удовлетворяют уравнению

$$y_j = S(t_j) + \xi_j\,, \quad t_j = \frac{j}{n} \quad \text{и} \quad 1 \leq j \leq n\,, \tag{1}$$

где $S(\cdot)$ — неизвестная неслучайная $[0,1] \to \mathbb{R}$ функция из $\mathcal{L}_2[0,1]$, а шум $(\xi_j)_{1\leq j\leq n}$ представляет собой набор независимых одинаково распределённых случайных величин с неизвестным распределением $\mathbf{p}$ (не обязательно гауссовским). Требуется оценить функцию S по наблюдениям $(y_j)_{j\geq 1}$. Данная задача для модели аддитивной регрессии (1) широко встречается в обработке сигналов, экономике и других прикладных областях. Также важной характеристикой современного подхода является робастность и эффективность разрабатываемых алгоритмов оценивания. Проблема эффективного оценивания для моделей (1) изучалась во многих работах, в том числе в [1–4]. Однако в этих работах свойство эффективности доказывается в случае гауссовских шумов или для схем с бесконечным числом наблюдений. Это существенно ограничивает применение полученных оценок на практике.

Цель настоящей работы – разработать робастный эффективный метод оценивания для модели (1), в которой распределение шума может выбираться из достаточно широкого класса.

Работа выполнена при поддержке РНФ, проект № 24-11-00191.

Прежде всего, для оценивания функции S в модели (1) необходимо зафиксировать в $\mathcal{L}_2[0,1]$ ортонормированный базис $(\psi_j)_{j\geq 1}$ на решетке $\mathcal{T}_n = t_1, ..., t_n$, т.е. для любых $i, j \geq 1$

$$(\psi_i, \psi_j)_n = \frac{1}{n}\sum_{k=1}^{n} \psi_i(t_k)\psi_j(t_k) = \mathbf{1}_{\{i=j\}}\,.$$

Далее обозначаем через $|\cdot|_n$ – норму, соответствующую этому скалярному произведению. Неизвестная функция S в (1) может быть представлена в виде

$$S(t) = \sum_{j=1}^{n} \theta_j\, \psi_j(t)\,, \quad t \in \mathcal{T}_n\,, \tag{2}$$

где $\theta_j = (S, \psi_j)_n$, $j = 1, ..., n$, – коэффициенты Фурье. Используя эти коэффициенты, определим квадратичный риск

$$\mathcal{R}_{\mathbf{p}}(\widehat{S}_n, S) \;=\; \mathbf{E}_{\mathbf{p},S}\|\widehat{S}_n - S\|_n^2 = \sum_{j=1}^{n} \mathbf{E}_{\mathbf{p},S}\,(\widehat{\theta}_j - \theta_j)^2\,, \tag{3}$$

где $\widehat{S}_n$ – оценка функции S, по наблюдениям $(y_j)_{1\leq j\leq n}$, а $\mathbb{E}_{\mathbf{p},S}$ — математическое ожидание относительно распределения $\mathbb{P}_{\mathbf{p},S}$ процесса $(y_j)_{1\leq j\leq n}$, соответствующего распределению шума $\mathbf{p}$ и функции S, и $\widehat{\theta}_j = (\widehat{S}, \psi_j)_n$ — соответствующие коэффициенты Фурье функции $\widehat{S}$. Более того, учитывая, что $\mathbf{p}$ неизвестно, введем робастный риск

$$\mathcal{R}^*(\widehat{S}_n, S) \;=\; \sup_{\mathbf{p}\in\mathcal{P}} \mathcal{R}_{\mathbf{p}}(\widehat{S}_n, S)\,.$$

Здесь

$$\mathcal{P} := \left\{\mathbf{p} \geq 0 \,:\, \int_{-\infty}^{+\infty} x\,\mathbf{p}(\mathrm{x}) \;=0 \quad \text{и} \quad \int_{-\infty}^{+\infty} x^2\,\mathbf{p}(\mathrm{x}) \leq \varsigma\right\}$$

представляет собой семейство вероятностных распределений с некоторым фиксированным неизвестным порогом дисперсии $\varsigma > 0$, который может быть медленно возрастающей функцией от числа наблюдений n, т.е. $\varsigma = \varsigma_*(n)$, такая, что для любого ≥ 0

$$\lim_{n\to\infty}\, n^{-\mathbf{b}}\,\varsigma_*(n) = 0\,.$$

Очевидно, что гауссовское распределение с параметрами $(0, \varsigma)$ принадлежит $\mathcal{P}$. Предположим, что коэффициенты Фурье $(\theta_1, \ldots, \theta_n)$ неиз-

вестной функции S принадлежат n-мерному эллипсу Θ в $\mathbb{R}^n$, определённому как

$$\Theta = \left\{ z = (z_1, \ldots, z_n) \in \mathbb{R}^n \, : \, \sum_{j=1}^{n} a_j \, z_j^2 \leq \mathbf{r} \right\}, \tag{4}$$

где $a_j = e^{2\varkappa j^\alpha}$, $\varkappa > 0$, $0 < \alpha < 1$, а $\mathbf{r} > 0$ — радиус, который может зависеть от n: $\mathbf{r} = \mathbf{r}_n$ такой, что для любого $\varepsilon > 0$

$$\mathbf{r}_n = \mathrm{o}(\ln^{1+\varepsilon} n) \quad \text{при} \quad n \to \infty \, .$$

2. Процедура оценивания

Для оценивания неизвестной функции S в (1) используется её разложение Фурье (2) для некоторых базисных функций $(\psi_j)_{1 \leq j \leq n}$. Например, можно использовать тригонометрический или сплайновый базис, определенный в [5]. Сначала оцениваем коэффициенты $(\theta_j)_{1 \leq j \leq n}$ с помощью метода наименьших квадратов

$$\widehat{\theta}_j = (\psi_j, y)_n = \frac{1}{n} \sum_{l=1}^{n} \psi_j(t_l) \, y_l \, .$$

Затем согласно методу взвешенных наименьших квадратов [6], определяем процедуру оценивания

$$S^*(t) = \sum_{j=1}^{n_*} \gamma_j \widehat{\theta}_j \psi_j(t) \, , \quad t \in \mathcal{T}_n \, , \tag{5}$$

где $n_* = [(2\varkappa)^{-1/\alpha} (\ln n)^{1/\alpha}] + 1$, $\gamma_j = 1 - e^{-\varkappa(n_*^\alpha - j^\alpha)}$ и параметры $\varkappa$ и α введены в (4).

3. Основные результаты

Сначала исследуется верхняя граница для риска (3) соответствующая оценке (5).

Теорема 1. Для любых $\varkappa > 0$ и $0 < \alpha < 1$ робастный риск оценки (5) допускает следующую верхнюю границу

$$\limsup_{n \to \infty} \upsilon_n \sup_{\theta \in \Theta} \mathcal{R}^*(S^*, S) \leq \ (2\varkappa)^{-1/\alpha} \, ,$$

где коэффициент $\upsilon_n = n(\ln n)^{-1/\alpha} / \varsigma_*$.

Для изучения свойства эффективности введем класс Ξ_n – всех оценок $\widehat{S}_n$ измеримых относительно $\sigma\{y_j \, , \, 1 \leq j \leq n\}$.

Теорема 2. Робастный риск (3) допускает следующую нижнюю границу

$$\liminf_{n\to\infty} \upsilon_n \inf_{\widehat{S}\in\Xi_n} \sup_{\theta\in\Theta} \mathcal{R}^*(\widehat{S},S) \geq (2\varkappa)^{-1/\alpha}\,.$$

Из этих теорем вытекает следующее свойство эффективности.

Теорема 3. Оценка (5) является асимптотически эффективной, то есть

$$\lim_{n\to\infty} \frac{\inf\limits_{\widehat{S}\in\Xi_n} \sup\limits_{\theta\in\Theta} \mathcal{R}^*(\widehat{S},S)}{\sup\limits_{\theta\in\Theta} \mathcal{R}^*(S^*,S)} = 1$$

и, более того,

$$\lim_{n\to\infty} \upsilon_n \sup_{\theta\in\Theta} \mathcal{R}^*(S^*,S) = (2\varkappa)^{-1/\alpha}\,.$$

Для модели (1) оптимальная скорость сходимости для параметрических задач равна n^{-1}, тут получена $n^{-1}(\ln n)^{1/\alpha}$, т.е. почти параметрическая скорость сходимости с точностью до логарифмически возрастающего коэффициента $(\ln n)^{1/\alpha}$, т.е. эта скорость быстрее, чем $n^{-\gamma}$ для любых $0<\gamma<1$. Поэтому процедура оценивания (5) называется *суперэффективной*.

Заключение

В статье для модели (1) предложена суперэффективная робастная непараметрическая оценка, для которой робастный квадратичный риск стремится к нулю с более высокой скоростью, чем любая степенная, в смысле непараметрической минимаксной скорости. Следует подчеркнуть, что для предложенной оценки риск вычислен в явной асимптотической форме, обеспечивая точную константу, которая является минимальной среди нормированных квадратичных рисков всех возможных процедур оценивания, основанных на наблюдениях $y=(y_j)_{1\leq j\leq n}$. Это означает, что для данной модели константа Пинскера найдена в явном виде.

СПИСОК ЛИТЕРАТУРЫ

1. *Nussbaum M.* Spline smoothing in Regression Models and Asymptotic Efficiency in L_2 // The Annals of Statistics. 1985. Vol. 13, № 3. P. 984–997.
2. *Speckman P.* Spline smoothing and optimal rates of convergence in nonparametric regression models // The Annals of Statistics. 1985. Vol. 13, № 3. P. 970–983.
3. *Golubev G. K., Nussbaum M.* Efficient estimation methods for non-Gaussian regression models in continuous time // Adaptive spline estimates for nonparametric regression models. 1993. Vol. 37, № 3. P. 521–529.

4. *Golubev Yu. K., Levit B. Y., Tsybakov A. B.* Asymptotically Efficient Estimation of Analytic Functions in Gaussian Noise // Bernoulli. 1996. Vol. 2, № 2. P. 167–181.
5. *Demmler A., Reinsch C.* Oscillation matrices with spline smoothing // Numer. Math. 1975. Vol. 24. P. 357–382.
6. *Pchelintsev E. A., Pergamenshchikov S. M., Povzun M. A.* Efficient estimation methods for non-Gaussian regression models in continuous time // Annals of the Institute of Statistical Mathematics. 2022. Vol. 74. P. 113–142.

Никифоров Никита Игоревич — аспирант, кафедра математического анализа и теории функций, механико-математический факультет, Томский государственный университет. E-mail: *Nikitanikiforov_97@bk.ru*

Пергаменщиков Сергей Маркович — доктор физико-математических наук, профессор, кафедра математического анализа и теории функций, механико-математический факультет, Томский государственный университет, лаборатория математики Рафаэля Салема, Руанский университет. E-mail: *serge.pergamenchtchikov@univ-rouen.fr*

Пчелинцев Евгений Анатольевич — доктор физико-математических наук, доцент, заведующий кафедрой математического анализа и теории функций, механико-математический факультет, Томский государственный университет. E-mail: *evgen-pch@yandex.ru*

СТОХАСТИЧЕСКИЕ МОДЕЛИ ЗАЩИТЫ ИНФОРМАЦИОННОЙ СИСТЕМЫ С УПРАВЛЕНИЕМ

О. В. Рыбкина[1], Е. В. Карачанская[1]

[1] *Дальневосточный государственный университет путей сообщения (ДВГУПС), г. Хабаровск, Россия*

Предложены математические модели информационной системы, защищенной с вероятностью 1 от угроз информационной безопасности на основе классической модели Ланчестера борьбы двух противников и классической SIR–модели. Для построения моделей использован метод построения систем стохастических дифференциальных уравнений с заданным набором инвариантов – первых интегралов. В качестве инвариантов рассматриваются свойства информационной системы, которые необходимы для ее работоспособности в условиях реализации различных сценариев атак на ИС. **Ключевые слова:** *модель Ланчестера, SIR-модель, информационная безопасность, стохастическая модель защиты информации, программное управление с вероятностью 1*

Введение

Рассматривая защиту ИС от угроз информационной безопасности (УИБ) как динамическую систему, можно предположить, что существуют показатели, сохранение значений которых говорит о том, что ИС находится в защищенном состоянии. С математической точки зрения данные показатели можно рассматривать как первые интегралы (инварианты) [1], на основе которых можно построить модель защиты ИС в виде СДУ Ито [2, 3]. Математическая модель ИС, построенная на основе сохранения с вероятностью 1 этих инвариантов, позволит в режиме реального времени отражать (блокировать) опасный входящий информационный поток. В работах [4, 5] рассматривались и исследовались модели построения рассматривались и исследовались «жесткая» и «мягкая» модели построения защиты от угроз ИБ, основанные на классической математической модели войны Ланчестера. В данной работе представлен общий подход к построению моделей защиты от угроз ИБ

Работа выполнена при частичной поддержке Минцифры (проект 22/23-К).

на основе классических моделей войны Ланчестера и SIR-модели, с использованием метода построения систем СДУ с заданным набором инвариантов. В качестве инвариантов рассматриваются свойства информационной системы, которые необходимы для ее работоспособности в условиях реализации различных сценариев атак на ИС.

1. Стохастическая модель защиты ИС от УИБ на основе модели Ланчестера

Под активными средствами защиты информации будем понимать такие средства защиты, которые в текущий момент времени способны противодействовать поступающей угрозе.

Используем классическую модель Ланчестера [6]. В рамках этой модели опишем исследуемую задачу. Пусть сторона – «нарушитель» в момент времени t осуществляет $u(t)$ сценариев угроз ИБ, используя уязвимости объектов информационного воздействия (ИВ). Сторона – «защитник» имеет z_0 объектов ИВ и $\alpha(t)$ активных средств защиты информации (СЗИ). Под воздействием осуществленных сценариев угроз число объектов ИВ $z(t)$ может изменяться [4, 5].

Предположим, что нарушителю известно количество объектов ИВ и используемые СЗИ, что соответствует ограничениям: $u(t) \leq z_0$ и $\alpha(t) \leq z_0$. Тогда система дифференциальных уравнений, учитывающая динамику изменения числа активных средств защиты информации под воздействием УИБ, будет иметь вид:

$$\begin{cases} \frac{du(t)}{dt} = -a \cdot \alpha(t) + b \cdot u(t), \\ \frac{dz(t)}{dt} = -u(t) + \beta \cdot (z_0 - z(t)), \\ \frac{d\alpha(t)}{dt} = \gamma \cdot (u(t) - \alpha(t)), \\ z(0) = z_0, \quad \alpha(0) = \alpha_0, \end{cases} \tag{1}$$

где a $(0 \leq a \leq 1)$ – коэффициент эффективности средств защиты информации, b – коэффициент интенсивности появления новых сценариев угроз, β – коэффициент интенсивности восстановления объектов информационного воздействия, γ – коэффициент интенсивности ввода новых активных средств защиты информации.

Нечувствительность ИС к внешнему воздействию будем сопоставлять с некоторыми инвариантами – функциями $s\,(t, u(t), z(t), \alpha(t))$ и $v\,(t, u(t), z(t), \alpha(t))$, которые будут сохранять постоянное значение. Учитывая случайность возникновения угроз ИБ, для более адекватного описания, внесем в модель (1) случайное слагаемое в виде винеровского процесса $w(t)$. Для сохранения необходимых инвариантов $s\,(t, u(t), z(t), \alpha(t))$ и $v\,(t, u(t), z(t), \alpha(t))$ модели со случайностью вве-

дем в систему уравнений функцию – компенсатор PCP1 (программное управление с вероятностью 1): $\hat{j}(t) = \hat{j}\,(t, u(t), z(t), \alpha(t))$. Учитывая, что на входящий информационный поток мы не можем воздействовать, управление будет внесено только во второе и третье уравнения системы. Тогда стохастическая модель, соответствующая (1), имеет вид:

$$\begin{cases} du(t) = (-a \cdot \alpha(t) + b \cdot u(t))\,dt + r_1(t)dw(t), \\ dz(t) = \Big(-u(t) + \beta \cdot (z_0 - z(t)) + \hat{j}_2(t)\Big)\,dt + r_2(t)dw(t), \\ d\alpha(t) = \Big(\gamma \cdot (u(t) - \alpha(t)) + \hat{j}_3(t)\Big)\,dt + r_3(t)dw(t). \end{cases} \tag{2}$$

Для определения управлений $\hat{j}_2(t)$ и $\hat{j}_3(t)$ необходимо потроить систему СДУ Ито [2] с первыми интегралами $s\,(t, u(t), z(t), \alpha(t))$ и $v\,(t, u(t), z(t), \alpha(t))$.

В качестве первых интегралов для системы СДУ – инвариантов модели – возьмем, напрмер, функции:

$$\begin{gathered} s\,(t, u(t), z(t), \alpha(t)) = \alpha(t) - u(t) + 1, \\ v\,(t, u(t), z(t), \alpha(t)) = z(t). \end{gathered} \tag{3}$$

Выбор первой функции объясняется тем, что для поддержания высокого уровня защиты ИС от УИБ, необходимо постоянно анализировать поступающие угрозы и вводить новые активные СЗИ. Вторая функция необходима в качестве инварианта по причине того, что требуется работоспособная ИС, т.е. неизменность количества объектов ИВ.

Согласно алгоритму [2], в качестве дополнительной и произвольной возьмем, например, функции

$$\begin{gathered} g\,(t, u(t), z(t), \alpha(t)) = \alpha(t) - z(t) + e^{-t}, \\ q_{00}(t, u(t), z(t), \alpha(t)) = u(t) - \alpha(t). \end{gathered} \tag{4}$$

Построенная система СДУ с инвариантами (3) и функциями (4) будет иметь вид:

$$\begin{cases} du(t) = e^{-t}dt + (\alpha(t) - u(t))\,dw(t), \\ dz(t) = 0, \\ d\alpha(t) = e^{-t}dt + (\alpha(t) - u(t))\,dw(t). \end{cases} \tag{5}$$

Получаем:

$$\begin{cases} \hat{j}_2(t) = u(t) - \beta \cdot (z_0 - z(t)), \\ \hat{j}_3(t) = \gamma \cdot (u(t) - \alpha(t))\,. \end{cases} \tag{6}$$

Таким образом, система СДУ (2) с начальными условиями $z(0) = z_0$, $\alpha(0) = \alpha_0$ и условиями (6) представляет стохастическую модель защиты ИС от УИБ, построенную на основе модели Ланчестера, в которой

сохраняются с вероятностью 1 инварианты (3). Меняя значения параметров a, β, γ, можно исследовать поведение модели при заданном значении z_0.

2. Модель информационной системы, защищенной от распространения компьютерных вирусов с вероятностью 1

Для моделирования эпидемии компьютерных вирусов воспользуемся классической биологической моделью распространения вирусов, предложенной в 1927 г. У. Кермаком и А. МакКендриком (SIR-модель) [7].

Для данной модели предполагается, что общая численность объектов информационной системы – величина постоянная:

$$S(t) + I(t) + R(t) = 1, \tag{7}$$

где $S(t)$ – доля уязвимых объектов; $I(t)$ – доля зараженных объектов; $R(t)$ – доля невосприимчивых или вылеченных объектов, обладающих иммунитетом.

SIR – модель распространения инфекции имеет вид:

$$\begin{cases} dS(t) = -\eta I(t)S(t)dt, \\ dI(t) = (\eta I(t)S(t) - \delta I(t))\,dt, \\ dR(t) = \delta I(t)dt, \end{cases} \tag{8}$$

с ограничением (7), где η – параметр интенсивности заражения, δ – параметр интенсивности выздоровления (скорость «иммунизации»).

Адаптируем модель (8) для описания информационной системы, которая была бы способна оставаться с вероятностью 1 в защищенном состоянии, не допускающем распространения эпидемии компьютерных вирусов. Аналогично модели, описанной выше, внесем в модель (8) случайность, определяющую внешние воздействия, на основе винеровского процесса, и PCP1 – функции управления $\hat{y}(t) := \hat{y}\,(t, S(t), I(t), R(t))$, обеспечивающие сохранение инвариантов модели. Так как в ИС имеем определенное количество объектов, то условие, порождаемое (7), является естественным ограничением при построении модели:

$$h\,(t, S(t), I(t), R(t)) = S(t) + I(t) + R(t) - 1. \tag{9}$$

В качестве функции, определяющей второй инвариант, используем

$$v\,(t, S(t), I(t), R(t)) = S(t), \tag{10}$$

которая позволит сохранить неизменность доли уязвимых объектов и не допустить их заражения, тем самым ликвидировать распространение эпидемии.

В начальный момент времени t_0 функции (9) и (10) принимают постоянные значения:

$$h\left(t_0, S\left(t_0\right), I\left(t_0\right), R\left(t_0\right)\right) = S(t) + I(t) + R(t) - 1 = C_1 = 0$$

$$v\left(t, S(t), I(t), R(t)\right) = S(t) = C_2,$$

которые должны сохраняться с вероятностью 1.

Пусть в ИС обнаружены вирусы в момент времени $t = 0$, в системе S_0 % уязвимых объектов, при этом система остается в работоспособном состоянии. Тогда модель ИС, защищенной от распространения компьютерных вирусов с вероятностью 1, будет иметь вид:

$$\begin{cases} dS(t) = (-\eta I(t)S(t) + \hat{y}_1(t))dt + r_1(t)dw(t), \\ dI(t) = (\delta I(t)S(t) - \eta I(t) + \hat{y}_2(t))\,dt + r_2(t)dw(t), \\ dR(t) = (\delta I(t) + \hat{y}_3(t))dt + r_3(t)dw(t), \\ S(0) = S_0. \end{cases} \tag{11}$$

Выбрав, например, в качестве дополнительной функцию

$$g\left(t, S(t), I(t), R(t)\right) = I(t) + e^t,$$

и, в качестве произвольной $q_{00}(t, S(t), I(t), R(t)) = S(t)$, построим систему СДУ с инвариантами (9) и (10):

$$\begin{cases} dS(t) = 0, \\ dI(t) = -e^t dt + S(t)dw(t), \\ dR(t) = e^t dt - S(t)dw(t). \end{cases} \tag{12}$$

с управлением РСР1 $\hat{y}(t) = \hat{y}(t, S(t), I(t), R(t))$ вида:

$$\begin{cases} \hat{y}_2(t) = -e^t - \eta I(t)S(t) + \delta I(t), \\ \hat{y}_3(t) = e^t - \delta I(t). \end{cases} \tag{13}$$

Поведение построенной модели (11) с управлением (13) можно исследовать при различных начальных условиях. Например, когда в начальный момент времени доли уязвимых объектов и зараженных одинакова, либо доля уязвимых значительно превышает долю зараженных, а также, когда доля зараженных значительно превышает долю уязвимых.

Заключение

Предложенные стохастические модели учитывают случайность внешних воздействий на информационную систему, которая может оказывать существенное значение на ее функционирование. Внесение управления РСР1 в построенные модели позволяет с вероятностью 1 противостоять этому внешнену воздействию. Модель защищенной от угроз ИБ

информационной системы, построенная на основе модели Ланчестера, учитывает текущее состояние защищенности ИС и способна мгновенно реагировать на изменение этого состояния. Предложенная модель защищенной ИС, построенная на основе модели SIR, позволит не допустить распространения эпидемии компьютерных вирусов при условии сохранения постоянных значений функций, взятых в качестве инвариантов. Анализ полученных моделей показал возможность их применения для защиты информационной безопасности.

СПИСОК ЛИТЕРАТУРЫ

1. *Дубко В. А.* Первый интеграл системы стохастических дифференциальных уравнений. Киев: Институт математики АН УССР, 1978. 21 с.
2. *Karachanskaya E. V.* Programmed control with probability 1 for stochastic dynamical systems // J. of Math. Sci. 2020. Vol. 248, № 1. P. 67–79.
3. *Карачанская Е. В.* Стохастизация классических моделей с динамическими инвариантами // Матем. заметки СВФУ. 2020. Т. 27, № 1. С. 69–87.
4. *Карачанская Е. В., Рыбкина О. В.* Построение математической модели защиты информационной системы // Информационные системы и технологии. 2021. № 6 (128). С. 113–121.
5. *Рыбкина О. В.* Построение модели угроз безопасности информации на основе математической модели Ланчестера // Научно-техническое и экономическое сотрудничество стран АТР в XXI веке. 2021. Т. 1. Хабаровск: Изд-во ДВГУПС, 2021. С. 255–260.
6. *Lanchester F.* Aircraft in warfare: The dawn of the fourth arm. London: Constable and Co., 1916. 243 p.
7. *Kermack W. O., McKendrick A. G.* Contributions to the Mathematical Theory of Epidemics // Proc. Royal Society. 1927. A 115. P. 700–721.

Рыбкина Олеся Викторовна — ст.преподаватель, кафедра ИТиС. E-mail: *ribkina_ol@mail.ru*

Карачанская Елена Викторовна — д.ф.–м.н., доцент, профессор, кафедра ИТиС. E-mail: *elena_chal@mail.ru*

ИСПОЛЬЗОВАНИЕ СВОЙСТВА ФРАКТАЛЬНОСТИ СЕТЕВОГО ТРАФИКА ДЛЯ ВЫЯВЛЕНИЯ АНОМАЛИЙ

Н. И. Соседова[1], Е. В. Карачанская[2]

[1] *Тихоокеанский государственный университет, г.Хабаровск, Россия*
[2] *Дальневосточный государственный университет путей сообщения, г.Хабаровск, Россия*

Целью данной работы является разработка статистического метода выявления аномалий сетевого трафика, основанного в первую очередь на его фрактальной природе. Решение о нормальности или аномальности трафика принимается исходя из попадания рассчитываемых параметров, в том числе и параметров, показывающих степень самоподобия, в необходимый диапазон значений. На основе данного метода разработано программное обеспечение для выявления аномалий на уровне устройства. Разработанное приложение прошло тестирование и успешно выявляет как краткосрочные, так и продолжительные изменения в поведении трафика. **Ключевые слова:** *фрактальность, показатель Хёрста, RS/анализ, аномалии в сетевом трафике*

Введение

Долгое время сетевой трафик рассматривали с помощью традиционной модели Пуассоновского процесса, основными свойствами которой являлись: случайность и независимость событий, постоянная интенсивность, экспоненциальные интервалы между событиями и отсутствие долгосрочной зависимости. Данная модель являлась достаточно простой, но не позволяла отразить реальное поведение трафика Ethernet-сетях.

В 1994 году Leland, Taqqu, Willinger и Wilson опубликовали статью «On the self-similar nature of Ethernet traffic», в которой показали, что сетевой трафик, в частности, Ethernet-трафик, обладает свойствами, характерными для фракталов: самоподобием и долгосрочной зависимостью. Исследователи проанализировали Ethernet-трафик в различных временных масштабах в реальных условиях в кампусной сети Bellcore, и зафиксировали следующие выводы:

1) Независимо от масштаба наблюдения (миллисекунды, секунды, минуты) структура колебаний трафика сохраняется.

2) Автокорреляция трафика не убывает быстро, как у пуассоновского процесса, а спадает медленно.

В ходе исследования данной темы другими учеными фрактальность была обнаружена не только в локальных сетях (LAN), но и в магистральных каналах передачи данных (WAN), трафике Интернета, в том числе в поведении пользователей веб-браузеров, сессиях загрузки файлов, потоковом видео и других формах цифровой активности. Открытие фрактальности сетевого трафика получило широкое применение в проектировании сетей, оценке качества обслуживании и многих других направлениях.

В данном исследовании фрактальность сетевого трафика используется для оценки нормальности и аномальности трафика с точки зрения вопросов информационной безопасности.

1. Фрактальная природа сетевого трафика

Трафик в современных компьютерных сетях обладает самоподобием и долгосрочной зависимостью. Эти свойства предполагают, что поведение трафика на больших временных интервалах сохраняет ту же статистическую структуру, что и на малых, и описываются с помощью формального понятия самоподобия по распределению.

Пусть $X(t)$ – стационарный стохастический процесс, моделирующий объем трафика за время t. Тогда процесс называется самоподобным с показателем Хёрста $H \in (0;1)$, если для любого $a > 0$ выполняется:

$$X(a,t) \sim a^H X(t), \tag{1}$$

где $\sim$ обозначает равенство по распределению.

Параметр Хёрста (H) - один из количественных способов оценки фрактальности. В зависимости от значения параметр Хёрста определяет классификацию случайного процесса:

1) Белый шум или некий независимый, случайный процесс.
2) Розовый шум или антиперсистентность (то есть ряд меняет направление чаще, чем ряд случайных независимых величин).
3) Черный шум или персистентность (то есть временной ряд характеризуется эффектом долговременной памяти и имеет склонность следовать трендам).

Для Ethernet-трафика этот показатель, как правило, находится в диапазоне 0,7–0,9, что указывает на выраженную фрактальную структуру.

Использование фрактальных свойств сетевого трафика стало важным направлением в задачах обнаружения сетевых аномалий и атак,

таких как DDoS, сканирование сети, утечка данных и т.д. Поскольку нормальный сетевой трафик обладает устойчивыми фрактальными характеристиками, они могут быть использованы как эталон поведения, отклонения от которого указывают на возможную аномалию. Так, например, в самом простом случае резкое снижение или повышение H может свидетельствовать о DDoS-атаке или передаче больших данных.

В данном исследовании фрактальность сетевого трафика используется для обнаружения аномалий на основе скользящего окна.

2. Расчет показателя Хёрста

В качестве показателя нормальности сетевого трафика, который бы мог в полной мере мере охарактеризовать его общее состояние, как уже упоминалось ранее, был выбран показатель Хёрста, расчёт которого производился методом нормированного размаха.

Расчёт параметра Хёрста можно разбить на следующие этапы.

Для исходного ряда S_t построим ряд приращений:

$$N_t = S_t - S_{t-1}. \tag{2}$$

Разделим полученный ряд приращений на A смежных периодов длинной n и отметим каждый период как I_a, где $a = 1, 2, ..., A$. Определим для каждого I_a среднее значение:

$$\overline{x}(n) = \frac{1}{n} \sum N_{i,a}. \tag{3}$$

Временной ряд накопленных отклонений от среднего значения $x^{cum}(t, n)$ для каждого I_a определим как:

$$x^{cum}(n, t) = \sum x(i) - \overline{x}(n). \tag{4}$$

Для каждого I_a определим размах кумулятивного ряда:

$$R_{Ia} = \max(x_k, a) - \min x_k, a. \tag{5}$$

Рассчитаем выборочное стандартное отклонения для каждого I_a:

$$S(n) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x(t) - \overline{x}(n))^2}. \tag{6}$$

Нормированный размах в течение каждого I_a равен $\frac{r(n)}{S(n)}$ Среднее значение нормированного размаха для подпериодов длины n определяется

как:

$$\frac{R}{S}(n) = \frac{1}{A}\sum_{i=1}^{A}\frac{R_a(n)}{S_a(n)}. \quad (7)$$

Значение вычисляется как тангенс угла наклона прямой аппроксимирующей зависимость $\log\frac{R(n)}{S(n)}$ от $\log n$.

Для построения аппроксимирующей зависимости линейной функции и нахождения тангенса угла наклона прямой используется метод наименьших квадратов.

После получения аппроксимированного параметра Хёрста для генеральной совокупности, выразим константу a из основной формулы RS-анализа:

$$a = e^{\frac{ln\frac{R}{S}}{H} - lnN}. \quad (8)$$

В дальнейшем расчет параметра Хёрста производится по формуле RS-анализа с использованием параметра a, рассчитанного для генеральной совокупности по формуле (8).

3. Краткое описание метода выявления аномалий

Метод состоит из трех основных этапов:

1) Выбор оптимальных параметров для расчета эталонных значений.
2) Расчет эталонных значений, с которыми будет производится сравнение сетевого трафика при выявлении аномалий.
3) Обнаружение аномалий сетевого трафика в режиме времени близком к реальному.

Поведение каждой системы индивидуально и поэтому подбор параметров является одним из самых важных шагов предлагаемого метода. В качестве параметров, подбор которых необходимо произвести выступают:

- количество профилей наблюдения;
- общее время наблюдения T;
- шаг наблюдения t;
- размер и количество окон наблюдения.

Опираясь на теоретические основы математической статистики, а также на серию проведенных экспериментов было выявлено, что нецелесообразно использовать интервалы и шаги таких размеров, при которых образуется менее 10 элементов ряда. Кроме того, при выборе параметров должна быть учтена интенсивность нормального сетевого трафика.

После выбора параметров наблюдения начинается этап обучения. На этапе обучения осуществляются следующие действия:

1) Разбиение на профили, для каждого из которых будут выполнены последующие шаги.
2) Сбор и анализ сетевого трафика, в ходе которого производится расчет числа флагов SYN, ACK в единицу времени, равную шагу, выбранному на этапе подбора параметров.
3) Группировка данных по окнам наблюдения и формирование выборок.
4) Для каждой полученной выборки осуществляется расчет параметра Хёрста методом RS-анализа.
5) Для каждой полученной выборки осуществляется расчет параметра Хёрста методом RS-анализа.
6) Расчет «доверительного интервала», попадание в который будет проверяться на этапе обнаружения аномалий.

В ходе выявления аномалий сетевого трафика производятся расчеты, аналогичные расчетам на этапе обучения и проверка вхождения рассчитанных значений в соответствующие доверительные интервалы.

Разработанный метод анализа аномалий сетевого трафика можно отнести к группе статистических методов, поскольку он использует оценку ряда статистических параметров, основан на полуконтролируемой технике выявления аномалий.

В рамках исследования было реализовано программное средство на языке C#, обеспечивающее применение предложенного метода на практике. Проведённые эксперименты подтвердили его эффективность и универсальность: система успешно выявляла аномалии различной продолжительности и структуры, что доказывает целесообразность использования фрактальных характеристик трафика в задачах информационной безопасности.

Заключение

В условиях постоянного появления новых и ранее неизвестных атак задача выявления аномалий в сетевом трафике остаётся актуальной. Несмотря на наличие множества существующих подходов, большинство из них ориентированы на сигнатурное или поведенческое распознавание уже известных угроз, что ограничивает их эффективность при обнаружении новых, нестандартных видов воздействия. В связи с этим особую ценность представляют методы, опирающиеся на статистические свойства трафика и не зависящие от конкретных сигнатур.

Одним из ключевых аспектов данной работы является учет фрактальной природы сетевого трафика. Фрактальность, выражающаяся в самоподобии и долгосрочной зависимости трафика, обеспечивает ста-

бильную основу для построения эталонного профиля нормального поведения. Отклонения от этого профиля становятся индикаторами потенциальных аномалий, вне зависимости от их специфики.

Разработанный метод обладает высокой адаптивностью и простотой реализации: путем настройки параметров наблюдения можно подстроить его под различные типы сетевых условий и ожидаемой активности. Это делает возможным обнаружение широкого спектра аномалий — от кратковременных, таких как сканирование портов, до продолжительных атак, например, распределённого отказа в обслуживании (DDoS).

СПИСОК ЛИТЕРАТУРЫ

1. *Leland, W. E., Taqqu, M. S., Willinger, W., Wilson, D. V.* On the self-similar nature of Ethernet traffic // ACM Transactions on Networking. 1994. Т. 2, № 1. С. 1–15.
2. *Соседова Н. И.* Выявление аномалий сетевого трафика путем анализа содержимого заголовков пакетов // Открытые эволюционирующие системы: вектор развития, Хабаровск: Дальневосточный государственный университет путей сообщения, 2024. С. 112–115.
3. `https://www.mql5.com/ru/articles/2930` – Вычисление коэффициента Хёрста. 2025.

Соседова Надежда Ивановна — аспирант. E-mail: *sosedowa.nadezhda@yandex.ru*

Карачанская Елена Викторовна — д.ф.-м.н., доцент, профессор каф. Информационные системы и технологии. E-mail: *elena_ chal@mail.ru*

ОБ ОЦЕНИВАНИИ ИНТЕРВАЛА ОХВАТА НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПО ВЫБОРОЧНЫМ ДАННЫМ

А. В. Степанов[1]

[1] *Всероссийский научно-исследовательский институт метрологии им. Д.И. Менделеева (ВНИИМ), г. Санкт-Петербург, Россия*

Рассматривается вопрос об оценивании интервала охвата непрерывной случайной величины по выборочным данным, как в общем случае, при отсутствии информации о законе ее распределения, так и в частных случаях, по коротким выборкам, для отдельных семейств распределений (при этом оценивание производится на основе выборочного размаха или среднеквадратического отклонения, с использованием метода Монте Карло).
Ключевые слова: *случайная величина, вероятность охвата, интервал вероятности, выборочный размах, метод Монте-Карло*

Введение

Оценивание интервалов охвата непрерывных случайных величин является важной задачей при статистической обработке данных, например, в метрологических приложениях, где интервал охвата представляет собой диапазон значений, в котором с заданной вероятностью находится истинное значение измеряемой величины [1, 2]. Дадим формальное определение: интервал (c_1, c_2) (в случае, когда носитель распределения ограничен, — отрезок $[c_1, c_2]$) будем называть интервалом охвата для заданного уровня вероятности P_0, если $F(c_2) - F(c_1) = P_0$, где F — функция распределения рассматриваемой случайной величины. В данной работе интервал охвата будет рассматриваться именно в таком «простом» смысле, по сути, совпадая с интервалом вероятности [3]; тонкости метрологической трактовки [4] не являются предметом рассмотрения. Очевидно, что интервал вероятности в такой формулировке определяется не единственным образом, поэтому, как правило, оговаривают дополнительные условия, считая его односторонним, или двусторонним вероятностно-симметричным (полученным отбрасыванием «хвостов» распределения равного веса), или кратчайшим из возможных. Тем не менее, ниже будут рассматриваться произвольные интервалы охвата (вероятности). Рассмотрим вопрос об их оценивании. Сначала рассмотрим общий случай, не конкретизируя вид закона распределения.

1. Об оценивании интервала охвата в общем случае

Предположим, что имеется распределение одномерной непрерывной случайной величины X, обозначим $F(x)$ соответствующую функцию распределения. Пусть имеется выборка $\{x_i\}_1^n$ значений из этого распределения, обозначим $x_{(1)} \leq \ldots \leq x_{(n)}$ соответствующий вариационный ряд, и $x_{min} = x_{(1)}$, $x_{max} = x_{(n)}$. Назовем отрезок $[x_{min}, x_{max}]$ опорным отрезком (интервалом) выборки.

Далее, пользуясь тем фактом, что значения $u_i = F(x_i)$ распределены равномерно на отрезке $[0, 1]$, можем сформулировать утверждение:

Утверждение 1. Для того, чтобы отрезок $[x_{min}, x_{max}]$ содержал некоторый ограниченный интервал охвата для заданного уровня вероятности P_0 с вероятностью, не меньшей P_1, значение n должно удовлетворять неравенству:

$$nP_0^{n-1} - (n-1)P_0^n \leq 1 - P_1. \tag{1}$$

Доказательство. Требуется доказать, что при выполнении неравенства (1) для случайных величин $X_{(i)}$ (порядковых статистик) выполнено неравенство:

$$P\left\{F\left(X_{(n)}\right) - F\left(X_{(1)}\right) \geq P_0\right\} = 1 - P\left\{F\left(X_{(n)}\right) - F\left(X_{(1)}\right) \leq P_0\right\} \geq P_1,$$

или $P\left\{U_{(n)} - U_{(1)} \leq P_0\right\} \leq 1 - P_1$, где $U_{(i)} = F\left(X_{(i)}\right)$. Величина $R_U = U_{(n)} - U_{(1)}$ представляет собой размах выборки из равномерного распределения $U([0, 1])$. Согласно [5], размах R_U имеет плотность распределения $f_R(r) = n(n-1)r^{n-2}(1-r)$ и функцию распределения $F_R(r) = \int_0^r f_R(r)dr = nr^{n-1} - (n-1)r^n$, следовательно, $P\left\{U_{(n)} - U_{(1)} \leq P_0\right\} = F_R(P_0) \leq 1 - P_1$, откуда и следует неравенство (1). ■

Замечание 1. Отметим также, что, так как математическое ожидание k-й порядковой статистики $EU_{(k)} = k(n+1)^{-1}$, то математичесое ожидание вероятности охвата, вычисленное для отрезка $[x_{min}, x_{max}]$, будет равно $E\left(U_{(n)} - U_{(1)}\right) = \frac{n-1}{n+1}$. Например, для $n = 39$ математическое ожидание вероятности охвата, вычисленной для данного отрезка, будет составлять 0.95.

Указанные тривиальные факты могут оказаться достаточно полезными при планировании экспериментов, так как позволяют сформулировать требования к минимальной длине выборки.

Ниже в табл. 1 приведены минимальные значения n, удовлетворяющие неравенству (1) для некоторых уровней вероятности P_0, P_1.

Таблица 1

Значения n, удовлетворяющие неравенству (1) при различных $P_{0,1}$.

P_0	P_1		
	0,95	0,99	0,999
0,90	46	64	89
0,95	93	130	181
0,99	473	662	920

Замечание 2. Используя понятие опорного отрезка, можно предложить простой способ проверки однородности выборок, также свободный от предположений о конкретном виде распределения. Допустим, имеется две выборки, первая из которых представляет собой достаточно длинный массив $\{x_i\}_1^n$ накопленных данных, а вторая — относительно короткую выборку свежих данных $\{y_j\}_1^m$ длины $m \ll n$, и требуется определить, могут ли данные быть объединены. Интуитивно понятно, что в случае, когда обе выборки взяты из одного распределения, то вероятность того, что значения y_j принадлежат отрезку $[x_{min}, x_{max}]$, достаточно велика. В противном случае (значительное число y_j находится вне опорного отрезка), гипотезу об однородности следует отклонить. Можно показать, что вероятность $P_{k\,out}$ того, что k из m значений второй выборки лежит вне отрезка $[x_{min}, x_{max}]$, выражается формулой

$$P_{k\,out} = \sum_{j=k}^{m} \binom{m}{j} \frac{n(n-1)(j+1)!\,(m+n-j-2)!}{(m+n)!} \approx$$

$$\approx \sum_{j=k}^{m} \binom{m}{j} \left(\frac{2}{n+1}\right)^j \left(1 - \frac{2}{n+1}\right)^{m-j} \qquad (\text{при } m, n \gg 1).$$

Например, при $n = 100$, $m = 10$, $k = 3$ значение $P_{k\,out}$ исчезающе мало (порядка сотых долей процента), поэтому, если три или более значений y_j из десяти лежат вне опорного отрезка первой выборки, то гипотеза об однородности отвергается с высокой степенью уверенности. В случае $k = 2$ и тех же m, n значение $P_{k\,out}$ составляет около 2% (т. е. также весьма невелико). Следует подчеркнуть, что данный критерий, очевидно, работает только на отклонение гипотезы об однородности.

В тривиальном случае, когда вместо второй выборки рассматриваем единичное значение x_{n+1}, $P\left\{X_{n+1} \notin [X_{(1)}, X_{(n)}]\right\} = \frac{2}{n+1}$ (что также следует из свойств порядковых статистик равномерного распределения).

Можно также показать, что

$$P\{[Y_{min}, Y_{max}] \not\subset [X_{min}, X_{max}]\} = \frac{2m}{m+n} - \frac{m^2}{(m+n)^2} \approx \frac{2m}{n}$$

(это следует из того, что $P\{Y_{min} < X_{min}\} = P\{Y_{max} > X_{max}\} = \frac{m}{m+n}$).

Полный вывод выражения для $P_{k,\,out}$ опустим (он достаточно громоздкий), отметив только отдельные его моменты. Снова перейдем к равномерному распределению, обозначив $u = U_{(1)}$, $v = U_{(n)}$, тогда

$$P\{Y_j \notin [X_{\min}, X_{\max}]\} = P\{F(Y_j) \notin [u, v]\} = u + (1 - v).$$

Число k элементов Y_j вне $[u, v]$ подчиняется биномиальному закону:

$$P\{k \text{ значений } Y_j \text{ лежит вне } [u, v]\} = \binom{m}{k}(u + 1 - v)^k (v - u)^{m-k}.$$

Совместная плотность u, v имеет вид:

$$f_{U_{(1)},U_{(n)}}(u, v) = n(n-1)(v-u)^{n-2}, \quad 0 < u < v < 1 \quad \text{и}$$

$$P_{k\,out} = \int_0^1 \int_u^1 \sum_{j=k}^{m} \binom{m}{j}(u + 1 - v)^j (v - u)^{m-j}\, n(n-1)(v-u)^{n-2}\, dv\, du,$$

откуда, меняя порядок суммирования и интегрирования и выражая интегралы через бета-функцию, после упрощения приходим к искомой формуле.

2. Оценивание интервалов охвата по коротким выборкам методом Монте-Карло

Заметим, что значения n, приведенные в табл. 1, достаточно велики. В качестве альтернативы оценке (1), можно также рассмотреть следующие методы оценивания интервалов охвата:

1) напрямую через выборочные квантили. В случае вероятностно-симметричного интервала охвата такая оценка будет иметь вид: $\left[x_{(k)},\, x_{(n-k+1)}\right]$, $k = \lfloor n\,(1 - P_0)/2 \rfloor + 1$;
2) в случае, когда имеется априорная информация о законе распределения, можно оценить его параметры, а далее выразить границы интервала охвата, обращая функцию распределения аналитически или численно.

Однако оба они (особенно первый), как правило, требуют еще более значительного объема выборочных данных. Оценки параметров распределения во втором случае при этом могут быть смещенными.

В то же время, зачастую возникает задача оценивания интервала охвата по коротким (порядка десятка значений) выборкам. В работе [6] был предложен способ оценки интервала охвата симметричного распределения с использованием выборочного размаха или выборочной дисперсии. Рассмотрим отрезки вида

$$I_R = [\bar{x} - C_R R,\ \bar{x} + C_R R], \quad I_S = [\bar{x} - C_S S,\ \bar{x} + C_S S],$$

где $\bar{x}$, R и S – это выборочные среднее, размах и среднеквадратическое отклонение, соответственно (в случае, если математическое ожидание μ известно, можно использовать его вместо $\bar{x}$). Коэффициенты C_R, C_S подбираются методом Монте-Карло, исходя из условия $P\{F(c_2) - F(c_1) \geq P_0\} = P_1$, где $c_{1,2}$ – верхняя и нижняя границы $I_{R,S}$. Последнее условие означает, что отрезок I_R (или I_S) содержит с вероятностью P_1 некоторый интервал охвата для заданного уровня вероятности P_0. Заметим, что применение данного метода требует априорной информации о законе распределения случайной величины (значения $C_{R,S}$ будут, очевидно, зависеть от него).

В работе [6] рассматривались случаи семейств обобщенных экспоненциальных распределений и двусторонних степенных распределений. Здесь, для разнообразия, рассмотрим в качестве примера случай распределения Стьюдента t_ν (ν – число степеней свободы), считая его математическое ожидание неизвестным. В табл. 2 приведены значения C_R, C_S для некоторых значений ν, длин выборки n и $P_0 = 0,95$. Длины отрезков I_R, I_S при этом, в среднем, сопоставимы.

Таблица 2

Значения C_R, C_S, для t_ν, $\nu = 5, 7, 20$; $P_0 = 0,95$; $P_1 = 0,95; 0,99$.

		$d = R$		$d = S$	
		P_1		P_1	
ν	n	0,95	0,99	0,95	0,99
5	7	1,67	2,33	4,62	6,42
	10	1,26	1,64	3,89	4,99
7	7	1,60	2,22	4,42	6,10
	10	1,21	1,56	3,73	4,75
20	7	1,50	2,07	4,14	5,69
	10	1,14	1,45	3,50	4,42

Для распределения с тяжелыми хвостами в качестве оценки центральной точки может оказаться предпочтительнее использовать меди-

ану выборки. При этом также следует отказаться от построения оценок на основе выборочного размаха, предпочитая среднеквадратическое отклонение.

Заключение

Предложенные в работе критерии обеспечивают достаточно простые алгоритмы оценивания интервалов охвата. Метод Монте-Карло, рассмотренный во второй части, допускает различные обобщения. В частности, для асимметричных распределений можно осуществлять раздельный подбор коэффициентов для левой и правой частей распределения. В случае подозрения на наличие выбросов, в качестве выборочной статистики, характеризующей разброс данных, можно применять робастные выборочные статистики (например, медианное абсолютное отклонение, MAD). Также можно производить предварительное цензурирование выборки. Полученные результаты (минимальные длины выборок и вероятности из первой части статьи, значения коэффициентов $C_{R,S}$ для различных видов распределений из второй) могут быть табулированы для дальнейшего использования при решении прикладных задач.

СПИСОК ЛИТЕРАТУРЫ

1. *Stoudt S., Pintar A., Possolo A.* Coverage intervals // Journal of Research of the National Institute of Standards and Technology. 2021. Vol. 126, Art. num. 126004.
2. *Joint Committee for Guides in Metrology.* Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement-– Propagation of distributions using a Monte Carlo method. BIPM, 2008.
3. *Ллойд Э., Ледерман У.* Справочник по прикладной статистике. М.: Финансы и статистика, 1989. Т. 1. 510 с.
4. *Kacker R.* Probability distributions and coverage probability in GUM, JCGM documents, and statistical inference // Measurement. 2015. Vol. 65. P. 61–70.
5. *Дейвид Г.* Порядковые статистики. М.: Наука, 1979. 336 с.
6. *Степанов А. В.* Об оценивании интервала охвата для некоторых распределений по выборочному размаху // Материалы международной конференции Воронежская зимняя математическая школа, Воронеж: ВГУ, 2025. С. 315–316.

Степанов Александр Владимирович — к.ф.м-н., вед. н. с., ВНИИМ им. Д.И. Менделеева. E-mail: *stepanov17@yandex.ru*

АППРОКСИМАЦИЯ ОБЪЕМА СУТОЧНОГО ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ В ЖИЛЫХ ПОМЕЩЕНИЯХ

В. В. Чураков[1], Е. А. Фёдорова[1], Я. А. Тюленина[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Работа посвящена анализу объема суточного потребления электроэнергии в жилых помещениях. На основе почасовых измерений построен график суточной нагрузки. Выявлены типичные поведенческие стратегии потребителей, образующие пиковые значения нагрузки электросети в утренние и вечерние часы. Построено эмпирическое распределение вероятностей объема суточного потребления электроэнергии. Предложена аппроксимация экспоненциальным распределением. Проведена оценка точности аппроксимации с использованием метрик расстояния Колмогорова и критерия Пирсона. **Ключевые слова:** *электропотребление, суточная нагрузка, аппроксимация, экспоненциальное распределение*

Введение

В условиях экспоненциального роста современных технологий, требующих все большего энергопотребления, дефицит энергоресурсов является одной из главных проблем мирового уровня. Развитие искусственного интеллекта, облачных вычислений, электромобилей и других энергоемких технологий создает высокое давление на существующие энергетические системы. Увеличение вычислительных мощностей, необходимых для обработки больших данных и обучения нейронных сетей, требует огромных объемов электроэнергии, а переход к электрическому транспорту ставит новые вызовы перед инфраструктурой электроснабжения. В связи с этим точное прогнозирование нагрузки на электрические сети становится критически важным для обеспечения надежности, эффективности и устойчивости энергосистем. Поэтому разработка и внедрение точных моделей прогнозирования нагрузки представляют собой актуальную и значимую задачу, направленную на оптимизацию управления энергосистемами и обеспечение их стабильного функционирования.

Исследование выполнено за счет гранта Российского научного фонда № 24-71-00022, https://rscf.ru/project/24-71-00022/

Проблема оценки среднего суточного энергопотребления и влияния на него экономических и демографических факторов изучалась в работах [1, 2]. Одними из наиболее информативных показателей запроса на электроэнергию в конкретном регионе или системе являются графики суточной нагрузки [3, 4]. В данной работе рассматривается исследование суточного потребления электроэнергии в жилых помещениях для осветительно-бытовых нужд на основе собственных измерений и анализа данных о потреблении в открытом доступе.

1. График суточной нагрузки

На первом этапе исследования был реализован сбор и анализ данных об объеме потребления электрической энергии в течение суток в жилом помещении. Мониторинг потребления электроэнергии осуществлялся с помощью цифрового электросчетчика с функцией фиксации почасовых показаний. Измерения проводились в течение нескольких дней, данные были усреднены по объему потреблений в соответствующие часы. На основе данных был построен график суточной нагрузки (рис. 1).

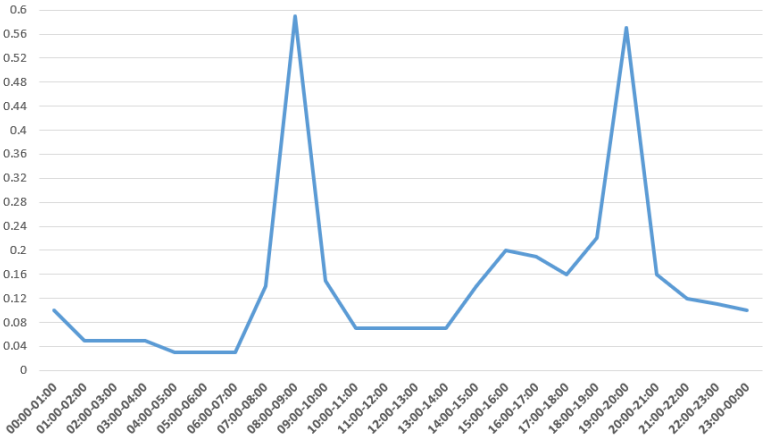


Рис. 1. График суточной нагрузки

Были выявлены следующие поведенческие особенности энергопотребления:

– пиковые значения потребления приходятся на утренние (07:00–09:00) и вечерние (18:00–21:00) часы;

– имеется минимальное ненулевое потребление, даже в ночное время (00:00–06:00);
– потребление электроэнергии в дневные часы имеет стохастический характер.

Для учета демографического фактора были применены следующие коэффициенты масштабирования базового уровня потребления [2, 5]:

– домохозяйство из 2 человек: $k_1 = 1,6$;
– домохозяйство из 3 человек: $k_2 = 2,0$;
– домохозяйство из 4 человек: $k_3 = 2,5$.

При этом демографический фактор оказывает слабое влияние на детерминированный объем минимального энергопотребления в ночное время, представляющий собой базовое потребление постоянно работающих электроприборов (например, холодильных установок) или устройств в режиме ожидания. Таким образом, можно построить график суточной нагрузки, демонстрирующий объемы потребления для домохозяйств разной численности (рис. 2).

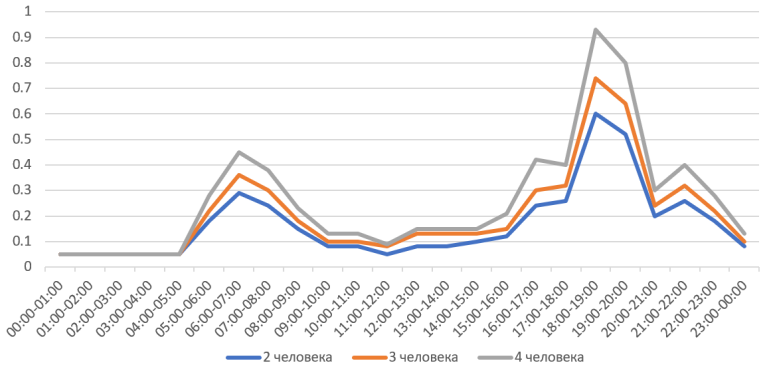


Рис. 2. График суточной нагрузки домохозяйств разной численности

Очевидно, что на график суточного энергопотребления также влияют время года, географические и политические факторы. Однако, построенные графики суточной нагрузки (рис. 1, 2) соответствуют известным типовым графикам [3, 4], а выявленные особенности закономерны для суточного потребления электроэнергии жилых зданий в любых условиях.

2. Анализ объема потребления

Внешний вид графиков суточной нагрузки напоминает реализацию случайного процесса, описывающего объем почасового энергопотребления в течение суток.

Как было замечено выше, электронагрузка складывается из детерминированной (постоянно включенные электроприборы) и стохастической составляющих (активное потребление). Будем исследовать случайную величину X, описывающую почасовой объем энергопотребления за вычетом минимального суточного значения.

Проведем анализ закона распределения вероятностей объема часового электропотребления по имеющимся данным. В качестве шага дискретизации было выбрано значение $\Delta = 0,1$ КВт·ч. На рис. 3 представлено эмпирическое распределение вероятностей $P(x)$ объема часового потребления.

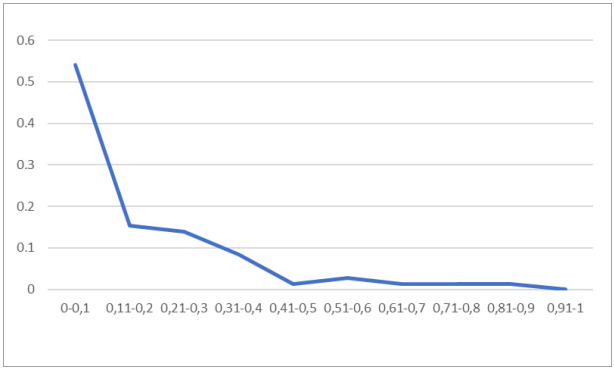


Рис. 3. Эмпирическое распределение вероятностей активного потребления

2.1. Аппроксимация

В работе предлагается аппроксимировать распределение вероятностей объема почасового потребления электроэнергии в течение суток экспоненциальным распределением с функцией плотности вероятности:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \tag{1}$$

где $\lambda = \dfrac{1}{\mathbb{E}[X]}$ – величина обратная среднему объему потребления в час. Эмпирическое математическое ожидание величины почасового потреб-

ления оценивается следующим образом:

$$\mathbb{E}[X] = \sum_i P(i) \cdot \frac{x_i + x_{i+1}}{2}.$$

На основе эмпирических данных получено значение: $\lambda = 5{,}54$. Учитывая условие нормировки, построим график распределения вероятностей и сравним аппроксимацию с эмпирическим распределением (рис. 4).

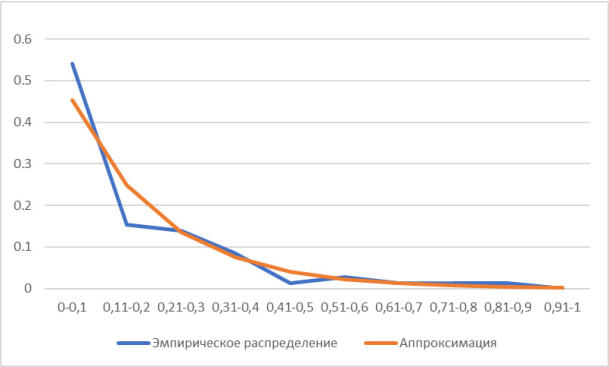


Рис. 4. Сравнение эмпирического распределения и аппроксимации

На рис. 5 представлено сравнение функций распределения объема почасового потребления.

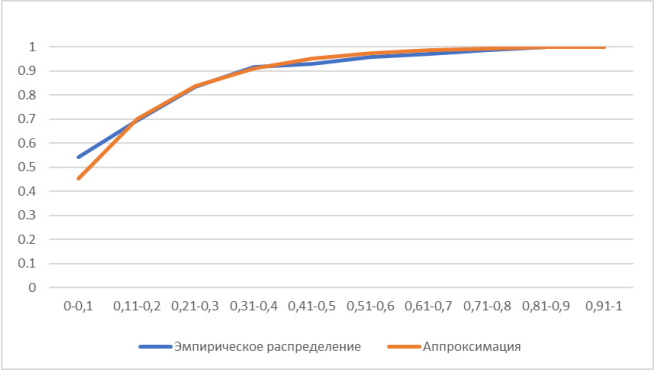


Рис. 5. Сравнение функций распределения вероятности

2.2. Оценка точности аппроксимации

Для оценки точности предложенной аппроксимации были использованы следующие метрики: критерий χ^2 и расстояние Колмогорова.

Расстояние Колмогорова вычисляется по формуле:

$$d = \max_i |F_n(x_i) - F(x_i)|, \tag{2}$$

где:

- $F_n(x_i)$ — эмпирическая функция распределения вероятности,
- $F(x_i)$ — теоретическая функция распределения вероятности.

Для предлагаемой аппроксимации расстояние Колмогорова составило $d = 0{,}0894$. Как видно на Рисунке 5 наибольшее отклонение между распределениями при низких объемах потребления (0–0,1 кВт·ч). При этом на остальной области значений погрешность аппроксимации не превышает $\varepsilon_{\max} = 0{,}022$.

Проверим гипотезу об экспоненциальном виде распределения, используя критерий Пирсона. Вычисленное значение $\chi^2 = 2,688$. Критическая точка для уровня значимости $\alpha = 0,05$ при количестве степеней свободы $k = 8$ равна 15,51. Таким образом, выдвинутую гипотезу можно принять.

Заключение

В результате исследования был проведен анализ распределения вероятностей объема почасового энергопотребления. Предложена аппроксимация экспоненциальным распределением, выполнено ее сравнение с эмпирической функцией распределения и проведена оценка точности. Установлено, что экспоненциальная аппроксимация демонстрирует достаточно высокую точность при объемах потребления $>$ 0,2 кВт·ч ($d < 3\%$.), тогда как оценка вероятности минимальных значений имеет меньшую точность ($d \approx 9\%$). Полученные решения могут быть использованы в системах энергомониторинга и прогнозирования нагрузок. В будущих исследованиях планируется аппробация предложенной аппроксимации на различных данных, в том числе и для прогнозирования объемов потребления нежилых зданий (производственных компаний, городских служб и т.д.).

СПИСОК ЛИТЕРАТУРЫ

1. *Liu Y., Wang M., Li W., Chen S., Chen L., Li X.* Modeling Residential Electricity Consumption from Public Demographic Data for Sustainable Cities // Energies. 2022. Т. 15, № 6. С. 2163.

2. *Wang X., Yang Y., Liu X., Lin B.* Nonlinear relationship between household composition and electricity consumption: optimal threshold models // Energy Economics. 2022. Т. 114. С. 106307.
3. *Демиденко А. С., Шведов Г. В.* Суточные графики электрических нагрузок многоквартирных домов в системах электроснабжения городов // Энергосбережение теория и практика: труды Десятой Международной школы-семинара молодых ученых и специалистов, Москва, 19–23 октября 2020 года. Курск: Университетская книга, 2020. С. 123–127.
4. *Хомичев В. А., Шведов Г. В.* Суточные графики электрической нагрузки отдельной квартиры многоквартирного дома в системах электроснабжения мегаполисов // Энергосбережение – теория и практика: Труды Одиннадцатой Всероссийской конференции с Международным участием, Москва, 10–14 октября 2022 года. М.: ООО ЦПУ РАДУГА, 2022. С. 42–48.
5. Методические рекомендации по определению потребности в электрической мощности жилых зданий // Министерство строительства и жилищно-коммунального хозяйства Российской Федерации. М., 2019. 48 с.

Чураков Валерий Вячеславович — студент Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *churakovvaler@yandex.ru*

Фёдорова Екатерина Александровна — кандидат физ.-мат.наук, доцент кафедры Теории вероятностей и математической статистики Томского государственного университета. E-mail: *moiskate@mail.ru*

Тюленина Яна Андреевна — аспирант Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *yanccch@gmail.ru*

A METHOD FOR EXPERIMENTAL CORRELATION DISTRIBUTION ESTIMATION FOR CORRELATION-BASED DIGITAL IMAGE WATERMARKS

V. Anzhin[1]

[1] *Tomsk State University, Tomsk, Russian Federation*

This paper proposes a method for experimentally estimating the distribution of correlation in correlation-based digital image watermarking. The proposed method evaluates how distortions affect the correlations between image and embedded and non-embedded watermarks. The method involves a computational experiment using an image dataset, where correlations are measured before and after filtering for embedded and non-embedded watermarks. Applying the proposed method to a DCT based watermarking algorithm demonstrates that filtering significantly reduces the correlation with the embedded watermark, while having minimal changes for non-embedded. **Keywords:** *watermark, robustness, correlation measure*

Introduction

The growth in digital multimedia production and consumption has increased the demand for technical solutions to protect copyrights and restrict unauthorized copying. Digital watermarks can be used for multimedia content protection. In the case of a content leak, digital watermarks allow to verify authorship and restrict the distribution of pirated copies.

In correlation-based watermarking [1], a pseudo-random signal is embedded into the image as watermark, and a correlation measure is used for watermark detection. The image is identified as watermarked when the correlation between it and the embedded watermark exceeds the selected detection threshold.

The image with embedded watermark can be modified by various distortions. These distortions can result from transmission losses and signal processing. They can also arise from attempts to remove the watermark. Watermark robustness [2] refers to the resilience of the watermark against various signal processing operations. Distortion of the image changes the value of correlation between the watermark and the distorted image. This can lead to the omission of embedded watermark (false-negative detection) or the false detection of non-embedded watermark (false-positive detection).

An important aspect of watermark robustness is the right detection threshold selection. Selecting an appropriate detection threshold requires an understanding of the distribution of correlation between the watermark and both watermarked and non-watermarked images. There are theoretical models which use statistical methods, typically developed for a specific watermarking method. For example, in [3], the correlation is estimated and detection threshold is selected through theoretical analysis for embedding method which uses discrete cosine transform (DCT). The authors of [3] also note that the selected threshold can lead to a significant number of false-negative detections under attack. Estimating the detection threshold theoretically is a complex task due to the wide variety of potential distortions that can affect the image. As an alternative, computational experiments can be used for modelling correlation distribution and detection threshold selection.

This paper proposes the method for experimentally estimating the distribution of correlation in correlation-based image watermarking algorithms. The method involves watermark embedding, applying filtering to introduce distortions, and calculating the correlation measure for both watermarked and non-watermarked images before and after filtering. The proposed method is applied to the watermarking algorithms described in [4].

The C++ implementation of the watermarking algorithm, along with the Python scripts of experiments and the obtained results, are available at [5].

1. DCT Based Watermarking

Within the DCT based watermarking algorithm [4] the original image is divided into blocks of 8x8 pixels. A DCT is applied to each block. The transformed block is then converted into a sequence of 64 elements using ZigZag scanning [6], which reorders the DCT coefficients from low to high frequencies. In each sequence, the most significant low-frequency DCT coefficients are discarded. The remaining DCT coefficients are concatenated into a single sequence c_o. The length of c_o is n. The watermark w_r is embedded into the sequence c_o by the equation (1). The watermark w_r elements are generated according to the normal distribution $N(0, 1)$. The modified blocks are transformed via inverse DCT to obtain the watermarked image.

$$c_w = c_o + \alpha w_r, \tag{1}$$

where $c_w, c_o, w_r \in R^n$, $n \in N$, $\alpha \in R$.

To detect the watermark w_r the normalized correlation (3) is calculated between w_r and the sequence of DCT coefficients c extracted from image.

$$\tilde{c}[i] = \frac{c[i]}{\sqrt{\sum_{j=1}^{n} c[j]^2}}, \qquad \tilde{w}_r[i] = \frac{w_r[i]}{\sqrt{\sum_{j=1}^{n} w_r[j]^2}}, \tag{2}$$

$$Z_{nc}(c, w_r) = \sum_{i=1}^{n} \tilde{c}[i]\tilde{w}_r[i], \tag{3}$$

where $c, w_r, \tilde{c}, \tilde{w}_r \in R^n$, $n \in N$.

2. Data collection method of evaluating the distribution of correlation

A computational experiment on a set of images is conducted to evaluate the distribution of correlation. Two random watermarks are generated for each image in the set. One watermark is embedded into the image. The correlation measures are calculated between the watermarked image and both the embedded and non-embedded watermarks. The watermarked image is processed with a filter. Correlations are calculated between the filtered image and both the embedded and non-embedded watermarks. The output of the algorithm is the arrays of correlation values. The method is described in Algorithm 1.

Algorithm 1. Data collection method

Require: Watermark embedding algorithm $W(c_o, w_r)$, correlation measure $Z(c, w_r)$, set of image filters F, images database I

Ensure: Z_e – array of correlations between the watermarked image and the embedded watermark, Z_{ne} – array of calculated between the image with embedded watermark and a different non-embedded watermark, Z_{ef} and Z_{nef} – arrays of correlations after filtering with $f \in F$

1: $j \leftarrow 0$
2: **for all** $i \in I$
3: Create random watermarks w and w' for image i, $w \neq w'$
4: $i_w = W(i, w)$
5: $Z_e[j] \leftarrow Z(i_w, w)$, $Z_{ne}[j] \leftarrow Z(i_w, w')$
6: **for all** $f \in F$
7: $i_{wf} = f(i_w)$
8: $Z_{ef}[j] \leftarrow Z(i_{wf}, w)$, $Z_{nef}[j] \leftarrow Z(i_{wf}, w')$
9: $j \leftarrow j + 1$

3. Testing results

The computational experiment, described in Algorithm 1 was conducted using the watermarking algorithm [4]. The Flickr8K dataset [7] with 8091 images was used. The following FFmpeg filters [8] were used to filter the images:

- Blurring using `unsharp`: `unsharp=3:3:-0.25:3:3:-0.25`;
- Sharping using `unsharp`: `unsharp=3:3:0.25:3:3:0.25`;
- DCT based denoising using `dctdnoiz`: `dctdnoiz=s=10`;
- Denoising using `nlmeans`: `nlmeans=s=10`.
- Pixelation using `pixelize`: `pixelize=w=3:h=3`;
- Denoising using `fftdnoiz`: `fftdnoiz=sigma=15`.

The Table 1 presents the mean and standard deviation values calculated for the array of correlation values generated by Algorithm 1. The first column indicates the α values used during watermark embedding. The second column specifies the filter applied to the images. Columns under the heading "With Embedding" contains the mean and standard deviation of correlation values between an image and an embedded watermark. The columns under "Without Embedding" show the results between an image and a non-embedded watermark.

The results of Algorithm 1 can be presented in a chart form. Fig. 1 shows the estimated distribution of correlation calculated using the Flickr8K dataset [7] with embedding performed at $\alpha = 10$. The chart on the left displays the correlation values without filtering, and the chart on the right shows the correlation values after filtering with dctdnoiz. The x-axis represents the correlation values, and the y-axis is the number of images with the corresponding correlation value. Orange color is used for cases where the detected watermark was not embedded into the image, and blue is used when the watermark was embedded. Dashed lines shows the mean values.

The results in Table 1 show that when a watermark is embedded into an image, the mean of correlation values is higher than the mean for a non-embedded watermark. Applying filters to the images generally decreases the mean of correlation values for the embedded watermark (with the exception of the sharpening filter), while producing minimal changes in the mean values for the non-embedded watermark. In the chart in Fig. 1, the orange distribution (correlation values for the non-embedded watermark) remains close to zero after filtering, whereas the blue distribution (correlation values for the embedded watermark) shifts toward lower correlations.

Table 1

Correlation estimation results

α	Filter	With Embedding		Wihtout Embedding	
		Mean	Std dev	Mean	Std dev
5	without filtering	9.84×10^{-7}	3.82×10^{-7}	2.25×10^{-11}	4.30×10^{-9}
	blur	9.32×10^{-7}	3.60×10^{-7}	2.73×10^{-11}	4.30×10^{-9}
	sharp	1.01×10^{-6}	3.93×10^{-7}	2.14×10^{-11}	4.31×10^{-9}
	dctdnoiz	1.14×10^{-7}	4.28×10^{-8}	3.64×10^{-11}	4.26×10^{-9}
	nlmeans	2.11×10^{-7}	8.83×10^{-8}	4.07×10^{-11}	4.31×10^{-9}
	pixelize	5.06×10^{-8}	1.62×10^{-8}	-4.22×10^{-12}	4.34×10^{-9}
	fftdnoiz	3.31×10^{-7}	1.09×10^{-7}	1.72×10^{-11}	4.25×10^{-9}
10	without filtering	1.55×10^{-6}	4.49×10^{-7}	-6.10×10^{-12}	4.35×10^{-9}
	blur	1.50×10^{-6}	4.40×10^{-7}	-1.08×10^{-11}	4.34×10^{-9}
	sharp	1.57×10^{-6}	4.53×10^{-7}	-7.70×10^{-12}	4.35×10^{-9}
	dctdnoiz	2.99×10^{-7}	9.59×10^{-8}	-2.35×10^{-13}	4.33×10^{-9}
	nlmeans	4.70×10^{-7}	1.66×10^{-7}	8.29×10^{-12}	4.30×10^{-9}
	pixelize	1.00×10^{-7}	3.17×10^{-8}	9.54×10^{-12}	4.38×10^{-9}
	fftdnoiz	1.02×10^{-6}	2.97×10^{-7}	-1.64×10^{-11}	4.31×10^{-9}
15	without filtering	1.87×10^{-6}	4.35×10^{-7}	-6.63×10^{-12}	4.33×10^{-9}
	blur	1.83×10^{-6}	4.32×10^{-7}	-4.53×10^{-12}	4.33×10^{-9}
	sharp	1.88×10^{-6}	4.36×10^{-7}	-8.24×10^{-12}	4.32×10^{-9}
	dctdnoiz	9.35×10^{-7}	3.09×10^{-7}	-5.13×10^{-11}	4.34×10^{-9}
	nlmeans	8.73×10^{-7}	2.41×10^{-7}	-4.07×10^{-11}	4.37×10^{-9}
	pixelize	1.50×10^{-7}	4.76×10^{-8}	-4.05×10^{-11}	4.29×10^{-9}
	fftdnoiz	1.60×10^{-6}	4.02×10^{-7}	-2.34×10^{-11}	4.33×10^{-9}

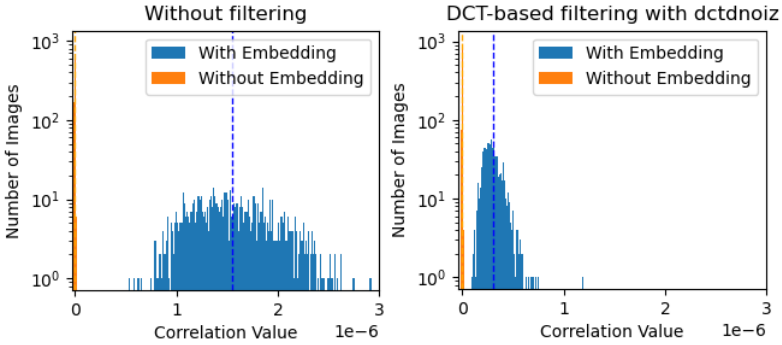


Figure 1. Chart of correlation without filtering (left) and after filtering with dctdnoiz (right)

4. Conclusion

This paper proposed a method for the experimental estimation of the correlation distribution in correlation-based image watermarking. The proposed approach enables analysis of how various distortions affect the corre-

lation between watermarked image and embedded and non-embedded watermarks. The method was applied to a DCT based watermarking algorithm [4] using the Flickr8K dataset [7] and a set of FFmpeg filters [8]. The implementation and experimental results are available at [5].

The results show that distortions introduced by filters significantly reduces the correlation values between the embedded watermark and the distorted watermarked image, while having minimal changes in the correlation between the distorted image and non-embedded watermarks.

REFERENCES

1. *Dixit A., Dixit R.* A review on digital image watermarking techniques // International Journal of Image, Graphics and Signal Processing. 2017. Vol. 9, № 4. P. 56.
2. *Asikuzzaman M., Pickering MR.* An overview of digital video watermarking. // IEEE Transactions on Circuits and Systems for Video Technology., 2017., Vol.28 № 9, P. 2131–2153.
3. *Barni M., Bartolini F., Cappellini V., Piva A* A DCT-domain system for robust image watermarking // Signal processing. 1998. Vol. 66, № 3. P. 357–372.
4. *Cox I., Kilian J., Leighton F.* Secure spread spectrum watermarking for multimedia // IEEE transactions on image processing. 1997. Vol. 6, № 12. P. 1673–1687.
5. `https://github.com/anjin-viktor/correlation\_estimation` — Watermark Correlation Distribution Estimation., 08.05.2025.
6. *Candra R.* The implementation of an efficient zigzag scan // Journal of Telecommunication, Electronic and Computer Engineering (JTEC). 2017. Vol. 9, № 2. pp. 95–98.
7. *Rashtchian C., Young P., Hodosh M., Hockenmaier J.* Collecting image annotations using amazon's mechanical turk // Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk, Los Angeles, 2010. P. 139–147.
8. `ffmpeg.org` — A complete, cross-platform solution to record, convert and stream audio and video., 08.05.2025.

Viktor Anzhin — Ph.D. student, Institute of Applied Mathematics and Computer Science, TSU. E-mail: *viktor.anjin@gmail.com*

COX MODEL-BASED SURVIVAL ANALYSIS FOR PREDICTING INFRASTRUCTURE FAILURES IN SMART CITIES UNDER RIGHT CENSORED DATA CONDITIONS

N. S. Nurmukhamedova[1], M. B. Berdimuradov[2]

[1] *Department of Econometrics & Economical Modelling, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan*

[2] *Department of Applied mathematics & Intellectual technologies, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan*

This study proposes the development of a survival-based reliability forecasting model for critical urban infrastructure systems, including power grid nodes, urban water networks, and composting units. Leveraging the Cox Proportional Hazards Model (PHM) under right-censoring conditions, the study aims to estimate failure probabilities over time using real-world environmental and operational covariates. The approach allows for dynamic risk assessment without requiring full specification of the baseline hazard function, making it well-suited for complex, evolving smart city environments. The proposed model will incorporate maximum likelihood and Bayesian estimation methods for robust inference and be supported by graphical tools such as survival curves and hazard function plots. This research contributes to the field of urban analytics by enabling early failure prediction and preventive maintenance planning, thereby supporting sustainability and resilience in smart infrastructure systems. **Keywords:** *Cox PHM, Right-Censored Data, Urban Infrastructure Reliability, Smart City Systems*

Introduction

Smart cities depend on resilient infrastructure systems to ensure uninterrupted services and sustainable urban living. However, failures in critical components such as power grid nodes, water distribution systems, and composting equipment can lead to significant operational and environmental consequences. Traditional reliability analysis methods often struggle to accommodate incomplete or right-censored data, which is common in real-world monitoring [1, 2].

To address this limitation, the present study proposes a survival-based reliability modeling framework employing Cox PHM. This approach enables the estimation of failure risk over time while incorporating relevant

covariates, ultimately enhancing the predictive accuracy and maintenance planning of urban infrastructure networks [3, 4].

Background and Related Work. Modern survival analysis for infrastructure reliability spans nonparametric, semiparametric, and parametric approaches [5]. The Kaplan-Meier (KM) estimator provides a model-free estimate of the survival function that is robust under right-censoring, yet it cannot directly adjust for multiple, time-varying risk factors typical of smart-city assets [1]. Parametric Accelerated Failure Time (AFT) models, such as the Weibull, Exponential, and Gompertz distributions, describe lifetimes by assuming a fully specified distribution; they enable direct time-scale interpretations and extrapolation but can be misspecified under heterogeneous operating regimes [2, 6].

The Cox PHM offers a pragmatic middle ground: it leaves the baseline hazard unspecified while estimating covariate effects through partial likelihood, thereby accommodating high-dimensional sensor and contextual data without committing to a parametric baseline [3, 4]. Methodological advances including tie handling and large-sample theory have made Cox PHM the workhorse for risk modeling under censoring [7–9]. In urban systems, recent studies have applied survival learning to power grids, water pipelines, and other composite assets, showing that hazard-based formulations can capture environmental load, usage intensity, and maintenance signals more faithfully than aggregate failure-rate summaries [3, 5, 7].

Positioning and Novelty. Unlike prior applications that either (a) report KM curves without covariate adjustment or (b) rely on single-family AFT assumptions, we develop a Cox PHM centered reliability workflow tailored to smart-city assets under right-censoring. Our contribution is twofold: (1) estimation strategy side-by-side MLE and Bayesian inference for Cox effects with uncertainty summaries suitable for operational decisions [3, 4]; (2) practical integration simulation assisted forecasting and a prototype dashboard that translate hazard-ratio insights into preventive maintenance policies under realistic data incompleteness [3, 5, 7].

1. Objectives

The primary objective of this study is to develop a Cox PHM based reliability modeling framework for smart-city infrastructure under right-censoring. We aim to estimate and interpret the effects of operational and environmental covariates on failure risks while maintaining flexibility with respect to the baseline hazard [2–4].

1.1. Specific objectives

Modeling under censoring Formulate time-to-failure T, censoring indicator δ, and covariate vector $\mathbf{x}$ for power grids, water systems, and composting units; adopt the Cox PHM model to quantify hazard ratios under right-censoring without specifying the baseline hazard [3].

Dual estimation strategies (i) *MLE* via Cox partial likelihood with appropriate tie handling (Efron/Breslow) and large-sample properties [7–9]; (ii) *Bayesian* inference to obtain posterior summaries and highest posterior density (HPD) intervals; report classical confidence intervals (CI) for MLE and HPD/credible intervals for Bayesian estimates [4].

Uncertainty and decision outputs Provide decision-oriented uncertainty summaries (HPD and CI) for hazard ratios and survival curves to support maintenance scheduling and risk ranking in operations [4].

Simulation design for evaluation Conduct Monte Carlo experiments across varying censoring levels and sample sizes to compare MLE and Bayesian performance. For each replication, we will compute accuracy metrics, including the Mean Squared Error (MSE) of the parameter estimates

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}\|^2,$$

as well as relative bias and interval coverage (CI vs. HPD). Runtime and convergence diagnostics will also be recorded.

Empirical validation Fit the Cox PHM model to real smart-city asset data, generate survival/hazard diagnostics, and assess out-of-sample risk predictions for preventive maintenance use cases.

Model diagnostics Check the proportional hazards assumption (e.g., Schoenfeld residuals and time-varying effects) and examine influential observations to ensure model adequacy [4].

Deliverables Produce survival curves, hazard function plots, and a lightweight prototype dashboard for reliability forecasting and decision support.

We estimate Cox PHM effects via maximum likelihood and Bayesian approaches and provide decision oriented uncertainty summaries (HPD and CI) under right-censoring [2–4].

2. Mathematical Formulation

Data includes lifetimes of infrastructure units, censoring indicators, and covariates such as temperature, load, usage, pressure, and maintenance frequency [2]. The survival time T of a smart city infrastructure component

is modeled using the Cox PHM, with the hazard function defined as:

$$h(t \mid \mathbf{x}) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{x}),$$

where $h(t \mid \mathbf{x})$ is the hazard rate at time t given covariates

$$\mathbf{x} = (x_1, x_2, \ldots, x_p),$$

$h_0(t)$ is the baseline hazard function (unspecified), $\boldsymbol{\beta}$ are the regression coefficients to be estimated. To estimate $\boldsymbol{\beta}$, we use the partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_j)},$$

where

$\delta_i \in \{0, 1\}$ is the censoring indicator: $\delta_i = 1$ if the event (e.g., failure) is observed; $\delta_i = 0$ if right-censored,

$R(t_i)$ denotes the risk set all components still under observation just prior to time t_i.

3. Conclusion

This project offers a methodological contribution to the reliability assessment of smart urban infrastructure by applying Cox PHM under right-censoring conditions. Through the estimation of hazard functions, generation of survival plots, and development of a dashboard prototype, the study aims to provide both theoretical insight and practical tools for infrastructure monitoring and failure prediction. The resulting analysis and outputs are expected to support more informed decision-making, enhance preventive maintenance strategies, and contribute to the sustainability and resilience of smart city systems.

Compared to classical nonparametric approaches such as the KM estimator and fully parametric AFT models (e.g., Weibull, Exponential, Gompertz), the Cox PHM framework provides a balance between flexibility and interpretability. It avoids baseline misspecification while retaining the ability to incorporate diverse operational and environmental covariates, which is crucial for evolving smart-city environments.

Furthermore, two estimation strategies were employed. MLE delivered efficient point estimates with relatively narrow confidence intervals, making it suitable for large-sample scenarios. In contrast, the Bayesian approach provided HPD intervals that better quantified parameter uncertainty, particularly under small-sample or high-censoring conditions. A comparison of results showed that hazard-ratio estimates were broadly consistent across

the two methods, but Bayesian inference offered a richer uncertainty characterization.

Finally, model accuracy was evaluated through Monte Carlo–based MSE comparisons between simulated and estimated parameters. The Bayesian approach tended to yield slightly lower MSE values under heavy censoring, indicating more stable inference in complex data settings, while MLE remained competitive in larger samples. This dual perspective demonstrates that MLE and Bayesian inference not only complement but also validate each other, thereby ensuring robust and reliable survival analysis for smart-city infrastructure under censoring.

The complete set of simulation outcomes, empirical findings, and detailed methodological extensions will be presented in the forthcoming extended paper.

REFERENCES

1. *Kaplan E.L., Meier P.* Nonparametric Estimation from Incomplete Observations // Journal of the American Statistical Association. 1958. Vol. 53. No. 282. P. 457–481.
2. *Kalbfleisch J.D., Prentice R.L.* The Statistical Analysis of Failure Time Data // 2nd ed. Wiley, 2002.
3. *Cox D.R.* Regression Models and Life-Tables // Journal of the Royal Statistical Society: Series B. 1972. Vol. 34. No. 2. P. 187–220.
4. *Therneau T.M., Grambsch P.M.* Modeling Survival Data: Extending the Cox Model // Springer, 2000.
5. *Nurmukhamedova N., Berdimuradov M., Murodov S.* Comparative Analysis of Unknown Parameters of the Normal Distribution under Right-Censoring Using MLE and Bayesian Methods // Mathematics and Statistics, Vol. 13, №. 4, P. 237-251, 2025. DOI: 10.13189/ms.2025.130408
6. *Wei L.J.* The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis // Statistics in Medicine. 1992. Vol. 11. P. 1871–1879.
7. *Efron B.* The Efficiency of Cox's Likelihood Function for Censored Data // Journal of the American Statistical Association. 1977. Vol. 72, №. 359. P. 557–565.
8. *Andersen P.K., Gill R.D.* Cox's Regression Model for Counting Processes: A Large Sample Study // Annals of Statistics. 1982. Vol. 10, №. 4. P. 1100–1120.
9. *Breslow N.* Covariance Analysis of Censored Survival Data // Biometrics. 1974. Vol. 30, №. 1. P. 89–99.

10. *Yang B., Liang X., Xu S., Wong M.S., Ma W.* TCNSurv: A TCN-based deep survival analysis model for failure prediction in urban infrastructure systems // SSRN Electronic Journal. 2023. doi: 10.2139/ssrn.4472264.
11. *Zhang Y., Zheng L., Chen Q., Zhang Y., He Y.* Time-dependent reliability analysis of smart city power systems under censored monitoring data // Applied Energy. 2023. Vol. 345. Art. ID 121154.
12. *Rauf A., Ali S., Aftab M., Javaid N., Awais M., Malik A.W.* Reliability prediction model for smart urban water pipelines using Cox proportional hazard model // Measurement. 2022. Vol. 200. Art. ID 111573.
13. *Maharela I.A., Fletcher L., Chen D.-G.* Modified Cox Models: A Simulation Study on Different Survival Distributions, Censoring Rates, and Sample Sizes // Mathematics. 2024. Vol. 12, № 2903. P. 1–10.

Nargiza Nurmukhamedova — Doctor of Philosophy in Physical and Mathematical Sciences, Associate Professor, Econometrics and Economic Modeling. E-mail: *rasulova_nargiza@mail.ru*

Mirkamol Berdimuradov — PhD student, Department of Applied mathematics and Intellectual technologies. E-mail: *mirkamol1403@gmail.com*

FAST NUMERICAL METHODS FOR STOCHASTIC MODELING BASED ON PROBABILISTIC EXTENSIONS

B. Dobronets[1], O. Popova[1]

[1] *Siberian Federal University, Krasnoyarsk, Russian Federation*

The article studies fast algorithms for numerical modeling of problems with random input data. The approaches under consideration are non-intrusive methods based on probabilistic extensions. They rely on existing numerical methods for solving deterministic problems and use them as solvers. Unlike polynomial chaos methods, in some cases it is possible to avoid exponential growth of the number of operations from the number of input parameters.

Keywords: *Computational probabilistic analysis, probabilistic extensions, uncertainty quantification*

Introduction

Real computational models usually contain many parameters.They are usually non-deterministic and can be represented either by random constants or random fields. If the input parameters of the model are represented by distribution functions, researchers tend to estimate the output data as distribution functions as well. An example of such a method is the stochastic finite element method for solving differential equations with random coefficients [2].

Methods for solving such problems are usually called Uncertainty Quantification (UQ) [3]. Uncertainty Quantification provides rigorous procedures for describing how the output of a model depends on the uncertainties of its input parameters.

Fig. 1 shows the UQ scheme, where the input parameters $x_1, ..., x_n$ of the model $y = f(x_1, ..., x_n)$ are represented by their distributions.

Uncertainty quantification methods play a key role in reducing the impact of uncertainties in both optimization and decision-making processes. The main goal of UQ is to construct output distributions of numerical simulations $\mathbf{y}$.

Monte Carlo (MC) methods can be attributed to UQ. Despite all its positive qualities, MC has an extremely low convergence rate. To obtain adequate estimates of the statistical characteristics of the modeling results, the original problem must be solved several thousand times.

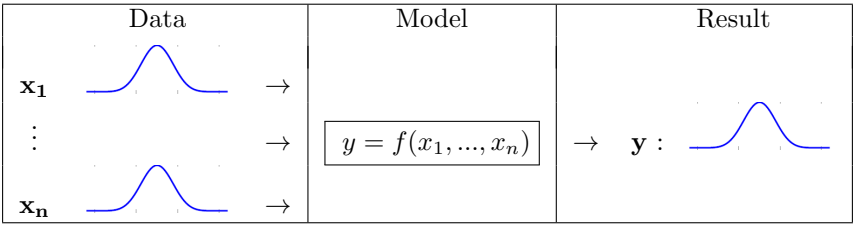


Figure 1. Uncertainty Quantification

Reducing the number of operations is one of the goals of stochastic modeling. In the field of fast computing, one way to actually reduce the number of operations is to use polynomial chaos (PC) methods in combination with stochastic collocations. A similar approach has been successfully applied to solving stochastic differential equations [2].

Polynomial chaos expansion seeks to approximate a random function via orthogonal polynomials of random variables. Thus, the Gaussian distribution corresponds to Hermitian polynomials, and the uniform distribution corresponds to the expansion in Legendre polynomials; for gamma distributions, it is convenient to use Laguerre polynomials [1, 3].

Sensitivity analysis quantifies how much of the uncertainty in the model output each uncertain parameter is responsible for. A sensitivity analysis provides a better understanding of the relationship between the parameters and output of a model. This can be useful in a model reduction context. For example, a parameter with a low sensitivity index can essentially be set to any fixed value (within the explored distribution), without affecting the variance of the model much [3].

It should be noted that the MC and polynomial chaos expansion are mainly focused on estimating the mathematical expectation and variance of the modeling results [5]. One of the limitations of using Polynomial chaos methods is the exponential growth of the number of operations from the number of input parameters [5].

1. Probabilistic extensions

A competitor of polynomial chaos methods is computational probabilistic analysis (CPA) [4]. The basis of CPA is the calculation of probability extensions, i.e. the construction of distribution laws for functions $f(x_1, \ldots, x_n)$ of random arguments, where the joint probability density function with respect to the random vector $(x_1, \ldots, x_n)$ is known. Unlike PC, CPA works with nonparametric representations of random arguments.

One of the problems considered by CPA is the problem to construct a probability density function of random variables. Let us consider the general case [5], when $(x_1, \ldots, x_n)$ is a system of continuous random variables with a joint probability density function $p(x_1, \ldots, x_n)$ and the random variable z is a function f of random arguments $x_1, \ldots, x_n$

$$z = f(x_1, \ldots, x_n).$$

We say that the random function $\mathbf{f} : \mathbf{R^n} \to \mathbf{R}$ is a probabilistic continuation of the deterministic function $f : \mathbb{R}^n \to \mathbb{R}$ on the set $D \subset \mathbb{R}^n$, if $\mathbf{f}(\mathbf{x}) = f(x)$ for all arguments $x \in D$ [4].

The random function $\mathbf{f} : \mathbf{R^n} \to \mathbf{R}$ is called the probabilistic extension of the deterministic function $f : \mathbb{R}^n \to \mathbb{R}$ on the set $D \subset \mathbb{R}^n$, if

(i) it is probabilistic continuation of f on D,

(ii) the probability density function $\mathbf{f}$ coincides with the probability density function $\mathbf{z}$ of the random variable $z = f(x_1, x_2, \ldots, x_n)$, where $(x_1, x_2, \ldots, x_n)$ is a system of continuous random variables with joint probability density functions $\mathbf{p}(x_1, x_2, \ldots, x_n)$ [4].

Thus, we can write

$$\mathbf{z} = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

In cases where it is necessary to directly indicate the value of $\mathbf{f}$ at some point ξ, we will use the notation

$$\mathbf{z}(\xi) = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n)(\xi).$$

Theorem 1. [4] Let $(x_1, \ldots, x_n)$ be independent continuous random variables, $\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is a probabilistic extension of the function $f(x_1, x_2, \ldots, x_n)$ and for all real t the function $\mathbf{f}(t, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is a probabilistic extension of the function $f(t, x_2, \ldots, x_n)$.

Then

$$\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)(\xi) = \int_{\mathrm{supp}(\mathbf{x}_1)} \mathbf{x}_1(t)\mathbf{f}(t, \mathbf{x}_2, \ldots, \mathbf{x}_n)(\xi)dt. \qquad (1)$$

Thus, applying the theorem successively to $\mathbf{f}(t, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, we can represent (1) as

$$\mathbf{f}(\mathbf{x}_1 \ldots \mathbf{x}_n)(\xi) =$$

$$= \int_D \mathbf{x}_1(t_1) \ldots \mathbf{x}_k(t_k)\mathbf{f}(t_1, \ldots, t_k, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n)(\xi)dt_1dt_2 \ldots dt_k, \qquad (2)$$

where $D = \mathrm{supp}(x_1) \times \ldots \times \mathrm{supp}(x_k)$, $t_1, \ldots, t_k$ are deterministic variables.

When calculating (2) using numerical quadratures, the number of operations can be estimated as $\sim Km^k$, where K is the number of operations for calculating the probabilistic extension $\mathbf{f}(t_1, \ldots, t_k, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n)(\xi)$, m is the number of quadrature nodes in one dimension. Thus, CPA is generally not inferior to polynomial chaos methods in the number of operations.

In a number of problems it is possible to construct algorithms where the number of operations depends linearly on the number of arguments.

Let $f(x_1, \ldots, x_n)$ be a rational function. We can obtained probabilistic extension $\mathbf{f}$ of real rational functions f by replacing (i) the real variables x_1, x_2, …, x_n with an probability density functions $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ and (ii) the real arithmetic operations with corresponding probabilistic operations. The result $\mathbf{f}$ is called a *natural probabilistic extension* [4]. In this case, the number of operations grows linearly from n.

Let us consider the question of calculating a probabilistic extension in the form (2) with as few operations as possible.

It is easy to see that we must strive to represent the construction of a probabilistic extension with the smallest k and the function $\mathbf{f}(t_1, \ldots, t_k, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n)$ falls into conditions when it is possible to construct a probabilistic extension for it with non-exponential (polynomial) growth in the number of operations.

2. Karhunen–Loéve expansions

Let's consider the Dirichlet problem:

$$Lu = f, \qquad x \in D, \tag{3}$$

$$u(x) = 0, \qquad x \in \partial\overline{D}, \tag{4}$$

D is a bounded open domain in R^2 with piecewise smooth boundary $\partial\overline{\Omega}$,

$$Lu = -\sum_{i=1}^{2} \frac{\partial}{\partial x_i}(a_i \frac{\partial}{\partial x_i} u).$$

In what follows, to solve (3), it is necessary to parameterize the input coefficients a_i and the right-hand side f. Without loss of generality (to simplify the presentation), we will assume that the coefficients a_i are determined by a set of random constants, and their probability density functions $\mathbf{a}_i, i = 1, ..., N_a$ are known.

To parameterize the right-hand side of f we will use the Karhunen–Loéve expansions [3, 5]

$$f(x, \omega) \approx f(x, \omega)_N = \sum_{n=1}^{N} \sqrt{\lambda_n} F_n(x) f_n(\omega),$$

$\{\lambda_n, F_n(x)\}_{n=1}^{\infty}$ denote pairs of eigenvalues and functions of a given covariance function, and $f_n(\omega)_{n=1}^{\infty}$ are independent random variables with zero mean and unit variance, represented by their probability density functions $\mathbf{f}_n$. Karhunen–Loéve expansions is also known as proper orthogonal decomposition (POD) and principal component analysis (PCA).

Moreover, the error decreases monotonically with an increase in the number of terms in the expansion.

Let it be necessary to find a probabilistic extension of a numerical solution at some point $\mathbf{u}^h(x) = \mathbf{u}^h(x, \mathbf{a}_1, ..., \mathbf{a}_{N_a})$.

The final probabilistic extension can be represented as [7]

$$\mathbf{u}^h(x) = \iint \mathbf{a}_1(t_1)...\mathbf{a}_{N_a}(t_{N_a})\mathbf{u}^h(x, t_1, ..., t_{N_a}, \mathbf{f}_1, ..., \mathbf{f}_{N_f})dt_1...dt_{N_a}. \quad (5)$$

The integral (5) will be calculated using numerical quadratures. The number of calculation operations (5) is proportional to $K\exp(N_a)$, where K depends linearly on N_f.

3. Using Taylor Series

Let us consider the case of approximate calculation of the probability extension of a smooth function $f(x_1, \ldots, x_n)$, when the random vector $x = (x_1, \ldots, x_n)$ has independent components.

In the neighborhood of some point $x_0 = (x_{01}, \ldots, x_{0n})$, we approximate the function $f(x_1, \ldots, x_n)$ by a segment of the Taylor series

$$f(x) \approx f(x_0) + \sum_{j=1}^{n} \frac{\partial f(x_0)}{\partial x_j}(x_j - x_{0j}).$$

In this case, the natural probability extension will look like this

$$\mathbf{f}(\mathbf{x}) \approx f(x_0) + \sum_{j=1}^{n} \frac{\partial f(x_0)}{\partial x_j}(\mathbf{x}_j - x_{0j}).$$

Due to the independence of $x = (x_1, \ldots, x_n)$, the number of operations for estimating the probabilistic extension $\mathbf{f}(\mathbf{x})$ will grow linearly with the dimension n.

This approach has much in common with interval sensitivity analysis [6].

$$y' = f(t, y, k), \quad (6)$$
$$y(0) = y0,$$

$y \in R^n$ is solution vector;

$y0 \in R^n$ is initial data vector;

$k \in R^m$ is a random vector of independent parameters with known probability densities. Similar problems arise when modeling complex nonlinear reactions in the kinetic region [8]. The solution can be represented as a function of the parameters

$$y = y(t, k).$$

To construct probability density functions $\mathbf{y}$, we approximate the function y by a segment of the series

$$y(t, k) \approx y(t, k_0) + \sum_j \frac{\partial y}{\partial k_j}(k_j - k_{0j}).$$

To find partial derivatives with respect to parameters, we use the original model and obtain a system of differential equations

$$\frac{\partial y'}{\partial k_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial y_i}\frac{\partial y_i}{\partial k_j} + \frac{\partial f}{\partial k_j}, \tag{7}$$

$$\frac{\partial y}{\partial k_j} = 0, j = 1, ..., m.$$

Simultaneously solving systems (7) and (6), we obtain

$$\mathbf{y}(t, \mathbf{k}) \approx y(t, k_0) + \sum_j \frac{\partial y}{\partial k_j}(t, k_0)(\mathbf{k}_j - k_{0j}).$$

Note that in this case the number of operations for constructing the probabilistic extension $\mathbf{y}(t, \mathbf{k})$ grows linearly with the dimension of the parameter vector $k \in R^m$.

4. Conclusion

Thus, it is shown that the algorithms for solving stochastic modeling problems proposed in the work allow in a number of cases to significantly reduce the number of operations. Thus, by dividing the input parameters in stochastic modeling into groups with different degrees of influence on the output distributions, it is possible to significantly reduce the number of operations. For this purpose, the work considers the Karhunen–Loéve expansions and Taylor Series.

REFERENCES

1. *Xiu D.* Fast Numerical Methods for Stochastic Computations: A Review // Commun. Comput. Phys. 2009. Vol. 5, №. 2–4. P. 242–272.
2. *Shalimova I. A., Sabelfeld K. K.* Solution to a stochastic Darcy equation by the polynomial chaos expansion // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. Novosibirsk, 2017. Vol. 20, №3. P. 313–327.
3. *Smith R.* Uncertainty Quantification. Theory, Implementation, and Applications. Philadelphia: SIAM, 2014.
4. *Dobronets B. S. Popova O. A.* Computational probabilistic analysis: models and methods. Krasnoyarsk: SFU, 2020. 236 p.
5. *Gunzburger M. D., Webster C. G., Zhang G.* Stochastic Finite element methods for partial differential equations with random input data // Acta Numerica. 2014. Vol. 23. P. 521–650.
6. *Dobronets B. S.* Interval Mathematics, Krasnoyarsk: Krasnoyarsk State University, 2004. 216 p.
7. *Dobronets B. S. Popova O. A., Merko A. M.* Numerical modeling of boundary value problems for differential equations with random coefficients // Journal of Physics: Conference Series. International Conference “Marchuk Scientific Readings 2021”. 2021.
8. *Bykov V. I., Dobronets B. S.* The forecast of nonlinear dynamics in kinetic region by interval analysis methods // Izvestiya VUZ. Applied Nonlinear Dynamics. 2004. Vol. 12. P. 18–25.

Dobronets Boris S. — Doctor of Physical and Mathematical Sciences, Professor, Professor of the Department of Artificial Intelligence Systems, IKIT SFU. E-mail: *BDobronets@yandex.ru*

Popova Olga A. — Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Artificial Intelligence Systems, IKIT SFU. E-mail: *OlgaArc@yandex.ru*

Математическая теория телетрафика и теория массового обслуживания

ЧИСЛЕННЫЙ АНАЛИЗ RQ-СИСТЕМЫ $M/M_k/1$ С ДЕГРАДАЦИЕЙ И ВОССТАНОВЛЕНИЕМ ПРИБОРА

А. А. Варакина[1], Е. А. Фёдорова[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В работе рассмотрена марковская однолинейная RQ-система, на вход которой поступает простейший поток заявок. Заявки, неполучившие обслуживание, переходят на орбиту и через случайное время, распределенное экспоненциально, реализуют повторные вызовы. Время обслуживания каждой заявки распределено экспоненциально с параметром, зависящем от текущей состояния прибора (степени деградации). Под деградацией прибора понимается уменьшение скорости обслуживания заявок в зависимости от суммарного количества обслуженных заявок с момента восстановления прибора. При достижении некоторого значения интенсивности обслуживание становится невозможным, и прибор уходит на восстановление, распределенное экспоненциально. Составлены уравнения Колмогорова и выполнен численный анализ системы на основе матричного метода.

Ключевые слова: *RQ-система, деградация, восстановление, матричный метод*

Введение

В теории массового обслуживания (ТМО) RQ-системы представляет класс математических моделей, особенностью которых являются повторные вызовы, то есть в таких системах заявки, неполучившие обслуживания не покидают систему, а продолжают пытаться получить доступ к обслуживающему устройству через случайное время. Такой класс моделей описывает многие технические или экономические системы, например сетей передачи данных, контакт-центры и др. [1, 4].

Системы массового обслуживания с восстановлением прибора исследуются в многих работах, однако обычно причиной восстановления являются поломки прибора, которые описываются, например, поступающим потоком отрицательных заявок. Модели ТМО с деградацией обслуживания исследованы слабо [3, 4].

Исследование выполнено за счет гранта Российского научного фонда № 24-21-00454, https://rscf.ru/project/24-21-00454/

Особенностью данной работы является исследование RQ модели, учитывающей изменение режима работы обслуживающего прибора в зависимости от количества обработанных заявок. В социально-экономических системах такие модели позволяют учесть человеческий фактор – степень усталости обслуживающего специалиста, а в технических системах – износ оборудования.

1. Описание математической модели

Рассмотрим RQ-систему, на вход которой поступает простейший поток заявок с интенсивностью λ. Если прибор свободен, то заявка занимает его на случайное время, распределённое по экспоненциальному закону с параметром μ_k, зависящим от степени деградации прибора (будет пояснено далее). В случае недоступности прибора входящая заявка отправляется на орбиту, откуда пытается получить обслуживание спустя случайное время, распределённое по экспоненциальному закону с параметром σ.

Под деградацией обслуживания понимается уменьшение скорости обслуживания в зависимости от числа обслуженных прибором заявок за рассматриваемый период времени. Пусть процесс деградации имеет конечное число состояний K. В начальный момент времени (или после периода восстановления) интенсивность обслуживания равна μ_0, а степень деградации прибора $k = 0$. Завершая обслуживание каждой последующей заявки, процесс деградации переходит из состояния k в состояние $k + 1$, то есть следующая заявка обслуживается с интенсивностью μ_{k+1}. В рамках данной работы предполагается, что $\mu_0 > \mu_1 > ... > \mu_k > \mu_{k+1} > ... > \mu_{K-1}$. В конечном итоге деградация прибора достигает состояния K, при котором обслуживание становится невозможным и необходимо восстановление прибора. То есть по завершению обслуживания K-й заявки прибор отключается для восстановления на время, распределенное экспоненциально с параметром μ_K.

Структура исследуемой модели представлена на рис. 1.

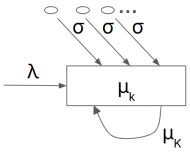


Рис. 1. RQ-система с деградацией прибора и восстановлением

Пусть $i(t)$ — число заявок на орбите ($i \geq 0$), $k(t)$ — состояние процесса деградации прибора ($k = \overline{0, K}$), $m(t)$ — состояние прибора:

$$m(t) = \begin{cases} 0, & \text{прибор свободен,} \\ 1, & \text{прибор занят обслуживанием заявки.} \end{cases}$$

Трехмерный случайный процесс $\{m(t), k(t), i(t)\}$ является марковским.

Обозначим $P_m(k, i, t) = P\{m(t) = m, k(t) = k, i(t) = i\}$ — вероятность того, что в момент времени t прибор находится в состоянии m, процесс деградации находится в состоянии k, а на орбите находится i заявок. Составим систему дифференциальных уравнений Колмогорова для распределения вероятностей $P_m(k, i, t)$ состояний рассматриваемой RQ-системы:

$$\begin{cases} \dfrac{\partial P_0(0,i,t)}{\partial t} = -(\lambda + i\sigma)P_0(0,i,t) + \mu_K P_0(K,i,t), \\ \dfrac{\partial P_1(0,i,t)}{\partial t} = -(\lambda + \mu_0)P_1(0,i,t) + \sigma(i+1)P_0(0,i+1,t) + \\ +\lambda P_0(0,i,t) + \lambda P_1(0,i-1,t), \\ \dfrac{\partial P_0(k,i,t)}{\partial t} = -(\lambda + i\sigma)P_0(k,i,t) + \mu_{k-1}P_1(k-1,i,t), \\ \dfrac{\partial P_1(k,i,t)}{\partial t} = -(\lambda + \mu_k)P_1(k,i,t) + \lambda P_0(k,i,t) + \\ +\lambda P_1(k,i-1,t) + (i+1)\sigma P_0(k,i+1,t), \\ \dfrac{\partial P_0(K,i,t)}{\partial t} = -(\lambda + \mu_K)P_0(K,i,t) + \lambda P_0(k,i-1,t) + \\ +\mu_{K-1}P_1(K-1,i,t). \end{cases} \tag{1}$$

В стационарном режиме система (1) имеет следующий вид:

$$\begin{cases} -(\lambda + i\sigma)P_0(0,i) + \mu_K P_0(K,i) = 0, \\ -(\lambda + \mu_0)P_1(0,i) + (i+1)\sigma P_0(0,i+1) + \\ +\lambda P_0(0,i) + \lambda P_1(0,i-1) = 0, \\ -(\lambda + i\sigma)P_0(k,i) + \mu_{k-1}P_1(k-1,i) = 0, \\ -(\lambda + \mu_k)P_1(k,i) + \lambda P_0(k,i) + \\ +\lambda P_1(k,i-1) + (i+1)\sigma P_0(k,i+1) = 0, \\ -(\lambda + \mu_K)P_0(K,i) + \lambda P_0(K,i-1) + \mu_{K-1}P_1(K-1,i) = 0. \end{cases} \tag{2}$$

Так как аналитическое исследование системы (2) довольно затруднительно, применим численный анализ для нахождения стационарных характеристик модели.

2. Численный анализ

Заметим, что система (2) имеет бесконечный размер, так как $i(t) = \overline{0,\infty}$. Для вычисления характеристик системы ограничим ёмкость орбиты некоторой границей N, тогда $i(t) = \overline{0,N}$. N будет определяться индивидуально в каждом числовом примере. В качестве критерия выбора N можно выбрать условие $1 - \sum\limits_{i=0}^{N}\sum\limits_{k=0}^{K}(P_0(k,i) + P_1(k,i)) \le 10^{-12}$. Краевые уравнения для $i = N$ имеют вид:

$$\begin{cases} -(\lambda + N\sigma)P_0(0,N) + \mu_K P_0(K,N) = 0, \\ -\mu_0 P_1(0,N) + \lambda P_0(0,N) + \lambda P_1(0,N-1) = 0, \\ -(\lambda + N\sigma)P_0(k,N) + \mu_{k-1} P_1(k-1,N) = 0, \\ -\mu_k P_1(k,N) + \lambda P_0(k,N) + \lambda P_1(k,N-1) = 0, \\ -\mu_K P_0(K,N) + \lambda P_0(K,N-1) + \mu_{K-1} P_1(K-1,N) = 0. \end{cases} \tag{3}$$

Применим матричный подход для нахождения стационарного распределения состояний усеченной системы. Представим систему уравнений (2)–(3) в матричном виде, добавив условие нормировки:

$$\mathbf{A}\mathbf{p} = \mathbf{b},$$

где $\mathbf{p}$ – вектор вероятностей $P_m(k,i)$, элементы которого будут записаны в следующей последовательности:

$$\begin{aligned} \mathbf{p} = [&P_0(0,0),\ P_1(0,0),\ P_0(1,0),\ P_1(1,0),\dots,P_0(K,0), \\ &P_0(0,i),\ P_1(0,i),\dots,\ P_0(k,i),\ P_1(k,i),\dots, \\ &P_0(K,i),\dots,P_0(k,N),\ P_1(k,N),,\dots,\ P_0(K,N)]^T, \end{aligned} \tag{4}$$

$\mathbf{b}$ – вектор размерности $[(N+1)(2K+1)+1] \times 1$ с нулевыми элементами и последним единичным, $\mathbf{A}$ – матрица коэффициентов размерности $[(N+1)(2K+1)+1] \times [(N+1)(2K+1)]$, полученная из системы (2), ниже приведён её фрагмент (для $i = 0$):

$$\mathbf{A} = \begin{bmatrix} -\lambda & 0 & 0 & 0 & \mu_2 & 0 & \dots \\ \lambda & -(\lambda+\mu_0) & 0 & 0 & 0 & \sigma & \dots \\ 0 & \mu_0 & -\lambda & 0 & 0 & 0 & \dots \\ 0 & 0 & \lambda & -(\lambda+\mu_1) & 0 & 0 & \dots \\ 0 & 0 & 0 & \mu_1 & -(\lambda+\mu_2) & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & -(\lambda+\sigma) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Тогда вектор вероятностей $\mathbf{p}$ можно вычислить с помощью обратной матрицы по следующей формуле:

$$\mathbf{p} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

3. Численный пример

Рассмотрим следующий пример. Пусть

$$\lambda = 0.5, K = 2, \mu_0 = 2, \mu_1 = 1.5, \mu_2 = 1, \sigma = 1.$$

Численное решение реализовывалось на языке программирования Python. Значение границы N выбиралось такое, что $\sum_{i=0}^{N-1} P(i) = 1$ с точностью порядка машинного нуля. Для указанного примера оказалось достаточным $N = 5$. Представим вычисленные значения вектора $\mathbf{p}$:

$$\begin{aligned}\mathbf{p} = [&P_0(0,0) = 0,14037298,\ P_0(0,1) = 0,04638731,\\ &P_0(0,2) = 0,01962463,\ P_0(0,3) = 0,00845963,\\ &P_0(0,4) = 0,00339800,\ P_0(0,5) = 0,00121942,\\ &P_0(1,0) = 0,18718546,\ P_0(1,1) = 0,04597572,\\ &P_0(1,2) = 0,01689008,\ P_0(1,3) = 0,00656049,\\ &P_0(1,4) = 0,00245602,\ P_0(1,5) = 0,00051931,\\ &P_0(2,0) = 0,06999987,\ P_0(2,1) = 0,06939663,\\ &P_0(2,2) = 0,04888143,\ P_0(2,3) = 0,02943646,\\ &P_0(2,4) = 0,01513299,\ P_0(2,5) = 0,00657444,\\ &P_1(0,0) = 0,04670355,\ P_1(0,1) = 0,03439040,\\ &P_1(0,2) = 0,02102414,\ P_1(0,3) = 0,01139776,\\ &P_1(0,4) = 0,00545251,\ P_1(0,5) = 0,00137103,\\ &P_1(1,0) = 0,06987627,\ P_1(1,1) = 0,04594238,\\ &P_1(1,2) = 0,02563314,\ P_1(1,3) = 0,01303566,\\ &P_1(1,4) = 0,00523055,\ P_1(1,5) = 0,00147094].\end{aligned}$$

На рис. 2 представлено маргинальное распределение вероятностей числа заявок на орбите такой системы $P(i)$.

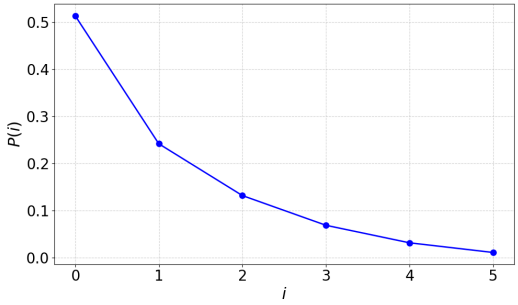


Рис. 2. График распределения вероятностей числа заявок на орбите

Из полученного распределения можно вычислить любые характеристики функционирования системы, например:

– вероятность восстановления

$$P_{restoration} = \sum_{i=0}^{N} \sum_{m=0}^{1} P_m(K, i) = 0.2394;$$

– доля времени простоя прибора

$$P_{idle} = \sum_{i=0}^{N} \sum_{k=0}^{K} P_0(k, i) = 0.7185.$$

Заключение

В ходе работы была исследована RQ-система $M/M_k/1$ с последовательной деградацией прибора и восстановлением. Выполнен численный анализ характеристик системы с помощью матричного подхода, реализованного в виде программы на языке Python.

СПИСОК ЛИТЕРАТУРЫ

1. *Falin G. I., Templeton J. G. C.* Retrial queues // London: Chapman and Hall. 1997. Vol. 7, № 1. P. 127–167.
2. *Artalejo J. R., Gomez-Corral A.* Retrial Queueing Systems. Berlin: Springer, 2008. 267 p.
3. *Choudhary A., Chakravarthy S. R., Sharma D. C.* Impact of the degradation in service rate in MAP/PH/1 queueing system with phase type vacations, breakdowns, and repairs // Annals of Operations Research. 2023. Vol. 331. P. 1207–1248.
4. *Ayyappan G., Nithya S.* Analysis of $M^{[X_1]}$,$M^{[X_2]}/G_1$, $G_2/1$ retrial queue with priority services, differentiate breakdown, repair, synchronized reneging and optional vacation // Reliability: Theory & Applications. 2023. № 2 (73). P. 376–391.

Варакина Анастасия Александровна — студент Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *varakina05@list.ru*

Фёдорова Екатерина Александровна — кандидат физ.-мат.наук, доцент кафедры теории вероятностей и математической статистики Томского государственного университета. E-mail: *moiskate@mail.ru*

СТАЦИОНАРНЫЕ ХАРАКТЕРИСТИКИ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ С ГРУППОВЫМ ОБСЛУЖИВАНИЕМ ТРЕБОВАНИЙ

Е. А. Васильева[1], И. Е. Тананко[1], Е. П. Станкевич[1]

[1] *Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, г. Саратов, Россия*

В одноприборную систему массового обслуживания с бесконечной очередью поступает пуассоновский поток однородных требований. Обслуживание требований начинается только тогда, когда в очереди накопится заданное количество требований, иначе прибор будет простаивать до поступления полной группы требований. Прибор обслуживает одновременно все требования, входящие в обслуживаемую группу. Длительность обслуживания группы требований является случайной величиной с экспоненциальной функцией распределения. После завершения обслуживания вся группа мгновенно возвращается в источник требований. Получены выражения для математических ожиданий длительности пребывания групп требований в очереди и длительности обслуживания групп требований прибором.

Ключевые слова: *система массового обслуживания, групповое обслуживание, стационарные характеристики*

Введение

Первые результаты исследования систем массового обслуживания с групповым обслуживанием представлены в работах [1, 2]. В настоящее время развитие методов анализа и оптимизации систем с групповым обслуживанием [3–7] также представляет большой интерес, так как системы этого класса могут использоваться в качестве моделей реальных систем, в которых объекты обрабатываются группами, например, транспортных и производственных систем, беспроводных сенсорных сетей и т.д. [5, 6]. Эта работа посвящена нахождению выражений для вычисления стационарных характеристик одноприборной системы массового обслуживания с групповым обслуживанием, что является продолжением результатов, опубликованных в статье [7], в которой, в частности, было получено выражение для вычисления математического ожидания (м. о.) длительности пребывания требований в системе.

1. Описание системы

Рассматривается система массового обслуживания с одним обслуживающим прибором и бесконечной очередью, функционирующая в непрерывном времени. Входящий поток требований в систему из источника является Пуассоновким. Интенсивность входящего потока λ. Требования прибором обслуживаются группой размера b, причем, если в очереди количество требований меньше b, то прибор ожидает до тех пор, пока в очередь поступит b требований, иначе из очереди на обслуживание случайным образом выбирается b требований. Длительность обслуживания группы, содержащей b требований, является экспоненциально распределенной случайной величиной с параметром μ. Полагается, что состояние n системы обслуживания – это количество требований, находящихся в ней.

В работе [7] были получены выражения для вычисления стационарных вероятностей состояний рассматриваемой системы:

$$\begin{cases} \pi(n) = \pi(0) \sum\limits_{i=0}^{n} \left(\frac{\lambda}{M}\right)^i, \ 1 \le n \le b-1, \\ \pi(n) = \pi(0) \left(\frac{\lambda}{M}\right)^{n-b} \frac{\lambda}{\mu}, \ n \ge b, \end{cases} \tag{1}$$

где

$$\pi(0) = \frac{M-\lambda}{bM}, \tag{2}$$

а M – единственный корень уравнения

$$M^{b+1} - (\lambda+\mu)M^b + \lambda^b \mu = 0,$$

принадлежащий интервалу

$$\left(\frac{b(\lambda+\mu)}{b+1}, \frac{(\lambda+\mu)^{b+1} - \lambda^b \mu}{(\lambda+\mu)^b} \right).$$

2. Стационарные характеристики

Утверждение 1. Математическое ожидание длительности пребывания групп требований в очереди системы массового обслуживания с групповым обслуживанием имеет вид

$$E[W_b] = \frac{\lambda^b}{\mu M^b} + \frac{\lambda^{b+1}}{b\,\mu\, M^b (M-\lambda)}.$$

Доказательство. Так как на обслуживание из очереди выбирается группа размера b, то ожидание групп требований начнется с момента когда в очереди накопится $2b$ требований, поэтому используя закон

Литтла и определение м. о. дискретной случайной величины, получим, что м. о. длительности пребывания групп требований в очереди:

$$E[W_b] = \frac{1}{\lambda} \sum_{n=2b}^{\infty} (n-b)\,\pi(n). \tag{3}$$

Подставим (1) и (2) в (3)

$$E[W_b] = \frac{1}{\lambda} \sum_{n=2b}^{\infty} (n-b)\,\pi(0) \left(\frac{\lambda}{M}\right)^{n-b} \frac{\lambda}{\mu} =$$

$$= \frac{(M-\lambda)\,M^b}{b\,M\,\lambda^b\,\mu} \sum_{n=2b}^{\infty} (n-b) \left(\frac{\lambda}{M}\right)^n =$$

$$= \frac{(M-\lambda)\,M^b}{b\,M\,\lambda^b\,\mu} \left(\sum_{n=2b}^{\infty} n \left(\frac{\lambda}{M}\right)^n - b \sum_{n=2b}^{\infty} \left(\frac{\lambda}{M}\right)^n \right).$$

Учитывая формулу для вычисления суммы бесконечной геометрической прогрессии и, что

$$\sum_{i=a}^{\infty} i x^i = \frac{x^a(-ax+a+x)}{(1-x)^2}, \text{ при } x < 1,$$

получим

$$E[W_b] = \frac{(M-\lambda)M^b}{b\,M\,\lambda^b\,\mu} \left(\frac{(\frac{\lambda}{M})^{2b}(-2b\frac{\lambda}{M} + 2b + \frac{\lambda}{M})}{\left(1-\frac{\lambda}{M}\right)^2} - b\frac{(\frac{\lambda}{M})^{2b}}{1-\frac{\lambda}{M}} \right) =$$

$$= \frac{(M-\lambda)M^b(\frac{\lambda}{M})^{2b}}{b\,M\,\lambda^b\,\mu} \frac{(-2b\frac{\lambda}{M} + 2b + \frac{\lambda}{M} - b + b\frac{\lambda}{M})}{\left(1-\frac{\lambda}{M}\right)^2} =$$

$$= \frac{(M-\lambda)M^b(\frac{\lambda}{M})^{2b}}{b\,M\,\lambda^b\,\mu} \frac{(-b\frac{\lambda}{M} + b + \frac{\lambda}{M})}{\left(1-\frac{\lambda}{M}\right)^2} =$$

$$= \frac{(M-\lambda)\lambda^b M^2}{b\,M\,\mu\,M^b} \frac{(-b\frac{\lambda}{M} + b + \frac{\lambda}{M})}{(M-\lambda)^2} = \frac{\lambda^b}{b\,\mu\,M^b} \frac{(-b\lambda + bM + \lambda)}{(M-\lambda)} =$$

$$= \frac{\lambda^b b(M-\lambda)}{b\,\mu\,M^b(M-\lambda)} + \frac{\lambda^{b+1}}{b\,\mu\,M^b(M-\lambda)} = \frac{\lambda^b}{\mu\,M^b} + \frac{\lambda^{b+1}}{b\,\mu\,M^b(M-\lambda)}.$$

■

Утверждение 2. Математическое ожидание длительности обслуживания групп требований в системе массового обслуживания с

групповым обслуживанием имеет вид

$$E[V_b] = \frac{1}{M}.$$

Доказательство. Так как в рассматриваемой системе прибор может обслуживать группу, состоящую только из b требований, то обслуживание группы требований может начинаться только, когда в системе находится число требований кратное заданному размеру группы b. Получим м. о. длительности обслуживания группы требований, используя закон Литтла и определение математического ожидания,

$$E[V_b] = \frac{1}{\lambda}\sum_{n=1}^{\infty} b\,\pi(nb) = \frac{b}{\lambda}\sum_{n=1}^{\infty}\pi(nb). \tag{4}$$

Подставим (1) и (2) в (4)

$$E[V_b] = \frac{b}{\lambda}\sum_{n=1}^{\infty}\pi(0)\left(\frac{\lambda}{M}\right)^{nb-b}\frac{\lambda}{\mu} = \frac{b\,\pi(0)\,M^b\,\lambda}{\lambda\,\lambda^b\,\mu}\sum_{n=1}^{\infty}\left(\frac{\lambda}{M}\right)^{nb} =$$

$$= \frac{b\,\pi(0)\,M^b}{\lambda^b\,\mu}\sum_{n=1}^{\infty}\left(\left(\frac{\lambda}{M}\right)^b\right)^n = \frac{b(M-\lambda)\,M^b}{b\,M\,\lambda^b\,\mu}\sum_{n=1}^{\infty}\left(\left(\frac{\lambda}{M}\right)^b\right)^n =$$

$$= \frac{(M-\lambda)\,M^b}{M\,\lambda^b\,\mu}\sum_{n=1}^{\infty}\left(\left(\frac{\lambda}{M}\right)^b\right)^n.$$

Применим формулу суммы бесконечной геометрической прогрессии

$$E[V_b] = \frac{(M-\lambda)\,M^b}{M\,\lambda^b\,\mu}\,\frac{(\frac{\lambda}{M})^b}{1-(\frac{\lambda}{M})^b} = \frac{(M-\lambda)\,M^b\lambda^b}{M\,\lambda^b\,\mu(M^b-\lambda^b)}.$$

В работе [7] было показано, что

$$M^{b-1} + M^{b-2}\lambda + \cdots + M\lambda^{b-2} + \lambda^{b-1} = \frac{M^b}{\mu}.$$

Поэтому

$$\lambda^b - M^b = (\lambda - M)(M^{b-1} + M^{b-2}\lambda + \cdots + M\lambda^{b-2} + \lambda^{b-1}) = (\lambda - M)\frac{M^b}{\mu}.$$

Таким образом,

$$E[V_b] = \frac{(M-\lambda)\,M^b\lambda^b\,\mu}{M\,\lambda^b\,\mu(M-\lambda)\,M^b} = \frac{1}{M}.$$

■

Заключение

Для системы с групповым обслуживанием получены выражения м. о. длительности пребывания групп требований в очереди и м. о. длительности обслуживания группы. В дальнейшем полученные результаты можно использовать для решения задач оптимизации систем массового обслуживания с групповым обслуживанием.

СПИСОК ЛИТЕРАТУРЫ

1. *Bailey N. T. J.* On Queueing Processes with Bulk Service // Journal of the Royal Statistical Society: Series B (Methodological). 1955. Vol. 17. P. 80–84.
2. *Downton F.* Waiting Time in Bulk Service Queues // Journal of the Royal Statistical Society. Series B (Methodological). 1955. Vol. 17. P. 256–261.
3. *Economou A.* An Alternative Model for Queueing Systems with Single Arrivals, Batch Services and Customer Coalescence // Queueing Systems. 2002. Vol. 40, №. 16. P. 407–432.
4. *Sasikala S., Indhira K.* Bulk Service Queueing Models – A Survey // International Journal of Pure and Applied Mathematics. 2016. Vol. 106, №. 6. P. 43–56.
5. *Hopp W. J., Spearman M. L., Chayet S., Donohue K. L., Gel E. S.* Using an Optimized Queueing Network Model to Support Wafer Fab Design // IIE Transactions. 2002. Vol. 4. P. 119–130.
6. *Kar S., Rehrmann R., Mukhopadhyay A., Alt B., Ciucu F., Koeppl H., Binnig C., Rizk A.* On the Throughput Optimization in Large-Scale Batch-processing Systems // Performance Evaluation. 2020. Vol. 144. 102142.
7. *Stankevich E., Tananko I., Pagano M.* Optimization of Open Queuing Networks with Batch Services // Mathematics. 2022. Vol. 10, №. 16. 3027.

Васильева Евгения Андреевна — студентка, факультет компьютерных наук и информационных технологий. E-mail: *yevgenia.vas@yandex.ru*

Тананко Игорь Евстафьевич — к.ф.-м.н., доцент, заведующий кафедрой, кафедра системного анализа и автоматического управления. E-mail: *TanannkoIE.sgu@gmail.com*

Станкевич Елена Петровна — к.ф.-м.н., доцент, кафедра системного анализа и автоматического управления. E-mail: *StankevichElena@mail.ru*

ПРИМЕНЕНИЕ УПРАВЛЯЕМОЙ СЕТИ МАССОВОГО ОБСЛУЖИВАНИЯ ДЛЯ МОДЕЛИРОВАНИЯ ПЕРЕВАЛКИ ГРУЗОВ В ПОРТУ

Х. З. Ву[1], М. Л. Жарков[2], А. Л. Казаков[1,2], А. А. Лемперт[2]

[1] *Иркутский национальный исследовательский технический университет, г. Иркутск, Россия*

[2] *Институт динамики систем и теории управления имени В.М. Матросова СО РАН, г. Иркутск, Россия*

В статье представлена обобщенная модель работы универсального морского порта, особенностью которого является изменение технологии и скорости перевалки груза в зависимости от его типа. Она имеет вид управляемой сети массового обслуживания, в которую поступает четыре неординарных потока заявок, описывающих прибытие грузов на морском, речном, автомобильном и железнодорожном транспорте. Для отображения работы причалов, складов и грузового фронта применяются разные типы систем массового обслуживания, в том числе управляемые, в которых параметры работы регулируются по определенным правилам. В результате, в модели учитываются изменение технологии перегрузки в подсистемах порта и несколько различных маршрутов движения грузов между ними.

Ключевые слова: *математическое моделирование, теория массового обслуживания, универсальный морской порт*

Введение

Теория массового обслуживания является эффективным инструментом сценарного моделирования работы технических систем, которые предназначены для массовой обработки материальных потоков различной природы и подвержены воздействию случайных факторов [1]. Такими свойствами обладают транспортные объекты, в частности, морские порты. Поэтому еще со второй половины XX века системы массового

Работа выполнена в рамках гранта № 075-15-2024-533 Министерства науки и высшего образования РФ на выполнение крупного научного проекта по приоритетным направлениям научно-технологического развития (проект «Фундаментальные исследования Байкальской природной территории на основе системы взаимосвязанных базовых методов, моделей, нейронных сетей и цифровой платформы экологического мониторинга окружающей среды»).

обслуживания (СМО) стали активно применяться для анализа их пропускной способности [2].

В настоящее время для описания работы морских портов с учетом их сложной структуры, как правило, применяется аппарат сетей массового обслуживания (СеМО) [1, 3]. Большая часть публикаций, в которых для анализа логистических процессов в портах используется данный аппарат, посвящена специализированным морским терминалам, предназначенным для перевалки одного типа груза, в частности, контейнеров [4, 5] или угля [6]. В таких системах параметры работы оборудования, как правило, слабо меняются со временем, поэтому для их исследования оказалось достаточно применения стационарных СеМО.

Отличительной особенностью универсальных морских портов (УМП) является обработка разнотипных грузов. Для каждого из них на причале применяется специальное оборудование, из-за чего существенно меняется технология и скорость перегрузки. В таком случае, требуется использование нестационарных СеМО. В частности, на их основе авторы предложили способ математического моделирования работы морских портов [7, 8].

В данной работе предложенный ранее подход усовершенствован за счет использования особого класса нестационарных моделей – управляемых СеМО. В них допустима регулировка параметров работы узлов по заданным правилам, а также изменение вероятностей переходов заявок между узлами в зависимости от заполненности их очередей. На основе данного аппарата предложена обобщенная модель работы УМП.

1. Объект исследования

Универсальный (многофункциональный) морской порт – механизированный комплекс, ключевой функцией которого является перевалка грузов разных типов (генеральных, навалочных, насыпных и лесных) в произвольных количественных сочетаниях, а также временное их хранение для координации работы транспорта.

В УМП стыкуются речной (баржи), морской (суда), автомобильный (грузовики) и железнодорожный транспорт. Они независимы между собой и имеют разные характеристики: интенсивность поступления и размеры перевозимых партий грузов. Точное время прибытия транспорта в порт и размер партии груза, как правило, заранее неизвестны [1].

В структуре типового УМП имеется три зоны. Причальная включает несколько причалов, которые могут быть специализированы на обработке одного типа груза или же многоцелевыми. В последнем случае причал оснащен разнородным оборудованием для обработки нескольких типов грузов. В складской зоне, как правило, имеется крытые и

открытые склады, предназначенные для разного типа товаров. Вблизи складов расположен грузовой фронт: железнодорожная станция и ворота для автомобилей. Также рядом с портом присутствуют автомобильная и якорные стоянки, где транспорт ожидает разрешение на заход в систему. Перевозка грузов между причалами и складами осуществляются портовым транспортом (внутренние грузовики).

Баржи и суда обрабатываются на различных специализированных (основных) причалах. Однако, если в системе имеется многоцелевой причал, то на нем может происходить погрузка/разгрузка обоих видов транспорта в случае, если основные причалы заняты. В этом случае в системе появляется альтернативный маршрут движения грузов.

2. Обобщенная модель универсального морского порта

Обобщенная модель работы УМП имеет вид управляемой СеМО, в которой часть узлов относится к управляемым СМО [9]. В них происходит регулировка следующих параметров: числа работающих каналов, их интенсивности обслуживания и вместимости очереди. Также в такой СеМО возможна смена маршрута движения заявок между узлами при наступлении определенных событий.

В УМП обрабатываются разные типы груза, поэтому заявкой на обслуживание считается одна единица массы груза, а транспорт – источником потока заявок. Разные виды транспорта независимы между собой и имеют отличную вместимость, следовательно в модели имеется четыре независимых неординарных потока заявок. Их параметры определяются на основе натурного обследования конкретного объекта.

Для моделирования работы структурных элементов (подсистем) порта применяется $S\,(S \geqslant 2)$ СМО (узлы) разных типов. В них канал соответствует техническому устройству, перегружающему грузы, или же самому транспортному средству с грузом, а очередь – доступной площади для временного размещения грузов. Заявки из очереди выбираются согласно FIFO (First In, First Out).

В табл. 1 для формального описания стационарных узлов используется классификация Кендалла-Башарина [3]: $A_i^X/B_i^Y/c_i/d_i$, где ^{X,Y} – законы распределения размеров поступающей и обслуживаемой группы заявок. Для описания управляемых СМО применяются следующие обозначения: $A_i(t)$ и $B_i(t)$ – параметры потока заявок и их обслуживания зависят от времени; $c_i(n_i)$, где c_i и n_i – минимальное и максимальное число работающих каналов; $d_i(X)$ – вместимость очереди меняется динамически и становится равной размеру поступающей или обслуживаемой группы заявок в узле.

Таблица 1

Модели работы структурных элементов порта в терминах ТМО

Элемент	Модель
Якорные и автомобильная стоянки, ворота и движение внутренних грузовиков	$A_i^X/B_i^X/c_i/0$
Склады, железнодорожная станция	$*/B_i^X/c_i/d_i$
Причал для загружаемых судов/барж	$*/B_i^Y/1/d_i(Y)$
Причал для разгружаемых судов/барж	$A_i^X/B_i^Y/c_i/d_i(X)$
Внешние транспортные средства, движущиеся между воторами и складами	$*/B_i^X/c_i(n_i)/0$

Маршруты движения грузов в порту описываются в маршрутной матрице P. Ее элементы $P_{i;j}, i, j \in S$ – вероятность перехода заявки из узла i в узел j – определяются как относительные частоты – отношение размера потока грузов, поступающих в определенную подсистему, к общему объему грузов в одном направлении.

В модели учитывается смена направления движения грузов с занятого специализированного причала (Узел a) на свободный многоцелевой (Узел b). Для этого значения элементов $P_{i,a}$ и $P_{i,b}$, описывающих движение по основному и альтернативному маршрутам, изменяются следующим образом:

$$(P_{i,a}, P_{i,b}) = \begin{cases} (1;0), E_a(t) \le F_a(t), \\ (0;1), E_a(t) > F_a(t), \end{cases}$$

где $E_a(t)$ – размер поступающей группы заявок и $F_a(t)$ – число свободных мест в очереди Узла a в момент времени t.

Между узлами также имеются обратные связи, которые заключаются во временной блокировке работы каналов текущего узла до тех пор, пока в следующем узле не освободится достаточно места для принятия заявок с текущего. Обратные связи применяются для описания поведения системы при перегрузках ее элементов.

Модель работы типового УМП на основе нестационарной СеМО имеет 18 узлов (рис. 1). Из них 4 узла фиктивные и описывают прибытие грузов с морских судов – O_1, речных барж – O_2, грузовиков – O_3 и поездов – O_4. Остальные узлы моделируют функционирование подсистем: 2, 5, 6, 10 и 14 – причальной зоны; 8 – складов; 3, 7, 9 и 13 – грузового фронта; 1, 4 – якорных стоянок, 11, 12 – автомобильной стоянки и же-

лезнодорожной станции. Узлы 2, 6, 13 являются управляемыми СМО, остальные – стационарными.

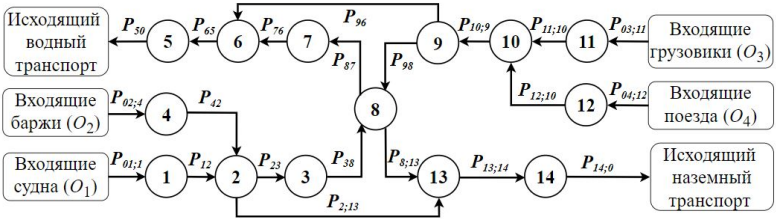


Рис. 1. Схема СеМО универсального морского порта

Полученная СеМО слишком сложна для аналитического исследования, поэтому ее анализ выполняется с помощью имитационной модели, которая основана на дискретно-событийном подходе моделирования и реализована в виде программы для ЭВМ [7]. Она предназначена для определения показатели эффективности СеМО, имеющей до 10 неординарных потоков заявок и до 100 узлов разных типов. В данном исследовании программа была усовершенствована для учета управляемых СМО в структуре выбранной СеМО.

В докладе будет более подробно представлена построенная обобщенная модель, а также ее идентификация для универсального порта Куангнинь, одного из крупнейших во Вьетнаме. Также будут освещены результаты численного исследования полученной модели и сделанные на их основе выводы о работе рассматриваемого объекта.

Заключение

В данном исследовании представлена обобщенная математическая модель функционирования универсального морского порта, в котором происходит смена технологии и скорости погрузки/разгрузки судов в зависимости от типа обрабатываемого груза. Модель построена на основе управляемой сети массового обслуживания, за счет чего в ней учитывается влияние гетерогенной среды на параметры и работу подсистем порта, а также возможен выбор маршрута движения товаров в зависимости от загрузки основных причалов.

Перспективным направлением дальнейших исследования является повышение детализации предложенной модели за счет учета различных сбоев в работе порта (выхода из строя оборудовани, отмены прибытия судна) и перерывов между сменами.

СПИСОК ЛИТЕРАТУРЫ

1. *Гнеденко Б. В., Коваленко И. Н.* Введение в теорию массового обслуживания. М.: ЛКИ, 2007. 400 с.
2. *Edmond E. D., Maggs R. P.* How useful are queue models in port investment decisions for container berths? // Journal of the Operational Research Society. 1978. Vol. 29, № 8. P. 741–750.
3. *Ивницкий В. А.* Теория сетей массового обслуживания. М.: ФИЗМАТЛИТ, 2004. 772 с.
4. *Zhang X., Zeng Q., Sheu J.* Modeling the productivity and stability of a terminal operation system with quay crane double cycling // Transportation Research Part E: Logistics and Transportation Review. 2019. Vol. 122. P. 181–197.
5. *Xiang X., Liu C.* Modeling and analysis for an automated container terminal considering battery management // Computers & Industrial Engineering. 2021. Vol. 156. P. 107258.
6. *Kazakov A., Lempert A., Zharkov M.* Modeling of a coal transshipment complex based on a queuing network // Applied Sciences. 2024. Vol. 14, № 16. P. 6970.
7. *Лемперт А. А., Жарков М. Л., Казаков А. Л., Ву Х. З.* Моделирование морского контейнерного терминала с использованием сети массового обслуживания // Управление большими системами. 2024. № 112. С. 310–337.
8. *Kazakov A., Lempert A., Vu G., Zharkov M.* On the application of a queuing network to simulate container transshipment in a sea container termina // In: Dudin A., Nazarov A., Moiseev A. (eds) Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM WRQ 2024. Communications in Computer and Information Science. Springer, Cham. 2024. Vol. 2472. P. 1–16.
9. *Назаров А. А.* Управляемые системы массового обслуживания. Томск: Национальный исследовательский Томский государственный университет, 1984. 234 с.

Хоанг Занг Ву — аспирант. E-mail: *giangtrang14022016@gmail.com*

Жарков Максим Леонидович — к.т.н., с.н.с., Лаб. 6.2. Логических и оптимизационных методов анализа сложных систем. E-mail: *zharkm@mail.ru*

Казаков Александр Леонидович — д.ф.-м.н., г.н.с., Отд. 2. Прикладных проблем математической физики и теории поля. E-mail: *kazakov@icc.ru*

Лемперт Анна Ананьевна — к.ф.-м.н., зав.лаб., Лаб. 6.2. Логических и оптимизационных методов анализа сложных систем. E-mail: *lempert@icc.ru*

ХАРАКТЕРИСТИКИ ОБЩЕЙ СУММЫ ОБРАЗУЮЩИХ НЕЗАТУХАЮЩИЙ ПОТОК ВЗНОСОВ НА НАКОПИТЕЛЬНЫЙ СЧЁТ

Д. Д. Даммер[1], Е. А. Гренц[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В работе рассматривается модель функционирования финансовой организации в виде системы массового обслуживания с бесконечным количеством приборов. Предполагается, что клиенты открывают накопительные счета и совершают на них взносы в случайные моменты времени. Поток взносов моделируется дополнительным потоком событий, сгенерированным каждым открывшим счет клиентом. Исследование направлено на изучение характеристик общей суммы денежных средств, поступивших на накопительные счета за фиксированный период времени. Полученные результаты могут быть использованы при прогнозировании денежных потоков и управлении рисками в пенсионных фондах, банках и других коммерческих организациях.

Ключевые слова: *система массового обслуживания, накопительный счет, финансовая организация, производящая функция, характеристики суммы денежных взносов*

Введение

Современные финансовые организации сталкиваются с необходимостью моделирования и прогнозирования денежных потоков, поступающих от клиентов. Эти процессы характеризуются высокой степенью неопределённости, связанной с моментами открытия накопительных счетов, частотой их пополнения, а также пополняемой суммой. Для анализа подобных систем всё шире применяются методы теории массового обслуживания [1–4], позволяющие учитывать случайный характер протекающих внутри организации процессов и влияющих на ее деятельность.

Особый интерес представляет исследование моделей, в которых помимо основного потока заявок (открытий счётов) существует дополнительный поток событий, соответствующий повторным взаимодействиям заявок с системой (пополнению накопительных счетов). Анализ характеристик суммарного объёма пополнений на счета позволяет оценивать

финансовую устойчивость организации. Системы, моделирующие работу различных коммерческих организаций с такого рода дополнительными потоками, рассматриваются в работах [5, 6, 7, 8], где находится, например, характеристическая функция числа событий такого потока и анализируется его поведение в рамках численных экспериментов.

1. Постановка задачи и описание модели

Рассмотрим модель работы финансовой организации, например, пенсионного фонда, коммерческого банка или инвестиционной платформы, в которой клиенты открывают накопительные счета на определённый срок. После открытия счёта клиент может совершать взносы любой величины в случайные моменты времени в течение этого срока. Задача состоит в нахождении числовых характеристик общей суммы пополнения на накопительный счет от всех клиентов, открывших счета в период $(-\infty, T]$, $T > 0$, и совершивших его пополнение на промежутке $[0, T]$. Стохастический характер рассматриваемых процессов позволяет моделировать работу финансовой организации в виде системы массового обслуживания.

Будем считать, что счет может открыть любой приходящий в организацию клиент, то есть система имеет бесконечное количество приборов; клиенты (заявки) образуют входящий поток, который является пуассоновским с параметром λ. Срок, на который может открываться счет, т.е. длительность обслуживания на приборе, является экспоненциально распределенной с параметром μ случайной величиной. Будем считать, что моменты пополнения накопительного счета одним клиентом образуют дополнительный поток событий с параметром γ; такой поток будем называть локальным d-потоком (рис. 1). Суммарным d-потоком будем называть дополнительный поток, сформированный всеми клиентами.

В нашей модели будем учитывать тех клиентов, которые открыли счета на промежутке времени $(-\infty, T]$, $T > 0$, и те события суммарного d-потока, которые наступили на промежутке $[0, T]$, $T > 0$. Эту ситуацию демонстрирует рис. 2: моментами времени t_i обозначены моменты прихода (открытия счета) клиентов в организацию, точками – моменты очередного взноса, пополняющего счет. Так как события d-потока учитываются на ограниченном промежутке $[0, T]$, то такой поток будем называть незатухающим.

2. Производящая функция общей суммы пополнений

Определим величину $n(t)$ – число событий дополнитиельного потока взносов на накопительный счет от клиентов, пришедших в организацию

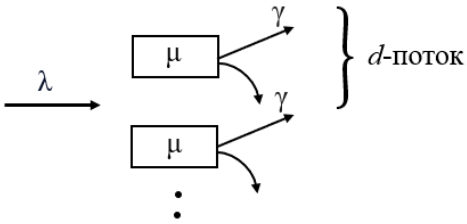


Рис. 1. Модель системы с неограниченным числом приборов и дополнительным потоком

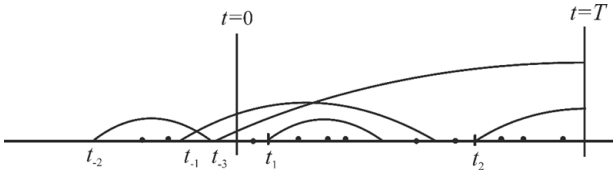


Рис. 2. Схема формирования потока взносов (d-потока) на накопительный счет

на промежутке $(-\infty, t]$. В работе [8] была рассмотрена система массового обслуживания по типу модели рис. 1, и с использованием метода марсковского суммирования найдена характеристическая функция числа $n(T)$ событий суммарного d-потока, которая имеет следующий вид:

$$H(u,T) = \exp\left\{\lambda\left[\left(e^{c(u)T}-1\right)\frac{1}{\mu}+\right.\right.$$

$$\left.\left.+\frac{\mu}{c^2(u)}\left(1-e^{c(u)T}\right)-\left(\frac{\mu}{c(u)}+1\right)T\right]\right\}, \qquad (1)$$

где $c(u) = \gamma(e^{ju} - 1) - \mu$.

Теперь введем производящую функцию $\Phi(z,t)$ числа взносов $n(t)$. Из (1) несложно записать ее выражение, предварительно обозначив $c_1(z) = \gamma(z-1) - \mu$:

$$\Phi(z,T) = \exp\left\{\lambda\left[\left(e^{c_1(z)T}-1\right)\frac{1}{\mu}+\right.\right.$$

$$\left.\left.+\frac{\mu}{c_1^2(z)}\left(1-e^{c_1(z)T}\right)\quad-\left(\frac{\mu}{c_1(z)}+1\right)T\right]\right\}. \qquad (2)$$

Пусть $S(t)$ есть общая сумма денежных средств, поступивших на накопительные счета от клиентов, открывших счета за период $[-\infty, t]$, а случайная величина ξ характеризует размер одного взноса. Введём характеристическую функцию (в смысле преобразования Лапласа) суммарной величины поступлений $S(t)$, определив её следующим образом: $\Psi(\alpha, t) = M\{e^{-\alpha S(t)}\}$. Найдем выражение для этой функции. Имеем:

$$\Psi(\alpha,t) = M\left\{e^{-\alpha S(t)}\right\} = M\left\{e^{-\alpha\sum_{i=1}^{n(t)}\xi_i}\right\} =$$

$$= \sum_{n=0}^{\infty} M\left\{e^{-\alpha\sum_{i=1}^{n(t)}\xi_i}\middle| n(t)=n\right\} P(n,t) =$$

$$= \sum_{n=0}^{\infty} M\left\{\prod_{i=1}^{n(t)} e^{-\alpha\xi_i}\middle| n(t)=n\right\} P(n,t) = \sum_{n=0}^{\infty}\varphi^n(\alpha)P(n,t),$$

где $\varphi(\alpha) = M\{e^{-\alpha\xi}\}$ – характеристическая функция величины ξ и $P(n,t) = \mathbf{P}\{n(t)=n\}$.Таким образом, можем записать характеристическую функцию величины $S(t)$ как производящую функция числа $n(t)$ аргумента $\varphi(\alpha)$:

$$\Psi(\alpha,t) = \sum_{n=0}^{\infty}\varphi^n(\alpha)P(n,t) = \Phi(\varphi(\alpha),t),$$

и с учетом (2) вида функции $\Phi(\varphi(\alpha),t)$ для $t=T$ получим

$$\Psi(\alpha,T) = \exp\left\{\lambda\left[\left(e^{(\gamma(\varphi(\alpha)-1)-\mu)T}-1\right)\cdot\frac{1}{\mu}+\right.\right.$$

$$+\frac{\mu}{(\gamma(\varphi(\alpha)-1)-\mu)^2}\cdot\left(1-e^{(\gamma(\varphi(\alpha)-1)-\mu)T}\right)-$$

$$\left.\left.-\left(\frac{\mu}{\gamma(\varphi(\alpha)-1)-\mu}+1\right)\cdot T\right]\right\}.$$

3. Числовые характеристики общей суммы пополнений

Пусть $M\{\xi\} = a_1$, $M\{\xi^2\} = a_2$. Найдем математическое ожидание величины общей суммы $S(T)$ дополнительных взносов на накопительные счета клиентов. Так как $\left.\frac{\partial\Psi(\alpha,T)}{\partial\alpha}\right|_{\alpha=0} = -M\{S(T)\}$, то после преоб-

разований получим выражение для математического ожидания исследуемой величины:

$$M\{S(T)\} = \frac{a_1 \cdot \gamma \cdot \lambda \cdot \left(\mu \cdot T + 2 - 2e^{-\mu \cdot T}\right)}{\mu^2}. \tag{3}$$

Так как $\left.\frac{\partial^2 \Psi(\alpha,T)}{\partial \alpha^2}\right|_{\alpha=0} = M\{S^2(T)\}$, то дисперсия $S(T)$ будет иметь вид:

$$D\{S(T)\} = \frac{\gamma\lambda}{\mu^3} \cdot \Big[Ta_2\mu^2 + 6a_1^2\gamma + 2\mu(Ta_1^2\gamma + a_2) - \\ -6a_1^2\gamma e^{-\mu T} - 2\mu(2Ta_1^2\gamma + a_2)e^{-\mu T}\Big]. \tag{4}$$

4. Коэффициент вариации

Для анализа поведения величины $S(T)$ рассмотрим также коэффициент вариации $V\{S(T)\}$, который показывает разброс значений случайной величины $S(T)$ относительно ее математического ожидания: $V\{S(T)\} = \frac{\sqrt{D\{S(T)\}}}{M\{S(T)\}}$. Рассмотрим поведение $V\{S(T)\}$ при $a_1 = 10$, $a_2 = 110$ с учетом полученных выражений (3) и (4).

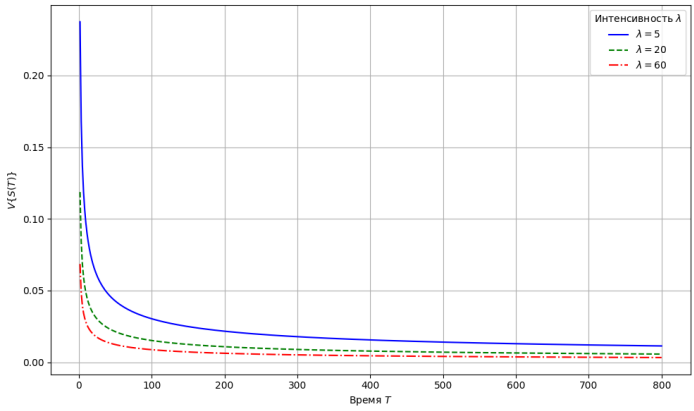


Рис. 3. Динамика изменения коэффициента вариации $V\{S(T)\}$ при различных значениях λ

На рис. 3 изображены кривые для значений λ, равных $5, 20, 60$, $\mu = 0,0002$, $\gamma = 0,01$. Можно заметить, что коэффициент вариации $V\{S(T)\}$ убывает со временем для всех значений λ, что указывает на

стабилизацию процесса накопления средств с течением времени. Другими словами, с увеличением T поведение суммарного объёма пополнения средств на счетах становится всё более предсказуемым, и разброс относительно среднего значения уменьшается. С увеличением λ стабилизация процесса происходит быстрее.

Заключение

В данной работе была исследована модель финансовой организации, функционирующая как система массового обслуживания с бесконечным числом приборов и дополнительным потоком событий. Такой подход позволил описать динамику суммарной величины денежных средств, поступающих на накопительные счета за конечный временной период. На основе полученной ранее характеристической функции числа событий дополнительного потока были найдены выражения для производящей функции, математического ожидания и дисперсии общей суммы пополнений. Представлена графическая визуализация поведения коэффициента вариации исследуемой величины при различных значениях параметров системы. Полученные результаты могут быть использованы для управления финансовой деятельностью коммерческих организаций.

СПИСОК ЛИТЕРАТУРЫ

1. *Forozandeh M.* Optimizing the banking service system using queue theory, fuzzy DEMATEL and TOPSIS approach: Case study // Annals of Human Resource Management Research. 2022. № 2. P. 87–104.
2. *Afeche P.* Incentive-compatible revenue management in queueing systems: Optimal strategic delay // Manufacturing and Service Operations Management. 2013. № 3. P. 423–443.
3. *Dhar R., Chakraborty A., Sarkar J.* Queuing Theory and Optimization in Banking Models // Cybernetics, Cognition and Machine Learning Applications. Algorithms for Intelligent Systems. Singapore: Springer, 2020. P. 205–213.
4. *Dammer D.* Research of Mathematical Model of Insurance Company in the Form of Queueing System in a Random Environment // Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2017. Communications in Computer and Information Science. Cham: Springer, 2017. Vol. 800. P. 204–214.
5. *Nazarov A., Dammer D.* Methods of Limiting Decomposition and Markovian Summation in Queueing System with Infinite Number of Servers // Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM WRQ 2018. Communications in Computer and Information Science. Cham: Springer, 2018. Vol. 912. P. 71–82.

6. *Даммер Д.* Исследование дополнительно формируемого потока на бесконечном интервале в системе с экспоненциальным обслуживанием // Информационные технологии и математическое моделирование (ИТММ-2023): материалы XXII Международной конференции имени А.Ф. Терпугова, Томск, 4–9 декабря 2023 г. Томск: Изд-во Том. ун-та, 2023. С. 89–94.
7. *Шкленник М., Моисеев А.* Метод марковского суммирования для исследования потока повторных обращений в двухфазных системах $M|GI|\infty \to GI|\infty$ // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2021. Т. 21, № 1. С. 125–137.
8. *Даммер Д., Гренц Е.* Исследование незатухающего дополнительного потока в системе с простейшим входящим потоком методом марковского суммирования // ИТ. Наука. Креатив: материалы II Междунар. форума (Россия, Омск, 13–15 мая 2025 г.) С. 441–451.

Даммер Диана Даммировна — канд. физ.-мат. наук, доцент, доцент, Институт прикладной математики и компьютерных наук. E-mail: *di.dammer@yandex.ru*

Гренц Елизавета Алексеевна — студент, Институт прикладной математики и компьютерных наук. E-mail: *grents.el@yandex.ru*

СВОЙСТВА ВЛОЖЕННОЙ ЦЕПИ МАРКОВА ДЛЯ ПРОЦЕССА ТИПА $M(t)/M(t)/1/\infty$ С ПЕРИОДИЧЕСКИМИ ИНТЕНСИВНОСТЯМИ

А. В. Зорин[1]

[1] *Национальный исследовательский Нижегородский государственный университет, г. Нижний Новгород, Россия*

Рассматривается однолинейная система массового обслуживания с неограниченным числом мест ожидания, в предположении что интенсивность входного потока и интенсивности обслуживания зависят от времени периодическим образом. Кроме того, интенсивность обслуживания разная для очереди из одиночного требования и очереди из двух и более требований. Рассматривается вложенная в моменты изменения длины очереди цепь Маркова для расширенного процесса, включающего число требований и фазу на периоде. Такая цепь Маркова будет иметь несчетное пространство состояний. В работе устанавливается ее неприводимость, находятся некоторые минорантные множества, доказывается достаточное условие существования стационарного распределения. **Ключевые слова:** *периодические нестационарные интенсивности, неоднородный марковский процесс, вложенная цепь, общая цепь Маркова, условие существования стационарного распределения, интегративно-мажорантный метод*

1. Проблема изучения неоднородных марковских моделей

Рассматривается однолинейная система обслуживания. Нестационарный, однородный входной поток без последействия задается мгновенной интенсивностью $\lambda(t)$ в момент $t \geqslant 0$. Требования помещаются в очередь неограниченной вместимости в порядке поступления. Если в очереди в момент t находится одно требование, то обслуживание осуществляется с интенсивностью $\mu_1(t)$, а если два или больше требований, то с интенсивностью $\mu_2(t)$. Обслуженное требование мгновенно покидает систему. По-видимому, одной из первых работ по теории систем массового обслуживания с переменными интенсивностями является работа А. Н. Колмогорова 1931 г. [1], в которой выписаны соответствующие дифференциальные уравнения. Решение уравнений для аналога системы $M/M/1$ даны в работе [2], однако при дополнительном предположении о характере зависимости входной интенсивности $\lambda(t)$ от времени.

В общем случае, задача определения вероятностей на переходном процессе является весьма трудной. Поэтому разрабатываются косвенные методы. Так, асимптотический метод анализа применялся в работе [4].

В данной работе описанная только что система массового обслуживания исследуется в предположении периодичности интенсивностей: $\lambda(t+1) = \lambda(t)$, $\mu_k(t+1) = \mu_k(t)$, $k = 1, 2$, для всех $t \geqslant 0$. Периодический характер интенсивности потоков является часто наблюдаемым явлением. Например, в работе [3] приводится график интенсивности прибытия самолетов в аэропорт. Наш интерес к данной задаче вызван тем, что эта модель является естественным обобщением одной из моделей образования транспортной пачки на автомагистрали [5], когда скорость обгона быстрой машиной медленной машины зависит от размера всей группы. Обобщение получается, если заметить, что интенсивность автотранспортного потока испытывает суточные колебания.

Основной математический аппарат для подобных задач — неоднородные по времени процессы Маркова со счетным числом состояний [6] или скачкообразные марковские процессы [7]. Оказывается, что поведение решений дифференциальных уравнений Колмогорова для состояний процесса может быть «хорошим», но уже в двух разных смыслах: либо существует предельное распределение (как и в однородном случае), либо пределы при всех t не существуют, но любые два решения сближаются друг с другом с ростом времени (так называемая, слабая эргодичность). Для периодических коэффициентов это может означать существование пределов у подпоследовательностей и наличие периодического решения исходной системы ОДУ. Однако разыскание этого периодического решения для процесса Маркова с бесконечным счетным числом состояний также может представлять трудность.

Обозначим через $X(t)$ число требований в системе в момент $t \geqslant 0$; тогда $\{X(t); t \geqslant 0\}$ будет неоднородным марковским процессом со счетным числом состояний. Будем считать его реализации непрерывными справа. Стандартным приемом устанавливается система Колмогорова обыкновенных дифференциальных уравнений для вероятностей $p_x(t) = \mathbb{P}(X(t) = x)$, $x = 0, 1, \ldots$:

$$\begin{aligned}
\frac{dp_0}{dt}(t) &= -\lambda(t)p_0(t) + \mu_1(t)p_1(t), \\
\frac{dp_1}{dt}(t) &= \lambda(t)p_0(t) - (\lambda(t) + \mu_1(t))p_1(t) + \mu_2(t)p_2(t), \\
\frac{dp_x}{dt}(t) &= \lambda(t)p_{x-1}(t) - (\lambda(t) + \mu_2(t))p_x(t) + \\
&\qquad + \mu_2(t)p_{x+1}(t), \quad x = 2, 3, \ldots
\end{aligned} \tag{1}$$

Введением дополнительной компоненты $J(t) = \{t\}$, где $\{\cdot\}$ обозначает дробную часть (по модулю 1) числа, указанного в скобках, мы можем перейти к *однородному* марковскому процессу $\{(X(t), J(t)); t \geqslant 0\}$ с фазовым пространством $\{0, 1, \ldots\} \times [0, 1)$. Обозначим

$$p_x(t, u)\, du = d_u \mathbb{P}(X(t) = x, J(t) < u), \qquad 0 \leqslant u < 1,$$

тогда получим следующую систему уравнений для плотности вектора $(X(t), J(t))$:

$$\begin{aligned}
&\frac{\partial p_0}{\partial t}(t,u) + \frac{\partial p_0}{\partial u}(t,u) = -\lambda(u)p_0(t,u) + \mu_1(u)p_1(t,u),\\
&\frac{\partial p_1}{\partial t}(t,u) + \frac{\partial p_1}{\partial u}(t,u) = \lambda(u)p_0(t,u) - (\lambda(u) + \mu_1(u))p_1(t,u) +\\
&\qquad\qquad + \mu_2(u)p_2(t,u),\\
&\frac{\partial p_x}{\partial t}(t,u) + \frac{\partial p_x}{\partial t}(t,u) = \lambda(u)p_{x-1}(t,u) - (\lambda(u) + \mu_2(u))p_x(t,x) +\\
&\qquad\qquad + \mu_2(u)p_{x+1}(t,u), \quad x = 2, 3, \ldots
\end{aligned} \tag{2}$$

Видно, что стационарное решение $p_x(u) = p_x(t, u)$ уравнений (2) также является решением уравнений (1). В то же время, процесс $\{(X(t), J(t)); t \geqslant 0\}$ принадлежит семейству однородных марковских процессов и выделяется тем условием, что его начальное распределение сосредоточено на множестве $\{0, 1, \ldots\} \times \{0\}$. Стационарное распределение, соответствующее переходной вероятности процесса $\{(X(t), J(t)); t \geqslant 0\}$, может приписывать некоторое невырожденное распределение компоненте $J(0)$. Тем не менее, эргодические свойства функционалов от процесса $\{X(t); t \geqslant 0\}$, будут определяться именно этим стационарным распределением расширенного однородного процесса $\{(X(t), J(t)); t \geqslant 0\}$.

2. Свойства вложенной цепи скачков

Эргодические свойства процесса $\{(X(t), J(t)); t \geqslant 0\}$ должны быть связаны с аналогичными свойствами вложенной цепи Маркова в моменты изменения длины очереди. Рассмотрим соответствующую вложенную цепь и применим для ее анализа итеративно-мажорантный метод [8–10]. Пусть $\lambda(t) \geqslant \delta > 0$, $\mu_1(t) \geqslant \delta > 0$, $\mu_2(t) \geqslant \delta > 0$ для некоторого $\delta > 0$. Введем последовательность моментов скачков процесса $\{X(t); t \geqslant 0\}$:

$$\tau_0 = 0, \qquad \tau_{n+1} = \inf\{t : t > \tau_n, X(t) \neq X(\tau_n)\}, \quad n = 0, 1, \ldots.$$

Введем случайные последовательности $X_n = X(\tau_n)$, $J_n = J(\tau_n)$, $n = 0, 1, \ldots$. Из строго марковского свойства процесса $\{(X(t), J(t)); t \geqslant 0\}$ следует, что последовательность

$$\{(X_n, J_n); n = 0, 1, \ldots\} \tag{3}$$

будет общей цепью Маркова [11]. Введем совместную (дискретно-непрерывную) плотность переходного распределения вероятностей

$$p_{x,w}(u, v)dv = \mathbb{P}(X_{k+1} = w, J_{k+1} \in dv \mid X_k = x, J_k = u).$$

Обозначим

$$\Lambda(u) = \int_0^u \lambda(t)\, dt, \qquad M_k(u) = \int_0^u \mu_k(t)\, dt,\ k = 1, 2.$$

Теорема 1. Переходная плотность цепи Маркова (3) имеет следующий вид: при $u < v$

$$p_{0,1}(u, v) = \frac{e^{-(\Lambda(v)-\Lambda(u))}}{1 - e^{-\Lambda(1)}},$$
$$p_{x,x-1}(u, v) = \frac{\mu_k(v)e^{-(\Lambda(v)+M_k(v)-\Lambda(u)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}},$$
$$p_{x,x+1}(u, v) = \frac{\lambda(v)e^{-(\Lambda(v)+M_k(v)-\Lambda(u)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}},$$

а при $u > v$

$$p_{0,1}(u, v) = \frac{\lambda(v)e^{-(\Lambda(1+v)-\Lambda(u))}}{1 - e^{-\Lambda(1)}},$$
$$p_{x,x-1}(u, v) = \frac{\mu_k(v)e^{-(\Lambda(1+v)+M_k(1+v)-\Lambda(u)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}},$$
$$p_{x,x+1}(u, v) = \frac{\lambda(v)e^{-(\Lambda(1+v)+M_k(1+v)-\Lambda(u)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}},$$

где $k = 1$ для $x = 1$ и $k = 2$ для $x \geqslant 2$. Данная общая цепь Маркова является ψ-неприводимой. Множества вида $\{x\} \times [0, 1)$, $x = 0, 1, \ldots$ являются минорантными [11].

Рассмотрим в некоторой области $|z| \leqslant 1 + \varepsilon$, $\varepsilon > 0$, производящие функции $\Psi_n(z) = \mathbb{M}(z^{X_n})$, $n = 0, 1, \ldots$. Введем обозначения:

$$q_k(z, u) = \int_u^{1+u} (\mu_k(v)z^{-1} + \lambda(v)z)\frac{e^{-(\Lambda(v)-\Lambda(u)+M_k(v)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}}\, dv, \quad k = 1, 2,$$

$$\rho_k(u) = \int_u^{1+u} (\lambda(v) - \mu_k(v)) \frac{e^{-(\Lambda(v)-\Lambda(u)+M_k(v)-M_k(u))}}{1 - e^{-\Lambda(1)-M_k(1)}}\, dv, \qquad k = 1, 2,$$

$$\rho_{2,n+1}(u) = \int_u^{1+u} \rho_{2,n}(v)(\lambda(v) + \mu_2(v)) \frac{e^{-(\Lambda(v)-\Lambda(u)+M_2(v)-M_2(u))}}{1 - e^{-\Lambda(1)-M_2(1)}}\, dv,$$

для $n = 1, 2, \ldots$ и $\rho_{2,1}(u) = \rho_2(u)$.

Теорема 2. Справедливы рекуррентные по $i = 0, 1, \ldots$ соотношения

$$\Psi_{n+1}(z) = \mathbb{M}(q_2(z, J_n) z^{X_n}) + \int_0^1 \mathbb{P}(X_n = 1, J_n \in du)(q_1(z, u) - q_2(z, u)) + \\ + \int_0^1 \mathbb{P}(X_n = 0, J_n \in du)(z - q_2(z, u)).$$

Теорема 3. Для существования стационарного распределения достаточно выполнения неравенства

$$\int_0^1 (\lambda(t) - \mu_2(t))\, dt < 0 \qquad (\Lambda(1) < M_2(1)).$$

При доказательстве существенно используются следующие утверждения.

Лемма 1. Для того, чтобы у цепи Маркова (3) не существовало стационарного распределения, необходимо, чтобы последовательность средних $\{\mathbb{M}(z^{X_n}); n = 0, 1, \ldots\}$ неограниченно возрастала.

Лемма 2. Имеет место равномерная по $u \in [0, 1]$ сходимость

$$\frac{1}{n}(\rho_2(u) + \rho_{2,2}(u) + \rho_{2,3}(u) + \ldots + \rho_{2,n}(u)) \to \Lambda(1) - M_2(1).$$

Заключение

Используя итеративно-мажорантный метод, получено легко интерпретируемое достаточное условие существования стационарного распределения у вложенной цепи в неоднородный марковский процесс. Это существенно расширяет области применения данного метода к исследованию моделей процессов обслуживания.

СПИСОК ЛИТЕРАТУРЫ

1. *Колмогоров А. Н.* Sur la problème d'attante // Математический сборник. 1931 Т. 38. № 1/2. С. 101–106.
2. *Knessl C., Yongzhi P. Y.* An exact solution for an $M(t)/M(t)/1$ queue with time-dependent arrivals and service // Queueing systems. 2002, № 40. P. 233–245.
3. *Koopman B. O.* Air-Terminal Queues under Time-Dependent Conditions // Operations Research. 1972. Vol. 20, № 6. P. 1089–1114.
4. *Keller J. B.* Time-dependent queues // SIAM Review. 1982. Vol. 24, № 4. P. 401–412.
5. *Федоткин М.А., Федоткин А. М., Кудрявцев Е. В* Динамические модели неоднородного потока транспорта на магистралях // Автоматика и телемеханика. 2020. Вып. 8. С. 149–164.
6. *Зейфман А. И., Бенинг В. Е., Соколов И. А.* Марковские цепи и модели с непрерывным временем. М.: ЭЛЕКС-КМ, 2008. 168 с.
7. *Гихман И. И.. Скороход А. В.* Теория случайных процессов. Т. II. М.: Наука, 1973. 640 с.
8. *Зорин А. В.* О циклическом обслуживании неординарных рекуррентных потоков // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с муждународным участием имени А. Ф. Терпугова (29—30 ноября 2013 г.). Ч. 2. Томск: Издательство Томского университете, 2013. С. 15—20.
9. *Zorine A. V.* On ergodicity conditions in a polling model with Markov modulated input and state-dependent routing // Queueing Systems. 2014. Vol. 76, № 2. P. 223–241.
10. *Zorine A. V.* Controlled $M/M/1 - RQ$ system with randomized acceptance from the orbit // Queueing Theory and Network Applications. QTNA 2019. Lecture Notes in Computer Science. V. 11688. Cham: Springer, 2019. P. 64–76.
11. *Meyn S. P., Tweedie R. L.* Markov chains and stochastic stability. London: Springer, 1993. 550 p.

Зорин Андрей Владимирович — доктор физико-математических наук, доцент, заведующий кафедрой теории вероятностей и анализа данных ННГУ им. Н.И. Лобачевского. E-mail: *andrei.zorine@itmm.unn.ru*

ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ RQ-СИСТЕМЫ M|M|1 С ИСКАЖЕНИЕМ ЗАЯВОК И ИХ ВОССТАНОВЛЕНИЕМ

Я. Е. Измайлова, З. Канылбай

Национальный исследовательский Томский государственный университет, г. Томск, Россия

В данной работе рассматривается RQ-система с искажениями заявок, орбитой восстановления и простейшим потоком входящих заявок. Под «искажением» подразумевается, что заявка может прерваться во время обслуживания, а орбита восстановления предназначена для искаженных заявок. Реализована имитационная модель функционирования системы и получено двумерное распределение вероятностей числа заявок на орбитах ожидания и восстановления. **Ключевые слова:** *система массового обслуживания, повторные вызовы, RQ-система , искажение заявок, дискретно-событийное моделирование*

Введение

RQ-системы массового обслуживания или системы с повторными вызовами представляют собой модели, в которых требования, не получившие обслуживания, уходят на так называемую «орбиту» – виртуальное пространство, из которого они предпринимают повторные попытки получить обслуживание через случайные промежутки времени. Такие системы широко применяются для анализа телекоммуникационных сетей, управления трафиком и передачей данных, где повторные обращения — обычное явление. Обзор подходов к изучению RQ-систем представлен в [1].

RQ-системы с отрицательными заявками расширяют базовую модель, добавляя заявки, которые не требуют обслуживания, но воздействуют на систему, например, прерывая или удаляя положительные заявки. Концепция отрицательных заявок была предложена Геленбе [2] и развита в работах X. Chao и др. [4], а также Gelenbe E. и др. [3]. Исследования Artalejo J. R. и Gomez-Corral A. [5] и [6] демонстрируют применение отрицательных заявок в RQ-системах.

1. Описание модели

В рассматриваемой RQ-системе предполагается, что в систему поступают два независимых простейших потока заявок: «положитель-

ных» с интенсивностью λ и «отрицательных» с интенсивностью γ (рис. 1). Обслуживание производится одним прибором. Время обслуживания одной заявки имеет экспоненциальное распределение с параметром μ. При поступлении положительной заявки, если прибор свободен, она приступает к обслуживанию. В противном случае заявка уходит на орбиту ожидания.

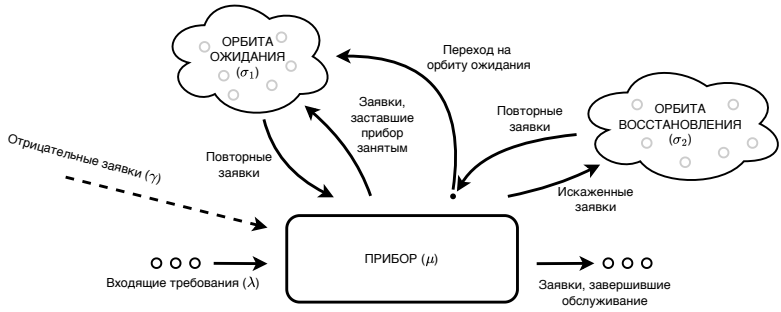


Рис. 1. RQ-система с искажениями заявок

Поступившая отрицательная заявка, если прибор занят, искажает обслуживание текущей положительной заявки, текущая заявка прерывается и уходит на орбиту восстановления, а отрицательная заявка покидает систему. В случае, когда прибор свободен, отрицательная заявка не обслуживается и также покидает систему.

Заявки, находящиеся на орбитах ожидания и восстановления, совершают независимые попытки захватить прибор через экспоненциально распределённые интервалы времени с параметром σ_1 и σ_2 соответственно. Если при попытке захвата прибор свободен, заявка занимает его и начинает обслуживание. Заявки из орбиты восстановления после первой попытки занять прибор переходят на орбиту ожидания. Также заявки, находящиеся на орбите ожидания и заставшие прибор занятым, остаются на той же орбите и осуществляют следующую экспоненциально распределённую задержку с параметром σ_1.

Цель исследования данной RQ-системы с искажениями заявок и орбитой восстановления – нахождение совместного распределения вероятностей числа заявок на орбитах восстановления и ожидания в стационарном режиме функционирования ($t \to \infty$).

Построенную математическую модель можно формализовать с помощью марковских цепей, которая описывает динамику работы систе-

мы. Построим граф вероятностей переходов, где номера, указанные на вершинах графа Рис. 2, соответствуют состоянию прибора: 0 – прибор свободен, 1 – прибор занят. Столбцы графа отражают количество заявок на орбите ожидания, а строки — число заявок на орбите восстановления.

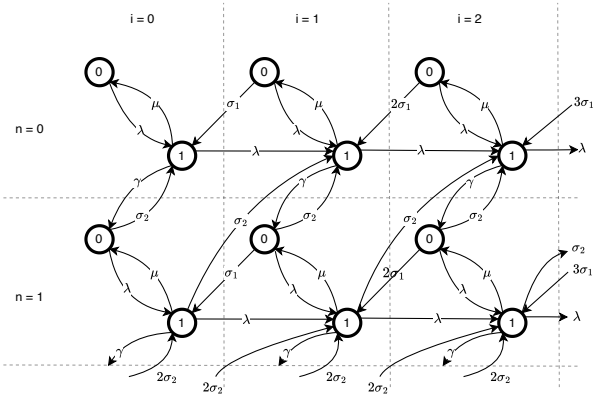


Рис. 2. Граф вероятностей перехода

В дальнейшем состояние системы будем обозначать в виде (i, n, k), где i – количество заявок на орбите ожидания, n – количество заявок на орбите восстановления, k – состояние прибора.

2. Имитационное моделирование

Для анализа применяется стохастическое дискретно-событийное моделирование (DES). DES представляет систему как последовательность мгновенных событий. Время в DES продвигается скачкообразно, переходя от одного события к следующему.

Реализованный алгоритм 1 основан на продвижении времени к моменту наступления ближайшего события. Для исследуемой системы основными событиями являются:

- Прибытие положительной заявки.
- Искажение обслуживаемой заявки.
- Завершение обслуживания.
- Захват прибора заявкой с орбиты ожидания/восстановления.
- Переход заявки из орбиты восстановления на орбиту ожидания.

Время обслуживания, интервалы времени между поступлением положительных и отрицательных заявок, а также интервалы между попытками захватами прибора заявками из орбит имеют экспоненциальное распределение. Случайное время T_i до наступления события вычисляется по формуле:

$$T_i = \frac{ln(R)}{\beta_i},$$

где R – равномерно распределенная случайная величина на интервале $(0, 1)$, β_i – параметр соответствующего события. При этом, если переход по соответствующему типу события из текущего состояния невозможен, полагается $T_i = +\infty$

Алгоритм 1. Дискретно-событийное моделирование СМО

Вход: Максимальное время моделирования $T_{\max}$ и параметры $\lambda, \gamma, \mu, \sigma_1, \sigma_2$

Выход: Матрица распределения $\hat{P}(i, n, k)$, массивы времени и состояний системы

Время моделирования: $t \leftarrow 0$

Текущее состояние $S \leftarrow (0, 0, 0)$

пока $t < T_{\max}$

Найти наиболее ранее событие $S_{\min}$ с временем:

$$T_{\min} \leftarrow \min\left\{\frac{ln(R)}{\lambda}, \frac{ln(R)}{\gamma}, \frac{ln(R)}{\mu}, \frac{ln(R)}{\sigma_1}, \frac{ln(R)}{\sigma_2}\right\}$$

Обновить время, проведенное в данном состоянии:

$$\hat{P}[S] \leftarrow \hat{P}[S] + T_{\min}$$

Обновить время моделирования: $t \leftarrow t + T_{\min}$

Выполнить переход до ближайшего выбранного события:

$$S \leftarrow S_{\min}$$

Сохранить статистику о состоянии S и времени t

$\hat{P} \leftarrow \hat{P}/t$

Как можно заметить, вероятность $\hat{P}(i, n, k)$ вычисляется как отношение времени, проведенного системой в данном состояний (i, n, k), к общему времени моделирования. А одномерные распределения числа заявок на орбитах определяются, как:

$$\hat{P}(i) = \sum_{n,k\geqslant 0} \hat{P}(i,n,k), \quad \hat{P}(n) = \sum_{i,k\geqslant 0} \hat{P}(i,n,k). \tag{1}$$

Имитационная модель была проведена с использованием следующих параметров: $\lambda = 0,8, \gamma = 2, \mu = 1, \sigma_1 = 0,08, \sigma_2 = 0,07$ при времени моделирования $T_{\max} = 10^8$ ед.. На рис. 3 представлены полученные результаты работы модели.

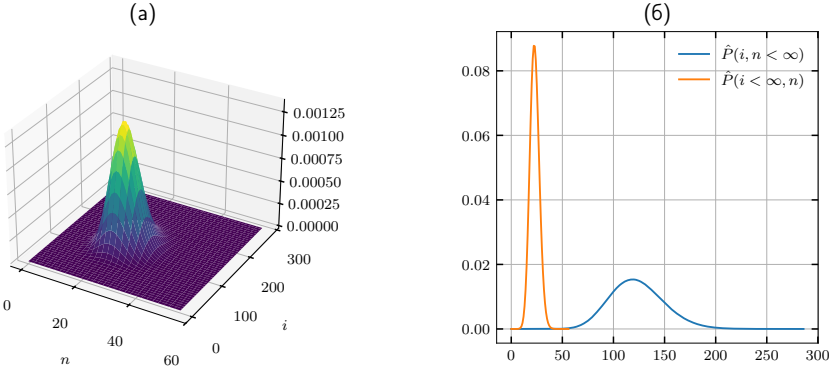


Рис. 3. Распределения числа заявок на орбитах ожидания и восстановления, полученные методом дискретно-событийного моделирования: (а) двухмерное распределение, показывающее совместную вероятность числа заявок на орбитах; (б) одномерные распределения, иллюстрирующие вероятности числа заявок на каждой орбите

3. Валидация модели

Для подтверждения достоверности результатов работы имитационной модели проведена валидация.

Проверка воспроизводимости результатов. Проведена оценка устойчивости результатов при повторных запусках модели с идентичными параметрами (табл. 1). В рамках проверки рассчитывалась величина максимального отклонения:

$$\Delta_{\max} = \max_{i,j,k} \left| \hat{P}^{(1)}(i,n,k) - \hat{P}^{(2)}(i,n,k) \right|,$$

где $\hat{P}^{(1)}$ и $\hat{P}^{(2)}$ — результаты двух независимых запусков модели.

Таблица 1

Значения максимальных отклонений при повторных запусках.

$T_{\max}$	10^4	10^5	10^6	10^7
$\Delta_{\max}$	0.0014	0.00046	0.00012	$3.317 \cdot 10^{-5}$

Сопоставление с аналитическим решением в частном случае. При установке значения параметра γ, равного нулю, система переходит в классическую СМО с повторными вызовами, простейшим потоком входящих требований, одной орбитой ожидания и известной аналитической формулой распределения числа заявок на орбите в стационарном режиме функционирования:

$$H(u) = (1 - \rho(e^{ju} - 1)) \left(\frac{1-\rho}{1-\rho e^{ju}} \right)^{\frac{\lambda}{\sigma_1}+1}, \tag{2}$$

$$P(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jui} H(u) du, \tag{3}$$

где $\rho = \lambda/\mu$ - пропускная способность системы.

Расстояние Колмогорова между статистическим и теоретическим распределениями числа заявок на орбите ожидания (табл. 2).

$$D = \max_i \left| P(i) - \hat{P}(i)|_{\gamma=0} \right|.$$

Таблица 2

Расстояние Колмогорова D при различных значениях параметров λ и μ при фиксированных $\sigma_1 = 0.08$, $T_{\max} = 10^7$.

μ \ λ	10	8	5	0,1
15	0,00011	0,00012	0,00018	0,00013
11	0,00009	0,00012	0,00025	0,00011

Заключение

В данной работе рассмотрена математическая модель RQ-системы с искажениями заявок и орбитой восстановления. На основе построен-

ной модели реализован алгоритм дискретно-событийного имитационного моделирования, позволяющий проводить численный анализ системы при различных параметрах. Проведена валидация модели, результаты которой демонстрируют возможность применения программной реализации для исследования характеристик системы и оптимизации процессов СМО.

СПИСОК ЛИТЕРАТУРЫ

1. *Falin. G.* A survey of retrial queues. Queueing Syst. // Theory Appl. 1990. Vol. 7,2. P. 127–167.
2. *Gelenbe E.* Product-form queueing networks with negative and positive customers. Journal of Applied Probability. 1991. Vol. 28(3). P. 656–663.
3. *Gelenbe, E., Glynn, P., Singman, K.* Queues with negative arrivals. Journal of Applied Probability. (1991). Vol. 28. P. 245–250.
4. *Daduna, Hans.* Queueing networks – customers, signals, and product form solutions, by X. Chao, M. Miyazawa, and M. Pinedo and Introduction to stochastic networks, by R.F. Serfozo. Journal of Applied Mathematics and Stochastic Analysis. 2001.
5. *Artalejo, J. R., Gomez-Corral, A.* Analysis of a stochasticclearing system with repeated attempts. Communications in Statistics. Stochastic Models, 1998. Vol. 14(3). P. 623–645.
6. *J. R. Artalejo, A. Gomez-Corral* Steady state solution of a single-server queue with linear repeated requests, // Journal of Applied Probability, 1997. Vol. 34, № 1. P. 223–233.
7. *Марголис Н. Ю.* Имитационное моделирование: учеб. пособие. Томск: М253 Издательский Дом Томского государственного университета, 2015. 130 с.
8. *Стародубов А. Н., Зиновьев В. В., Николаев П. И., Кузнецов И. С.* Верификация и валидация имитационных моделей систем: методические указания к лабораторной работе; КузГТУ. Электрон. дан. Кемерово, 2018.
9. *Назаров А. А., Терпугов А. Ф.* Теория массового обслуживания. 2-е изд., исправ. Томск : Издательство научно-технической литературы, 2010. 228 с.

Измайлова Яна Евгеньевна — кандидат физико-математических наук; доцент; каф. теории вероятностей и математической статистики.. E-mail: *evgenevna.92@mail.ru*

Канылбай Зангар — студент 4 курса; институт прикладной математики и компьютерных наук.. E-mail: *zzangar72@gmail.com*

ЭКСТРЕМАЛЬНЫЙ ИНДЕКС ВРЕМЕНИ ОЖИДАНИЯ В СИСТЕМЕ ОБСЛУЖИВАНИЯ С СУПЕРПОЗИЦИЕЙ ВХОДНЫХ ПОТОКОВ ПАРЕТО

Н. С. Калинин[1], И. В. Пешкова[1,2]

[1] *Институт прикладных математических исследований КарНЦ РАН, г. Петрозаводск, Россия,*

[2] *Петрозаводский государственный университет, г. Петрозаводск, Россия*

Применение теории экстремальных значений к системам массового обслуживания позволяет определить некоторые экстремальные свойства, например, предельное поведение максимального времени ожидания или максимальной длины очереди. Важным понятием в теории экстремальных значений является экстремальный индекс. С помощью данного индекса можно определить степень кластеризации экстремальных значений. В статье методы теории экстремальных значений применяются для вычисления оценки экстремального индекса времени ожидания системы $\sum_i G_i/M/1$, в которой входной поток представляет собой суперпозицию независимых потоков, заданных распределением Парето. С помощью метода экспоненциального расщепления строятся моменты регенерации, на основе которых далее вычисляется регенеративная оценка экстремального индекса. **Ключевые слова:** *система обслуживания, экстремальный индекс, суперпозиция независимых потоков*

Введение

Теория экстремальных значений зародилась в 20-х годах прошлого столетия и с тех пор нашла широкое применение в различных областях науки, в том числе для исследования систем массового обслуживания. Например, в работе [5] применяются методы данной теории для определения предельного поведения максимального времени ожидания и максимальной длины очереди для $GI/GI/1$ системы.

Классическая теория экстремальных значений не учитывает зависимости между данными, что влияет на точность вычислений. Эту проблему можно решить с помощью так называемого экстремального индекса, который часто интерпретируется как мера тенденции к кластеризации экстремальных значений [3–5, 7]. На практике экстремальный

индекс применяется для оценки размера кластеров или средней длины промежутков между превышениями высоких пороговых значений. Методы вычисления оценок экстремального индекса обсуждаются в работах [1, 6].

В статье предлагается метод экспоненциального расщепления для регенеративного оценивания экстремального индекса стационарного времени ожидания в системе с суперпозицией независимых входных потоков, заданных распределением Парето.

1. Общие положения

Строго стационарная последовательность $\{X_n,\ n \geq 1\}$ с общей ф. р. F имеет *экстремальный индекс* θ $(\theta \in [0,1])$, если для каждого $\upsilon > 0$ существует такая последовательность вещественных чисел $\{u_n = u_n(\upsilon)\}$, что выполнены соотношения

$$n\overline{F}(u_n) \to \upsilon \text{ при } n \to \infty \tag{1}$$

и

$$P(M_n \leq u_n(\upsilon)) \to e^{-\theta\upsilon} \text{ при } n \to \infty. \tag{2}$$

Поскольку параметр θ для стационарной последовательности $\{X_n,\ n \geq 1\}$ можно вычислить по формуле [2]:

$$\theta = \lim_{n\to\infty} \frac{\log P(M_n \leq u_n)}{n \log F(u_n)}, \tag{3}$$

то регенеративную оценку экстремального индекса можно записать в виде

$$\tilde{\theta}_\Delta(n) = \frac{1}{\tilde{\Delta}} \frac{\log\left(1 - \dfrac{m_\Delta(u_n)}{k_\Delta}\right)}{\log\left(1 - \dfrac{N(u_n)}{n}\right)}, \tag{4}$$

где $m_\Delta(u_n) = \#(i \leq k_\Delta : M_i > u_n)$, — число циклов регенерации, на которых превышено пороговое значение u_n, $N(u_n) = \#(i \leq n : X_i > u_n)$ – число превышений случайной последовательностью $\{X_n,\ n \geq 0\}$ порогового значения u_n, k_Δ — число циклов регенерации и $\tilde{\Delta}$ — оценка $M\widehat{\Delta}$.

Теперь рассмотрим систему $\sum_{i=1}^{m} G_i/M/1$, в которой входной поток представляет собой m независимых входных потоков, i-й поток имеет распределение Парето с параметрами $x_0, \alpha_i > 1$. Время обслуживания имеет показательное распределение с параметром μ. В такой системе интервалы между приходами заявок зависимы, поэтому классическая

регенерация не применима. Для получения моментов искусственной регенерации предлагается применить метод экспоненциального расщепления

Согласно [4] метод заключается в фиксации одного из потоков и замене всех остальных входных потоков на стохастически эквивалентные. При этом плотность распределения g исходной случайной величины представляется в виде смеси двух плотностей распределений, одно из которых показательное. Такое экспоненциальное расщепление плотности g возможно, если выполнено условие

$$g(x) \geq \delta g_0(x), \quad x \geq 0$$

где $\delta \in (0,1)$ и $g_0(x) = \lambda e^{-\lambda x}, \lambda > 0, x \geq 0$. В этом случае исходная величина T заменяется на стохастически эквивалентную T^* вида

$$T^* = IT_0 + (1-I)T_1, \tag{5}$$

где T_0 — с.в с плотностью g_0, I — индикатор расщепления, такой, что $P(I=1) = \delta$, T_1 — с.в. с плотностью

$$g_1(x) = \frac{g(x) - \delta g_0(x)}{1-\delta}.$$

В частности, если T имеет распределение Парето с параметрами x_0 и α с функцией распределения

$$G(x) = 1 - \left(\frac{x_0}{x_0+x}\right)^{\alpha}, x_0 > 0, x \geq 0,$$

то плотность случайной величины T_1 равна

$$g_1(x) = \frac{\alpha}{x_0(1-\delta)}\left(\frac{x_0}{x_0+x}\right)^{\alpha+1} - \frac{\delta\lambda}{1-\delta}e^{-\lambda x}, \quad x \geq 0. \tag{6}$$

Моментами регенерации для стационарного времени ожидания являются моменты времени, в которые все входные потоки (кроме фиксированного) попадают в экспоненциальную фазу (т.е. для которых $\delta = 1$), и одновременно система пуста.

На рис. 1 приведены результаты регенеративной оценки экстремального индекса для системы с двумя входными потоками, заданными распределением Парето при $\alpha_1 = 2, x_0 = 1, \alpha_2 = 3, \mu = 4$ и при различных значениях перцентиля порогового значения.

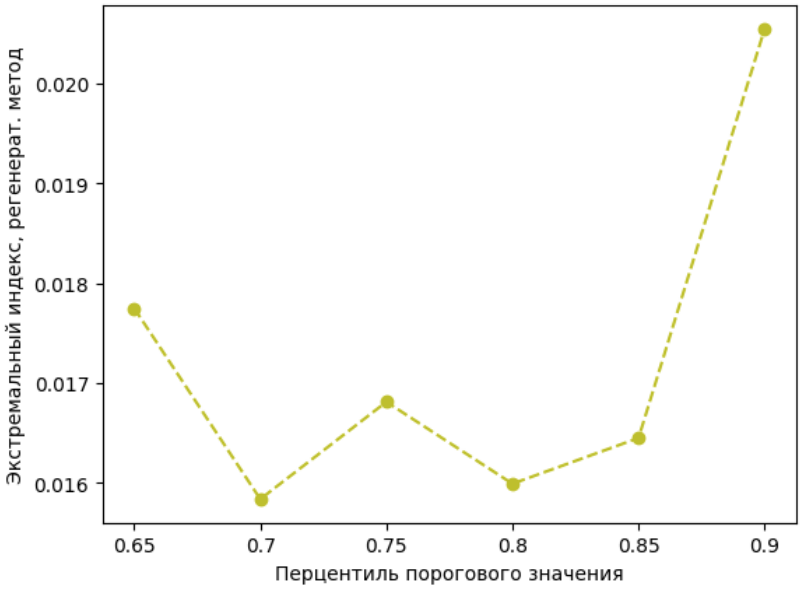


Рис. 1. Регенеративная оценка экстремального индекса для системы $G_1 + G_2/M/1$ при $\alpha_1 = 2, x_0 = 1, \alpha_2 = 3, \mu = 4$

Результаты экспериментов показывают, что для данной системы характерна высокая степень кластеризации, поскольку экстремальный индекс близок к 0.

Заключение

В ходе данной работы были применены методы теории экстремальных значений и метод экспоненциального расщепления для расчёта регенеративной оценки экстремального индекса стационарного времени ожидания системы $\sum_i G_i/M/1$ с суперпозицией независимых входных потоков, имеющих распределение Парето.

СПИСОК ЛИТЕРАТУРЫ

1. Bertail P. Extreme values statistics for Markov chains via the (pseudo-) regenerative method / P. Bertail, S. Clémençon, J. Tressou // Extremes. 2009. Vol. 12. P. 327–360.
2. Embrechts P. Modelling Extremal Events for Insurance and Finance / P. Embrechts, C. Klüppelberg, T. Mikosch. London: Springer, 1997.
3. Leadbetter M. Extremes and Related Properties of Random Sequences and Processes / M. Leadbetter, G. Lindgren, H. Rootzen. New York: Springer-Verlag, 1983. 336 p.
4. Resnick S. I. Extreme Values, Regular Variation and Point Processes / S. I. Resnick. New York: Springer,1987. 320 p.
5. Smith R. L. The extremal index for a Markov chain / R. L. Smith // Journal of Applied Probability. 1992. Vol. 29(1). P. 37–45.
6. Smith L. Estimating the extremal index / L. Smith, I. Weissman // Journal of the Royal Statistical Society. Series B: Methodological. 1994. Vol. 56(3). P. 515–528.
7. Weissman I. The Extremal Index and Clustering of High Values for Derived Stationary Sequences / I. Weissman, U. Cohen // Journal of Applied Probability. 1995. Vol. 32, №. 4. P. 972–981.
8. *Janos Galambos* Extreme value theory for applications. USA: Temple University, 1993. 248 p.
9. Peshkova I. Regenerative analysis and approximation of queueing systems with superposed input processes / I. Peshkova, E. Morozov, M. Pagano // Mathematics. 2024. Vol. 12, №. 14. Art. no. 2202. P. 1–22.
10. *Saralees Nadarajah* Extreme Value Theory, Models and Simulation // India, 2003. 691 p.

Калинин Никита Сергеевич — аспирант, Институт прикладных математических исследований КарНЦ РАН. E-mail: *kalinin.nikita344@yandex.ru*

Пешкова Ирина Валерьевна — к.ф.-м.н., доцент, Петрозаводский государственный университетж ведущий научный сотрудник, Институт прикладных математических исследований КарНЦ РАН .

ИССЛЕДОВАНИЕ RQ-СИСТЕМЫ С НЕИСПРАВНЫМИ ЗАЯВКАМИ И ИХ ВОССТАНОВЛЕНИЕМ

К. М. Костарева[1], Я. Е. Измайлова[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В данной работе исследуется система обработки заявок с повторными попытками, основанная на RQ-модели. Рассматривается система с простейшим потоком заявок, который с помощью заданных вероятностей разделяется на исправные и неисправные заявки. Неисправные заявки переходят на орбиту для восстановления, после чего обслуживаются. Представлена функциональная модель, на основе которой построена соответствующая математическая модель. Исследование системы проводится методом имитационного моделирования, воспроизводящего поведение системы массового обслуживания с повторными попытками. **Ключевые слова:** *RQ-система, неисправные заявки, исправные заявки*

Введение

Современные системы электронной коммерции сталкиваются с необходимостью эффективной обработки заказов, включая случаи с ошибочными или неполными данными. В данной работе рассматривается система обработки заказов с механизмом повторных попыток, основанная на RQ-модели (Retrial Queue). Предлагаемое решение предусматривает два уровня обработки проблемных заявок: автоматическое исправление технических ошибок и интерактивное взаимодействие с клиентом при ошибках в данных.

1. Функциональная модель

Анализ систем обработки заказов с повторными попытками (рис. 1).

1) Отправка нового запроса (может содержать ошибки);
2) при обнаружении ошибок происходит передача запросов в модуль исправления;
3) корректные запросы сразу обрабатываются при доступности системы;

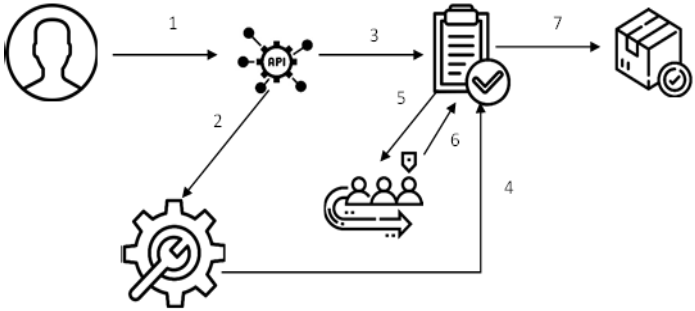


Рис. 1. Система обработки заказов с повторными попытками

4) исправленные запросы поступают на обработку при доступности;
5) при перегрузке системы запросы помещаются в блок ожидания;
6) запросы из блока ожидания обрабатываются по мере освобождения ресурсов;
7) завершающий этап – формирование и отправка результата.

Описание системы:

Клиентская часть: отправляет запросы на создание заказа через HTTP. Получает ответы: успех, ошибка системы или ошибка данных.

Входной шлюз (API Gateway):

- проверяет и классифицирует запросы;
- корректные запросы (валидные данные, успешный платеж) направляет на обработку;
- ошибочные запросы отправляет в модуль исправления.

Обработчик заказов:

- выполняет основные операции: запись в БД, резервирование товара, уведомления;
- при перегрузке помещает запросы в блок повторной обработки.

Модули обработки ошибок:

- автоматическое исправление системных ошибок;
- интерактивное исправление ошибок клиента (с уведомлением);

– после исправления – повторная отправка на обработку или в блок ожидания.

Блок ожидания повторной обработки: временное хранилище для запросов, ожидающих доступности системы.

2. Математическая модель

Рассмотрим RQ-систему (Retrial Queue) с одним обслуживающим прибором и двумя орбитами (рис. 2).

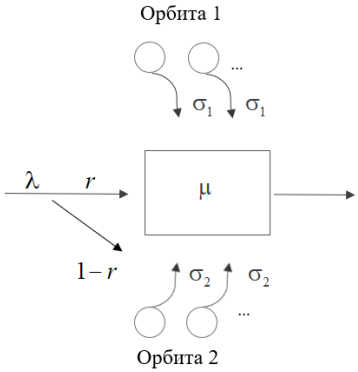


Рис. 2. Математическая модель RQ-системы с неисправными заявками и их восстановлением

На вход RQ-системы поступает простейший поток заявок с интенсивностью λ. С вероятностью r заявка является исправной. Если прибор свободен, то исправная заявка начинает обслуживание. Время обслуживания имеет экспоненциальное распределение с параметром μ. Если прибор занят, то заявка уходит на орбиту 1 (орбита ожидания обслуживания исправных заявок). На орбите заявка осуществляет случайную задержку, имеющую экспоненциальное распределение с параметром σ_1. После случайной задержки заявка вновь обращается к прибору с повторной попыткой его захвата. Если прибор занят, то мгновенно возвращаются обратно и осуществляют следующую случайную задержку. После окончания обслуживания заявка покидает систему.

С вероятностью $1 - r$ заявка неисправная. Неисправная заявка переходит для восстановления на орбиту 2 (блок восстановления неисправных заявок). Время исправления имеет экспоненциальное распределение с параметром σ_2. После восстановления заявка обращается к

прибору. Если прибор свободен, то она встает на обслуживание. Если занят, то переходит на орбиту для исправных заявок.

Пространство состояний:

$k(t)$ – состояние прибора (0 – свободен, 1 – занят);

$n(t) \in \mathbb{N}$ – число заявок на орбите 2;

$i(t) \in \mathbb{N}$ – число заявок на орбите 1.

Тогда $P\{k(t) = k,\ i(t) = i,\ n(t) = n\} = P_k(i, n, t)$ – вероятность того, что в момент времени t прибор находится в состоянии k, на орбите 1 находится i заявок, на орбите 2 находится n заявок.

Возможные события системы с переходами за малый интервал времени Δt:

- приход исправной заявки $\lambda r \Delta t$;
- приход неисправной заявки $\lambda(1 - r)\Delta t$;
- завершение обслуживания $\mu \Delta t$;
- поступление заявок с орбит:
  - исправная заявка с орбиты ожидания $i\sigma_1 \Delta t$;
  - восстановленная заявка с орбиты восстановления $n\sigma_2 \Delta t$.

3. Имитационная модель

Для исследования поведения RQ-системы с двумя орбитами была разработана имитационная модель на языке Python.

Алгоритм моделирования

1) Инициализация:
   Функция exponential(rate): генерирует случайное время с экспоненциальным распределением;
   Параметры основной функции: λ, r, μ, σ_1, σ_2, $T_{\max}$ – максимальное время моделирования;
   Инициализация переменных:
   - k – состояние прибора (0 – свободен, 1 – занят);
   - i – число заявок на орбите ожидания, n – число заявок на орбите восстановления, t – текущий момент времени. В начальный момент все переменные равны нулю;
   - history – список всех событий и состояний системы;
2) Настройка файлов вывода: имеется три файла для записи результатов;
3) Основной цикл моделирования:
   Выполняется пока $t < T_{\max}$:

– **Генерация событий и выбор ближайшего**:

$$t_{arrival} = -\frac{\ln(R(0,1))}{\lambda}, \quad t_{service} = -\frac{\ln(R(0,1))}{\mu},$$
$$t_{retry} = -\frac{\ln(R(0,1))}{i\sigma_1}, \quad t_{recovery} = -\frac{\ln(R(0,1))}{n\sigma_2},$$
$$t_{next} = \min(t_{arrival}, t_{service}, t_{retry}, t_{recovery}).$$

– **Обработка события**:
  – Поступление заявки:
    – с вероятностью r поступившая заявка является исправной и она либо обслуживается, либо переходит на орбиту ожидания;
    – с вероятностью $1 - r$ поступившая заявка является неисправной и она уходит на орбиту восстановления;
  – Завершение обслуживания: освобождение прибора ($k = 0$);
  – Повторная попытка обслужиться: происходит обслуживание заявки или возврат на орбиту ожидания, в случае, если прибор занят;
  – Завершение восстановления: переход на обслуживание или орбиту ожидания.
– **Обновление времени**:

$$t = t + t_{\text{next}}.$$

4) Расчет вероятностей. Рассчитываем трехмерное распределение вероятностей состояний прибора, числа заявок на орбите ожидания, числа заявок на орбите восстановления по формуле

$$P(k,i,n) = \frac{N(k,i,n)}{T_{total}},$$

где $N(k,i,n)$ – суммарное время прибывания системы в состоянии (k,i,n). От трехмерного переходим к одномерным распределениям по формулам

$$P(i,n) = P(0,i,n) + P(1,i,n), \quad P(i) = \sum_{n=0}^{\infty} P(i,n), \quad P(n) = \sum_{i=0}^{\infty} P(i,n).$$

5) Завершение моделирования.
  – сохранение результатов в файлы Excel;
  – возврат истории событий.

На рисунках 3–4 представлены файлы с результатами работы имитационной модели.

Время	k	i	n	Событие
0	0	0	0	START
0,815159158	1	0	0	Прибытие исправной заявки
0,817451445	0	0	0	Завершение обслуживания
5,42553404	0	0	1	Прибытие неисправной заявки
6,161648905	1	0	0	Восстановление заявки
6,473628466	0	0	0	Завершение обслуживания
7,338271357	1	0	0	Прибытие исправной заявки
7,521049804	0	0	0	Завершение обслуживания
7,927134421	1	0	0	Прибытие исправной заявки
8,7148256	0	0	0	Завершение обслуживания
8,954519734	1	0	0	Прибытие исправной заявки
8,970792141	0	0	0	Завершение обслуживания
13,05490119	0	0	1	Прибытие неисправной заявки
13,3413402	1	0	1	Прибытие исправной заявки
13,41743288	0	0	1	Завершение обслуживания
15,22715264	1	0	1	Прибытие исправной заявки
15,53075801	1	0	2	Прибытие неисправной заявки

Рис. 3. Поведение системы массового обслуживания с течением времени

i	n	P(i,n)
0	0	3,32143E-06
0	1	1,50974E-06
0	2	3,01948E-07
1	2	9,05845E-07
2	2	3,01948E-07
2	3	6,03897E-07
2	4	1,50974E-06
2	5	3,01948E-07
2	6	6,03897E-07
3	6	3,01948E-07
3	7	9,05845E-07
3	8	3,01948E-07
4	4	1,50974E-06
4	5	1,50974E-06
4	6	6,03897E-07
4	7	6,03897E-07

Рис. 4. Двумерное распределение вероятностей $P(i, n)$

Заключение

Построена математическая модель системы обработки заказов в виде RQ-системы с блоком восстановления. Исследование системы проводилось методом имитационного моделирования, воспроизводящего поведение RQ-системы. В ходе исследования моделируются различные сценарии работы системы, что в итоге дает два ключевых распределения: распределение вероятностей количества заявок на орбите ожидания (первая орбита) и орбите восстановления (вторая орбита).

СПИСОК ЛИТЕРАТУРЫ

1. *Назаров А. А., Терпугов А. Ф.* Теория массового обслуживания: учебное пособие. 2-е изд. Томск, 2010. 228 с.
2. *Галажинская О. Н., Моисеева С. П.* Теория случайных процессов : учеб. пособие. Ч. 1. Томск, 2015. 128 с.
3. *Назаров А. А., Терпугов А. Ф.* Теория вероятностей и случайных процессов: учебное пособие. 2-е изд. Томск, 2010. 204 с.

Измайлова Яна Евгеньевна — кандидат физико-математических наук, доцент кафедры теории вероятностей и математической статистики Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *evgenevna.92@mail.ru*

Костарева Ксения Максимовна — студент Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *ksenia.kostareva@gmail.com*

АЛГОРИТМЫ ПОДБОРА ПАРАМЕТРОВ MAP-ПОТОКА ДЛЯ МОДЕЛИРОВАНИЯ ON-OFF СОСТОЯНИЙ МНОГОЯДЕРНЫХ ПРОЦЕССОРОВ

И. Л. Лапатин[1], Н. А. Луценко[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Модели случайных потоков класса MAP обладают рядом преимуществ, таких как возможность аналитического исследования и возможность учета корреляции между длинами интервалов. В работе предлагается использовать коррелированность MAP-потока для моделирования on-off процессов при работе многоядерных процессоров. Эти процессы определяют активную работу процессоров и периоды их простоя. Моделирование этих процессов является одним из шагов создания определенных типов вычислительной нагрузки при симуляции производительности микроархитектуры. В работе предлагаются алгоритмы подбора параметров MAP-потока по некоторым числовым характеристикам, включая коэффициент корреляции, выборки длин интервалов между событиями в некотором случайном потоке событий. **Ключевые слова:** *MAP-поток, коэффициент корреляции, подбор параметров*

Введение

Системы массового обслуживания, используемые для моделирования процессов функционирования современных телекоммуникационных сетей и компьютерных вычислительных систем предъявляют высокие требования к моделированию случайных потоков событий. Одним из наиболее эффективных инструментов для описания таких потоков являются марковские процессы поступления (MAP, Markovian Arrival Processes). Эти процессы обеспечивают высокую гибкость и точность при моделировании потоков, характеризующихся зависимостями и коррелированностью между событиями.

В данной работе мы предлагаем использовать MAP-потоки для моделирования корреляции между длинами интервалов в активной фазе работы отдельных ядер многоядерных процессоров. Это отличается от подхода параметрической задачи оценки параметров в рамках конкретного закона распределения, который никак не учитывает корреляцию

между длинами интервалов. А с точки зрения процессов активности ядер процессора учет корреляции безусловно необходим.

Одной из ключевых задач в работе с MAP-потоками является подбор параметров, которые обеспечивают точное соответствие модели заданным числовым характеристикам, таким как средняя интенсивность поступления событий, дисперсия, коэффициенты автокорреляции и другие статистические показатели. Правильный выбор параметров необходим для построения адекватной модели, способной точно описывать реальные процессы, при этом сохраняя вычислительную эффективность.

В работе предлагается описание двух алгоритмов подбора параметров MAP-потока на основании значений числовых характеристик длин интервалов между наступлениями событий. То есть, по числовым характеристикам некоторой временной последовательности наступления некоторых событий (исходные данные) подбираются параметры MAP-потока, числовые характеристики длин интервалов которого были бы близки к числовым характеристикам исходных данных.

1. Алгоритм последовательного перебора

Данный алгоритм будет исходить от начального решения, которое подбирается случайно с заданной точностью. Для оценки точности решения будет использоваться функция среднеквадратической ошибки [1]:

$$\frac{1}{N}\sum_{i=0}^{N}(x_i - y_i)^2,$$

где N – количество параметров, которые участвуют в оценке точности; x_i – получившееся значение параметра в ходе подбора параметров; y_i – целевое значение.

Основными характеристиками, на основе которых будут находиться параметры, были выбраны коэффициенты корреляции ($CCor$) и вариации ($CVar$). Будем полагать, что $CCor$ и $CVar$ известны. Чтобы данная функция подходила под задачу, её нужно отнормировать, поскольку коэффициент вариации может быть достаточно большим, а коэффициент корреляции изменяется от -1 до 1. Поэтому финальный вариант функции ошибки выглядит так:

$$fit = \frac{(\frac{CVar_T - CVar}{CVar})^2 + (\frac{CCor_T - CCor}{1+|CCor|})^2}{2}.$$

Задача сводится к тому, чтобы найти такие матрицы $\mathbf{Q}$, Λ, $\mathbf{D}$ [6], определяющие MAP-поток, чтобы вычисленные по ним теоретические

числовые характеристики $(CVar_T, CCor_T)$ длин интервалов [3] были близки к заданным.

Начальное решение выбирается случайно с прямой генерацией всех значений матриц $\mathbf{Q}$, Λ и $\mathbf{D}$ разом с заданным соотношением параметров. Поиск начального решения продолжится до тех пор, пока его значение ошибки не достигнет результата менее, чем $\varepsilon = 0,001$. Шаги генерации начального решения:

Шаг 1. Генерируются 100 MAP-потоков;

Шаг 2. Для каждого сгенерированного MAP-потока вычисляется значение ошибки (fit);

Шаг 3. Если $fit < 0,001$, то начальное решение найдено. Иначе идем на шаг 1;

Когда начальное решение было получено, то у нас имеются матрицы $\mathbf{Q}_0$, Λ_0 и $\mathbf{D}_0$. На следующем шаге алгоритма фиксируются матрицы Λ_0 и $\mathbf{D}_0$. Для матрицы $\mathbf{Q}_0$ происходит перебор значений до тех пор, пока значение ошибки не будет менее $\varepsilon = 10^{-4}$, либо не превысит заданное количество итераций. После того, как значение ошибки будет меньше ε, у нас получится матрица $\mathbf{Q}_1$.

Следующим действием фиксируются матрицы $\mathbf{Q}_1$ и $\mathbf{D}_0$. Для матрицы Λ_0 происходит перебор значений до тех пор, пока значение ошибки не будет менее $\varepsilon = 10^{-5}$, либо не превысит заданное количество итераций. Когда значение ошибки будет менее ε, то получится матрица Λ_1.

Последним шагом алгоритма будет фиксация матриц $\mathbf{Q}_1$ и Λ_1. Значения для матрицы $\mathbf{D}_0$ также перебираются все вместе. Перебор происходит до тех пор, пока значение ошибки не будет менее $\varepsilon = 10^{-6}$, либо не превысит заданное количество итераций. После получается матрица $\mathbf{D}_1$.

После того, как все матрицы были перебраны, у нас получается MAP-поток с искомыми значениями коэффициента корреляции и коэффициента вариации, которые отклоняются от искомого не более, чем $\varepsilon = 10^{-6}$, при условии, что перебор значений матриц не останавливался по заданному лимиту итераций. Тем самым удается максимально близко приблизиться к искомым характеристикам.

2. Перебор параметров в заданной окрестности определенных значений

Следующий алгоритм, который использовался для поиска параметров MAP-потока, это алгоритм, который основан на переборе каждого параметра в матрицах $\mathbf{Q}$ и Λ в некоторой заданной окрестности. Для этого алгоритма нужно также начальное решение, от которого будет

начинаться сам перебор параметров. Начальное решение выбирается также, как и в предыдущем алгоритме.

После того, как найдено начальное решение, начинается постепенный перебор всех параметров для матриц $\mathbf{Q}$ и Λ. Сначала копируется начальное решение в другую переменную, чтобы была возможность откатиться назад и начать поиск сначала. Это необходимо потому, что решение может уйти далеко от искомого. Далее запускается цикл, в котором происходит проход по всем элементам этих матриц и прибавляется или вычитается к уже имеющимся значениям случайное число в заданной окрестности (по умолчанию 2%). После того, как элементы матриц $\mathbf{Q}$ и Λ были изменены, вычисляется значение ошибки для MAP-потока. Если решение за 25 тыс. итераций не было найдено, то происходит откат к начальному решению и поиск начинается заново. Когда значение ошибки получается меньше заданного ε (задается вручную), то перебор для матриц $\mathbf{Q}$ и Λ завершается. Потом происходит перегенерация матрицы $\mathbf{D}$.

Перегенерация матрицы $\mathbf{D}$ нужна для потенциального улучшения решения. Запускается цикл, где целиком генерируется сама матрица, вычисляется значение ошибки и если это значение меньше, чем $\varepsilon/100$, то решение считается найденным. В ходе генерации матрицы $\mathbf{D}$ сохраняется наилучшее решение и если за 50 тыс. итераций не будет достигнута точность в $\varepsilon/100$, то возвращается наилучшее найденное решение.

3. Численные эксперименты

В ходе численного эксперимента осуществлялся поиск параметров MAP-потоков с разными коэффициентами корреляции и вариации. Данный эксперимент даст понимание того, сколько по времени работают алгоритмы и средняя достигнутая точность.

В первом эксперименте в качестве искомых данных использовались коэффициент вариации в диапазоне [1; 4] и коэффициент корреляции в диапазоне [0; 0,3]. Бралось случайное значение из этого диапазона, с округлением до 2 знаков после запятой. Далее, эти случайные значения подавались в алгоритм как целевые значения. Нужно было подобрать такой MAP-поток, чтобы у него были коэффициенты корреляции и вариации близки к искомым. Для эксперимента бралось 100 попыток найти случайные варианты корреляции и вариации. Такое малое число попыток обусловлено тем, что тесты проводились не на мощном оборудовании, и даже 100 попыток занимали достаточное количество времени. В ходе этого эксперимента значения матрицы $\mathbf{Q}$ брались из диапазона [0; 10], а для матрицы Λ из диапазона [0; 1000].

Эксперимент проводился на компьютере с 4-ядерным процессором Intel Core i5 9300H с максимальной частотой 4 ГГц.

Второй эксперимент проводился в таких же условиях, но коэффициент вариации брался в диапазоне [1; 8] и коэффициент корреляции в диапазоне [0; 0,4]. Значения матрицы $\mathbf{Q}$ брались из диапазона [0; 10], а для матрицы Λ из диапазона [0; 10000].

В табл.1 приведены результаты первого эксперимента.

Таблица 1

Результаты первого эксперимента

Алгоритм	Среднее время поиска, сек.	Среднее значение ошибки
Алгоритм последовательного перебора	5,3	0,0001
Перебор параметров в заданной окрестности определенных значений	10,67	0,0003

На основе этих результатов можно сказать, что алгоритм последовательного перебора работает быстрее, но в нем не так много настраиваемых параметров для более точного поиска.

Алгоритм перебора параметров в заданной окрестности по времени показывает результат в 2 раза хуже алгоритма последовательного перебора. Среднее значение ошибки тоже немного больше, но при этом имеет одинаковый порядок.

В табл.2 приведены результаты второго эксперимента

Таблица 2

Результаты второго эксперимента

Алгоритм	Среднее время поиска, сек.	Среднее значение ошибки
Алгоритм последовательного перебора	8,1	0,0019
Перебор параметров в заданной окрестности определенных значений	10,68	0,0044

Можно наблюдать, что скорость и точность работы выше при небольших значениях коэффициента корреляции и вариации, но и для расширенного диапазона их значений удается достигать достаточно хорошей точности.

Заключение

В работе предложены алгоритмы подбора параметров MAP-потока по известным числовым характеристикам длин интервалов реального процесса. В качестве числовых характеристик предлагается брать среднюю длину, коэффициент вариации длин интервалов и коэффициент корреляции соседних длин интервалов. Для определенного диапазона значений коэффициента корреляции и вариации алгоритм подбирает параметры с хорошей точностью до шестого знака после запятой. Предложенный алгоритм может быть применен для моделирования on-off процессов при работе многоядерных процессоров. При таком подходе удается учитывать корреляцию длин интервалов.

СПИСОК ЛИТЕРАТУРЫ

1. `https://www.geeksforgeeks.org/mean-squared-error/` — GeeksForGeeks. 2025.
2. *Гарайшина И.Р., Моисеева С.П., Назаров А.А.* Методы исследования коррелированных потоков и специальных систем массового обслуживания. Томск: Изд-во НТЛ, 2010. 204 с.
3. *Вишневский В.М., Дудин А.Н., Клименок В.И.* Стохастические системы с коррелированными потоками. Теория и применение в телекоммуникационных сетях. М.: Рекламно-издательский центр «ТЕХНОСФЕРА», 2018. 564 с.

Лапатин Иван Леонидович — канд. физ.-мат. наук, доцент кафедры ТВиМС ИПМКН Томского государственного университета. E-mail: *ilapatin@mail.ru*

Луценко Никита Андреевич — студент ИПМКН Томского государственного университета. E-mail: *itname@bk.ru*

АНАЛИЗ ВОЗРАСТА ИНФОРМАЦИИ В СИСТЕМАХ С ВЫРАЖЕННОЙ КОРРЕЛЯЦИЕЙ ВО ВРЕМЕНИ

В. Е. Манаева[1], Е. В. Маркова[1,2]

[1] *Российский университет дружбы народов имени Патриса Лумумбы, г. Москва, Россия*

[2] *Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук, г. Москва, Россия*

Возраст информации – количественная метрика, измеряющая актуальность данных, – может быть использована, как показатель качества обслуживания (quality of service, QoS) в многочисленных приложениях современных мобильных сетей связи, в частности, в сетях промышленных приложений Интернета вещей (Internet of Things, IoT). В таких приложениях конечным устройствам (датчикам и/или пользовательским устройствам) свойственно собираться в группы у одного источника данных (как правило, на одном приборе IoT установлены несколько датчиков), при этом интенсивность поступления информации от них может иметь высокий разброс и корреляцию во времени. Для анализа QoS в модели беспроводной сети промышленного IoT, в работе построена математическая модель технической системы в виде системы массового обслуживания в дискретном времени с групповым поступлением и обслуживанием, с несколькими входящими потоками, процесс поступления каждого из которых имеет несколько состояний. В связи со сложностями аналитического подхода к анализу системы, для системы была построена имитационная модель, учитывающая особенности математической модели. Данную модель предлагается использовать при дальнейшем анализе системы. **Ключевые слова:** *5G, Интернет вещей, возраст информации, корреляция трафика*

Введение

Современные беспроводные сети активно используются в различных сферах жизни. Интернет вещей (IoT) – это тип построения сети связи, при котором «вещи» связываются с помощью беспроводных сетей между собой и с внешними центрами управления (ЦУ). Термин «вещи»

Публикация выполнена в рамках проекта №.021937-2-000 Системы грантовой поддержки научных проектов РУДН.

в IoT может описывать любой физический объект, к которому присоединены датчики и/или устройства связи для обеспечения удаленного мониторинга и контроля. Так IoT находит свои приложения при создании «умных домов», «умных бытовых приборов» и подключенного здравоохранения.

IoT имеет множество применений в промышленности, выделяемых в отдельную группу услуг промышленного Интернета вещей (industrial Internet of Things, IIoT), например, в энергосистемах, для контроля сетевой нагрузки, на «умных заводах», для мониторинга урожая и во множестве других отраслей. Устройства IIoT собирают и анализируют данные с подключенного оборудования, датчиков и устройств, которые могут стоять в самых труднодостижимых местах. Полученные данные могут быть использованы для мониторинга, поддержки или управления работой устройств удаленно и/или автоматически.

Для адекватной оценки информации в удаленных ЦУ собранная информация должна быть как можно более актуальной. Для измерения актуальности информации используется показатель возраста информации (AoI), который количественно измеряет «свежесть» информации, полученной в ЦУ. Пиковый возраст информации (PAoI) определяется как верхняя граница AoI. Средний PAoI, а также статистический ряд AoI за достаточно большой промежуток времени в системе может быть использован в качестве показателя оптимизации.

Средний PAoI для любой системы можно разделить на две составляющие: (i) интервал между поступлениями заявок на передачу информации и (ii) время пребывания заявки в системе [1]. Отметим, что численный анализ вероятностно-временных характеристик, связанных с задержкой (которая является компонентой времени пребывания), в реальных системах затруднен режимом ортогонального частотного мультиплексирования (OFDMA) [2,3], что приводит к необходимости использования пакетной обработки заявок в системе.

В связи с этим для учета особенностей беспроводных систем с поддержкой технологии OFDMA, в данной работе предлагается модель с групповым обслуживанием нескольких устройств, которые генерируют группы заявок. Каждое из устройств может пребывать в одном из нескольких состояний, в каждом из которых отличается интенсивность генерации заявок. Например, для устройства с двумя состояниями одно можно описать, как «обычное», а второе – «аварийное», которое будет иметь более высокую интенсивность генерации заявок.

1. Особенности технической системы

Рассмотрим модель сети, в которой на нескольких роботизированных устройствах имеются датчики, которые отправляют информацию о состоянии устройств в удаленный ЦУ. Описанная система является моделью сетевого приложения IIoT и рассматривает восходящий канал связи на базовой станции (BS). Информация о состоянии N устройств отправляется в виде групп обновлений и собирается в буферном пространстве на BS. Максимальное число обновлений, которое может находиться в буферном пространстве на BS, обозначим R. Каждое из N устройств может находиться в любом из M состояний, в каждом из которых отличается интенсивность генерации обновлений. Важно отметить, что в состоянии с большим i интенсивность генерации обновлений на устройстве будет выше, чем в состоянии с меньшим i.

Обновления, поступающие в буфер, формируют единый поток и упорядочиваются во времени. При поступлении обновления могут быть заблокированы, если для них недостаточно свободного места в буфере. Обозначим K максимальный размер группы обслуживания. На обслуживание из буфера отбирается не более K обновлений, если в буфере есть обновления, в противном случае группа обслуживания остается пуста. Отметим, что время в рассматриваемой системе дискретно. Время обслуживания одной группы заявок составляет один временной интервал (слот) равный 1 мс.

2. Математическая модель

Опишем рассматриваемую техническую систему с помощью аппарата теории массового обслуживания. Перейдем к системе массового обслуживания в дискретном времени с групповым поступлением и обслуживанием заявок, генерируемых несколькими источниками. Отметим, что в математической модели вместо обновлений будем использовать термин заявки, так как для модели неважно, какое именно содержание у обновления, а важен лишь факт его отправки. Отметим, что каждый из источников генерирует заявки согласно переключающемуся пуассоновскому процессу (switched Poisson process, SPP). В отличие от обычного пуассоновского потока, у потока SPP изменяется интенсивность поступления заявок во времени. Изменяющееся состояние каждого источника в рамках SPP описывается марковской цепью $\mathrm{S}_n(t), n = \overline{1,N}, \mathrm{S}_n(t) \in \{1,\ldots,M\}$. Тогда состояние всех источников в слот n можно задать вектором $\mathbf{S}(t)^\top = (\mathrm{S}_1(t),\ldots,\mathrm{S}_N(t))$.

В конце временного слота каждый источник генерирует группу заявок размером $W_n(t)$, $n = \overline{1,N}$. $W_n(t)$ выбирается случайно согласс-

но ограниченному пуассоновскому распределению со средним λ_{ni}, $n = \overline{1,N}$, $i = \overline{1,M}$, которое зависит от состояния n-го источника. При поступлении в систему все сгенерированные заявки объединяются в единую группу размером $W(t)$. В зависимости от числа заявок в буфере $Q(T)$ на момент начала слота в него поступает $(R - Q(t))$ заявок из единой группы $W(t)$, отобранных случайным образом. В случае, если $(R-Q(t))$ меньше числа заявок в группе поступления, $(W(t)-R+Q(t)$ заявок считаются заблокированными.

В работе рассматривается задержанное обслуживание, то есть, даже при простаивающем приборе заявки сначала поступают в буфер и лишь оттуда на обслуживание. По завершении нового временного слота завершается предыдущее обслуживание, и в соответствии с заданной дисциплиной обслуживания из буфера отбираются заявки на предстоящее обслуживание. Процедура обслуживания следует порядку, заданному технической системой.

2.1. Дисциплины обслуживания

При исследовании возраста информации и пикового возраста информации интерес представляет анализ влияния на эти параметры выбранных дисциплин обслуживания. Рассмотрим следующие четыре: (i) «первый пришел – первый обслужился» (FCFS), (ii) «последний пришел – первый обслужился» (LCFS), (iii) случайный выбор и (iv) приоритетное обслуживание. Отдельное внимание уделяется приоритетному обслуживанию: разным способам его реализации и влиянию этих способов на поведение пикового возраста информации. Далее для краткости обозначения выбранной дисциплины обслуживания введем три компоненты:

1) классическая дисциплина обслуживания буфера (SD): FCFS, LCFS и RAND;
2) номер приоритетного источника заявок, P;
3) процент группы обслуживания, отведенный на заявки от приоритетного источника, ν.

Отметим, что при наборе группы обслуживания система заполняет до $\lceil \frac{\nu K}{100} \rceil$ мест заявками от приоритетного источника, после чего оставшиеся места в буфере заполняются согласно заданной классической дисциплине обслуживания. Классические дисциплины обслуживания в системе можно задать, установив значение процента приоритетного обслуживания равным 0.

2.2. Сложности анализа математической модели

В связи с анализом поведения возраста информации, как в системе в целом, так и для отдельных источников, пространство состояний должно иметь размерность $2N$: N элементов состояний для отслеживания состояния источников и N элементов состояний для отслеживания числа заявок от n-го источника в буфере. Большая размерность пространства состояний и сложность перехода между каждым из состояний приводят к тому, что в работе для вычисления QoS характеристик используется имитационное моделирование.

3. Имитационная модель

Имитационная модель в работе построена методом дискретно-событийного моделирования. Моделирование ограничивается заданным максимальным временем, t_{max}. Так как в математической модели время дискретно, за время $(t_{max}+1)$ происходит t_{max} циклов обслуживания (в первый цикл происходит только поступление).

В качестве входящих данных для модели используются следующие параметры: N, M, R и K, описанные в секции 1; λ_{ni}, $n=\overline{1,N}$, $i=\overline{1,M}$, SD, SD $\in \{\mathrm{FCFS}, \mathrm{LCFS}, \mathrm{RAND}\}$, P и ν, описанные в секции 2; переходные вероятности между состояниями источников, $\alpha_{nij}, n=\overline{1,N}, i,j=\overline{1,M-1}, j=i+1$ для увеличения интенсивности поступления заявок, $\beta_{nij}, n=\overline{1,N}, i=\overline{2,M}, j=i-1$ для уменьшения интенсивности поступления заявок; t_{max}.

Каждый запуск имитационного моделирования собирает свою глобальную статистику. Глобальная статистика имитационного моделирования включает в себя: число сгенерированных заявок, U; число сгенерированных заявок от источника n-го типа, U_n; число принятых в систему заявок, E; вероятность потери заявки, B; средний интервал между поступлениями заявок в систему $M[Y]$ и между поступлениями заявок от определенного источника n, $M[Y]_n$; среднее время пребывания произвольной заявки в системе, $M[T]$, и среднее время пребывания в системе заявки от определенного источника n, $M[T]_n$; средний пиковый возраст информации для системы, $M[A]$, и для заявок от определенного источника n, $M[A]_n$.

В моделировании есть 3 вида промежуточных событий: поступление заявок, обслуживание заявок, выбор заявок из буфера согласно дисциплине обслуживания.

Поступление заявок создает объекты, имитирующие заявки, со следующими данными: глобальный порядковый номер заявки, p; номер источника, n_p; временной слот, в который заявка была сгенерирована, G_p;

булеановая переменная о состоянии блокировки заявки, E_p; временной слот, в который было завершено обслуживание заявки, F_p; временной слот предыдущего поступления без блокировки заявки от того же источника n_p, G_p^{prev}; интервал между поступлениями для p-й заявки, Y_p; время пребывания в системе для p-й заявки, T_p; PAoI для p-й заявки, A_p. Также в глобальной статистике обновляются значения U; U_n; E; B. Заявки поступают в буфер согласно процедуре, описанной в секции 1.

Во время обслуживания заявок обновляются данные заявки, касающиеся обслуживания: T_p, A_p. Также обновляются следующие элементы глобальной статистики: $M\left[Y\right]$ и $M\left[Y\right]_n$; $M\left[T\right]$ и $M\left[T\right]_n$; $M\left[A\right]$ и $M\left[A\right]_n$.

Выбор новой группы обслуживания проходит по процедуре, описанной в секции 2.1. Глобальная статистика при этом не обновляется.

Заключение

Рассмотренная математическая модель позволяет описать беспроводные сети связи в приложениях Интернета вещей с большей точностью, чем классические модели. В связи со сложностями математического анализа модели для оценки возраста информации и среднего пикового возраста информации было предложено использовать метод имитационного моделирования. В работе построена имитационная модель рассматриваемой системы, в дальнейшей работе предполагается анализ построенной модели на основании интересующих характеристик.

СПИСОК ЛИТЕРАТУРЫ

1. *Kosta A., Pappas N., Angelakis V.* Age of Information: A New Concept, Metric, and Tool // Foundations and Trends in Networking. 2017 Vol. 12, № 3. P. 162–259.
2. *Akar N., Dogan O.* Discrete-time queueing model of age of information with multiple information sources // IEEE Internet of Things Journal. 2021. Vol. 8, № 19. P. 14531-–14542.
3. *Abbas Q., Hassan S.A., Qureshi H.K., Dev K., Jung H.* A comprehensive survey on age of information in massive IoT networks // Computer Communications. 2023. Vol. 197, P. 199–213.

Манаева Варвара Евгеньевна — лаборант-исследователь, кафедра теории вероятностей и кибербезопасности. E-mail: *manaeva_ve@pfur.ru*

Маркова Екатерина Викторовна — к.ф.-м.н., доцент, доцент, кафедра теории вероятностей и кибербезопасности. E-mail: *markova_ev@pfur.ru*

АСИМПТОТИЧЕСКИЙ АНАЛИЗ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ M|M|∞ С MMPP-ПОТОКОМ ОТРИЦАТЕЛЬНЫХ ЗАЯВОК

В. М. Монгуш[1], Д. А. Королев[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В статье рассматривается бесконечнолинейная система массового обслуживания с пуассоновским потоком положительных и марковски модулированным потоком отрицательных заявок. Особенностью модели является тот факт, что положительные заявки начинают свое обслуживание, если в системе нет отрицательных. Поступившие отрицательные не реагируют на присутствующие в системе положительных, а ждут поступления новых положительных заявок, после поступления которых они вместе покидают систему. Обслуживание заявок из очереди происходит с экспоненциальным распределением времени обслуживания. Для описания системы построена система уравнений Колмогорова, учитывающая переходы состояний цепи Маркова, поступление и обслуживание заявок. **Ключевые слова:** *системы массового обслуживания, отрицательные заявки, асимптотический анализ, MMPP-поток*

Введение

Системы и сети с отрицательными заявками были введены Е. Геленбе [2, 3], где он представил новый класс сетей массового обслуживания с двумя типами заявок. Первый тип клиентов — это постоянные клиенты, и сервер обслуживает их обычным образом. Таких клиентов называют положительными или постоянными клиентами. Положительный клиент подчиняется установленным правилам обслуживания и маршрутизации, которые определяют динамику рассматриваемой сети. С другой стороны, второй тип отрицательных требований действует как сигнал, побуждающий положительный запрос, если таковой имеется в узле, немедленно покинуть узел. Изначально такие сети применялись для моделирования биофизических нейронных сетей. В этом контексте узел представляет собой нейрон. Положительные и отрицательные запросы, курсирующие по сети, представляют собой сигналы возбуждения и торможения, которые увеличивают или уменьшают на единицу потенциал

нейрона (узла сети), к которому они приходят. Обобщения такой модели сети, предложенной Эролом Геленбе, привели к возникновению класса моделей, называемого в литературе G-сетями [1, 4].

Отрицательные заявки в системах обслуживания могут не только потреблять ресурсы, но и влиять на обработку положительных заявок, то есть вызывать задержки или даже блокировки процессов обслуживания, что отрицательно сказывается на общей эффективности системы. Например, в работе [5] рассматривается RQ-система с разноприоритетными положительными потоками отрицательных заявок и прогулками прибора. Отрицательные заявки инициируют сбои в работе сервера, после которых сервер уходит на ремонт. Исследуются условия стационарности и получены выражения для распределения вероятностей состояний системы как в переходном, так и в стационарном режиме функционирования. В работе [6] авторы проводят анализ системы M/M/1 с катастрофами и отказами серверов; обобщение этих результатов на дополнительное условие блокировки прибора, когда размер очереди превышает пороговое значение приведено в статье [7]. В данной работе исследуется модель бесконечнолинейной системы массового обслуживания с отрицательными заявками, которые накапливаются в системе, уничтожая вновь пришедшие, но при этом не взаимодействуют с теми положительными заявками,которые уже находяься в системе.

1. Математическая модель и постановка задачи

Пусть на вход системы поступают два потока заявок (рис. 1). Поток положительных заявок является простейшим с интенсивностью λ. Если в системе нет отрицательных заявок положительная заявка сразу начинает обслуживание. Время обслуживания на приборе есть экспоненциально распределенная случайная величина с параметром μ. Также на вход поступает MMPP-поток отрицательных заявок с матрицей условных интенсивностей $\mathbf{A}$ и инфинитезимальным генератором $\mathbf{Q}$. Отрицательная заявка, приходящая в систему, не взаимодействуя с обслуживающимися положительными заявками, ожидает прихода новой положительной заявки, «уничтожает» её и обе покидают систему.

Будем полагать, что MMPP-поток, определяется эргодической цепью Маркова $k(t) = 1, 2, ..., K$, заданной матрицей $\mathbf{Q}$ — её инфинитезимальных характеристик с элементами $q_{k_1 k_2}$, набором неотрицательных величин $\alpha_k > 0$, $k(t) = 1, 2, ..., K$.

В работе исследуется трехмерный процесс $\{i(t), l(t), k(t)\}$, где $i(t)$ — число положительных заявок в системе в момент времени t, $l(t)$ – число отрицательных заявок и $k(t)$ – состояние цепи Маркова, управляющей MMPP-потоком.

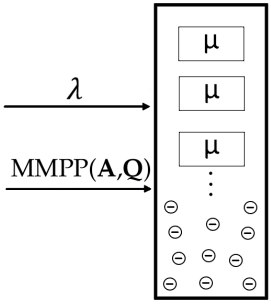


Рис. 1. Система с накопленными отрицательными заявками

2. Система дифференциальных уравнений Колмогорова

Для стационарного распределения вероятностей

$$\Pi(k,i,l) = P\{k(t)=k, i(t)=i, l(t)=l\},\ k=\overline{1,K},\ i \geqslant 0,\ l \geqslant 0$$

построим систему линейных алгебраических уравнений Колмогорова:

$$-\Pi(k,0,0)(\lambda+\alpha_k)+\Pi(k,0,1)\lambda+\Pi(k,1,0)\mu+\sum_{\nu=k}\Pi(\nu,0,0)q_{\nu k}=0,$$

$$-\Pi(k,i,0)(\lambda+\alpha_k+i\mu)+\Pi(k,i-1,0)\lambda+\Pi(k,i+1,0)(i+1)\mu+$$
$$+\Pi(k,i,1)\lambda+\sum_{\nu=k}\Pi(\nu,i,0)q_{\nu k}=0,$$

$$-\Pi(k,0,l)(\lambda+\alpha_k)+\Pi(k,0,l+1)\lambda+\Pi(k,0,l-1)\alpha_k++\Pi(k,1,l)$$
$$+\sum_{\nu=k}\Pi(\nu,0,l)q_{\nu k}=0,$$

$$-\Pi(k,i,l)(\lambda+\alpha_k+i\mu)+\Pi(k,i,l+1)\lambda+\Pi(k,i,l-1)\alpha_k+$$
$$+\Pi(k,i+1,l)(i+1)\mu+\sum_{\nu=k}\Pi(\nu,i,l)q_{\nu k}=0.$$

Введем частичные характеристические функции следующего вида:

$$h(k,u,l)=\sum_{i=0}^{\infty}e^{jui}\Pi(k,i,l),\quad l=0,1,2,...$$

Тогда, система уравнений для частичных характеристических функций будет иметь вид:

$$\begin{cases} j\mu(1-e^{-ju})\dfrac{\partial h(k,u,0)}{\partial u} + (\lambda e^{ju}(\lambda+\alpha_k))h(k,u,0) + \lambda h(k,u,1) + \\ + \sum\limits_{\nu} h(\nu,i,0)q_{\nu k} = 0, \\ \\ j\mu(1-e^{-ju})\dfrac{\partial h(k,u,l)}{\partial u} - (\lambda+\alpha_k)h(k,u,l) + \alpha_k h(k,u,l-1) + \\ + \lambda h(k,u,l+1 + \sum\limits_{\nu} h(\nu,i,l)q_{\nu k} = 0. \end{cases} \quad (1)$$

3. Асимптотическое решение системы уравнений Колмогорова

Поставленную задачу будем решать методом асимптотического анализа. В качестве предельного условия возьмем условие эквивалентно высокой интенсивности входящих потоков заявок (отрицательные и положительные).

Введем обозначения:

$$\lambda_1 = \lambda N, \quad \mathbf{Q_1} = \mathbf{Q}N, \quad \alpha_{k,1} = \alpha_k N, \quad k = \overline{1,K}, \quad \text{где} \quad N \to \infty.$$

Введем матричные обозначения:

$$\mathbf{h}(u,l) = [h(1,u,l)...h(k,u,l)],$$

$$\mathbf{\Lambda} = \lambda I = \lambda_1 N I, \quad \text{где} \quad I - \text{единичная матрица},$$

$$\mathbf{A} = diag[\alpha_{1,1}N, ..., \alpha_{k,1}N] = N \cdot diag[\alpha_{1,1}, ..., \alpha_{k,1}].$$

Тогда система уравнений примет вид:

$$\begin{cases} \dfrac{1}{N} j\mu(1-e^{-ju})\dfrac{d\mathbf{h}(u,0)}{\partial u} + \mathbf{h}(u,0)\left(\mathbf{\Lambda}(e^{ju}-1) - \mathbf{A}\right) + \\ + \mathbf{h}(u,1)\mathbf{\Lambda} + \mathbf{h}(u,0)\mathbf{Q} = 0, \\ \\ \dfrac{1}{N} j\mu(1-e^{-ju})\dfrac{d\mathbf{h}(u,l)}{\partial u} - \mathbf{h}(u,l)(\mathbf{\Lambda}+\mathbf{A}) + \mathbf{h}(u,l-1)\mathbf{A} + \\ + \mathbf{h}(u,l+1)\mathbf{\Lambda} + \mathbf{h}(u,1)\mathbf{Q} = 0. \end{cases} \quad (2)$$

В результате асимптотического анализа была получена характеристическая функция совместного распределения числа положительных

и отрицательных заявок:

$$h(u,l) = \exp\left\{jk_1uN + \frac{j^2k_2uN}{2}\right\}\left(1 - \frac{\alpha_1}{\lambda_1}\right)\left(\frac{\alpha_1}{\lambda_1}\right)^l.$$

Здесь

$$k_1 = \frac{\lambda_1 N}{\mu}d(0) = \frac{\lambda_1 N}{\mu}\left(1 - \frac{\alpha_1 N}{\lambda_1 N}\right) = \frac{N(\lambda_1 - \alpha_1)}{\mu},$$

$$k_2 = \frac{2\lambda_1 N - \alpha_1 N}{\mu} + \frac{\alpha_1 N}{\mu}\left(1 - \frac{\alpha_1 N}{\lambda_1 N}\right) =$$

$$= \frac{(2\lambda_1 - \alpha_1)N}{\mu} + \frac{\alpha_1 N}{\mu}\left(1 - \frac{\alpha_1}{\lambda_1}\right),$$

где $\alpha_1 = \mathbf{rAe}$.

Вектор стационарного распределения вероятностей состояний цепи Маркова $\mathbf{r}$ определяется из системы линейных уравнений.

$$\begin{cases}\mathbf{rQ_1} = 0\\ \mathbf{re} = 1.\end{cases} \tag{3}$$

Численный анализ

Для оценки точности полученного приближения и установления границ его применимости была проведена серия численных экспериментов, в которых асимптотические распределения численно сравнивались с эмпирическими, полученными в результате имитационного моделирования. Разработана программа для симуляции работы системы с ожиданием отрицательных заявок на языке Python. Для экспериментов были выбраны следующие значения параметров:

$$\mathbf{A} = \begin{bmatrix}5 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 4\end{bmatrix},\ \mathbf{Q} = \begin{bmatrix}-1 & 0.5 & 0.5\\ 1 & -2 & 1\\ 0.3 & 0.3 & 0.6\end{bmatrix},\ \lambda = 7,\ \mu = 3.$$

На рис. 2 представлены графики асимптотических распределений вероятностей числа положительных заявок и распределения вероятностей, полученные с помощью имитационной модели для различных значений параметра N.

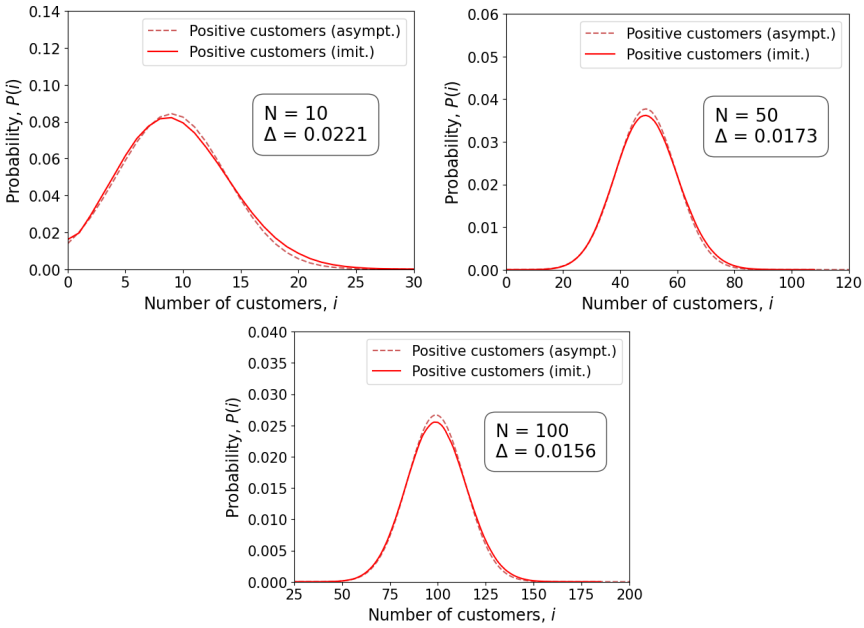


Рис. 2. Сравнение эмпирических и асимптотических распределений вероятностей числа положительных заявок при различных значениях параметра N

Заключение

В статье рассматривается математическая модель системы массового обслуживания с ожиданием отрицательных заявок. Проведен асимптотический анализ этой системы, получена характеристическая функция совместного распределения вероятностей числа положительных и отрицательных заявок. Доказано, что при условии высокой интенсивности входящих потоков асимптотическое распределениея вероятностей числа положительных является гауссовским, а число отрицательных заявок можно аппроксимировать геометрическим. Проведена серия численных экспериментов и сравнение с результатами имитационного моделирования для распределения вероятностей числа положительных заявок в системе. На основании этих результатов оценена точность и область применимости асимптотических приближений. Численный анализ показал, что аппроксимация геометрическим распределением вероятностей числа отрицательных заявок в системе не всегда применима. Например, в случаях, когда в одном из состояний управляющая потоком цепь Маркова находится в 10 и более раз дольше, чем в других. В

ходе дальнейших исследований мы планируем построить аппроксимацию распределения вероятностей числа отрицательных заявок в виде взвешенной суммы геометрических распределений.

СПИСОК ЛИТЕРАТУРЫ

1. *Gelenbe E., Labed A.* G-networks with multiple classes of signals and positive customers // European Journal of Operational Research. 1998. Vol. 108, № 2. P. 293–305.
2. *Gelenbe E., Glynn P., Sigman K.* Queues with Negative Arrivals // Journal of Applied Probability. 1991. Vol. 28, № 1. P. 245–250.
3. *Fourneau J. N., Gelenbe E., Suros R. G.* G-networks with multiple classes of negative and positive customers // Theoret. Comput. Sci. 1996. Vol. 155. P. 141–156.
4. *Artalejo J. R.* A versatile approach for work removal in queueing networks // European Journal of Operational Research 2000. № 2. P. 233–249.
5. *Dimitriou I. A.* A Mixed Priority Retrial Queue with Negative Arrivals, Unreliable Server and Multiple Vacations // Appl. Math. Model. 2013. Vol. 37, № 2., P. 1295–1309.
6. *Мелошникова Н.П., Назаров А.А., Фёдорова Е.А.* Исследование RQ-системы M/M/1 с катастрофами // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2024. № 68. С. 28–37.
7. *Krishna Kumar B., Krishnamoorthy A., Pavai Madheswari S.* Transient analysis of a single server queue with catastrophes, failures and repairs. // Queueing Syst 2007. № 56. P. 133–141.

Монгуш Веронеса Мергеновна — студент магистратуры ИПМКН, НИ ТГУ. E-mail: *aysulu.mongush.2000@mail.ru*

Королев Данил Анатольевич — аспирант ИПМКН, НИ ТГУ. E-mail: *danilkorolev9999@gmail.com*

ПОСТРОЕНИЕ МОДЕЛИ КОМПОЗИЦИИ ВЕБ-СЕРВИСОВ С НЕНАДЕЖНЫМИ СЕРВИСАМИ В ВИДЕ СЕТИ МАССОВОГО ОБСЛУЖИВАНИЯ

Р. С. Назаренко[1], И. Е. Тананко[1], Е. С. Рогачко[1]

[1] *Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, г. Саратов, Россия*

Рассматривается композиция веб-сервисов, используемая при реализации процессов в информационных системах с сервис-ориентированной архитектурой, описание которой выполняется на языке BPEL. Предлагается математическая модель композиции веб-сервисов с ненадежными сервисами в виде открытой экспоненциальной сети массового обслуживания с одним классом требований. Все поступающие в сеть потоки требований предполагаются пуассоновскими. С помощью данной модели могут быть оценены характеристики производительности композиции веб-сервисов такие, как среднее время отклика, надежность, загрузка отдельных сервисов и их композиции. **Ключевые слова:** *композиция веб-сервисов, моделирование, открытая сеть массового обслуживания, ненадежность*

Введение

Распространенной архитектурной моделью как программного обеспечения, так и различных распределенных систем является сервис-ориентированная архитектура, благодаря которой приложения разделяются на независимые функциональные блоки — веб-сервисы, которые взаимодействуют между собой через стандартизированные протоколы [1]. Композиция веб-сервисов представляет собой подход к интеграции функциональных возможностей отдельных сервисов в единую систему, которая обладает свойствами масштабируемости, гибкости и высокой отказоустойчивости.

Большое число исследований композиции веб-сервисов связано с вычислением характеристик ее производительности. Для решения этой задачи используются математические модели, например, в виде сетей Петри, сетей массового обслуживания, цепей Маркова и других стохастических моделей [2–4]. Одной их характеристик производительности композиции веб-сервисов является ее надежность, определяемая в зависимости от надежности отдельных веб-сервисов [5, 6].

В данной работе предлагается математическая модель композиции веб-сервисов с ненадежными сервисами в виде открытой сети массового обслуживания и описывается метод вычисления ее характеристик.

1. Описание композиции веб-сервисов

Выделяют два подхода к построению композиции веб-сервисов — оркестровка и хореография [1]. Оркестровка является исполняемым бизнес-процессом и определяет автоматизированную координацию и управление сложными системами и сервисами. В данной работе рассматривается интеграция сервисов с помощью оркестровки.

Для решения задач оркестровки используется стандарт BPEL, который представляет собой стандарт проектирования и исполнения бизнес-процессов и протоколов взаимодействия процессов между собой. Данный язык позволяет определить последовательность выполнения функционала веб-сервисов и координировать их работу. На рис. 1 приведен пример фрагмента BPEL-кода исполнения некоторого процесса.

Данный процесс можно рассматривать, например, как последовательность операций при проведении банковского перевода. В блоке <receive .../> клиент инициирует перевод через мобильное приложение. В блоке <invoke name="WS1".../> происходит верификация данных (сервис WS1). В блоке <invoke name="WS2".../> проверяется баланс и блокируется сумма на счете отправителя (сервис WS2). В блоке <if ...> ... </if> определяется тип перевода: внутрибанковский перевод по любым реквизитам (WS3), перевод через СБП (WS4), перевод по номеру счета в сторонний банк (WS5). В блоке <while> ... </while> происходит циклическая проверка статуса перевода (сервис WS6). В блоках <invoke name="WS7".../> и <reply .../> отправляется чек клиенту (сервис WS7) и обновляется история операций.

Наиболее важным показателем качества обслуживания композиции веб-сервисов является время отклика, которое зависит от числа поступивших запросов к сервисам и возможности сбоев в работе веб-сервисов.

2. Модель композиции веб-сервисов с надежными сервисами

Построим модель композиции веб-сервисов в виде открытой сети массового обслуживания. Последовательность выполнения и взаимодействия веб-сервисов, представленная логикой соответствующего кода BPEL, отображается топологией сети массового обслуживания и ее маршрутной матрицей. Запросы, получаемые каждым веб-сервисом, эквивалентны требованиям, поступающим в систему массового обслуживания. Отдельные веб-сервисы композиции представляются системами

```
<process...>
<sequence>
  <receive .../>
  <invoke name="WS1".../>
  <invoke name="WS2".../>
  <if ...>
    <condition>"condition1"</condition>
    <invoke name="WS3"... />
  <elseif>
    <condition>"condition2"</condition>
    <invoke name="WS4"... />
  </elseif>
  <else>
    <invoke name="WS5"... />
  </else>
  </if>
  <while> <condition>"condition3"</condition>
    <invoke name="WS6"... />
  </while>
  <invoke name="WS7" ... />
  <reply .../>
</sequence>
</process...>
```

Рис. 1. Пример фрагмента BPEL-кода

массового обслуживания S_i, $i = 1, \ldots, N$, типа $M/M/k_i$. Описанная сеть массового обслуживания по сути является сетью Джексона [7].

Вычислительные возможности сервера определяют время обработки запроса на веб-сервисе, что соответствует в модели длительности обслуживания требований в системе обслуживания сети. Предположим, что длительность обслуживания требований в системе S_i, $i = 1, \ldots, N$, имеет экспоненциальное распределение с параметром μ_i.

Внешние потоки запросов в модели отображаются входящими потоками требований из источника S_0 в системы обслуживания S_i, $i = 1, \ldots, N$, сети. Предположим, что все входящие в сеть потоки требований — пуассоновские с интенсивностями $\lambda_{0,i}$, $i = 1, \ldots, N$.

Переходы требований между системами обслуживания в сети и источником определяются маршрутной матрицей $P = (p_{ij})$, $i, j = 0, \ldots, N$, где p_{ij} — вероятность перехода требований из системы S_i в систему S_j. Очевидно, что $p_{00} = 0$, $p_{N0} = 1$. Вероятности p_{0j}, $j = 1, \ldots, N$, определяются как $p_{0j} = \frac{\lambda_{0,j}}{\lambda_0}$, где $\lambda_0 = \sum_{j=1}^{N} \lambda_{0,j}$.

С помощью математической модели можно оценить производительность композиции веб-сервисов. В частности, среднему времени выполнения запросов соответствует математическое ожидание длительности пребывания требований в сети.

3. Модель композиции веб-сервисов с ненадежными сервисами

Предположим, что среди веб-сервисов существуют ненадежные сервисы, которые вследствие аппаратных или программных сбоев не могут обслуживать запросы в течение случайного интервала времени, называемого длительностью восстановления веб-сервиса. В модели обозначим через $U = \{j_1, \ldots, j_K\}$ — множество номеров систем массового обслуживания, соответствующих ненадежным веб-сервисам, K — число таких систем. Для системы S_j, $j \in U$, вводится дополнительная система массового обслуживания F_j типа $M/M/1$ с интенсивностью обслуживания μ_j^F, моделирующая восстановление веб-сервиса. Вероятность перехода требований из системы S_i, $i = 0, 1, \ldots, N$, в систему S_j, $j \in U$, полагается равной $p_{ij}(1 - q_j)$, а из системы S_i в систему F_j — $p_{ij}q_j$, где q_j отображает вероятность сбоя соответствующего веб-сервиса, $0 < q_j \ll 1$ ($q_j = 0$, если $j \notin U$). После завершения обслуживания в системе F_j, $j \in U$, требования переходят в систему S_j.

Для удобства выполним перенумерацию систем в полученной сети массового обслуживания: $S'_i = S_i$, $i = 0, 1, \ldots, N$, $S'_{N+l} = F_{j_l}$, $l = 1, \ldots, K$, $j_l \in U$, и перейдем к сети, включающей $N + K$ систем обслуживания. Для этой сети интенсивности обслуживания $\mu'_i = \mu_i$, $i = 1, \ldots, N$, $\mu'_{N+l} = \mu^F_{j_l}$, $l = 1, \ldots, K$, $j_l \in U$; интенсивности поступления требований в сеть не меняются. Элементы маршрутной матрицы $P' = (p'_{ij})$, $i, j = 0, 1, \ldots, N + K$, определяются следующим образом: для $i = 0, 1, \ldots, N$ $p'_{ij} = p_{ij}$, если $j \notin U$, $p'_{ij} = p_{ij}(1 - q_j)$, если $j \in U$, $p'_{i,N+l} = p_{ij}q_j$, если $l = 1, \ldots, K$, $j = j_l \in U$; для $i = N + 1, \ldots, N + K$ $p'_{ij} = 1$, если $j = j_{i-N} \in U$; $p'_{ij} = 0$ во всех остальных случаях.

На рис. 2 изображена сеть массового обслуживания, представляющая модель процесса банковского перевода, описанного ранее, при условии, что веб-сервисы WS3, WS4, WS5 являются ненадежными.

4. Метод анализа модели

Рассмотрим метод вычисления характеристик сети Джексона, содержащей $N+K$ систем массового обслуживания. В стационарном режиме интенсивность поступления требований в систему S'_i, $i = 1, \ldots, N + K$, $\lambda_i = \frac{\omega_i}{\omega_0}\lambda_0$. Вектор $\omega = (\omega_i), i = 0, 1, \ldots, N + K$, находится как решение системы уравнений $\omega = \omega P'$ с условием $\sum_i \omega_i = 1$.

Затем вычисляются $\overline{n}_i$, $\overline{n}$ — математические ожидания числа требований в системе S'_i и в сети обслуживания:

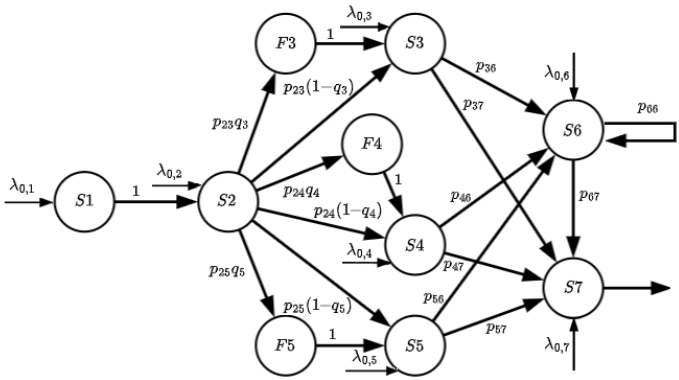


Рис. 2. Пример отображения кода BPEL в виде сети массового обслуживания

$$\overline{n}_i = \frac{\rho_i}{1-\rho_i}, \quad \overline{n} = \sum_{i=1}^{N+K} \overline{n}_i,$$

где $\rho_i = \lambda_i/\mu'_i$; $\overline{u}_i$, $\overline{u}$ — математические ожидания длительности пребывания требований в системе S'_i, $i = 1, \ldots, N+K$, и в сети обслуживания:

$$\overline{u}_i = \frac{1}{\mu'_i(1-\rho_i)}, \quad \overline{u} = \frac{1}{\omega_0} \sum_{i=1}^{N+K} \omega_i \overline{u}_i.$$

5. Численные результаты

Рассмотрим сеть Джексона, изображенную на рис. 2, с параметрами: $k_i = 1$, $i = 1, \ldots, 7$, $(\mu_i) = (200, 260, 300, 400, 100, 300, 300)$, $(\lambda_{0,i}) = (28, 8, 40, 8, 19, 2, 50, 4, 40, 8, 9, 6, 50, 4)$; $p_{23} = 0,2$, $p_{24} = 0,5$, $p_{25} = 0,3$, $p_{36} = 0,7$, $p_{37} = 0,3$, $p_{46} = p_{47} = 0,5$, $p_{56} = 0,2$, $p_{57} = 0,8$, $p_{66} = p_{67} = 0,5$; $q_j = 0,01$, $\mu_j^F = 1$, $j = 3, 4, 5$.

Таблица 1

Зависимость $\overline{u}$ от p_{24}

p_{24}	0,1	0,193	0,287	0,38	0,473	0,567	0,66	0,753
$\overline{u}^{(1)}$	0,138	0,062	0,046	0,039	0,036	0,033	0,032	0,031
$\overline{u}^{(2)}$	0,144	0,067	0,050	0,043	0,040	0,038	0,037	0,037

В табл. 1 показано изменение математического ожидания $\overline{u}$ при увеличении вероятности p_{24}. Как видно из таблицы, величина $\overline{u}^{(2)}$ для модели композиции веб-сервисов с ненадежными сервисами имеет большее значение, чем $\overline{u}^{(1)}$ для модели с надежными сервисами.

Заключение

В данной работе была описана математическая модель композиции веб-сервисов с ненадежными сервисами в виде сети Джексона. В дальнейшем предполагается рассмотреть случай неэкспоненциальных длительностей обработки запросов веб-сервисами.

СПИСОК ЛИТЕРАТУРЫ

1. *Meftah M.* A comparative study of formal approaches for Web service oriented architecture // Network and communication technologies. 2020. Vol. 5, № 2. P. 15–33.
2. *Khalifa F., Bouchiba G.* Aspect oriented Web service composition based Petri net model // Advanced computational techniques for renewable energy systems. Cham: Springer, 2023. P. 148–159.
3. *Gao A., Yang D., Tang S., Zhang M.* Mining models of composite Web services for performance analysis // Database systems for advanced applications. Berlin, Heidelberg: Springer, 2006. P. 828–837.
4. *Zheng Z., Li C., Liu Y., Xi Z.* A phase-type expansion approach for the performability of composite Web services // IEEE Transactions on reliability. 2022. Vol. 71, № 2. P. 579–589.
5. *Deng Z., Chen L., He T., Meng T.* A reliability calculation method for Web service composition using fuzzy reasoning colored Petri nets and its application on supercomputing cloud platform // Future Internet. 2016. Vol. 8, № 4. Article 47. P. 1–21.
6. *Dahan F.* Neighborhood search based improved bat algorithm for Web service composition // Computer systems science and engineering. 2023. Vol. 45, № 2. P. 1343–1356.
7. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. М.: Изд-во РУДН, 1995. 529 с.

Назаренко Регина Сергеевна — магистрант, кафедра системного анализа и автоматического управления. E-mail: *grest.till@bk.ru*

Тананко Игорь Евстафьевич — кандидат физико-математических наук, доцент, заведующий кафедрой, кафедра системного анализа и автоматического управления. E-mail: *TanankoIE@info.sgu.ru*

Рогачко Екатерина Сергеевна — кандидат физико-математических наук, доцент, кафедра системного анализа и автоматического управления. E-mail: *RogachkoES@info.sgu.ru*

АСИМПТОТИЧЕСКИ-ДИФФУЗИОННЫЙ АНАЛИЗ RQ-СИСТЕМЫ С ОТРИЦАТЕЛЬНЫМИ ЗАЯВКАМИ И ВОССТАНОВЛЕНИЕМ ИСКАЖЕННЫХ ПОЛОЖИТЕЛЬНЫХ ЗАЯВОК

А. А. Назаров[1], Я. Е. Измайлова[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В работе рассматривается система с повторными вызовами с двумя входящими потоками заявок: поток положительных заявок и поток отрицательных заявок. Отрицательные заявки искажают обслуживание положительных заявок. Искаженные положительные заявки переходят в блок восстановления, после чего повторяют попытку обслужиться. Система исследуется методом асимптотически-диффузионного анализа. Получены распределение вероятностей числа заявок на орбите для положительных заявок и распределение вероятностей числа заявок в блоке восстановления. **Ключевые слова:** *RQ-система, отрицательные заявки, блок восстановления, асимптотически-диффузионный анализ*

Введение

Модель RQ-системы с отрицательными заявками и с восстановлением искаженных положительных заявок отражает важные аспекты функционирования реальных процессов, где возможны нарушения, сбои работы из-за кибератак требующие повторной обработки или восстановления данных. Такие ситуации характерны для многих реальных систем: в телекоммуникациях – это сбои и помехи в передаче данных, в банковской сфере – это DDoS-атаки, в сфере услуг – ошибки операторов или технические неисправности, что приводит к необходимости восстановления данных и их дальнейшей защите. Например, в компьютерных сетях, если вирус проникает в узел, файл может быть заражен, и системному администратору придется выполнить несколько операций резервного копирования, чтобы восстановить зараженные файлы.

Понятие отрицательных заявок было введено Gelenbe E. [1]. Система массового обслуживания (СМО) M/M/1 с отрицательными заявками была исследована Harrison P. и Pitel E. [2].Yang W. и Chae K. [3] расширили ее до СМО GI/M/1. Применение отрицательных заявок в сетях

массового обслуживания можно найти в обзоре Artalejo J. [4]. Liu Z. и др. [5] проанализировали систему с повторными вызовами (RQ-систему) M/G/1 с отрицательными заявками и обратной связью.

Особенностью рассматриваемой модели является возможность восстановления искаженных положительных заявок.

1. Математическая модель и постановка задачи

Рассмотрим однолинейную систему массового обслуживания с повторными вызовами (рис.1) на вход которой поступают два простейших потока заявок: поток положительных заявок, поток отрицательных заявок. Поток положительных заявок поступает с интенсивностью λ, поток отрицательных заявок поступает с интенсивностью γ.

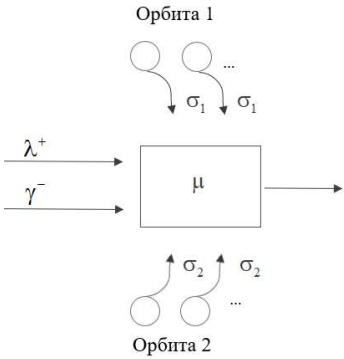


Рис. 1. Математическая модель RQ-системы с отрицательными заявками и блоком восстановления искаженных положительных

Обозначим состояние прибора $k(t)$. Данная величина может принимать два значения: прибор свободен $k = 0$, прибор занят $k = 1$.

Поступившая новая положительная заявка, обнаружив прибор в состоянии $k = 0$, встает для обслуживания. Время обслуживания имеет экспоненциальное распределение. Параметр распределения обозначим через μ. Если в процессе обслуживания не случилось искажения, то после обслуживания заявка покидает систему.

Если положительная заявка, обнаруживает прибор в состоянии $k = 1$, то она переходит на орбиту ожидания (первая орбита). На орбите

ожидания заявки прибывают некоторое случайное время, которое имеет экспоненциальное распределение с параметром σ_1. После задержки заявка вновь пытается обслужиться и также в зависимости от значения k, либо начинает обслуживание, либо мгновенно возвращается обратно и делает следующую задержу.

Если в момент обслуживания положительной заявки в систему приходит отрицательная заявка, то она искажает обслуживание положительной и покидает систему. Искаженная положительная заявка переходит на орбиту для восстановления искаженных положительных заявок (вторая орбита, блок восстановления). На второй орбите заявки осуществляют случайные задержки. Время восстановления имеет экспоненциальный закон распределения с параметром σ_2, после чего заявки обращаются к прибору для обслуживания. Если состояние прибора $k = 1$, то заявки переходят на орбиту для ожидания обслуживания (первая орбита), если состояние прибора $k = 0$, то заявки начинают обслуживание. Если отрицательная заявка приходит в систему и обнаруживает прибор свободным, то она не влияет на состояние системы и покидает ее. Введем следующие обозначения в табл. 1. Тогда распределение вероятностей трехмерного случайного процесса обозначим $P_k(i, n, t)$.

Таблица 1

Обозначения случайных процессов, описывающих состояние системы

Случайный процесс	Обозначение	Возможные значения
Число положительных неискаженных заявок на первой орбите	$i(t)$	$i = 0, 1, 2, ...$
Число восстанавливаемых заявок на второй орбите	$n(t)$	$n = 0, 1, 2, ...$

Цель исследования: найти распределения вероятностей числа заявок на орбите ожидания и на орбите восстановления.

2. Система уравнений Колмогорова

Для распределения вероятностей $P_k(i,n,t)$, запишем систему уравнений Колмогорова:

$$\begin{gathered}\frac{\partial P_0(i,n,t)}{\partial t} = -(\lambda + i\sigma_1 + n\sigma_2)P_0(i,n,t) + \mu P_1(i,n,t) + \gamma P_1(i,n-1,t),\\ \frac{\partial P_1(i,n,t)}{\partial t} = -(\lambda+\gamma+\mu+n\sigma_2)P_1(i,n,t) + \lambda P_0(i,n,t) + \\ +(i+1)\sigma_1 P_0(i+1,n,t) + (n+1)\sigma_2 P_0(i,n+1,t) + \\ +(n+1)\sigma_2 P_1(i-1,n+1,t) + \lambda P_1(i-1,n,t).\end{gathered} \tag{1}$$

Введем частичные производяще-характеристические функции

$$H_k(x,u,t) = \sum_{i=0}^{\infty} x^i \sum_{n=0}^{\infty} e^{jun} P_k(i,n,t), \quad k = 0,1.$$

Тогда система (1) запишется в виде

$$\begin{gathered}\frac{\partial H_0(x,u,t)}{\partial t} = -\lambda H_0(x,u,t) + (\mu + \gamma e^{ju}) H_1(x,u,t) + \\ +j\sigma_2 \frac{\partial H_0(x,u,t)}{\partial u} - \sigma_1 x \frac{\partial H_0(x,u,t)}{\partial x},\\ \frac{\partial H_1(x,u,t)}{\partial t} = -(\lambda(1-x)+\gamma+\mu)H_1(x,u,t) + \lambda H_0(x,u,t) + \\ +\sigma_1 \frac{\partial H_0(x,u,t)}{\partial x} - je^{-ju}\sigma_2 \frac{\partial H_0(x,u,t)}{\partial u} + j\sigma_2(1-xe^{-ju})\frac{\partial H_1(x,u,t)}{\partial u}.\end{gathered} \tag{2}$$

В допредельной ситуации найти решение системы (2) не представляется возможным. Предлагается проводить исследование методом асимптотически-диффузионного анализа в условии длительного восстановления заявок ($\sigma_2 \to 0$).

3. Асимптотически-диффузионный анализ

Обозначим $\sigma_2 = \varepsilon$. Сделаем замены

$$\tau = \varepsilon t, \quad u = \varepsilon w, \quad H_k(x,u,t) = F_k^{(1)}(x,w,\tau,\varepsilon), k = 0,1.$$

Тогда из системы уравнений (2) получим систему уравнений для нахождения функций $F_k^{(1)}(x,w,\tau,\varepsilon)$.

Обозначим решение системы $\lim\limits_{\varepsilon\to 0} F_k^{(1)}(x,w,\tau,\varepsilon) = F_k^{(1)}(x,w,\tau)$, $k = 0,1$. Сформулируем следующее утверждение.

Теорема 1. Решение $F_k^{(1)}(x,w,\tau)$ имеет вид

$$F_k^{(1)}(x,w,\tau)=F_k^{(1)}(x,s,w,\tau)=G_k(x,s)e^{jws(\tau)},k=0,1,$$

где $G(x,s)=G_0(x,s)+G_1(x,s)$ определяется в виде

$$G(x,s)=G_0(1,s)\Big(\frac{\mu+\gamma-(\lambda+s)}{\mu+\gamma-x(\lambda+s)}\Big)^{\frac{\lambda+s}{\sigma_1}}\frac{\mu+\gamma+(1-x)(\lambda+s)}{\mu+\gamma-x(\lambda+s)}, \qquad (3)$$

а $s=s(\tau)$ является решением уравнения

$$s'(\tau)=\gamma G_1(1,s)-s. \qquad (4)$$

Здесь

$$G_0(1,s)=1-\frac{\lambda+s}{\mu+\gamma},\quad G_1(1,s)=\frac{\lambda+s}{\mu+\gamma}.$$

Обозначим правую часть уравнения (5) через $a(s)$. Тогда коэффициент диффузии некоторого диффузионного процесса имеет вид

$$a(s)=\gamma G_1(1,s)-s. \qquad (5)$$

Следствие 1. В RQ-системе с отрицательными заявками и блоком восстановления для искаженных положительных заявок производящая функция распределения вероятностей числа заявок на орбите ожидания имеет вид

$$G(x)=\frac{\mu-\lambda}{\mu}\Big(\frac{\mu-\lambda}{\mu-x\lambda}\Big)^{\frac{\lambda(\mu+\gamma)}{\mu\sigma_1}}\frac{\mu+(1-x)\lambda}{\mu-x\lambda}.$$

Сделаем замены в системе уравнений (2):

$$H_k(x,u,t)=H_k^{(2)}(x,u,t)e^{j\frac{u}{\sigma_2}s(\sigma_2 t)},k=0,1.$$

Обозначим $\sigma_2=\varepsilon^2$. Выполним замены

$$\tau=\varepsilon^2 t,\quad u=\varepsilon w,\quad H_k^{(2)}(x,u,t)=F_k^{(2)}(x,w,\tau,\varepsilon),k=0,1$$

и получим систему уравнений для нахождения функций $F_k^{(2)}(x,w,\tau,\varepsilon)$. Обозначим решение системы через $\lim\limits_{\varepsilon\to 0}F_k^{(2)}(x,w,\tau,\varepsilon)=F_k^{(2)}(x,w,\tau)$. Сформулируем следующее утверждение.

Теорема 2. Функции $F_k^{(2)}(x,w,\tau), k=0,1$ имеют вид

$$F_k^{(2)}(x,w,\tau) = G_k(x,s)\Phi(w,\tau), k=0,1,$$

где $G(x,s)$ определяется (4), функция $\Phi(w,\tau)$ является характеристической функцией случайного процесса $y(\tau) = \lim\limits_{\sigma_2\to 0}\sqrt{\sigma_2}(n(\frac{\tau}{\sigma_2}) - \frac{s(\tau)}{\sigma_2})$, и удовлетворяет дифференциальному уравнению

$$\frac{\partial\Phi(w,\tau)}{\partial\tau} = a'(s)w\frac{\partial\Phi(w,\tau)}{\partial w} - \frac{w^2}{2}b(s)\Phi(w,\tau),$$

где $a(s)$ имеет вид (6), а $b(s) = a(s) + 2(g_1(1,s)\gamma + s)$. Здесь

$$g_1(1,s) = \frac{(a(s)+2s)G_1(1,s) + sG_0(1,s)}{-(\gamma+\mu)} + \\ + \frac{(a(s)+s-\gamma)G_1'(1,s) + (a(s)+s)G_0'(1,s)}{-(\gamma+\mu)}.$$

Функция $b(s)$ имеет смысл коэффициента переноса некоторого диффузионного процесса. Зная коэффициент переноса и диффузии, нетрудно построить дискретное распределение вероятностей

$$P_1(n) = \frac{C}{b(\sigma_2 n)}e^{\frac{2}{\sigma_2}\int\limits_0^{\sigma_2 n}\frac{a(s)}{b(s)}ds}, \quad P_{dif}^{(2)}(n) = \frac{P_1(n)}{\sum\limits_{n=0}^{\infty}P_1(n)},$$

которое и будем называть диффузионной аппроксимацией распределения вероятностей числа заявок на орбите восстановления.

Заключение

В работе исследована RQ-система с отрицательными заявками и с восстановлением искаженных положительных заявок. Найдены распределения вероятностей числа заявок на орбите ожидания и числа заявок на орбите восстановления. Предложенные модель и методы расширяют класс RQ-систем с отрицательными заявками, а также могут быть применены для проектирования систем безопасности данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Gelenbe E.* Random neural networks with negative and positive signals and product form solution // Neural Comput. 1989. № 11. P. 502–510.
2. *Harrison P. G., Pitel E.* Sojourn times in singleserver queues with negative customers // J. Appl. Prob. 1993. № 30. P. 943–963.

3. *Yang W.S., Chae K.C.* A note on the GI/M/1 queue with Poisson negative arrivals // J. Appl. Prob. 2001. № 38. P. 1081–1085.
4. *Artalejo J.* G-networks: a versatile approach for work removal in queueing networks // Eur. J. Oper. Res. 2000. № 126. P. 233–249.
5. *Liu Z., Wu J., Yang G.* An M/G/1 Retrial G – queue with Preemptive Resume And Feedback under N-policy Subject to Server Breakdowns and Repairs // Computers and Mathematics with Applications. 2009. № 58. P. 1792–1807.

Назаров Анатолий Андреевич — доктор технических наук, профессор, профессор кафедры теории вероятностей и математической статистики Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *nazarov.tsu@gmail.com*

Измайлова Яна Евгеньевна — кандидат физико-математических наук, доцент кафедры теории вероятностей и математической статистики Института прикладной математики и компьютерных наук Томского государственного университета. E-mail: *evgenevna.92@mail.ru*

АНАЛИЗ ПРОИЗВОДИТЕЛЬНОСТИ СИСТЕМ DECT-2020 НА ОСНОВЕ МОДЕЛИ ПОЛЛИНГА СО ШЛЮЗОВОЙ ДИСЦИПЛИНОЙ

Д. И. Николаев[1,2], Е. А. Гайдамака[1], А. А. Живцова[1,2]

[1] *Кафедра теории вероятностей и кибербезопасности, Российский университет дружбы народов имени Патриса Лумумбы, г. Москва, Россия*

[2] *Московский институт электроники и математики им. А.Н. Тихонова, Научно-исследовательский институт телекоммуникаций, г. Москва, Россия*

Стандарт DECT-2020 является одной из ключевых технологий 5G для массового межмашинного взаимодействия (mMTC), однако его производительность в условиях высокой нагрузки требует детального исследования. Цель данной работы заключается в разработке и анализе математической модели для оценки ключевых показателей производительности, таких как время доставки данных, в сетях DECT-2020. Для описания процесса доступа множества устройств к общему радиоресурсу используется математическая модель поллинговой системы массового обслуживания (СМО) со шлюзовой дисциплиной и прерыванием обслуживания. Анализ проводился методами вложенных цепей Маркова, а для численной оценки применялось имитационное моделирование. В результате получены аналитические выражения для переходных вероятностей вложенной цепи Маркова, описывающей динамику системы. С помощью имитационного моделирования проведён численный анализ зависимости времени обслуживания заявки от нагрузки в системе. Численные результаты демонстрируют нелинейную деградацию производительности при приближении нагрузки к предельной, что необходимо учитывать при планировании сетей DECT-2020.

Ключевые слова: *DECT-2020, поллинг, система массового обслуживания, теория массового обслуживания, шлюзовая дисциплина, 5G, mMTC*

Введение

Стандарт DECT-2020 NR (New Radio), ратифицированный Международным союзом электросвязи как технология 5G, представляет собой

Исследование выполнено за счет гранта Российского научного фонда № 24-19-00804, https://rscf.ru/project/24-19-00804/.

перспективное решение для построения локальных и частных беспроводных сетей. Он нацелен на два ключевых сценария использования: сверхнадёжная связь с низкой задержкой (URLLC) и массовое межмашинное взаимодействие (mMTC). В режиме mMTC, предназначенном для приложений Интернета вещей (IoT), большое количество автономных устройств (сенсоров, актуаторов) периодически передают небольшие объёмы данных.

Механизм доступа к среде в DECT-2020, описанный в [1] (с. 25, Scheduled access data transfer), предполагает, что базовая станция координирует работу сети, выделяя временные ресурсы подключенным устройствам. Этот процесс можно эффективно смоделировать с помощью систем массового обслуживания с циклическим опросом (поллинговых систем), где сервер (базовая станция) последовательно посещает очереди (устройства) для обслуживания накопившихся заявок (пакетов данных). Анализ таких моделей, являющийся классической задачей теории телетрафика [3], позволяет оценить ключевые показатели производительности сети, такие как среднее время доставки пакета и пропускная способность, в зависимости от числа устройств и интенсивности трафика.

1. Описание модели и методы анализа

Для анализа производительности сети DECT-2020 в режиме mMTC мы используем поллинговую систему массового обслуживания (СМО), состоящую из M очередей с бесконечными буферами и одного обслуживающего прибора. Заявки в i-ю очередь поступают в соответствии с пуассоновским процессом с интенсивностью λ_i. Прибор посещает очереди в циклическом порядке (см. рис. 1). Модель характеризуется следующими ключевыми особенностями:

- **Шлюзовая дисциплина:** прибыв к очереди, прибор обслуживает только те заявки, которые находились в ней в момент подключения. Новые заявки, поступившие во время обслуживания, будут ожидать следующего посещения прибора.
- **Ограниченное время посещения:** фиксированное время (время визита), которое прибор проводит у i-й очереди, является случайной или детерминированной величиной T_i.
- **Прерывание обслуживания (preemptive repeat):** если по истечении времени визита T_i обслуживание заявки не завершено, оно прервётся и при следующем посещении начнётся заново.

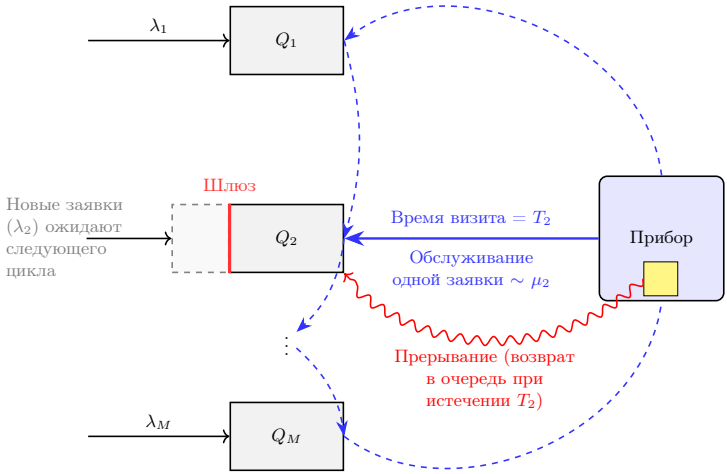


Рис. 1. Схема моделируемой поллинговой системы со шлюзовой дисциплиной и прерыванием обслуживания по времени визита

Такая модель адекватно отражает работу систем с временным разделением доступа (TDMA, Time Division Multiple Access), где каждому устройству выделяется фиксированный временной слот.

Общий теоретический фундамент для исследования поллинговых систем представлен в работе [4]. В случае конечного пространства состояний (например, при ограниченных буферах) эффективны матрично-аналитические методы, позволяющие найти стационарное распределение системы путём решения системы линейных алгебраических уравнений [2]. В данной работе мы фокусируемся на выводе переходных вероятностей вложенной цепи Маркова, моменты которой соответствуют началу визитов к очередям.

2. Формула переходных вероятностей

Рассмотрим важный частный случай, в котором общий ресурс фрейма равномерно распределяется между устройствами. В этом случае время визита T_i к каждой очереди детерминировано, а время обслуживания заявки (передачи пакета) является случайной величиной. Для экспоненциального распределения времени обслуживания с параметром μ_i, вероятность перехода системы из состояния $(\mathbf{n}, i)$ в состояние $(\mathbf{m}, j)$ за

один шаг (один визит) вычисляется по формуле:

$$q_{ij}(\mathbf{n},\mathbf{m}) = \sigma_{ij} e^{-\left(\mu_i + \sum\limits_{k=1}^{M} \lambda_k\right) T_i} \cdot \left(\prod_{\substack{k=1 \\ k \neq i}}^{M} u(\Delta_k) \frac{(\lambda_k T_i)^{\Delta_k}}{(\Delta_k)!} \right) \times$$

$$\times \left[\sum_{r=0}^{n_i - 1} u(\Delta_i + r) \frac{(\mu_i T_i)^r}{r!} \frac{(\lambda_i T_i)^{\Delta_i + r}}{(\Delta_i + r)!} + u(m_i) \left(\sum_{k=n_i}^{\infty} \frac{(\mu_i T_i)^k}{k!} \right) \frac{(\lambda_i T_i)^{m_i}}{m_i!} \right],$$

где $\mathbf{n} = (n_1, \ldots, n_M)$ и $\mathbf{m} = (m_1, \ldots, m_M)$ — векторы числа заявок в очередях до и после визита, i и j — номера обслуживаемой и следующей очереди, $\Delta_k = m_k - n_k$, $u(x)$ — функция Хевисайда, а σ_{ij} — вероятность переключения с очереди i на очередь j. Далее рассматривается случай строгого циклического обхода, при котором вероятность переключения равняется $\sigma_{ij} = 1$ для $j = (i \pmod M) + 1$ и $\sigma_{ij} = 0$ в остальных случаях.

3. Программный симулятор и результаты

Для верификации модели и анализа производительности была разработана имитационная модель, реализующая прямое моделирование событий (event-based simulation) исходного стохастического процесса. Для моделирования были выбраны параметры, характерные для фреймовой структуры DECT-2020 и представленные в табл. 1. В таблице D обозначает число устройств в кластере, что соответствует параметру M в математической модели. Длительность визита T_i в модели равна длине слота T_{slot}.

Таблица 1

Параметры имитационного моделирования

Величина	**Обозн.**	**Значение**
Длина кадра, мс	T_{frame}	10 мс
Длина слота, мс	T_{slot}	$T_{frame}/N_{slot}^{frame} = 10/24$ мс
Число устройств	D	$0, 1, 2, \ldots$
Среднее время обслуживания	$1/\mu$	$1/\mu = T_{slot} = 10/24$ мс
Интенсивность поступления	λ	$\lambda \in (0, \mu/D) = (0, 0, 48)$ мс$^{-1}$

В ходе симуляции оценивалась эмпирическая функция распределения числа циклов опроса, необходимых для полного обслуживания одной случайно выбранной заявки. Результаты для различных значений коэффициента загрузки системы $\rho = D \cdot \lambda/\mu$ приведены на рис. 2, где $\lambda_1 = \lambda_2 = \ldots = \lambda_D = \lambda$, т.е. рассматривается симметричная система.

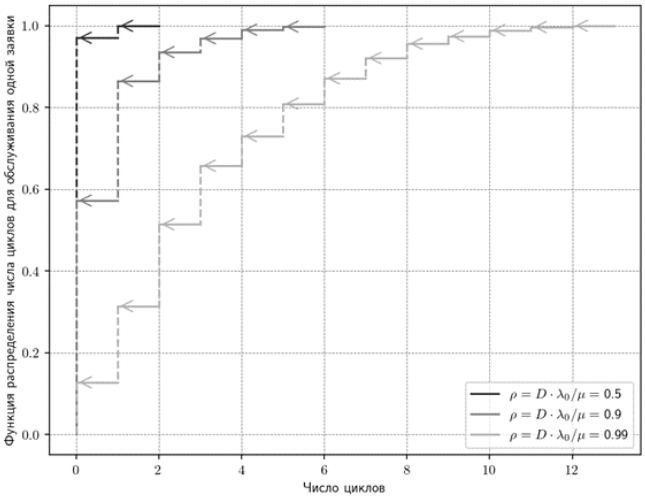


Рис. 2. Эмпирическая функция распределения числа циклов, необходимых для обслуживания одной заявки, при различных коэффициентах загрузки системы ρ

Из графика видно, что с ростом нагрузки производительность системы существенно падает. Если при $\rho = 0,5$ большинство заявок обслуживаются за 1–2 цикла, то при $\rho = 0,9$ это значение возрастает до 4–6 циклов, а при $\rho = 0,99$ для обслуживания заявки может потребоваться более 10 циклов.

Полученные результаты имеют ясную теоретическую интерпретацию. Коэффициент загрузки ρ характеризует, насколько интенсивно используются ресурсы системы. При $\rho \to 1$ система приближается к границе устойчивости. Это означает, что за время одного цикла опроса всех очередей в систему в среднем поступает почти столько же заявок, сколько может быть обслужено.

В результате длина очередей начинает расти, что приводит к увеличению времени ожидания заявкой своего обслуживания. Шлюзовая дисциплина вносит дополнительную задержку: заявка, прибывшая в очередь, когда прибор уже находится у неё, будет ждать следующего полного цикла.

4. Заключение

В работе предложена и проанализирована математическая модель для оценки производительности сетей стандарта DECT-2020 в режиме

mMTC. Модель на основе поллинговой системы со шлюзовой дисциплиной и прерыванием обслуживания позволила получить аналитические выражения для характеристик системы и провести имитационное моделирование.

В такой системе наблюдается сильная нелинейная зависимость времени обслуживания заявок от общей нагрузки в сети. Этот факт необходимо учитывать при планировании сетей, чтобы обеспечить требуемое качество обслуживания (QoS), особенно для критически важных приложений Интернета вещей.

Дальнейшие исследования будут направлены на расширение предложенной модели: рассмотрение альтернативных дисциплин обслуживания (например, исчерпывающей), учёт гетерогенного трафика с различными приоритетами, а также анализ систем с более общими, неэкспоненциальными распределениями времени обслуживания.

СПИСОК ЛИТЕРАТУРЫ

1. ETSI TS 103 636-4 V2.1.1 (2024-10). DECT-2020 New Radio (NR); Part 4: MAC layer. [Электронный ресурс]. URL: https://www.etsi.org/deliver/etsi_ts/103600_103699/10363604/02.01.01_60/ts_10363604v020101p.pdf (дата обращения: 15.06.2025).
2. *Naumov V., Samouylov K., Gredinarova A., Pechinkin A.* Matrix and Analytical Methods for Performance Analysis of Telecommunication Systems. Springer Cham, 2022. 343 p.
3. *Bertsekas D., Gallager R.* Data Networks. Prentice-Hall, 1992. 556 p.
4. *Vishnevskii V. M., Semenova O. V.* Mathematical methods to study the polling systems // Autom Remote Control. 2006. Vol. 67. P. 173–220.

Николаев Дмитрий Иванович — студент, кафедра теории вероятностей и кибербезопасности, Российский университет дружбы народов им. П. Лумумбы; м.н.с., Московский институт электроники и математики им. А. Н. Тихонова, Научно-исследовательский институт телекоммуникаций. E-mail: *nikolaev-di@rudn.ru, dinikolaev@hse.ru*

Гайдамака Елизавета Александровна — студент, кафедра теории вероятностей и кибербезопасности, Российский университет дружбы народов им. П. Лумумбы. E-mail: *gaydamaka-ea@rudn.ru*

Живцова Анна Александровна — студент, кафедра теории вероятностей и кибербезопасности, Российский университет дружбы народов им. П. Лумумбы; м.н.с., Московский институт электроники и математики им. А.Н. Тихонова, Научно-исследовательский институт телекоммуникаций. E-mail: *zhivtsova-aa@rudn.ru, azhivtsova@hse.ru*

МОДЕЛЬ СИСТЕМЫ СВЯЗИ СО СКАНИРОВАНИЕМ АБОНЕНТСКОЙ ОБЛАСТИ ОБСЛУЖИВАЮЩИМИ ПРИБОРАМИ

С. Б. Пичугин[1]

[1] *ПАО РКК «Энергия», г. Королёв, Россия*

построена математическая модель системы связи спутник с многолучевой антенной, сканирующей в надир. Оценивалась вероятность занятости заданного числа лучей обслуживанием абонентов. Для построения оценки использовался математический аппарат теории массового обслуживания и комбинаторики. Полученные результаты позволяют судить о перспективах использования фазированных антенных решёток и иных многолучевых антенн с узкими диаграммами направленности на борту спутников связи.

Ключевые слова: *спутник, многолучевая антенна, система массового обслуживания, обслуживающий прибор, комбинаторика*

Введение

Системы сканирующей связи — это разновидность технологий, использующих автоматическое сканирование и переключение каналов или частот для эффективной передачи данных или мониторинга радиоволн. Они варьируются от портативных радиосканеров (для прослушивания радиоволн) до сложных электронно-сканирующих антенных решеток (фазированных решеток) и даже сетевых сканеров для аудита ИТ-инфраструктуры, где «сканирование» означает проверку портов или протоколов. В данной работе предлагается математическая модель системы связи со сканированием спутника - ретранслятора [1]). Модель системы связи со сканированием представлена На рис. 1.

Приборы из множества (1) обслуживающих приборов сканируют (с заданной скоростью сканирования) область, в которой в случайные моменты появляются и исчезают абоненты из сообщества абонентов 2. Появившийся абонент занимает ближайший к нему прибор 4, чья траектория сканирования пересекается с точкой его появления. Занятый прибор 4 перестаёт сканировать область до момента завершения обслуживания. С ростом числа абонентов в области сканирования число

занятых приборов увеличивается, а свободных приборов уменьшается. Соответственно, увеличивается период сканирования области свободными приборами. С ростом числа занятых приборов время ожидания абонентом свободного прибора увеличивается. Наибольшее число точек появления абонентов случайно и может превышать число обслуживающих приборов. Требуется оценить вероятности занятости заданного числа приборов 1. .

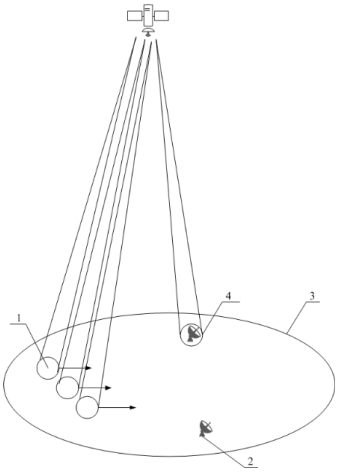


Рис. 1. Модель системы связи со сканированием

1. Построение модели объекта исследования

Для оценки работы системы связи со сканированием предлагается математическая модель на базе аппарата теории массового обслуживания (рис. 2).

Построена модель объекта вида, приведенная на рис. 2:

Введем в рассмотрение случайные процессы $i_1(t)$ — число незанятых приборов на фазе поиска, $i_2(t)$ — число занятых приборов на фазе обработке. Обозначим число приборов $n = i_1 + i_2 \to \infty, \quad i_2 = n - i_1$. Данный процесс является двумерным марковским. Граф переходов по состояниям представлен на рис. 3.

Схема переходов системы, как объекта, из состояния в состояние, приведена на рис. 3. Система дифференциальных уравнений Колмого-

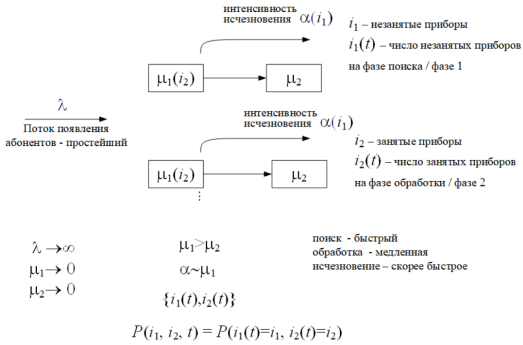


Рис. 2. Математическая модель системы связи со сканированием в виде СМО

рова [5] принимает вид

$$\frac{\partial p(i,t)}{\partial t} = (-\lambda + \mu_{1i} + \mu_2 - i\alpha_i)p(i,t) + (1-\lambda)p(i-1,t) + \lambda p(i+1,t). \quad (1)$$

Для произвольных i_1 и i_2 равных i, при $i \to \infty$ имеем

$$\frac{\partial p(0,t)}{\partial t} = -(\lambda + \mu_{10} + \mu_2 + 0)p(0,t) + (1-\lambda)p(-1,t) + \lambda p(1,t)$$

$$\frac{\partial p(1,t)}{\partial t} = -(\lambda + \mu_{10} + \mu_2 + 0)p(1,t) + (1-\lambda)p(0,t) + \lambda p(2,t) \quad (2)$$

...

$$\frac{\partial p(i,t)}{\partial t} = -(\lambda + \mu_{10} + \mu_2 + i\alpha_i)p(i,t) + (1-\lambda)p(i-1,t) + \lambda p(i+1,t)$$

...

$$\frac{\partial p(n,t)}{\partial t} = -(\lambda + \mu_{10} + \mu_2 + i\alpha_i)p(n,t) + (1-\lambda)p(n-1,t) + \lambda p(n+1,t).$$

Для стационарного распределения, когда p, и $i(t) = i$ от времени не зависят, при произвольных $i = 1...n$, система уравнений (2) преобразуется в следующий вид:

$$-(\lambda + \mu_{10} + \mu_2 + 0)p(0,t) + (1-\lambda)p(-1,t) + \lambda p(1,t) = 0,$$

$$-(\lambda + \mu_{10} + \mu_2 + \alpha_i)p(1,t) + (1-\lambda)p(0,t) + \lambda p(2,t) = 0,$$

...

$$-(\lambda + \mu_{10} + \mu_2 + i\alpha_i)p(i,t) + (1-\lambda)p(i-1,t) + \lambda p(i+1,t) = 0, \quad (3)$$

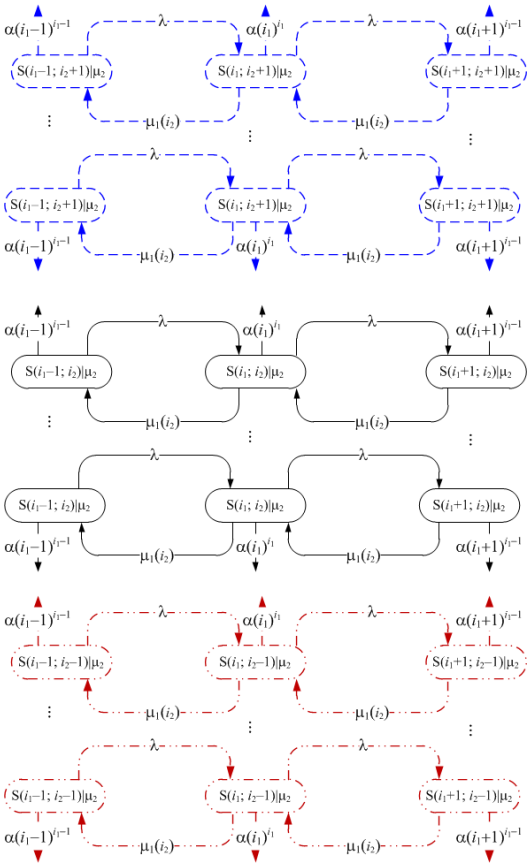


Рис. 3. Схема переходов системы из состояния в состояние

...

$$-(\lambda+\mu_{10}+\mu_2+i\alpha_i)p(n,t)+(1-\lambda)p(n-1,t)+\lambda p(n+1,t)=0.$$

Обозначим:

$$p(i-1,t)=x_i;\quad 1-\lambda=m;\quad (\lambda+\mu_{10}+\mu_2)=k;$$

Дополним (3) уравнением полной вероятности и перепишем в виде

$$-\lambda x_0-(k+0)x_1=0,$$

$$mx_0-\lambda x_1-(k+\alpha_i)x_3=0,$$

$$\begin{gathered} \ldots \\ mx_i - \lambda x_{i+1} - (k + i\alpha_i)x_{i+2} = 0, \\ \ldots \\ mx_{n-2} - \lambda x_{n-1} - (k + i\alpha_i)x_n = 0, \\ x_{i-1} + x_i + \ldots + x_n = 1. \end{gathered} \tag{4}$$

2. Результаты исследования построенной модели

Систему (4) решаем методом обращения матриц. Имеем

$$A * X = B, \tag{5}$$

где A – матрица коэффициентов системы,является прямоугольной размерности $[(n+1)n]$ имеет вид, приведенный в виде табл. 1: X – вектор-

Таблица 1

Табличный вид матрицы A

$-l$	$-(k+0)$	0	0	0	0	...	...	0
m	$-l$	$-(k+a_1)$	0	0	0	...	...	0
0	m	$-l$	$-(k+2a_1)$	0	0	...	...	0
0	0	m	$-l$	$-(k+3a_1)$	0	...	...	...
...	...	...	...	...	...	...	...	...
0	0	0	...	...	m	$-l$	$-(k+(n-2)a_{(n-2)}$	0
0	0	0	...	...	0	m	$-l$	$-(k+(n-1)a_{(n-1)}$
0	0	0	...	...	0	0	m	$-l$
1	1	...	...	...	...	...	...	1

строка вероятностей состояний системы B — вектор-столбец свободных членов ситемы уравнений (4). Так как система (4) дополнена уравнением полной вероятности и не является квадратной, то и матрица коэффициентов $\mathbf{A}$ (см. табл. 1) квадратной не является. Чтобы получить её решение, воспользуемся известным методом, предполагающим умножение обеих частей матричного уравнения на транспонированную A^T, получая в результате квадратную матрицу $(\mathbf{A}^T * \mathbf{A}) * \mathbf{X} = \mathbf{A}^T * \mathbf{B}$ и, имея её, решим систему (4)

$$(\mathbf{A}^T * \mathbf{A})^{-1} * \mathbf{A}^T\mathbf{B} = X. \tag{6}$$

На базе решения (6) были получены значения вероятностей нахождения исследуемой системы в состояниях от «все приборы системы свободны» до «все приборы системы заняты», что позволяет оценить вероятность доставки сообщения по вероятности нахождения системы в заданном состоянии. Это позволяет определить закон распределения, медианное время ожидания и среднее квадратическое отклонение времени ожидания абонентом системы времени обслуживания одним из приборов этой системы.

Заключение

В ходе исследования работы спутника с многолучевой антенной, сканирующей в надир, были получены соотношения для оценки вероятности занятости обслуживанием абонентов заданного числа лучей. Работа спутника моделировалась системой массового обслуживания с множеством обслуживающих приборов и вариацией числа обслуживающих приборов в зависимости от состояния системы, при этом использовался математический аппарат систем массового обслуживания и комбинаторики. Модель включала систему линейных уравнений Колмогорова, дополненную уравнением полной вероятности, решение которой было получено методом обращения матриц. Полученные результаты позволяют судить о перспективах использования фазированных антенных решёток и иных многолучевых антенн с узкими диаграммами направленности на борту спутников связи.

СПИСОК ЛИТЕРАТУРЫ

1. *Пичугин С. Б.* Спутниковый ретранслятор: пат. 2793898 Рос. Федерация: МПК H 04 B 7/185. Заявитель и патентообладатель: ПАО РКК «Энергия» им. С.П. Королёва. № 2022127471; заявл. 21.10.2022; опубл. 07.04.2023, Бюл. № 12. 17 с.: ил.
2. *Назаров А. А., Терпугов А. Ф.* Теория массового обслуживания: учеб. пособие Томск: Изд-во НТЛ, 2004. 228 с.
3. *Вечтомов Е. М., Широков Д. В.* Математика: логика, множества, комбинаторика: учебник для вузов 2-е изд. М.: Юрайт, 2025. 176 с. URL: `https://urait.ru/bcode/563780` (дата обращения: 06.06.2025).

Пичугин Сергей Борисович — доктор технических наук. E-mail: *sergey.pichugin@rsce.ru*

ОЦЕНКА ДЛИТЕЛЬНОСТИ СБОРКИ ТРЕБОВАНИЙ ИЗ ФРАГМЕНТОВ ДЛЯ СЕТИ МАССОВОГО ОБСЛУЖИВАНИЯ С ДЕЛЕНИЕМ И СЛИЯНИЕМ ТРЕБОВАНИЙ

О. С. Постнова[1], И. Е. Тананко[1], Е. С. Рогачко[1]

[1] *Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, г. Саратов, Россия*

Рассматривается открытая сеть массового обслуживания с делением и слиянием требований. При поступлении в сеть требования делятся на фиксированное число родственных фрагментов. Фрагменты независимо друг от друга обслуживаются в системах экспоненциальной ациклической сети массового обслуживания. Сбор требований из родственных фрагментов осуществляется в выделенной системе обслуживания — системе сборки. Предлагается метод вычисления математического ожидания длительности сборки требований из фрагментов. Знание данной характеристики позволяет определить математическое ожидание длительности пребывания требований в рассмотренной сети массового обслуживания с делением и слиянием требований. Сети данного класса могут использоваться в качестве математических моделей информационно-вычислительных систем и сетей с параллельной и распределенной обработкой данных.
Ключевые слова: *ациклическая сеть массового обслуживания, деление и слияние требований, фрагменты требований, длительность сборки требований*

Введение

Сети массового обслуживания с делением и слиянием требований являются удобными математическими моделями высокопроизводительных вычислительных систем, систем многопутевой маршрутизации, а также многих технических систем, в которых возможна параллельная обработка. Особенностью таких сетей обслуживания является мгновенное разделение требования на фрагменты, одновременное начало обслуживания с последующим процессом сборки требований. Наиболее простыми моделями являются сети параллельных систем массового обслуживания [1–3]. Сеть обслуживания с произвольной структурой и с бесконечноприборными системами обслуживания рассмотрена в работе [4].

В этой сети отображены процессы разделения, передачи и слияния родственных фрагментов требований. В качестве математической модели сети передачи данных с пакетной коммутацией может быть использована эффективная в вычислительном отношении открытая ациклическая сеть массового обслуживания вместе с системой сборки требований [5].

В данной работе предлагается метод вычисления математического ожидания длительности пребывания фрагментов требований в системе сборки в сети массового обслуживания с делением и слиянием требований.

1. Описание модели

Рассматривается открытая сеть массового обслуживания N^f с делением и слиянием требований, состоящая из L систем массового обслуживания S_i, $i = 1, \ldots, L$, типа $M/M/1$. Интенсивности обслуживания в системах S_i обозначим через μ_i. Сеть N^f также включает систему обслуживания S_{L+1} — систему сборки [5]. В сеть из источника S_0 поступает пуассоновский поток требований одного класса с интенсивностью λ_0. Каждое поступающее в сеть N^f требование разделяется на K фрагментов, $K \geq 2$. Фрагменты одного требования называются родственными. Они независимо друг от друга направляются в системы S_i, $i = 1, \ldots, L$, с которыми связан S_0.

Родственные фрагменты переходят между системами сети и системой сборки независимо друг от друга и независимо от фрагментов других требований в соответствии с маршрутной матрицей $\boldsymbol{\Theta} = (\theta_{ij})$, $i, j = 0, 1, \ldots, L+1$, где θ_{ij} — вероятность того, что после пребывания в системе S_i фрагмент перейдет в систему S_j. В систему сборки фрагменты поступают после того, как завершится их обслуживание в системах S_i, $i = 1, \ldots, L$. Как только в систему S_{L+1} поступает последний из родственных фрагментов, они мгновенно объединяются в исходное требование, которое покидает систему сборки и сеть N^f.

Обозначим через N описанную ранее сеть массового обслуживания, но без системы сборки. Предположим, что сеть N является ациклической. Элементы маршрутной матрицы $\bar{\boldsymbol{\Theta}} = (\bar{\theta}_{ij})$, $i, j = 0, 1, \ldots, L$, сети N определяются как $\bar{\theta}_{ij} = \theta_{ij}$, и $\bar{\theta}_{i,0} = \theta_{i,L+1}$. Предполагается, что для систем S_i, $i = 0, 1, \ldots, L$ (кроме тех, для которых $\bar{\theta}_{i,0} = 1$), число систем S_j, в которые возможен переход фрагментов из S_i, намного больше K, и вероятности $\bar{\theta}_{ij} > 0$, $j = 0, 1, \ldots, L$, являются сравнимыми.

Определим λ_i, $i = 1, \ldots, L$, — интенсивности потоков фрагментов требований в системы сети N: $\lambda_i = K\lambda_0\omega_i/\omega_0$, где вектор $\boldsymbol{\omega} = (\omega_i)$, $i = 0, 1, \ldots, L$, находится как решение уравнения $\boldsymbol{\omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Theta}}$ с условием $\sum_{i=0}^{L} \omega_i = 1$. Необходимым условием существования стационарного ре-

жима сети N является выполнение неравенств $\lambda_i < \mu_i$, $i = 1, \ldots, L$, или

$$\lambda_0 < \frac{\omega_0}{K} \min_{i=1,\ldots,L} \frac{\mu_i}{\omega_i} = a. \tag{1}$$

2. Вычисление математического ожидания длительности сборки требований

Будем называть «первым» фрагментом требования первый из K поступающих в систему сборки родственных фрагментов. В момент поступления «последнего» из K родственных фрагментов мгновенно происходит объединение фрагментов в единое требование, которое покидает систему сборки. Таким образом, длительность сборки требования из родственных фрагментов определяется как длительность интервала времени между моментами выхода первого и последнего родственных фрагментов из сети N.

Обозначим $\bar{r} = (r_1, \ldots, r_K)$ — маршрут требования в сети N, r_k — маршрут k-го фрагмента требования, $k = 1, \ldots, K$, $\bar{r} \in R$, R — конечное множество всех маршрутов требований. Длительность маршрута r_k, определяемая как длительность времени, в течение которого k-ый фрагмент требования пребывает в сети N, является случайной величиной Y_k, $k = 1, \ldots, K$.

Математическое ожидание длительности сборки требований $E\tau^s$ определяется выражением

$$E\tau^s = \sum_{\bar{r} \in R} E[\max\{Y_1, \ldots, Y_K\} - \min\{Y_1, \ldots, Y_K\}] P(\bar{r}), \tag{2}$$

где $P(\bar{r})$ — вероятность маршрута $\bar{r}$ требования, $\sum_{\bar{r} \in R} P(\bar{r}) = 1$. Так как маршруты фрагментов требований независимы друг от друга, то $P(\bar{r}) = \prod_{k=1}^{K} P(r_k)$, $P(r_k)$ — вероятность маршрута r_k фрагмента требования.

Предположим, что длина маршрута r_k равна m_k, $k = 1, \ldots, K$, тогда $r_k = (i_1, \ldots, i_{m_k})$ представляет собой последовательность номеров систем обслуживания, соответствующую маршруту r_k, и

$$P(r_k) = \bar{\theta}_{0,i_1} \bar{\theta}_{i_1,i_2} \ldots \bar{\theta}_{i_{m_k-1},i_{m_k}}.$$

Случайные величины Y_k, $k = 1, \ldots, K$, имеют обобщенное распределение Эрланга порядка m_k с различными параметрами $\gamma_{i_1}, \ldots, \gamma_{i_{m_k}}$. Параметры определяются выражением $\gamma_l = \mu_l - \lambda_l$, $l = i_1, \ldots, i_{m_k}$. Для удобства записи при фиксированном r_k положим $\alpha_1 = \gamma_{i_1}, \ldots, \alpha_m = \gamma_{i_{m_k}}$. Тогда плотность распределения и функция распределения

вероятностей случайной величины Y_k имеют вид соответственно [6]:

$$f_k(y) = (-1)^{m-1} \prod_{i=1}^{m} \alpha_i \sum_{j=1}^{m} \frac{e^{-\alpha_j y}}{\prod_{l=1, l\neq j}^{m} (\alpha_j - \alpha_l)}, \quad y > 0;$$

$$F_k(y) = (-1)^{m-1} \prod_{i=1}^{m} \alpha_i \sum_{j=1}^{m} \frac{1 - e^{-\alpha_j y}}{\alpha_j \prod_{l=1, l\neq j}^{m} (\alpha_j - \alpha_l)}, \quad y > 0.$$

Если $\alpha_1 = \cdots = \alpha_m = \alpha$, то получим распределение Эрланга порядка m с параметром α, для которого

$$f_k(y) = \frac{\alpha(\alpha y)^{m-1} e^{-\alpha y}}{(m-1)!}, \quad F_k(y) = 1 - \sum_{l=0}^{m-1} \frac{(\alpha y)^l}{l!} e^{-\alpha y}, \quad y > 0.$$

Так как $Y_1, \ldots, Y_K$ — независимые случайные величины, то плотности распределения и функции распределения случайных величин $\min\{Y_1, \ldots, Y_K\}$ и $\max\{Y_1, \ldots, Y_K\}$ имеют вид соответственно [6]:

$$g_1(y) = \sum_{k=1}^{K} f_k(y) \prod_{l=1}^{K} \frac{1 - F_l(y)}{1 - F_k(y)}, \quad G_1(y) = 1 - \prod_{k=1}^{K} (1 - F_k(y)), \quad y > 0;$$

$$g_2(y) = \sum_{k=1}^{K} \frac{f_k(y)}{F_k(y)} \prod_{l=1}^{K} F_l(y), \quad G_2(y) = \prod_{k=1}^{K} F_k(y), \quad y > 0.$$

Следовательно,

$$E[\max\{Y_1, \ldots, Y_K\} - \min\{Y_1, \ldots, Y_K\}] = \int_0^\infty y(g_2(y) - g_1(y))dy.$$

Если случайные величины $Y_1, \ldots, Y_K$ имеют одинаковое распределение с плотностью $f(y)$ и функцией распределения $F(y)$, то

$$g_1(y) = Kf(y)(1 - F(y))^{K-1}, \quad G_1(y) = 1 - (1 - F(y))^K, \quad y > 0;$$

$$g_2(y) = Kf(y)(F(y))^{K-1}, \quad G_2(y) = (F(y))^K, \quad y > 0.$$

Рассмотрим частный случай $m_k = 2$, $k = 1, \ldots, K$. Тогда

$$f_k(y) = \frac{\alpha_1 \alpha_2}{\alpha_2 - \alpha_1} (e^{-\alpha_1 y} - e^{-\alpha_2 y}), \quad y > 0;$$

$$F_k(y) = 1 - \frac{1}{\alpha_2 - \alpha_1} (\alpha_2 e^{-\alpha_1 y} - \alpha_1 e^{-\alpha_2 y}), \quad y > 0.$$

Если $\alpha_1 = \alpha_2 = \alpha$ для всех $k = 1, \ldots, K$, то

$$f(y) = \alpha^2 y e^{-\alpha y}, \quad F(y) = 1 - e^{-\alpha y} - \alpha y e^{-\alpha y}, \quad y > 0.$$

При этом

$$g_1(y) = K\alpha^2 y e^{-\alpha K y}(1 + \alpha y)^{K-1}, \quad y > 0;$$
$$g_2(y) = K\alpha^2 y e^{-\alpha y}(1 - e^{-\alpha y} - \alpha y e^{-\alpha y})^{K-1}, \quad y > 0.$$

3. Численные результаты

Рассматривается сеть N^f с параметрами: $L = 30$, $K = 3$; маршрутные вероятности $\bar{\theta}_{0i} = \bar{\theta}_{ij} = 1/15, \bar{\theta}_{j0} = 1, i = 1, ..., 15, j = 16, ..., 30$. Приводятся результаты вычислительных экспериментов при двух вариантах векторов интенсивностей обслуживания в системах сети N: вектор одинаковых интенсивностей обслуживания $\mu^1 = (\mu_i^1)$, $\mu_i^1 = 4$, $i = 1, ..., 30$; вектор различных интенсивностей обслуживания $\mu^2 = (\mu_i^2)$, $\mu_i^2 = 4 + 0{,}1(i-1), i = 1, ..., 30$. При таких параметрах сети N согласно условию (1) существования стационарного режима $a = 20$.

В табл. 1 приводятся значения математического ожидания длительности сборки, вычисленные аналитически (АМ) по формуле (2) и с помощью имитационной модели (ИМ) [7], при различных значениях λ_0. Результаты имитационного моделирования получены с доверительной вероятностью 0,95 и шириной доверительного интервала 0,01. Из таблицы видно, что результаты аналитического вычисления и имитационного моделирования различаются не более чем на 3%.

Таблица 1

Математическое ожидание длительности сборки требований

λ_0		0,1	1,0	3,0	7,0	11,0	15,0	19,0
μ^1	АМ	1,769	1,689	1,511	1,156	0,800	0,444	0,089
	ИМ	1,725	1,653	1,486	1,148	0,783	0,451	0,091
μ^2	АМ	2,325	2,241	2,056	1,679	1,290	0,873	0,350
	ИМ	2,268	2,185	2,010	1,661	1,287	0,879	0,343

Заключение

В работе предложен метод вычисления математического ожидания длительности сборки требований из фрагментов для сети массового обслуживания с делением и слиянием требований. Данный метод является

вычислительно затратным и применим только для сетей массового обслуживания небольшой размерности, поэтому в дальнейшем представляет интерес разработка более эффективного в вычислительном отношении метода нахождения рассмотренной характеристики.

СПИСОК ЛИТЕРАТУРЫ

1. *Flatto L., Hahn S.* Two parallel queues created by arrivals with two demands I // SIAM Journal on Applied Mathematics. 1984. Vol. 44, № 5. P. 1041—1053.
2. *Ko S. S., Serfozo R.* Response times in $M/M/s$ fork–join networks // Advances in Applied Probability. 2004. Vol. 36, № 3. P. 854—871.
3. *Nelson R., Tantawi A. N.* Approximate analysis of fork/join synchronization in parallel queues // IEEE Transactions on Computers. 1988. Vol. 37, № 6. P. 739–743.
4. *Осипов О. А., Тананко И. Е.* Сети массового обслуживания произвольной топологии с делением и слиянием требований: случай бесконечноприборных систем обслуживания // Вестник Тверского государственного университета. Серия: Прикладная математика. 2017. № 4. С. 43–58.
5. *Постнова О. С., Тананко И. Е., Рогачко Е. С.* Приближенный анализ длительности пребывания требований в сети массового обслуживания с делением и слиянием требований // Управление большими системами. 2025. № 115. С. 33-51.
6. *Вентцель Е. С., Овчаров Л. А.* Теория вероятностей и ее инженерные приложения. М: Наука, 1988. 480 с.
7. *Постнова О. С., Тананко И. Е.* Свидетельство № 2025660299. Имитационная модель сетей массового обслуживания с делением и слиянием требований, 22.04.2025.

Постнова Оксана Сергеевна — ассистент, кафедра системного анализа и автоматического управления. E-mail: *popova.np0062@gmail.com*

Тананко Игорь Евстафьевич — кандидат физико-математических наук, доцент, заведующий кафедрой, кафедра системного анализа и автоматического управления. E-mail: *TanankoIE@info.sgu.ru*

Рогачко Екатерина Сергеевна — кандидат физико-математических наук, доцент, кафедра системного анализа и автоматического управления. E-mail: *RogachkoES@info.sgu.ru*

ОЦЕНКА ПАРАМЕТРОВ РЕКУРРЕНТНЫХ ПОТОКОВ ДЛЯ АНАЛИЗА ПРОИЗВОДИТЕЛЬНОСТИ МНОГОЯДЕРНЫХ ПРОЦЕССОРОВ

Д. Д. Салимзянова[1], С. А. Самойлов[1], Е. Ю. Лисовская[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Современные методы анализа производительности процессоров требуют точного моделирования потоков инструкций и корректной оценки параметров вычислительной нагрузки. Особое внимание уделяется рекуррентным шаблонам, возникающим при выполнении циклов и оптимизированных фрагментов кода. В работе исследованы методы оценки параметров потоков с применением градиентного бустинга. Для анализа были сгенерированы синтетические выборки моментов наступления событий и интервалов между ними, соответствующие различным распределениям (Леви, Парето, Фреше, Ломакса и др.). В качестве признаков использовались как исходные интервалы, так и их статистические характеристики: математическое ожидание, дисперсия, коэффициент вариации, квантили. Качество моделей оценивалось с помощью коэффициента детерминации (R^2) и средней абсолютной процентной ошибки ($MAPE$). Результаты показали, что использование числовых характеристик значительно повышает точность оценки параметров по сравнению с исходными интервалами.

Ключевые слова: *оценка параметров, рекуррентный поток, градиентный бустинг, тяжелые хвосты, идентификация потоков*

Введение

Современные многоядерные процессоры обеспечивают значительный рост вычислительной мощности, однако эффективность использования этих процессоров во многом определяется особенностями распределения вычислительной нагрузки. Для анализа производительности уже недостаточно традиционных метрик, так как ключевую роль начинают играть микроархитектурные характеристики — организация кэш-памяти, механизмы предсказания переходов, глубина конвейера и уровни параллелизма на уровне инструкций, ILP и потоков, TLP [1, 2, 3].

Современные методы анализа производительности процессора требуют точного моделирования и параметризации потоков инструкций, особенно в контексте повторяющихся (рекуррентных) шаблонов [4, 5]. Эти шаблоны могут возникать из-за наличия циклов, рекурсий или оптимизированных участков кода, и оказывают значительное влияние на поведение микросхем на уровне микроархитектуры – таких, как предсказание переходов, кеширование и конвейерная обработка.

Важной задачей является корректная оценка параметров, характеризующих как программные потоки (например, частота инструкций загрузки/выгрузки данных, интенсивность обращений к памяти), так и аппаратные метрики (CPI, IPC, latency, Cache Miss Rate). Применение методов для оценки этих параметров позволяет выявлять узкие места в архитектуре, прогнозировать производительность и оптимизировать распределение нагрузки между ядрами.

1. Описание исследуемых данных и используемых архитектур

Для решения задачи оценки параметров были использованы данные о моментах наступления события в потоке t_i, которые были получены путем моделирования [6]. При формировании наборов данных для обучения было принято решение использовать не моменты наступления событий t_i, а интервалы между этими моментами времени $\tau_i = t_i - t_{i-1}$. Число наблюдений и число интервалов в итоговой выборке было выбрано 10000. Итоговая выборка была поделена на обучающую и тестовую в соотношении 80/20. Затем 30% от полученной тестовой выборки было выделено для формирования валидационной выборки. В выборку для обучения также была добавлена информация о параметрах распределений.

При моделировании использовались следующие функции распределения длин интервалов между моментами наступления событий в рекуррентном потоке: Леви, Фишера, Парето, Обратное Гамма, Ломакса, Фреше, Бура XII. В табл. 1 представлены значения, с которыми были сгенерированы выборки.

Для решения задач был использован алгоритм градиентного бустинга в реализации библиотеки catboost.

2. Метрики оценки качества

В качестве метрик качества оценки параметров обученной моделью, были использованы следующие метрики:

Таблица 1

Значения параметров функций распределения длин интервалов

Параметр	**Значение параметра**
Количество моментов	10000
Число наблюдений	10000
Леви	
Параметр смещения	$\mu = \text{unif}(0{,}01; 0{,}5)$
Параметр масштаба	$c = \text{unif}(0{,}1; 10)$
Фишера	
Параметр числа степеней свободы числителя	$d_1 = \text{unif}(0{,}01; 1)$
Параметр числа степеней свободы знаменателя	$d_2 = \text{unif}(1{,}5; 3)$
Парето	
Параметр формы	$a = \text{unif}(1{,}5; 3)$
Коэффициент масштаба	$x_m = \text{unif}(0{,}01; 1)$
Обратное Гамма	
Параметр формы	$\alpha = \text{unif}(1{,}5; 3)$
Параметр масштаба	$c = \text{unif}(0{,}01; 1)$
Ломакса	
Параметр формы	$\alpha = \text{unif}(1{,}5; 3)$
Параметр масштаба	$\lambda = \text{unif}(0{,}01; 1)$
Фреше	
Параметр формы	$\alpha = \text{unif}(1{,}5; 3)$
Параметр масштаба	$s = \text{unif}(0{,}01; 1)$
Параметр положения	$m = \text{unif}(0; 0{,}5)$
Бура XII	
Первый параметр формы	$c = \text{unif}(1{,}5; 3)$
Второй параметр формы	$k = \text{unif}(1{,}5; 3)$
Параметр масштаба	$\lambda = \text{unif}(0{,}01; 1)$

1. R^2 – коэффициент детерминации

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2},$$

где n – число наблюдений, y_i – действительное значение переменной, $\hat{y}_i$ – оцененное значение, $\bar{y}$ – выборочное среднее значение y_i.

2. $MAPE$ – средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

3. Оценка параметров

Для каждой функции распределения длин интервалов было обучено по две модели градиентного бустинга, которые обучались на двух наборах данных:

1. Интервалы между моментами времени наступления событий τ_i.
2. Числовые характеристики этих интервалов: математическое ожидание, среднее квадратичное отклонение, дисперсия, коэффициент вариации, 6 квантилей (0,1; 0,25; 0,5; 0,75; 0,9; 0,95).

В табл. 2 представлены значения метрик для моделей, обученных на 1-ом и 2-ом на наборах данных.

Результаты сравнительного анализа демонстрируют значимое преимущество моделей, использующих числовые характеристики, над моделями, обученными на исходных реализациях интервалов. Наибольшая эффективность наблюдалась для распределения Леви, где коэффициент детерминации R^2 для оценки параметра локализации μ увеличился с 0,032 до 0,851, при одновременном снижении средней абсолютной процентной ошибки $MAPE$ в 3,6 раза.

В случае многопараметрических распределений (Фреше с параметрами (α, s, m), Бура XII с (c, k, λ)) использование второго набора данных показало улучшение предсказательной способности: для параметра положения m распределения Фреше достигнуто снижение $MAPE$ с 235,35 до 52,19.

Заключение

В работе проведено исследование оценки параметров рекуррентных потоков с различными законами распределения вероятностей. Анализ полученных результатов позволяет сделать следующий вывод: модели, построенные на числовых признаках интервалов обеспечивают лучшую точность оценки параметров распределений, чем модели, обученные на интервалах между моментами наступления событий τ_i. Это может быть связано с высокой вариативностью интервалов для распределений, которые характеризуются тяжелыми хвостами. Обычно при работе с такими распределениями традиционные подходы к оценке параметров часто

Таблица 2

Значения метрик для моделей, обученных на разных наборах данных (Н.д.)

Распределение	**Н. д.**	**Параметр**	R^2	$MAPE$
Леви	1	μ	0,032	128,981
		c	0,978	11,698
	2	μ	**0,851**	**35,601**
		c	**0,998**	**2,461**
Парето	1	x_m	0,961	12,923
		a	0,863	5,436
	2	x_m	**0,989**	**1,457**
		a	**0,999**	**1,234**
Фишера	1	d_1	0,836	46,751
		d_2	0,844	40,001
	2	d_1	**0,998**	**3,038**
		d_2	**0,951**	**2,971**
Обратное Гамма	1	α	0,582	10,907
		c	0,931	11,732
	2	α	**0,988**	**1,566**
		c	**0,998**	**2,256**
Ломакса	1	α	0,325	13,768
		λ	0,860	23,124
	2	α	**0,954**	**3,045**
		λ	**0,990**	**4,526**
Фреше	1	α	0,050	11,737
		s	0,900	23,256
		m	0,646	235,351
	2	α	**0,907**	**4,107**
		s	**0,989**	**6,842**
		m	**0,964**	**52,188**
Бура XII	1	α	0,717	8,807
		s	0,137	15,845
		m	0,943	11,121
	2	α	**0,982**	**1,864**
		s	**0,799**	**6,682**
		m	**0,990**	**4,734**

оказываются неустойчивыми. Использование числовых характеристик сглаживает шум и обеспечивает более устойчивое представление признаков. Полученные результаты открывают перспективы для развития методов идентификации входящих потоков с различных типов.

Статья подготовлены при поддержке платформы Yandex Cloud, предоставившей доступ к вычислительным ресурсам в рамках Программы поддержки курсов и исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Hennessy J. L., Patterson D. A.* Computer Architecture: A Quantitative Approach. 6th ed. San Francisco: Morgan Kaufmann, 2019. 1527 p.
2. *Eyerman S., Eeckhout L.* Modeling critical sections in Amdahl's law and its implications for multicore performance // ACM SIGARCH Computer Architecture News. 2010. Vol. 38, № 3. P. 362–370.
3. *Hill M. D., Marty M. R.* Amdahl's law in the multicore era // Computer. IEEE, 2008. Vol. 41, № 7. P. 33–38.
4. *Kaxiras S., Martonosi M.* Computer Architecture Techniques for Power-Efficiency // Synthesis Lectures on Computer Architecture. 2008.
5. *Mutlu O., Gómez-Luna J., Ghose S., Ausavarungnirun R.* A Modern Primer on Processing in Memory // Emerging Computing: From Devices to Systems. Computer Architecture and Design Methodologies. Springer, Singapore. 2022. P. 171–243.
6. *Марголис Н. Ю.* Имитационное моделирование. Томск: Издательский Дом Томского государственного университета, 2015. 130 с.

Салимзянова Дарья Дмитриевна — ассистент кафедры теории вероятностей и математической статистики. E-mail: *darya.salimzyanova@mail.ru*

Самойлов Сергей Антонович — студент института прикладной математики и компьютерных наук. E-mail: *sergei.samoilov244@gmail.com*

Лисовская Екатерина Юрьевна — к.ф.-м.н., доцент кафедры теории вероятностей и математической статистики. E-mail: *ekaterina_lisovs@mail.ru*

МОДЕЛИРОВАНИЕ ЖЕЛЕЗНОДОРОЖНЫХ СОРТИРОВОЧНЫХ СИСТЕМ СЕТЬЮ МАССОВОГО ОБСЛУЖИВАНИЯ

Н. В. Сергеева[1], Д. Д. Шабанова[1], И. Е. Тананко[1]

[1] *Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, г. Саратов, Россия*

В качестве математической модели железнодорожной сортировочной станции используется открытая сеть массового обслуживания, состоящая из систем массового обслуживания трех типов. Простейшая система массового обслуживания отражает в модели приемо-отправочные парки, система массового обслуживания с групповым входящим потоком – сортировочную горку, системы массового обслуживания с групповым обслуживанием моделируют сортировочно-отправочный парк. В сеть поступает групповой пуассоновский поток требований. Длительности обслуживания в системах сети являются независимыми случайными величинами с экспоненциальным распределением. Для данной сети получены основные стационарные характеристики. **Ключевые слова:** *сеть массового обслуживания, групповой входящий поток, групповое обслуживание*

Введение

Имитационное и математическое моделирование транспортных систем позволяет анализировать и предсказывать поведение систем, разрабатывать стратегии оптимизации, и проводить эксперименты без реальной эксплуатации. Железнодорожные сортировочные станции (ЖСС) являются важной частью железнодорожной транспортной системы, от слаженной работы которых зависит надежность всего перевозочного процесса.

В данной работе построена математическая модель железнодорожной сортировочной станции открытой сетью массового обслуживания.

Типовые сортировочные станции выполняют следующие действия: прием поездов в парк прибытия, отцепление локомотива, осмотр вагонов, расформирование состава на сортировочной горке, накопление вагонов в сортировочном парке, выставление состава в парк отправления, подключение локомотива, отправление поезда со станции.

1. Математическая модель

Прием поездов в парк прибытия, отцепление локомотива, установка ограждения, оформление документов и осмотр вагонов, снятие ограждения отобразим в модели простейшей системой массового обслуживания. Если в процессе осмотра состава выявлены вагоны с неисправностями, то они направляются в ремонтное депо для восстановления. Этот процесс также представим простейшей системой. Расформирование состава на сортировочной горке (СГ) представим в модели системой массового обслуживания с групповым входящим потоком требований. Процесс накопления вагонов в сортировочном парке опишем в модели системой обслуживания с ординарным потоком и групповым обслуживанием требований. Процессы завершения обслуживания: выставление состава в парк отправления, подключение локомотива и отправление поезда со станции также отразим простейшей системой массового обслуживания.

Таким образом, сеть массового обслуживания (СеМО) состоит из систем массового обслуживания (СМО) трех типов.

Система S_1 представляет собой систему с групповым входящим потоком и групповым обслуживанием. В систему из источника поступает групповой пуассоновский поток требований с интенсивностью λ_0. В каждой группе находится случайное число требований b_1, равномерно распределенное на отрезке $[\alpha; \beta]$. Требования обслуживаются группой. Длительности обслуживания требований независимы и имеют экспоненциальное распределение с параметром μ_1. Группы требований обслуживаются в порядке их поступления. При обслуживании требование может быть забраковано с вероятностью p, $0 < p < 1$. В системе имеется один обслуживающий прибор.

После обслуживания в системе S_1 группа требований переходит в систему S_2. Если требование было забраковано в системе S_1, то оно переходит систему S_3 для восстановления, в противном случае начинает обслуживаться в системе S_2. Система S_2 представляет собой систему массового обслуживания типа $M^{b_2}/M/1$ с групповым входящим потоком. Требования обслуживаются по одному. Длительности обслуживания требований независимы и имеют экспоненциальное распределение с параметром μ_2.

Система S_3 – простейшая система типа $M/M/1$. После завершения обслуживания в системе S_3 требование возвращается в систему S_2 и завершает обслуживание с вероятностью 1.

Если требование обслужилось в системе S_2, то оно переходит в систему S_i, $i = 4, ...L-1$, где L – число СМО в СеМО, согласно вероятностям

перехода θ_{2i}, причем $\sum_{i=4}^{L-1} \theta_{2i} = 1$. Системы S_i представляют собой системы типа $M/M^{b_i}/1$ с групповым обслуживанием требованием. Размер группы b_i определяется для каждой системы постоянной целой величиной из отрезка $[\alpha, \beta]$. Длительности обслуживания групп заявок имеют экспоненциальное распределение с параметром μ_i.

После завершения обслуживания в системе S_i, $i = 4, ..., L-1$, группа требований переходит с вероятностью $\theta_{iL} = 1$ в систему S_L. Система S_L представляет собой систему с групповым входящим потоком заявок и групповым обслуживанием. Для упрощения модели будем считать, что системы S_1 и S_L – системы типа $M/M/1$.

Системы S_1 и S_L отображают в модели приемо-отправочные парки, система S_2 – СГ, система S_3 – вагонную часть депо по ремонту (ВЧДР), системы S_i, $i = 4, ..., L-1$ отображают сортировочно-отправочный парк.

Схема СеМО представлена на рис. 1.

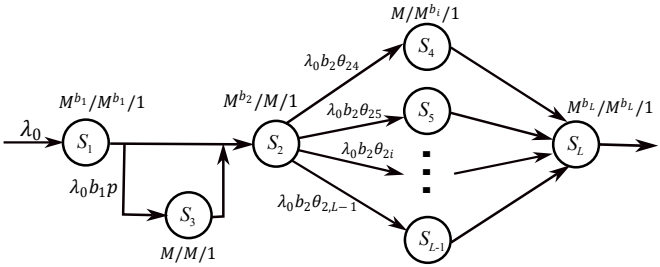


Рис. 1. Схема СеМО

Данную сеть можно рассматривать как сеть Джексона. В этом случае, для вычисления теоретических характеристик систем сети можно рассматривать системы обслуживания с групповым входящим потоком, групповым обслуживанием. Основной характеристикой сети массового обслуживания является математическое ожидание (м.о.) длительности пребывания требований в сети. Построенная модель предполагает тандемное обслуживание требований, поэтому для нашей модели данная характеристика может быть вычислена по формуле

$$\tau = \sum_{i=1}^{2} \bar{u}_i + p\bar{u}_3 + \sum_{i=4}^{L-1} \bar{u}_i \theta_{2i} + \bar{u}_L, \qquad (1)$$

где $\bar{u}_i$ – м.о. длительности пребывания требований в системе S_i, $i = 1, ..., L$. В формуле (1) первое слагаемое представляет собой сум-

му м.о. длительностей пребывания требований в системах S_1 и S_2, поскольку, согласно модели, все требования должны обслужиться в этих системах. Второе слагаемое отвечает за м.о. длительности пребывания в ВЧДР. Туда попадают только забракованные требования. Поскольку требование обслуживается только в одной из систем S_i, $i = 4, ..., L-1$, в которую попадает с вероятностью θ_{2i}, то третье слагаемое представляет собой м.о. длительности пребывания требования в сортировочно-отправочном парке. Последнее слагаемое характеризует м.о. длительности пребывания требований в системе S_L, в которой обслуживаются все требования.

Характеристики $\bar{u}_i$ вычисляются по различным формулам в зависимости от типа системы. Для системы типа $M^{b_2}/M/1$ [1]:
$\bar{u}_2 = \dfrac{l^{(2)} + l}{2\mu_2 l(1-\rho_2)}$, где $\rho_2 = \dfrac{l\lambda_0}{\mu_2}$ – коэффициент использования системы, $l = \sum\limits_{k=1}^{\infty} k l_k$ - математическое ожидание числа требований в группе, l_k – вероятности поступления в систему группы размера k, $l^{(2)} = \sum\limits_{k=1}^{\infty} k^2 l_k$ – второй начальный момент числа требований в группе.

Для системы типа $M/M^{b_i}/1$ [2]:

$$\bar{u}_i = \frac{b_i - 1}{2\lambda_i} + \frac{1}{M - \lambda_i}, \quad i = 4, ..., L-1,$$

где M - корень уравнения: $M^{b_i+1} - (\lambda_i + \mu_i)M^{b_i} + \lambda_i^{b_i}\mu_i = 0$, принадлежащий интервалу $\left(\dfrac{b_i(\lambda_i+\mu_i)}{b_i+1}, \dfrac{(\lambda_i+\mu_i)^{b_i+1} - \lambda_i^{b_i}\mu_i}{(\lambda_i+\mu_i)^{b_i}}\right)$.

Для системы типа $M/M/1$ [3]:

$$\bar{u}_i = (\mu_i - \lambda_i)^{-1}, \quad i = 1, 3, L.$$

2. Численный пример

Рассмотрим ЖСС, которая работает на 3 направления. Станция имеет три парка: два парка приемо-отправочных (парки А, В) и сортировочно – отправочный парк Б. Парки А и Б расположены параллельно, последовательно с ними расположен парк В. Приемо-отправочный парк А имеет 11 путей. Приемо-отправочный парк В – 12 путей. Сортировочно – отправочный парк Б включает 24 пути, разделенных на три пучка по 8 путей в каждом. ЖСС нацелена на расформирование грузовых поездов на отдельные группы вагонов, их обработку и накопление, а также формирование новых поездов из них. Поэтому в модели будем

учитывать только вагоны с переработкой. Транзитные пассажирские или грузовые поезда не рассматриваются. Поезда прибывают с трех направлений. Для упрощения модели будем рассматривать только одно направление.

В результате, СеМО содержит $L = 28$ СМО. Система S_1 моделирует приемо-отправочный парк А, система S_2 моделирует СГ, система S_3 – ВЧДР, системы $S_4, ..., S_{27}$ – сортировочно-отправочный парк Б, система S_{28} – приемо-отправочный парк В. В каждой группе приходит случайное число вагонов равномерно распределенное на отрезке $[50; 70]$. Железнодорожная сортировочная станция перерабатывает в среднем 1500 вагонов в сутки. Это соответствует 62,5 вагонов в час. Примем $\lambda_0 = 1$ поезд в час. Время проведенное составом в системе S_1 в среднем составляет 47 минут (оформление документов, осмотр вагонов), поэтому интенсивность обслуживания $\mu_1 = 1.3$ поезда в час. Технологический цикл горки 19.8 минут (заезд, надвиг, роспуск, окончание формирование и осаживание выезд из парка, постановка и закрепление барьерной группы). Интенсивность обслуживания в системе S_2 примем равную $\mu_1 = 179$ вагонов в час. После обслуживания вагоны распределяются на сортировочные пути равновероятно. Вместимость каждого сортировочного пути 85 вагонов, поэтому на каждом пути примем размер группы равной целому случайному числу равномерно распределенному из отрезка $[50; 70]$. Длительность обслуживания в каждой из систем $S_4, ..., S_{27}$ имеет экспоненциальное распределение с параметром μ_i поездов в час для $i = 4, ..., 27$, причем для каждой системы μ_i выбирается целым числом равномерно распределенным из отрезка $[4; 6]$. Интенсивность входного потока в систему S_3 равна $\lambda_0 b_1 p = 0,6$ вагонов в час, $p = 0,01$, интенсивность обслуживания $\mu_3 = 1,4$ вагонов в час. Для системы S_{28} интенсивность входного потока $\lambda_{28} = 1$ поезд в час, интенсивность обслуживания $\mu_{28} = 4$ поезда в час.

На основе теоретической модели были вычислены м.о. длительности пребывания вагонов на различных этапах обслуживания:
$\bar{u}_i = (3,33; 0,26; 1,25; 13,53; 10,46; 14,33; 12,08; 14,58; 12,54; 12,67; 12,13; 13,71; 13,08; 13,33; 11,00; 14,16; 13,53; 10,79; 13,17; 11,87; 12,70; 13,33; 14,16; 11,87; 11,83; 12,04; 11,45; 0,33)$.

Среднее время пребывания вагона на ЖСС, согласно формуле (1), составило $\tau = 16,1$ часа.

Для проверки теоретических результатов была создана имитационная модель функционирования рассмотренной ЖСС. Натурный эксперимент подтвердил гипотезу о том, что входящий поток на ЖСС можно считать пуассоновским, однако длительности обслуживания не

являются экспоненциальными. Поэтому в имитационной модели был использован равномерный закон распределения длительностей обслуживания. Были вычислены оценки м.о. длительностей пребывания вагонов на различных этапах обслуживания:
$\tilde{u}_i = (3,01; 0,27; 1,54; 12,88; 10,23; 14,23; 11,96; 13,68; 12,18; 12,28; 11,85; 13,11; 13,43; 12,93; 10,72; 14,08; 13,42; 10,49; 12,64; 11,24; 12,05; 12,97; 14,49; 11,51; 11,28; 11,43; 11,41; 0,38)$,
и оценка среднего времени пребывания вагона на ЖСС $\tilde{\tau} = 15,83$ часа.

Анализ результатов показал, что применение в имитационной модели равномерного распределения длительностей обслуживания в системах не дает существенных различий в характеристиках систем, поэтому для анализа подобных систем можно использовать теоретическую модель.

Заключение

Подобные модели могут быть использованы для оптимизации работы железнодорожной сортировочной станции, снижения времени ожидания составов на сортировочных горках, повышения общей пропускной способности системы, прогнозирования работы станции при изменении входного потока.

СПИСОК ЛИТЕРАТУРЫ

1. *Бочаров П. П., Печенкин А. В.* Теория массового обслуживания. М.: Изд-во РУДН, 1995. 529 с.
2. *Stankevich E., Tananko I., Pagano M.* Optimization of Open Queuing Networks with Batch Services // Mathematics. 2022. Vol. 10, No 16. 3027.
3. *Митрофанов Ю. И.* Анализ сетей массового обслуживания. Саратов: Научная книга, 2005. 175 с.

Сергеева Надежда Викторовна — ст. преподаватель, кафедра системного анализа и автоматического управления. E-mail: *SergeevaNV@info.sgu.ru*

Шабанова Дарья Дмитриевна — студентка 3 курса механико-математического факультета. E-mail: *shabanovadaria04@gmail.com*

Тананко Игорь Евстафьевич — к. ф.-м. н., доцент, заведующий кафедрой, кафедра системного анализа и автоматического управления. E-mail: *TanankoIE.sgu@gmail.com*

СМО С НЕНАДЕЖНЫМ ПРИБОРОМ И НЕЭКСПОНЕНЦИАЛЬНЫМ ВРЕМЕНЕМ ЕГО ВОССТАНОВЛЕНИЯ

А. В. Старовойтова[1], А. Н. Моисеев[1], А. А. Назаров[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В работе рассматривается однолинейная система массового обслуживания с ненадежным прибором. На вход поступает простейший поток заявок. Если прибор включен и свободен, заявка обслуживается; если выключен или занят, — переходит в неограниченную очередь. Продолжительность одного периода работы прибора до выключения является экспоненциальной случайной величиной, а случайное время пребывания в состоянии отключения (время восстановления прибора) задается произвольной функцией распределения. В результате исследования получено стационарное распределение вероятностей числа заявок в системе. **Ключевые слова:** *система массового обслуживания, ненадежный прибор*

Введение

Системы массового обслуживания (СМО) с ненадежными приборами находят широкое применение в различных областях, включая телекоммуникации, транспорт и вычислительные сети. Исследование таких систем началось в середине XX века с работ Г. П. Башарина [1] и Б. В. Гнеденко [2], которые рассматривают однолинейные системы с отказами и восстановлением прибора, заложив основы теории надежности в СМО. В последующие годы были разработаны различные модели и методы анализа СМО с ненадежными обслуживающими устройствами.

В данной работе рассматривается однолинейная СМО с ненадежным прибором, функционирование которого описывается двумя состояниями: прибор может находиться в состоянии «включен» или «выключен» (восстанавливаться). В отличие от работы [3], переход из состояния «включен» в состояние «выключен» происходит через экспоненциально распределенное время, а случайное время нахождения в состоянии «выключен» (время восстановления) задается произвольной функцией распределения. Входящие заявки формируются простейшим потоком, и при занятом или выключенном приборе заявки становятся в очередь,

размер которой не ограничен. Целью работы является нахождение стационарного распределения вероятностей числа заявок в системе.

1. Математическая модель и постановка задачи

Рассматривается однолинейная система массового обслуживания (Рис. 1), на вход которой поступает простейший поток заявок с параметром λ. Время обслуживания прибора зависит от состояния прибора: если прибор включен, то время обслуживания есть экспоненциальная случайная величина с параметром μ, если же прибор выключен, то заявки не обслуживаются. Изменение состояний прибора моделируется как случайный процесс $k(t)$ с двумя состояниями:

$$k(t) = \begin{cases} 1, & \text{прибор включен,} \\ 0, & \text{прибор отключен.} \end{cases}$$

Этот процесс находится в состоянии $k(t) = 1$ экспоненциально распределенное время с параметром q_{10}, а продолжительность его пребывания в состоянии $k(t) = 0$ задается произвольной функцией распределения $B_0(t)$. Интервалы времени, на которых $k(t) = 0$ будем называть периодами восстановления.

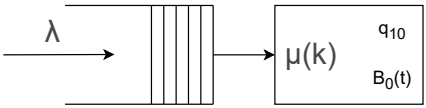


Рис. 1. Модель СМО с ненадежным прибором

При поступлении заявки из входящего потока, если прибор занят или отключен, то заявка становится в очередь. Размер очереди не ограничен. По окончании обслуживания заявки или по окончании периода восстановления прибор мгновенно берет на обслуживание первую заявку из очереди, если очередь не пуста.

Рассмотрим случайный процесс $i(t)$ — число заявок в системе в момент времени t. Ставится задача нахождения стационарного распределения вероятностей числа заявок в системе, то есть процесса $i(t)$.

2. Система уравнений Колмогорова

Введем вспомогательный случайный процесс $z(t)$, который имеет смысл времени, оставшегося до переключения прибора из состояния 0 в состояние 1 в момент времени t. Этот процесс определен только для периодов восстановления.

Рассмотрим марковский процесс с переменным числом состояний $\{k(t), i(t), z(t)\}$. Для его распределения вероятностей $P_0(i,z) = \mathrm{P}\{k(t)=0, i(t)=i, z(t)<z\}$ и $P_1(i) = \mathrm{P}\{k(t)=1, i(t)=i\}$ запишем систему уравнений Колмогорова в стационарном режиме:

$$\begin{cases} \dfrac{\partial P_0(0,z)}{\partial z} - \dfrac{\partial P_0(0,z)}{\partial z}\Big|_{z=0} - \lambda P_0(0,z) + q_{10}P_1(0)B_0(z) = 0, \\ -(\lambda+q_{10})P_1(0) + \mu P_1(1) + \dfrac{\partial P_0(0,z)}{\partial z}\Big|_{z=0} = 0, \\ i>0: \\ \dfrac{\partial P_0(i,z)}{\partial z} - \dfrac{\partial P_0(i,z)}{\partial z}\Big|_{z=0} - \lambda P_0(i,z) + q_{10}P_1(i)B_0(z) + \lambda P_0(i-1,z) = 0, \\ -(\lambda+q_{10}+\mu)P_1(i) + \mu P_1(i+1) + \lambda P_1(i-1) + \dfrac{\partial P_0(i,z)}{\partial z}\Big|_{z=0} = 0. \end{cases}$$

Введем частичные производящие функции

$$G_0(x,z) = \sum_{i=0}^{\infty} x^i P_0(i,z), \quad G_1(x) = \sum_{i=0}^{\infty} x^i P_1(i).$$

Получим следующую систему уравнений:

$$\begin{cases} \dfrac{\partial G_0(x,z)}{\partial z} + \lambda(x-1)G_0(x,z) + q_{10}G_1(x)B_0(z) - \dfrac{\partial G_0(x,0)}{\partial z} = 0, \\ -(\lambda+q_{10})G_1(x) - \mu(G_1(x) - P_1(0)) + \dfrac{\mu}{x}(G_1(x) - P_1(0)) + \\ +\lambda x G_1(x) + \dfrac{\partial G_0(x,0)}{\partial z} = 0. \end{cases} \tag{1}$$

Найдем решение этой системы при $z \to \infty$:

$$\begin{cases} \lambda(x-1)G_0(x) + q_{10}G_1(x) - \dfrac{\partial G_0(x,0)}{\partial z} = 0, \\ -(\lambda+q_{10})G_1(x) - \mu(G_1(x) - P_1(0)) + \dfrac{\mu}{x}(G_1(x) - P_1(0)) + \\ +\lambda x G_1(x) + \dfrac{\partial G_0(x,0)}{\partial z} = 0. \end{cases} \tag{2}$$

Можно заметить, что второе уравнение в (2) не зависит от $G_0(x)$, поэтому, если найти неизвестные константу $\dfrac{\partial G_0(x,0)}{\partial z}$ и вероятность $P_1(0)$, то можно будет выразить из этого уравнения $G_1(x)$ и подставить в первое для нахождения $G_0(x)$, что и будет сделано в следующем разделе.

3. Стационарное распределение числа заявок

Лемма 1. Пусть $\varphi(x,s)$ и $\beta_0(s)$ — преобразования Лапласа от $G_0(x,z)$ и $B_0(z)$ соответственно:

$$\varphi(x,s) = \int_0^\infty e^{-sz} G_0(x,z)dz,$$

$$\beta_0(s) = \int_0^\infty e^{-sz} B_0(z)dz.$$

Тогда

$$\frac{\partial G_0(x,0)}{\partial z} = \lambda(1-x)q_{10}G_1(x)\beta_0(\lambda(1-x)). \tag{3}$$

Доказательство. Применим преобразование Лапласа к первому уравнению системы (1), получим

$$(s-\lambda[1-x])\,\varphi(x,s) = q_{10}G_1(x)\beta_0(s) - \frac{1}{s}\frac{\partial G_0(x,0)}{\partial z}.$$

При $s=\lambda(1-x)$ левая часть обращается в ноль, откуда:

$$\frac{1}{\lambda[1-x]}\frac{\partial G_0(x,0)}{\partial z} = q_{10}G_1(x)\beta_0(\lambda[1-x]),$$

что и доказывает лемму. ■

Лемма 2. Пусть R_0 и R_1 — стационарное распределение вероятностей состояний прибора. Тогда стационарная вероятность $P_1(0)$, выражается через параметры модели как

$$P_1(0) = R_1 - \frac{\lambda}{\mu}. \tag{4}$$

Доказательство. Выполним алгебраические преобразования системы (2): сложим её уравнения, умножим результат на x, а затем разделим на $x-1$. В результате получим:

$$\lambda x\,(G_0(x)+G_1(x)) - \mu G_1(x) = -P_1(0)\mu.$$

Отсюда при $x=1$ имеем

$$G_1(1) - \frac{\lambda[G_0(1)+G_1(1)]}{\mu} = P_1(0)$$

и с учетом условия нормировки $G_0(1) + G_1(1) = 1$ получаем

$$P_1(0) = G_1(1) - \frac{\lambda}{\mu}. \tag{5}$$

Найдем $G_1(1)$ из системы (1) при $x = 1$:

$$\begin{cases} \dfrac{dG_0(1,z)}{dz} + q_{10}G_1(1)B_0(z) - \dfrac{dG_0(1,0)}{dz} = 0, \\ -(\lambda + q_{10})G_1(1) - \mu(G_1(1) - P_1(0)) + \mu(G_1(1) - P_1(0)) + \\ +\lambda G_1(1) + \dfrac{dG_0(1,0)}{dz} = 0. \end{cases}$$

Обозначим:

$$G_1(1) = R_1, \quad G_0(1,z) = R_0(z), \quad R_0'(0) = \frac{dG_0(1,0)}{dz}.$$

Решение системы (1) в новых обозначениях приводит к выражению:

$$R_1 = G_1(1) = \left(1 + q_{10}\int_0^\infty [1 - B_0(s)]ds\right)^{-1}.$$

Подставляя это выражение в (5), получаем (2). ■

Следствие 1. Условие существования стационарного режима работы системы имеет вид:

$$\frac{\lambda}{\mu} < R_1.$$

Теорема 1. Производящая функция $G(x)$ стационарного распределения вероятностей количества заявок в системе имеет вид:

$$G(x) = \frac{[q_{10} + \lambda(x-1)\,(q_{10}\beta_0(\lambda(1-x)) - 1)]\,(\lambda - R_1\mu)}{\lambda\,\{x\,[q_{10} + \lambda(x-1)(q_{10}\beta_0(\lambda(1-x)) - 1)] + \mu(x-1)\}},$$

где

$$R_1 = \left(1 + q_{10}\int_0^\infty [1 - B_0(z)]dz\right)^{-1},$$

$\beta_0(s)$ — преобразование Лапласа функции $B_0(z)$.

Доказательство. Подставляем результаты лемм 1 и 2 в систему уравнений (2). После алгебраических преобразований получаем приведенную формулу для $G(x)$. ■

Замечание 1. Для вычисления стационарного распределения вероятностей числа заявок в системе $\Pi(i)$ можно использовать обратное преобразование Фурье от характеристической функции:

$$\Pi(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jzi} G(e^{jz}) dz,$$

где $j = \sqrt{-1}$.

Заключение

В работе рассмотрена однолинейная система массового обслуживания с ненадежным прибором и простейшим входящим потоком. Время работы прибора распределено экспоненциально, время восстановления — по произвольному закону. С помощью методов дополнительной переменной и производящих функций получено стационарное распределение числа заявок в системе. Найдено условие существования стационарного режима. Результаты применимы для анализа систем связи и вычислительных узлов с произвольным временем восстановления, например, серверов с задержками включения.

СПИСОК ЛИТЕРАТУРЫ

1. *Башарин Г. П.* Один прибор с конечной очередью и заявки нескольких видов // Теория вероятностей и ее применения. 1965. Т. 10, № 2. С. 282–296.
2. *Гнеденко Б. В., Коваленко И. Н.* Введение в теорию массового обслуживания. М.: Наука, 1966. 312 с.
3. *Афанасьева Л. Г., Булинская Е. В.* Математические модели транспортных систем, основанные на теории очередей // Труды Московского физико-технического института. 2010. № 2 (4). С. 6–21.
4. *Вишневский В. М.* Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 506 с.

Старовойтова Александра Владимировна — студентка магистратуры ИПМКН, НИ ТГУ. E-mail: *starovoytova.tsu@gmail.com*

Моисеев Александр Николаевич — доктор физико-математических наук, доцент, заведующий кафедрой программной инженерии ИПМКН, НИ ТГУ. E-mail: *moiseev.tsu@gmail.com*

Назаров Анатолий Андреевич — доктор технических наук, профессор, профессор кафедры теории вероятностей и математической статистики ИПМКН НИ ТГУ. E-mail: *nazarov.tsu@gmail.com*

СХЕМЫ ДОСТУПА К РАДИОРЕСУРСАМ БЕСПРОВОДНОЙ СЕТИ С ДВУМЯ СЕТЕВЫМИ СЛОЯМИ

А. А. Тазаева[1], Е. В. Маркова[1,2]

[1] *Российский университет дружбы народов им. Патриса Лумумбы, г. Москва, Россия*

[2] *Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия*

В данной работе исследуются две модели доступа к радиоресурсам сетей нового поколения 5G/6G. В моделях рассмотрены два типа услуг, генерирующих потоковый и эластичный трафик. Услуги, генерирующие потоковый трафик, отдают приоритет непрерывному резервированию ресурсов для приложений реального времени, обеспечивая стабильное качество обслуживания (Quality of Service, QoS). В отличие от них, услуги, генерирующие эластичный трафик, адаптируются динамически к изменяющимся условиям сети, оптимизируя использование ресурсов для некритичных по времени служб. Для удовлетворения потребностей к QoS в условиях непрерывного роста объема трафика в сети в настоящее время широко используется технология сетевой нарезки (Network Slicing, NS). В работе рассмотрена сеть с двумя сетевыми слоями, один из которых выделен для обслуживания потокового трафика, второй - эластичного. Проведен сравнительный анализ ключевых показателей производительности предложенных схем доступа. Результаты подчеркивают потенциал гибридных стратегий для балансировки строгих требований к качеству обслуживания с гибким управлением ресурсами в развивающихся системах связи. **Ключевые слова:** *управление радиоресурсами, схемы доступа, потоковый трафик, эластичный трафик, QoS, 5G, NS, приоритетное обслуживание*

Введение

В условиях стремительной цифровизации общества и развития систем беспроводной связи, особенно с внедрением технологий 5G и последующих поколений, возникает острая необходимость в адаптации су-

Публикация выполнена в рамках проекта № 021937-2-000 Системы грантовой поддержки научных проектов РУДН

ществующих моделей управления радиоресурсами. Современные приложения реального времени, генерирующие потоковый трафик (потоковое видео, VoIP, промышленная автоматизация), предъявляют строгие требования к качеству обслуживания (Quality of Service, QoS), в то время, как приложения с эластичным трафиком (загрузка файлов, просмотр веб-страниц) выигрывают от динамического перераспределения ресурсов.

Традиционные методы статического разделения ресурсов оказываются недостаточно эффективными для удовлетворения разнородных требований к обслуживанию. В связи с этим возникает потребность в разработке адаптивных моделей доступа, обеспечивающих баланс между надежностью обслуживания и эффективным использованием сетевых ресурсов.

В работе рассмотрена модель управления радиоресурсами беспроводной сети с применением технологии сетевой нарезки (Network Slicing, NS), предусматривающая разделение доступной полосы пропускания на зарезервированные зоны для потокового (чувствительного к задержкам) и эластичного трафика (ориентированного на пропускную способность), что позволяет обеспечить разнородные требования к QoS.

Для удовлетворения требований к QoS рассмотрены две схемы реализации сетевой нарезки:

1. Схема доступа с частичным вытеснением: предусматривает прерывание обслуживания одного или нескольких запросов на передачу одного типа трафика для освобождения минимально необходимого (в случае эластичного трафика) или гарантированного (в случае потокового трафика) числа ресурсов для передачи другого типа трафика.

2. Схема доступа с полным вытеснением: предполагает освобождение всех ресурсов, занимаемых запросами на передачу потокового трафика в гарантированной емкости, выделенной для обслуживания запросов на передачу эластичного трафика.

Данные подходы к распределению ресурсов представляют собой фундаментальный компромисс: частичное вытеснение способствует сохранению качества обслуживания для приложений реального времени, но может приводить к недоиспользованию ресурсов, в то время, как полное вытеснение обеспечивает более эффективное использование ресурсов для передачи эластичного трафика, но может приводить к перебоям в обслуживании. Работа направлена на оценку производительности предложенных схем доступа по таким характеристикам, как вероятность блокировки запроса, вероятность прерывания обслуживания запроса и вероятность приема запроса.

1. Системная модель

Рассмотрим модель беспроводной сети общей емкостью C доступных единиц ресурса. Пусть данная сеть разделена на два логических сегмента (слоя): первый сетевой слой предназначен для обслуживания запросов на передачу потокового трафика (тип 1), а второй – запросов на передачу эластичного трафика (тип 2). Для обеспечения эффективного разделения и управления ресурсами в системе предусмотрены отдельные пулы ресурсов для каждого типа трафика. Выделенный пул C_1 предназначен для обслуживания запросов, генерирующих потоковый трафик, в то время, как пул C_2 обеспечивает потребности запросов на передачу эластичного трафика. Отметим, что суммарная емкость выделенных пулов $C_1 + C_2$ может превышать общее количество доступных ресурсов C, что создает потенциал для перекрытия пулов и требует внедрения механизмов управления конкуренцией за ресурсы. Для обслуживания одного запроса k-го типа необходимо количество ресурсов, равное $b_k, k = 1, 2$, причем для запроса 1-го типа это гарантированное значение, а для запроса 2-го типа – минимальное.

Для обеспечения требований, предъявляемых к качеству обслуживания, в модели предусмотрено резервирование ресурсов Q_k, $Q_k \leq C_k$, $k = 1, 2$ для каждого типа трафика. Механизм резервирования гарантирует, что для обоих типов трафика будет доступно минимум $\lfloor \frac{Q_k}{b_k} \rfloor, k = 1, 2$ ресурсов, даже при конкуренции за них.

В рамках рассматриваемой модели реализовано дифференцированное управление доступом к ресурсам. Запросы на передачу потокового трафика, предъявляющие строгие требования к качеству обслуживания, преимущественно используют свои выделенные ресурсы C_1. При недостатке ресурсов в C_1 система может выделять дополнительные ресурсы из пула C_2, обеспечивая таким образом гибкость в распределении пропускной способности. Аналогичный подход применяется к обработке запросов на передачу эластичного трафика. Будучи адаптируемыми к доступным ресурсам, они преимущественно используют выделенные ресурсы C_2, но при необходимости могут обращаться и к ресурсам из C_1.

2. Управление доступом к ресурсам

Для эффективного управления ресурсами в условиях высокой загрузки сети в модели реализованы два различных механизма вытеснения обслуживаемых запросов. Важно отметить, что процесс вытеснения запросов на передачу эластичного трафика из гарантированной емкости Q_1 всегда происходит по единому сценарию, в то время, как вытесне-

ние запросов на передачу потокового трафика из Q_2 осуществляется по-разному в зависимости от выбранной схемы обслуживания.

1. Первая схема доступа, реализующая механизм минимального (частичного) вытеснения, базируется на консервативном подходе к управлению ресурсами. В случае недостаточного количества свободных ресурсов механизм анализирует возникший дефицит и осуществляет вытеснение минимально необходимого числа запросов на передачу потокового трафика из гарантированной емкости Q_2. Такой подход позволяет минимизировать число прерываний обслуживания потокового трафика, однако может привести к неоптимальному использованию ресурсов при колебаниях сетевой нагрузки.

2. Вторая схема доступа, реализующая механизм полного вытеснения, применяет более агрессивную стратегию управления ресурсами. При поступлении запроса на передачу эластичного трафика система осуществляет полное вытеснение всех запросов на передачу потокового трафика, находящихся в процессе обслуживания в гарантированной емкости Q_2. Данный подход обеспечивает максимальную пропускную способность для эластичного трафика, хотя и связан с повышенным влиянием на качество обслуживания потокового трафика. Ключевым преимуществом этой модели является простота реализации, поскольку она исключает необходимость проведения сложных расчетов дефицита ресурсов в режиме реального времени.

Схема обработки запросов представлена на рис. 1, где λ_k – интенсивность поступления пуассоновских потоков запросов k-го типа, μ_k^{-1} – среднее время обслуживания запроса k-го типа, распределенное по экспоненциальному закону, Π_k – вероятность вытеснения обслуживаемого запроса, B_k – вероятность блокировки поступающего запроса, $k = 1, 2$. Заметим, что для обеих схем доступа возможны следующие случаи при поступлении запроса на передачу трафика любого типа.

1. Запрос может быть принят на обслуживание без вытеснения запросов другого типа. Это возможно, если число обслуживаемых системой запросов не превышает максимально допустимое число запросов данного типа, $n_i < N_i^{max}, i = 1, 2$, и количества свободных ресурсов системы хватает для обслуживания поступающего запроса,

$$(n_i + 1)b_i + n_k b_k \leq C, i \neq k, i = 1, 2, k = 1, 2.$$

2. Запрос будет принят на обслуживание за счет вытеснения запроса/запросов другого типа, если число обслуживаемых системой запросов меньше гарантированного числа запросов данного типа, $n_i < N_i^g$, и количество свободных ресурсов меньше необходимого числа ресурсов

для обслуживания 1-го запроса данного типа, $(n_i + 1)b_i + n_k b_k > C$, $i \neq k$, $i = 1, 2, k = 1, 2$.

3. В остальных случаях запрос будет заблокирован.

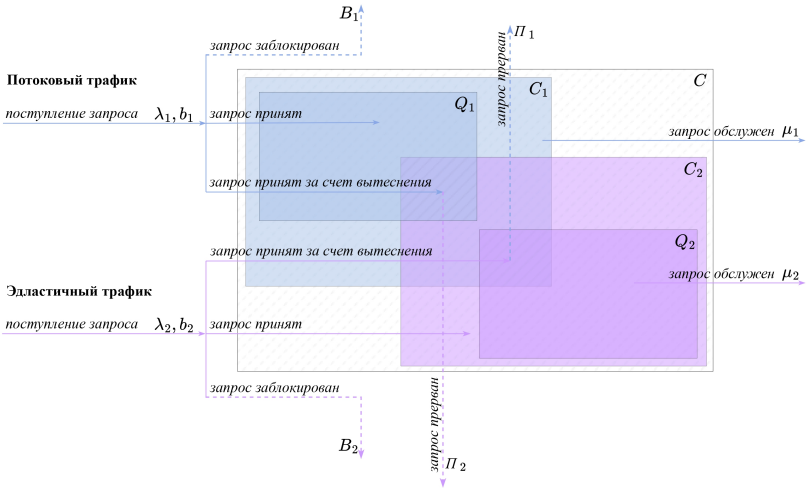


Рис. 1. Схема модели

Заключение

Анализ двух схем распределения ресурсов для беспроводных сетей выявляет фундаментальное противоречие в современной системной архитектуре: как сбалансировать строгие требования к качеству обслуживания для потокового трафика с гибкими требованиями для эластичного трафика. Подход с минимальным вытеснением выступает как более консервативная стратегия, отдавая предпочтение надежности для приложений реального времени за счет освобождения только необходимого минимума ресурсов. В то же время схема с механизмом полного вытеснения принимает более агрессивную позицию, отдавая приоритет освобождению максимальной пропускной способности для адаптируемых приложений, что обеспечивает высокую эффективность системы при доминировании эластичного трафика, но может негативно влиять на работу сервисов, чувствительных к временным задержкам.

Эти выводы показывают, что ни одна из стратегий не может оптимально справляться с изменчивостью реальных сетевых условий. Буду-

щие системы выиграют от адаптивных реализаций, способных сочетать эти подходы.

СПИСОК ЛИТЕРАТУРЫ

1. *Башарин Г. П.* Введение в теорию вероятностей: учеб. пособие для студентов II-II курсов специальностей «Математика», «Прикладная математика». М.: Изд-во РУДН, 1990. 228 с.
2. *Башарин Г. П.* Лекции по математической теории телетрафика: Учебное пособие. М.: Изд-во РУДН, 2009. 342 с.
3. *Бронштейн О. И.* Модели приоритетного обслуживания в информационно-вычислительных системах: научное издание. М.: Наука, 1976. 220 с.
4. *Алиев Т. И.* Основы моделирования дискретных систем. СПб: СПбГУ ИТМО, 2009. 363 с.
5. *Haneul Ko, Jaewook Lee, Sangheon Pack* Priority-based dynamic resource allocation scheme in Network Slicing. // International Conference on Information Networking (ICOIN), 2021. P. 62–64.
6. *Buyakar Tulja Vamshi Kiran, Agarwal Harsh, Tamma Bheemarjuna Reddy, Franklin A. Antony* Resource Allocation with Admission Control for GBR and Delay QoS in 5G Network Slices. // International Conference on COMmunication Systems and NETworkS (COMSNETS), 2020. P. 213–220.

Тазаева Анастасия Анатольевна — студент, кафедра Теории вероятностей и кибербезопасности, Российский университет дружбы народов им. П. Лумумбы. E-mail: *tazaaeva081@gmail.com*

Маркова Екатерина Викторовна — к.ф.-м.н., доцент, доцент кафедры Теории вероятностей и кибербезопасности, Российский университет дружбы народов им. П. Лумумбы, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. E-mail: *markova-ev@rudn.ru*

СРЕДНЕЕ ЧИСЛО ЗАЯВОК В СМО M/M/∞ С ПЕРЕКЛЮЧЕНИЕМ В НЕСТАЦИОНАРНОМ РЕЖИМЕ

Я. А. Тюленина [1], Е. А. Фёдорова[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В работе рассмотрена математическая модель гибридной энергетической системы в виде бесконечнолинейной системы массового обслуживания сс переключением между двумя блоками приборов. Система исследована в нестационарном режиме с помощью метода асимптотического анализа в условии высокой интенсивности входящего потока. Получено выражение для асимптотического среднего числа заявок в системе. **Ключевые слова:** *математическое моделирование, система массового обслуживания, асимптотический анализ*

Введение

Математическое моделирование играет ключевую роль в анализе процессов в различных прикладных областях [1, 2]. Одним из направлений моделирования является теория массового обслуживания, применяющаяся для описания стохастических изменений в технических системах.

В работе исследуется система массового обслуживания с двумя блоками приборов и переключением в качестве математической модели гибридной энергетической системы, сочетающей несколько источников энергии [4]. Блоки обслуживания описывают источники электроэнергии, один из которых постоянный (например, топливное центральное энергоснабжение), а второй непостоянный (альтернативная энергетика). В данной работе предлагается метод асимптотического анализа в условии высокой интенсивности входящего потока для исследования системы в переходном (нестационарном) режиме.

1. Описание модели

Рассмотрим бесконечнолинейную систему массового обслуживания с двумя блоками приборов, один из которых имеет непостоянное время

Исследование выполнено за счет гранта Российского научного фонда № 24-71-00022, https://rscf.ru/project/24-71-00022/

работы. На вход системы поступает простейший поток заявок с интенсивностью λ. Время нахождения первого блока в работающем состоянии распределено по экспоненциальному закону с параметром γ_1, а в неработающем – с параметром γ_0. Когда первый блок приборов не работает, обслуживание происходит на втором блоке. В любой момент времени работает только один из блоков обслуживания. При поступлении в систему заявка начинает обслуживание на работающем блоке, время обслуживания распределено по экспоненциальному закону с параметром μ_1 для первого блока и μ_2 для второго. Схема модели представлена на рис. 1.

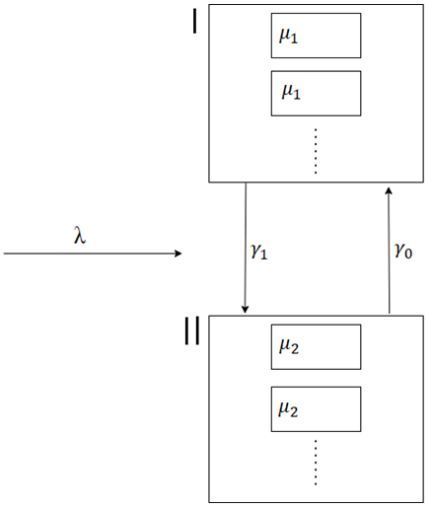


Рис. 1. Математическая модель гибридной системы с переключением

Введем случайные процессы: $i(t)$ – число заявок в системе, $i = \overline{0, \infty}$, $k(t)$ – состояние блоков обслуживания ($k = 1, 2$). Обозначим:

- $P(1, i, t)$ – вероятность того, что в момент времени t в системе работает первый блок и находится i заявок;
- $P(2, i, t)$ – вероятность того, что в момент времени t в системе работает второй блок и находится i заявок.

Двумерный процесс $\{k(t), i(t)\}$ является марковским. Запишем систему дифференциальных уравнений Колмогорова:

$$\begin{cases} \dfrac{\partial P(1,i,t)}{\partial t} = -P(1,i,t)(\lambda + i\mu_1 + \gamma_1) + P(2,i,t)\gamma_0 + \\ +P(1,i-1,t)\lambda + P(1,i+1,t)(i+1)\mu_1, \\ \dfrac{\partial P(2,i,t)}{\partial t} = -P(2,i,t)(\lambda + i\mu_2 + \gamma_0) + P(1,i,t)\gamma_1 + \\ +P(2,i-1,t)\lambda + P(2,i+1,t)(i+1)\mu_2. \end{cases} \tag{1}$$

2. Метод асимптотического анализа

Для исследования системы в переходном режиме, т.е. получения нестационарных характеристик, применим метод асимптотического анализа в условии высокой интенсивности входящего потока [3].

Введём характеристические функции следующего вида:

$$H_1(u,t) = \sum_{i=0}^{\infty} e^{jui} P(1,i,t), \quad H_2(u,t) = \sum_{i=0}^{\infty} e^{jui} P(2,i,t), \quad \text{где } j = \sqrt{-1}.$$

Переобозначим интенсивность входящего потока $\lambda = \lambda_1 N$, где N – параметр высокой интенсивности входящего потока ($N \to \infty$). Тогда из системы уравнений (1), запишем систему дифференциальных уравнений для частичных характеристических функций:

$$\begin{cases} \dfrac{\partial H_1(u,t)}{\partial t} = H_1(u,t)(N\lambda_1(e^{ju} - 1) - \gamma_1) - \\ -j\mu_1 \dfrac{\partial H_1(u,t)}{\partial u}(e^{-ju} - 1) + \gamma_0 H_2(u,t), \\ \dfrac{\partial H_2(u,t)}{\partial t} = H_2(u,t)(N\lambda_1(e^{ju} - 1) - \gamma_0) - \\ -j\mu_2 \dfrac{\partial H_2(u,t)}{\partial u}(e^{-ju} - 1) + \gamma_1 H_1(u,t). \end{cases} \tag{2}$$

Выполним следующие асимптотические замены:

$$N = 1/\varepsilon,\ \gamma_k = \varepsilon\delta_k u = w\varepsilon, H_k(u,t) = F_k(w,t,\varepsilon).$$

Получим асимптотические уравнения:

$$\begin{cases} \varepsilon \dfrac{\partial F_1(w,t,\varepsilon)}{\partial t} = F_1(w,t,\varepsilon)\left(\lambda_1(e^{jw\varepsilon} - 1) - \delta_1\right) - \\ -j\mu_1 \dfrac{\partial F_1(w,t,\varepsilon)}{\partial w}(e^{-jw\varepsilon} - 1) + \delta_0 F_2(w,t,\varepsilon), \\ \varepsilon \dfrac{\partial F_2(w,t,\varepsilon)}{\partial t} = F_2(w,t,\varepsilon)\left(\lambda_1(e^{jw\varepsilon} - 1) - \delta_0\right) - \\ -j\mu_2 \dfrac{\partial F_2(w,t,\varepsilon)}{\partial w}(e^{-jw\varepsilon} - 1) + \delta_1 F_1(w,t,\varepsilon). \end{cases} \tag{3}$$

Согласованное уравнение:

$$\varepsilon\frac{\partial F_1(w,t,\varepsilon)}{\partial t}+\varepsilon\frac{\partial F_2(w,t,\varepsilon)}{\partial t}=(F_1(w,t,\varepsilon)+F_2(w,t,\varepsilon))\lambda_1(e^{jw\varepsilon}-1)- \\ -\left(j\mu_1\frac{\partial F_1(w,t,\varepsilon)}{\partial w}+j\mu_2\frac{\partial F_2(w,t,\varepsilon)}{\partial w}\right)(e^{-jw\varepsilon}-1). \tag{4}$$

Обозначим $\lim\limits_{\varepsilon\to 0}F_k(w,t,\varepsilon)=F_k(w,t)$. Переходя к пределу $\varepsilon\to 0$ в уравнения (5), получаем

$$-\delta_1F_1(w,t)+\delta_0F_2(w,t)=0.$$

Откуда можно сделать вывод, что

$$F_k(w,t)=R_k\Phi(w,t), \tag{5}$$

где R_k удовлетворяют системе

$$\begin{cases}\delta_1R_1-\delta_0R_2=0,\\ R_1+R_2=1.\end{cases}$$

Откуда

$$R_1=\frac{\delta_0}{\delta_0+\delta_1},\quad R_2=\frac{\delta_1}{\delta_0+\delta_1}.$$

Подставляем (5) в уравнение (6):

$$\frac{\partial\Phi(w,t)}{\partial t}=\lambda_1jw\Phi(w,t)+jw(j\mu_1R_0+\mu_2R_1)\frac{\partial\Phi(w,t)}{\partial w}. \tag{6}$$

В стационарном режиме имеем уравнение

$$\Phi(w)\lambda_1+j(\mu_1R_0+\mu_2R_1)\Phi'(w)=0,$$

решение которого имеет вид

$$\Phi(w)=C\cdot\exp\left\{\frac{jw\lambda_1}{\mu_1R_0+\mu_2R_1}\right\}.$$

Пусть $\Phi(w,t)=e^{jwax(t)}$, где $a=\dfrac{\lambda_1}{\mu_1R_0+\mu_2R_1}$. Тогда из уравнения (6), получим линейное неоднородное дифференциальное уравнение первого порядка для функции $x(t)$:

$$ax'(t)=\lambda_1-(\mu_1R_0+\mu_2R_1)ax(t),$$

решение которого можно записать в виде:

$$x(t) = 1 + \left(\frac{x(0)}{a} - 1\right) \cdot e^{-(\mu_1 R_0 + \mu_2 R_1)t}. \tag{7}$$

Возвращаясь к начальным заменам, получим, что асимптотическая характеристическая функция первого порядка имеет вид

$$h_1(u,t) = \exp\{ju \cdot Na \cdot x(t)\} = \exp\left\{ju \cdot \frac{\lambda \cdot x(t)}{\mu_1 R_0 + \mu_2 R_1}\right\},$$

то есть нестационарное асимптотическое математическое ожидание числа заявок в системе описывается выражением:

$$m(t) = \frac{\lambda \cdot x(t)}{\mu_1 R_0 + \mu_2 R_1}. \tag{8}$$

3. Численный анализ

Рассмотрим численный пример. Пусть $\lambda = 100, \mu_1 = 1, \mu_2 = 1,5$, $\gamma_1 = 5$, $\gamma_0 = 10$. Пусть начальное условие $m(0) = 0$. На рис. 2 представлен график $m(t)$ и его схождения к стационарному математическому ожиданию, выражение для которого получено с помощью метода моментов в [5].

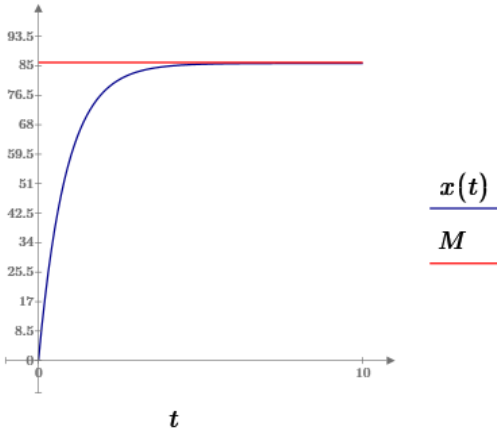


Рис. 2. Среднее число заявок в системе $m(t)$

Заключение

Для исследования СМО $M/M/\infty$ с переключением был применен метод асимптотического анализа в условии высокой интенсивности входящего потока. В результате получено выражение для нестационарного математического ожидания числа заявок в системе.

СПИСОК ЛИТЕРАТУРЫ

1. *Вишневский В. М., Семенова О. В., Шаров С. Ю.* Моделирование и анализ гибридного канала связи на базе лазерной и радио технологий // Управление большими системами. 2011. № 35. С. 237–249.
2. *Марченко О. В., Соломин С. В., Лебедев А. В.* Математическое моделирование энергосистем с возобновляемыми источниками энергии // Информационные и математические технологии в науке и управлении. 2017. № 2. С. 57–64.
3. *Вавилов В. А., Назаров А. А.* Асимптотический анализ системы $M(T)|M(T)|\infty$ // Информационные технологии и математическое моделирование (ИТММ-2024) : материалы XXIII Международной конференции им. А. Ф. Терпугова, 20–26 октября 2024 г. Томск: Изд-во Том. ун-та, 2024. С. 184–189.
4. *Тюленина Я. А., Фёдорова Е. А.* Математическая модель гибридной энергосистемы в виде бесконечнолинейной СМО с переключением // Информационные технологии и математическое моделирование (ИТММ-2023): материалы XXII Международной конференции имени А. Ф. Терпугова (4–9 декабря 2023 г.). Ч. 2. Томск: Изд-во Том. ун-та, 2024. С. 12–18.
5. *Тюленина Я. А., Фёдорова Е. А.* Метод моментов для исследования СМО $M/M/\infty$ с переключением // Материалы I Международного форума «ИТ. Наука. Креатив» (14-16 мая 2024 года, Омск). Т. 5: Системы управления, информационные технологии и математическое моделирование. М.: Колос-с, 2024. С. 219–226.

Фёдорова Екатерина Александровна — кандидат физ.-мат.наук, доцент кафедры теории вероятностей и математической статистики Томского государственного университета. E-mail: *moiskate@mail.ru*

Тюленина Яна Андреевна — аспирант, младший научный сотрудник лаборатории теории массового обслуживания и теории телетрафика Томского государственного университета. E-mail: *yanccch@gmail.ru*

ХАРАКТЕРИСТИКИ ПОКАЗАТЕЛЯ КАЧЕСТВА ОБСЛУЖИВАНИЯ СИСТЕМ С АБСОЛЮТНЫМ И ОТНОСИТЕЛЬНЫМ ПРИОРИТЕТАМИ

Н. Г. Хаустова[1], С. П. Моисеева[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Предлагается модель обработки заявок двух типов — приоритетных и неприоритетных. Время обслуживания заявок случайное с экспоненциальным распределением вероятностей, причем его параметры соответствуют типу заявки. У каждого типа заявок имеется свой бункер ожидания, ограниченный по объему. Информация, находящаяся в очереди, обладает сроком жизни, после истечения которого ее передача может стать неактуальной. Срок жизни — случайная величина, также имеющая экспоненциальное распределение. Для рассматриваемой системы реализован алгоритм поиска технических характеристик путем вычисления стационарного распределения вероятностей, что, в свою очередь, достигается автоматическим построением матриц коэффициентов. Приводится сравнение характеристик системы при абсолютном и относительном приоритетах. **Ключевые слова:** *системы массового обслуживания с приоритетами, абсолютный приоритет, относительный приоритет, построение матриц коэффициентов, распределение вероятностей системы, срок жизни информации*

Введение

Современные системы массового обслуживания играют ключевую роль в управлении бизнес-процессами, обеспечивая минимизацию потерь данных. Одной из важнейших задач в этой области остается исследование систем с приоритетным обслуживанием [1, 2], позволяющих оптимизировать обработку разнородных запросов [3, 4]. Однако при моделировании таких систем возникают сложности, связанные с ограниченной емкостью буферов ожидания [5] и необходимостью учета нетерпеливости заявок. Традиционные методы анализа стационарных распределений вероятностей для малых очередей становятся неприменимыми при масштабировании, что требует разработки новых алгоритмических решений. Такой алгоритм был представлен авторами настоящей статьи для системы с относительным приоритетом [6]. Обновленный алгоритм

для системы с абсолютным приоритетом и сравнение ее характеристик с относительным представлены в данной статье.

1. Математическая модель системы

Пусть $i(t)$ – число приоритетных заявок в бункере в момент времени t, $j(t)$ – число неприоритетных заявок в бункере, $k(t)$ – состояние прибора в момент времени t: $k(t) = 0$ – прибор свободен, $k(t) = 1$ – занят приоритетной заявкой, $k(t) = 2$ – занят неприоритетной заявкой.

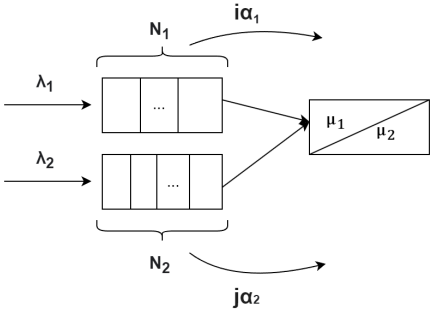


Рис. 1. Математическая модель обработки информации с приоритетами и нетерпеливыми заявками

Ставится задача нахождения стационарного распределения вероятностей трехмерного марковского процесса $\{k(t), i(t), j(t)\}$: $P_k(i,j) = \lim\limits_{t\to\infty} P\{k(t) = k, i(t) = i, j(t) = j\}$ при следующих параметрах системы: λ_p – интенсивности потоков, μ_p – интенсивности обслуживания, α_p – интенсивности выхода из бункеров, N_p – количество мест для ожидания в бункере, $p = 1, 2$.

2. Формализация событий в системе

Так же, как и для алгоритма с относительным приоритетом [6], необходимо начать с формализации основных событий системы. Ключевым событием, определяющим абсолютность приоритета, будет являться приход приоритетной заявки.

События в системе для приоритетного потока:

– **Пришла заявка приоритетного потока (λ_1)**

1. Если прибор свободен, то состояние прибора равно 1.
2. Если прибор занят приоритетной заявкой, и в 1-м бункере есть место, то число заявок в 1-м бункере увеличивается на 1.

3. Если прибор занят неприоритетной заявкой, и 1-й бункер свободен, то пришедшая приоритетная заявка вытесняет неприоритетную, и если во 2-м бункере есть место, то число заявок во 2-м бункере увеличивается на 1, состояние прибора равно 1.

– **Закончилось время ожидания заявки приоритетного потока** ($i\alpha_1$)

3. Если в 1-м бункере есть заявки, то число заявок в 1-м бункере уменьшается на 1.

– **Завершила обслуживание заявка приоритетного потока** (μ_1)

4. Если в 1-м бункере есть заявка, то число заявок в 1-м бункере уменьшается на 1.

5. Если в 1-м бункере нет заявок, во 2-м бункере есть заявка, то состояние прибора равно 2, а число заявок во 2-м бункере уменьшается на 1.

6. Если в 1-м бункере нет заявок и во 2-м бункере нет заявок, то состояние прибора равно 0.

События в системе для неприоритетного потока:

– **Пришла заявка неприоритетного потока** (λ_2)

7. Если прибор свободен, то состояние прибора равно 2.

8. Если прибор занят, и во 2-м бункере есть место, то число заявок во 2-м бункере увеличивается на 1.

– **Закончилось время ожидания заявки неприоритетного потока** ($j\alpha_2$)

9. Если во 2-м бункере есть заявки, то число заявок во 2-м бункере уменьшается на 1.

– **Завершила обслуживание заявка неприоритетного потока** (μ_2)

10. Если в 1-м бункере нет заявок, во 2-м бункере есть заявка, то число заявок во 2-м бункере уменьшается на 1.

11. Если в 1-м бункере есть заявка, то состояние прибора равно 1, а число заявок в 1-м бункере уменьшается на 1.

12. Если в 1-м бункере нет заявок, и во 2-м бункере нет заявок, то состояние прибора равно 0.

На рис. 2 и 3 представлены соответствующие графы переходов состояний рассматриваемого трехмерного марковского процесса.

3. Алгоритм поиска технических характеристик

Основные вычислительные трудности связаны с формированием матрицы коэффициентов, поскольку рост емкости буфера всего на одно

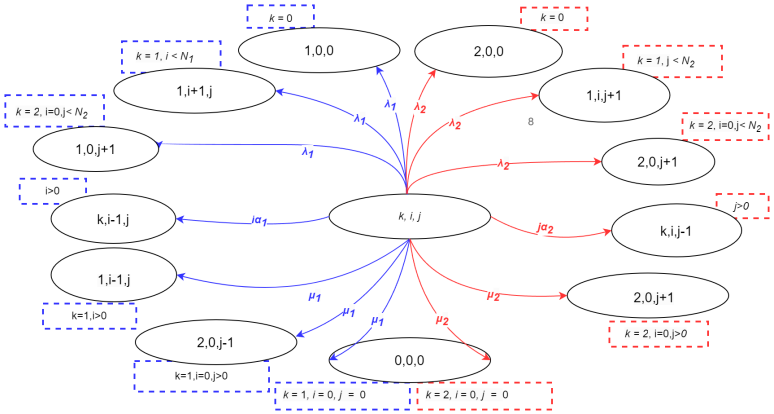


Рис. 2. Граф переходов для диагональных элементов матрицы процесса $\{k(t), i(t), j(t)\}$

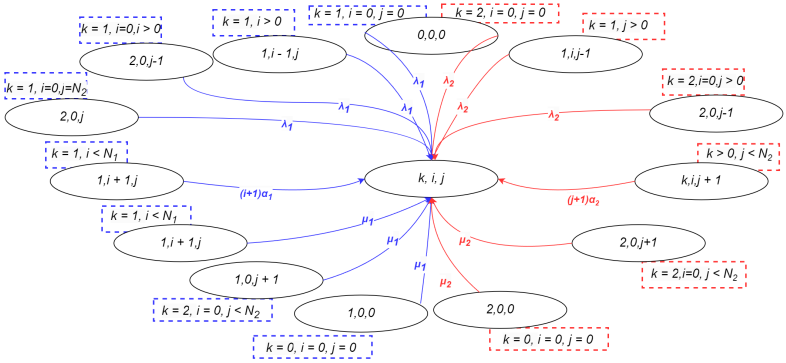


Рис. 3. Граф переходов для определенных состояний системы $\{k(t), i(t), j(t)\}$

место вызывает резкое увеличение числа уравнений, что также увеличивает размер матрицы коэффициентов. Рассмотрим обновленный [6] алгоритм заполнения данной матрицы (табл. 1).

4. Сравнение характеристик систем

В данном разделе приводятся характеристики показателя качества обслуживания систем с входными параметрами $N_1 = 5, N_2 = 8, \lambda_1 = 0,8,\ \lambda_2 = 0,5,\ \mu_1 = 1,7,\ \mu_2 = 1,6,\ \alpha_1 = 0,09,\ \alpha_2 = 0,03$ для работы с относительным/абсолютным приоритетами (табл. 2).

Таблица 1

Алгоритм заполнения матрицы коэффициентов

Условие	**Действие**
$k=0,\ i=0,\ j=0$	$A[n,n]-=(\lambda_1+\lambda_2);\ A[n,(1,0,0)]+=\mu_1;$ $A[n,(2,0,0)]+=\mu_2$
$k=1,\ i=0,\ j=0$	$A[n,n]-=\mu_1;\ A[n,(0,0,0)]+=\lambda_1$
$k=2,\ i=0,\ j=0$	$A[n,n]-=\mu_2;\ A[n,(0,0,0)]+=\lambda_2$
$k=1,\ i=0,\ j>0$	$A[n,n]-=\mu_1$
$k=2,\ i=0,\ j>0$	$A[n,n]-=\mu_2;\ A[n,(2,0,j-1)]+=\lambda_2$
$k=1,\ i>0$	$A[n,n]-=\mu_1;\ A[n,(1,i-1,j)]+=\lambda_1$
$k=1,\ j>0$	$A[n,(1,i,j-1)]+=\lambda_2$
$i>0$	$A[n,n]-=i\,\alpha_1$
$j>0$	$A[n,n]-=j\,\alpha_2$
$k=1,\ i<N_1$	$A[n,n]-=(\lambda_1+\lambda_2);\ A[n,(1,i+1,j)]+=\mu_1$
$k=1,\ j<N_2$	$A[n,n]-=\lambda_2$
$k=1,\ i=0,\ j=N_2$	$A[n,(2,0,j)]+=\lambda_1$
$k=2,\ i=0,\ j<N_2$	$A[n,n]-=\lambda_1;\ A[n,(1,0,j+1)]+=\mu_1;$ $A[n,(2,0,j+1)]+=\mu_2$
$k>0,\ j<N_2$	$A[n,(k,i,j+1)]+=(j+1)\,\alpha_2$

Таблица 2

Характеристики показателя качества обслуживания систем

Характеристика	**Значение**
Вероятность поступления заявки на моментальное обслуживание	0,264/0,259
Вероятность отказа в обслуживании приоритетных заявок	0,005/0,003
Вероятность отказа в обслуживании неприоритетных заявок	0,005/0,009
Вероятность того, что прибор занят обслуживанием приоритетной заявки	0,438/0,447
Вероятность того, что прибор занят обслуживанием неприоритетной заявки	0,288/0,282

Способность приоритетной заявки вытеснять неприоритетную при абсолютном приоритете дает хоть и незначительное, но снижение ве-

роятности отказа в обслуживании приоритетных заявок. В сценариях реальной жизни, моделируемых системами с абсолютным приоритетом, счет может идти на секунды — например, оказание медицинской помощи в экстренных случаях.

Заключение

В представленной статье описан расширенный алгоритм поиска технических характеристик системы с абсолютным приоритетом. Приведен анализ полученных характеристик в сравнении с характеристиками системы с относительным приоритетом. Снижение вероятности отказа в обслуживании приоритетных заявок может играть ключевую роль в работе высокоприоритетных систем массового обслуживания, или же в системах, моделирующих сценарии реальной жизни человека, к примеру, экстренная помощь.

СПИСОК ЛИТЕРАТУРЫ

1. *Jaiswal N. K.* Priority Queues// Academic Press, 1968. 240 p.
2. *Бронштейн О. И, Духовный И. М* Модели приоритетного обслуживания в информационно-вычислительных системах. М.: Наука, 1976. 220 с.
3. *Ронжин А. Л, Карпов А. А, Ли И. В.* Речевой и многомодальный интерфейсы. М.: Наука, 2006. 173 с.
4. *Степанов С. Н.* Теория телетрафика: концепции, методы, приложения. М.: Горячая линия — Телеком, 2015. 868 с.
5. *Назаров А. А., Змеев О. А.* Ограничение нагрузки в телекоммуникационных сетях // Научное творчество молодежи. Математика. Информатика. Материалы XVIII Всерос.науч.-практ. конф. 2014. С. 11–15.
6. *Хаустова Н. Г., Моисеева С. П., Пакулова Е. А.* Алгоритм вычисления стационарного распределения вероятностей системы с приоритетами для произвольного числа мест в очереди // Информационные технологии и математическое моделирование (ИТММ-2024): материалы XXIII Международной конференции имени А. Ф. Терпугова, Томск, 20–26 октября 2024. 159-165 с.

Хаустова Наталья Глебовна — студент Национального исследовательского Томского государственного университета. E-mail: *khaustovan@inbox.ru*

Моисеева Светлана Петровна — д.ф.-м.н., профессор Национального исследовательского Томского государственного университета. E-mail: *smoiseeva@mail.ru*

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С ПОВТОРНЫМИ ОБРАЩЕНИЯМИ И КОНЕЧНЫМ БУНКЕРОМ ПЕРЕД ПРИБОРОМ

И. И. Хомичков[1]

[1] *Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь*

В работе рассматривается экспоненциальная система массового обслуживания (СМО) с повторными заявками, одним прибором для обслуживания и конечным буфером для ожидания. Получены дифференциальные уравнения (ДУ) относительно производящих функций стационарных вероятностей состояний системы для конечного буфера. В качестве примера, когда буфер имеет одно место для ожидания, получено решение данного ДУ в виде гипергеометрического ряда. **Ключевые слова:** *система массового облуживания, конечный буфер для ожидания, гипергеометрический ряд*

Введение

В систему массового обслуживания СМО поступает пуассоновский поток заявок с интенсивностью λ. На приборе обслуживается только одна заявка с интенсивностью μ. Перед прибором обслуживания имеется $N-1$ мест для ожидания (буфер). Если при поступлении заявки в систему прибор занят, то заявка попадает в буфер. Если буфер заполнен, то заявка становится повторной. Каждая повторная заявка через экспоненциальное время с интенсивностью ν снова поступает на прибор обслуживания. Заявка покидает систему после обслуживания на приборе. Повторные заявки поступают на прибор до тех пор, пока не будут обслужены. Данная СМО является адекватной математической моделью системы передачи информации с одним каналом связи при ограниченном числе мест для ожидания настойчивых заявок. Настоящая СМО рассматривалась, например, в [1], где асимптотическим метом исследовались основные вероятностные характеристики в случае низкой интенсивности повторных заявок. Нас в данной работе будет интересовать метод нахождения производящих функций стационарных вероятностей состояний СМО в общем случае.

1. Система уравнений для произвольного объема буфера

Введем производящие функции

$$P_k(x) = \sum_{n=0}^{\infty} p_k(n) x^n, \quad k = \overline{0, N},$$

где $p_k(n)$ стационарные вероятности случайного процесса,

$\xi_t = (k(t), n(t))$, где $k(t)$ – суммарное число заявок в буфере и на приборе, $k(t) = \overline{0, N}$, $n(t)$ – число повторных заявок в системе, $n(t) = \overline{0, \infty}$.

Стационарные вероятности $p_k(n)$, $k = \overline{0, N}$, $n = \overline{0, \infty}$ случайного процесса $\xi_t = (k(t), n(t))$ удовлетворяют системе уравнений:

$$(\lambda + n\nu) p_0(n) = \mu\ p_1(n), n = \overline{0, \infty},$$

$$(\lambda + \mu + n\nu) p_k(n) = \lambda\ p_{k-1}(n) + \mu\ p_{k+1}(n) + (n+1)\nu\ p_{k-1}(n+1),$$

$$n = \overline{0, \infty}, k = \overline{1, N-1},$$

$$(\lambda + \mu) p_N(n) = \lambda\ p_{N-1}(n) + (n+1)\nu\ p_{N-1}(n+1) + \lambda\ p_{N-1}(n),$$

$$n = \overline{0, \infty},$$

$$\sum_{k=0}^{N} \sum_{n=0}^{\infty} p_k(n) = 1.$$

Система ДУ относительно производящих функций $P_k(x)$, $k = \overline{0, N}$, имеет вид:

$$\lambda\ P_0(x) + \nu\ x\ P_0'(x) = \mu\ P_1(x),$$

$$(\lambda+\mu)\ P_k(x) + \nu\ x\ P_k'(x) = \lambda\ P_{k-1}(x) + \mu\ P_{k+1}(x) + \nu\ P_{k-1}'(x),\ k = \overline{1, N-1},$$

$$(\lambda + \mu)\ P_N(x) + \nu\ x\ P_{N-1}'(x) = \lambda\ \ P_{N-1}(x) + \nu\ P_{N-1}'(x) + \lambda\ x\ P_N(x).$$

После суммирования этих ДУ и сокращения на множитель $(1 - x)$ последнее уравнение заменим на

$$\lambda\ P_N(x) = \nu \sum_{k=0}^{N-1}\ P_k'(x).$$

Введем обозначения $a = \frac{\lambda}{\mu},\ b = \frac{\nu}{\mu}$.

Теорема 1. Производящие функции $P_n(x)$, $n = \overline{0, N}$, выражаются через функции $P_0(x)$ по формулам:

$$P_n(x) = \sum_{k=0}^{n} b^k \left(\sum_{i=0}^{k} C_{k,i}^{(n)} x^i \right) P_0^{(k)}(x),\ n = \overline{1, N},$$

где коэффициенты $C_{k,i}^{(n)}$, $n = \overline{1,N}$, $k = \overline{0,n}$, $i = \overline{0,n}$ образуют нижнюю треугольную $(n+1) \times (n+1)$ матрицу с элементами:

при $n = 1$

$$C_{0,0}^{(1)} = a,$$
$$C_{1,0}^{(1)} = 0, \quad C_{1,1}^{(1)} = 1,$$

при $n = 2$

$$C_{0,0}^{(2)} = a^2,$$
$$C_{1,0}^{(2)} = -1, \quad C_{1,1}^{(2)} = b + 2a + 1,$$
$$C_{2,0}^{(2)} = 0, \quad C_{2,1}^{(2)} = 0, \quad C_{2,2}^{(2)} = 1,$$

при $n = \overline{3,N}$

$$C_{0,0}^{(n)} = a^n, \quad C_{n-1,i}^{(n)} = 0, \quad i = \overline{0,n-3},$$
$$C_{n,i}^{(n)} = 0, \quad i = \overline{0,n-1}, \quad C_{n,n}^{(n)} = 1$$

а ненулевые элементы матрицы определяются с помощью рекуррентных формул:

$$C_{n,n-1}^{(n+1)} = C_{n-1,n-2}^{(n)} - 1,$$
$$C_{n,n}^{(n+1)} = (1 + a + nb)\, C_{n-1,n-1}^{(n)},$$

где $n = \overline{2,N-1}$,

$$C_{k,k}^{(n+1)} = (1 + a + kb)\, C_{k,k}^{(n)} + C_{k-1,k-1}^{(n)} - a\, C_{k,k}^{(n-1)},$$
$$C_{k,i}^{(n+1)} = (1 + a + ib)\, C_{k,i}^{(n)} + C_{k-1,i-1}^{(n)} - a\, C_{k,i}^{(n-1)} - b(i+1)\, C_{k,i+1}^{(n-1)} - C_{k,i}^{(n-1)},$$

где $n = \overline{2,N-1}, \quad k = \overline{1,n-1}, \quad i = \overline{0,k-1}.$

Теорема 2. Производящая функция $P_0(x)$ является решением однородного дифференциального уравнения N-го порядка:

$$b^N x^{N-1}(ax-1)\,P_0^{(N)}(x)+ \\ + \sum_{k=2}^{N-1} b^k \Bigg\{ a \sum_{i=0}^{k} C_{k,i}^{(N)} x^i - \Bigg[\sum_{i=0}^{k-1} \Big(b(i+1) \sum_{n=k}^{N-1} C_{k,i+1}^{(n)} + \sum_{n=k}^{N-1} C_{k-1,i}^{(n)} \Big) x^i + \\ + x^{k-1} \Bigg] \Bigg\} P_0^{(k)}(x) + b\Big\{ a(C_{1,0}^{(N)} + C_{1,1}^{(n)} x) - 1 - \sum_{n=1}^{N-1} (bC_{1,1}^{(n)} + a^n) \Big\} P_0^{(1)}(x) \\ + a^{N+1} P_0(x) = 0.$$

Замечание 1. ДУ имеет две особые точки $= 0$ и $= \frac{1}{a} = \frac{\mu}{\lambda} > 1$. Последнее неравенство следует из условия существования стационарного распределения $\frac{\lambda}{\mu} < 1$. Таким образом, на отрезке $|x| < 1$ ДУ имеет единственную особую точку $= 0$, которая является иррегулярной.

2. Система с одним местом для ожидания

Рассмотрим частный случай СМО одним местом для ожидания, т.е. когда $N = 2$. В этом случае ДУ относительно производящей функции $P_0(x)$ имеет вид:

$$b^2 x(ax-1)P_0''(x) + b(b+2a+1)(ax-1)P_0'(x) + a^3 P_0(x) = 0.$$

Сделаем замену $P_0(x) = W(z),\ z = ax$. Придем к уравнению Гаусса:

$$z(z-1)\ W''(z) + [-\gamma + (1+\alpha+\beta)z]\ W'(z) + \alpha\beta\ W(z) = 0,$$

где

$$\alpha = \frac{2a+1+\sqrt{4a+1}}{2b},\ \beta = \frac{2a+1-\sqrt{4a+1}}{2b},\ \gamma = \frac{2a+b+1}{b}.$$

В окрестности особой точки $z = 0$, которая в данном случае будет регулярной, ДУ имеет два независимых решения:

$$W_1(z) = F(\alpha, \beta, \gamma; z),$$
$$W_2(z) = z^{1-\gamma} F(\alpha+1-\gamma, \beta+1-\gamma, 2-\gamma; z),$$

где $F(\alpha,\beta,\gamma;z)$ – гипергеометрический ряд:

$$F(\alpha,\beta,\gamma;z)=\sum_{n=0}^{\infty}\left(\prod_{k=0}^{n-1}\frac{(\alpha+k)(\beta+k)}{\gamma+k}\right)\frac{z^n}{n!}.$$

При $z=1$ этот ряд сходится при условии $\gamma-\alpha-\beta>0$. В нашем случае $\gamma-\alpha-\beta=\frac{1}{b}>0$ это условие выполняется. Производящая функция $P_0(x)$ в точке $x=0$ принимает конечное значение $P_0(x)=p_0(0)$. Первое решение $W_1(z)$ удовлетворяет этому условию. Второе решении $W_2(z)$, содержит множитель $z^{1-\gamma}=z^{-\frac{2a+1}{b}}$ с отрицательной степенью, а значит оно в окрестности точки $x=0$ будет неограниченным, т.е. оно нам не подходит. Окончательно с точностью до константы имеем:

$P_0(x)=W_1(ax)=F(\alpha,\beta,\gamma;ax)$, где C произвольная константа. С учетом теоремы 1 находим:

$$P_1(x)=bx\ P_0'(x)+a\ P_0(x),$$

$$P_2(x)=b^2x^2P_0''(x)+b\left((b+2a+1)x-1\right)P_0'(x)+a^2P_0(x).$$

Учитывая свойства функции $F(\alpha,\beta,\gamma;z)$,

$$F'(\alpha,\beta,\gamma;z)=\frac{\alpha\beta}{\gamma}F(\alpha+1,\beta+1,\gamma+1;z),$$

$$F''(\alpha,\beta,\gamma;z)=\frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}F(\alpha+2,\beta+2,\gamma+2;z),$$

получим выражения для производящих функций $P_1(x),\ P_2(x)$. Константу C находим из условия нормировки $P_0(1)+P_1(1)+P_2(1)=1$.

Заключение

В работе рассмотрена экспоненциальная система массового обслуживания с повторными вызовами и конечным числом мест для ожидания заявок перед обслуживающим прибором. В случае произвольного числа мест для ожиданий (конечный буфер) получены дифференциальные уравнения для расчета производящих функций стационарных вероятностей состояний системы. В случае одного места для ожидания найдены в явном виде решения данных систем дифференциальных уравнений.

СПИСОК ЛИТЕРАТУРЫ

1. *Подгайнов А. В, Назаров А. А.* Исследование RQ-системы с ожиданием заявок в бункере и на орбите методом асимптотического диффузионного анализа // Материалы XXVII Международной научной конференции «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN 2024)». С. 82–87.
2. *Смирнов В. И.* Курс высшей математики. М.: Наука, 1957. Т. III, ч. 2. 674 с.

Хомичков Иван Иванович — д.ф.-м.н., профессор БГУИР. E-mail: *iv_hom@tut.by*

ОБОБЩЕНИЕ МУЛЬТИПЛИКАТИВНОЙ ТЕОРЕМЫ ДЛЯ СЕТИ С ЗАВИСИМЫМИ ОТ ЧИСЛА ЗАЯВОК ПЕРЕХОДНЫМИ ИНТЕНСИВНОСТЯМИ

Г. Ш. Цициашвили[1]

[1] *Институт прикладной математики Дальневосточного отделения Российской акакдемии наук, г. Владивосток, Россия*

В работе построена мультипликативная формула для стационарного распределения открытой сети массового обслуживания при условии, что интенсивности обслуживания зависят от числа заявок в узлах сети. Эта формула базируется на выделении в графе переходных интенсивностей дискретного марковского процесса обслуживания базовых подграфов. В каждом базовом подграфе стационарные уравнения Колмогорова-Чепмена при выбранной мультипликативной форме стационарного распределения сводятся к балансовым соотношениям сети. Приводится пример сети массового обслуживания с несколькими потоками заявок для имитационнной модели работы морского транспорта. **Ключевые слова:** *балансовые соотношения, стационарное распределение, базовые подграфы*

Введение

Наряду с классическими мультипликативными теоремами для открытых сетей массового обслуживания (СеМО) и систем массового обслуживания (СМО) (см., например, [1]) существует большое число их обобщений. Эти обобщения представляют интерес, если они связаны с интересными практическими приложениями, например, с введением в модели СеМО блокирующих вероятностей (см., например, [2]). Настоящая работа вызвана появлением новых интересных имитационных моделей СеМО в морском транспорте (см., например, [3]).

В этих моделях появляется необходимость определения стационарных характеристик нескольких разнородных потоков и учета зависимостей интенсивностей обслуживания от числа заявок в разных узлах или от числа заявок разных потоков в одном узле. Это требование целесообразно использовать при рассмотрении узлов сети, выполняющих роль

Работа выполнена в рамках государственного задания ИПМ ДВО РАН № 075-00459-25-00.

складов, в которых помещаются грузы, двигающиеся с судов на сушу, и грузов, направляющихся с суши на суда. Обычно для таких сетей определяются стационарные характеристики, что связано с так называемым линейным судоходством, когда суда работают по строго оговоренному маршруту и расписанию.

В настоящей работе предлагается обобщение мультипликативной теоремы для открытой сети, в которой допускается зависимость интенсивностей обслуживания от числа заявок в различных узлах сети или от числа заявок различных потоков в отдельном узле. Это обобщение требует новой формы выражения для стационарного распределения числа заявок в узлах сети. В частности, такое обобщение вызвано необходимостью рассматривать сети с многоканальными узлами.

1. Теорема эргодичности дискретного марковского процесса

Рассмотрим дискретный марковский процесс $X(t)$, определяемый взвешенным связным ориентированным графом переходных интенсивностей Г. Пусть орграф Г имеет счетное множество вершин V и счетное множество E ориентированных ребер без петель. Здесь ребро $(v, v') \in E$ направлено из вершины v в вершину v' и имеет вес $L(v, v') > 0$, характеризующий интенсивность перехода марковского процесса $X(t)$ из состояния v в состояние v'. Из любой вершины $v \in V$ в любую вершину $v' \in V$, $v' \neq v$ в орграфе Г существует путь с ребрами, имеющими положительные веса.

Теорема 1. Пусть существует набор положительных чисел $f(v)$, $v \in V$, удовлетворяющих равенствам

$$\sum_{(v',v)\in E} f(v')L(v', v) = \sum_{(v,v')\in E} f(v)L(v, v'),\ v \in V \tag{1}$$

и условию $\sum\limits_{v\in V} f(v) < \infty$. Тогда марковский процесс $X(t)$ является эргодическим, а функция

$$p(v) = \frac{f(v)}{\sum_{v'\in V} f(v')},\ v \in V \tag{2}$$

является его предельным распределением [4, Теорема 2.4, с. 77].

2. Открытая СеМО

Рассмотрим открытую сеть Джексона G с пуассоновским входным потоком интенсивности λ, состоящую из m одноканальных систем массового обслуживания с экспоненциальными временами обслуживания

интенсивности $\mu_i,\ i=1,...,m$ (могущими зависеть от числа заявок в узлах СеМО). Динамика перемещения заявки в сети задается маршрутной матрицей $\Theta=||\theta_{ij}||_{i,j=0}^{m}$, где θ_{ij} – вероятность перехода после обслуживания из i-го узла в j-ый, $\theta_{00}=0$, узел с номером 0 – внешний источник. Предполагается, что маршрутная матрица неразложимая: $\forall i,j\in\{0,1,...,m\}\ \exists\ i_1,i_2,...,i_r\in\{1,...,m\}:\quad \theta_{ii_1}>0,\theta_{i_1i_2}>0,...,\theta_{i_rj}>0$. Далее для удобства изложения материала статьи (но без ограничения общности) рассмотрим СеМО с двумя обслуживающими узлами $m=2$, полагая $\theta_{ii}=0,\ i=0,1,2$. Тогда вектор $\Lambda=(\lambda,\lambda_1,\lambda_2)$ является единственным решением системы балансовых соотношений

$$\lambda=\lambda_1\theta_{10}+\lambda_2\theta_{20},\ \lambda_1=\lambda\theta_{01}+\lambda_2\theta_{21},\ \lambda_2=\lambda\theta_{02}+\lambda_1\theta_{12}. \tag{3}$$

Функционирование СеМО (число заявок в узлах) описывается дискретным марковским процессом $y(t)$ с множеством состояний $\mathbf{N}=\{\boldsymbol{n}=(n_1,n_2):\ n_1,n_2\geq 0\}$ и переходными интенсивностями:

$$L[(n_1,n_2),(n_1+1,n_2)]=\lambda\theta_{01},\ L[(n_1,n_2),(n_1,n_2+1)]=\lambda\theta_{02},$$

$$L[(n_1+1,n_2),(n_1,n_2)]=\mu_1(n_1+1,n_2)\theta_{10},$$

$$L[(n_1,n_2+1),(n_1,n_2)]=\mu_2(n_1,n_2+1)\theta_{20},$$

$$L[(n_1+1,n_2),(n_1,n_2+1)]=\mu_1(n_1+1,n_2)\theta_{12},$$

$$L[(n_1,n_2+1),(n_1+1,n_2)]=\mu_2(n_1,n_2+1)\theta_{21},$$

которые определяют взвешенный граф переходных интенсивностей Γ. Будем искать стационарное распределение процесса $y(t)$ в виде

$$\pi(n_1,n_2)=C^{-1}\prod_{k_1=0}^{n_1}\frac{\lambda_1}{\mu_1(k_1,n_2)}\cdot\prod_{k_2=0}^{n_2}\frac{\lambda_2}{\mu_2(n_1,k_2)},$$

$$\frac{\lambda_1}{\mu_1(0,n_2)}=\frac{\lambda_2}{\mu_2(n_1,0)}=1, \tag{4}$$

полагая $C<\infty$.

Выделим в графе переходных интенсивностей дискретного марковского процесса $y(t)$ базовые подграфы $\Gamma(n_1,n_2)$ с множествами вершин $V(n_1,n_2)=\{(n_1,n_2),\ (n_1+1,n_2),\ (n_1,n_2+1)\}$ и ребрами, соединяющими эти вершины. Назовем вершину (n_1,n_2) базовой вершиной графа $\Gamma(n_1,n_2)$, Тогда в базовом подграфе $\Gamma(n_1,n_2)$ стационарное распределение $\pi(n_1,n_2)$ удовлетворяет стационарным уравнениям Колмогорова-Чепмена, которые совпадают с балансовыми соотношениями (4). А поскольку базовые графы с различными базовыми вершинами не имеют

общих ребер, то стационарное распределение $\pi(n_1, n_2)$ удовлетворяет стационарным уравнениям Колмогорова–Чепмена на всем графе Γ переходных интенсивностей процесса $y(t)$, причем граф

$$\Gamma = \bigcup_{(n_1,n_2)\in\mathbf{N}} \Gamma((n_1, n_2)).$$

Замечание 1. Каждому направленному ребру базового графа $\Gamma(n_1, n_2)$ соответствует ребро направленное в противоположную сторону. Любые два базовых графа не имеют общих ребер.

3. СМО с несколькими входными потоками

Предложенная конструкция выделения в графе Γ (без петель и кратных ребер) базовых подграфов допускает следующее обобщение. Пусть дискретный марковский процесс $y(t)$ с множеством состояний $\mathbf{N}$ определяется переходными интенсивностями

$$L[(n_1, n_2), (n_1 + 1, n_2)] = \lambda_1,\ L[(n_1, n_2), (n_1, n_2 + 1)] = \lambda_2,$$

$$L[(n_1 + 1, n_2), (n_1, n_2)] = \mu_1(n_1 + 1, n_2),$$
$$L[(n_1, n_2 + 1), (n_1, n_2)] = \mu_2(n_1, n_2 + 1).$$

Такой процесс описывает СМО с двумя входными потоками интенсивности $\lambda_1,\ \lambda_2$ и интенсивностями обслуживания $\mu_1(n_1, n_2),\ \mu_2(n_1, n_2)$. Тогда состояние (n_1, n_2) процесса $z(t)$ характеризует число заявок первого и второго потока в СМО. В графе переходных интенсивностей Γ процесса $y(t)$ можно выделить базовые подграфы $\Gamma(n_1, n_2)$, соединяющие вершину (n_1, n_2) с вершинами $(n_1 + 1, n_2),\ (n_1, n_2 + 1)$. В этом случае стационарное распределение $\pi(n_1, n_2)$, определяемое равенствами (4) также удовлетворяет стационарным уравнениям Колмогорова-Чепмена, которые в вершинах графа $\Gamma(n_1, n_2)$ преобразуются в тождества $\lambda_1 = \lambda_1,\ \lambda_2 = \lambda_2$, являющиеся балансовыми соотношениями для СМО с двумя входными потоками. А поскольку базовые графы с различными базовыми вершинами не имеют общих ребер, то стационарное распределение $\pi(n_1, n_2)$ удовлетворяет стационарным уравнениям Колмогорова-Чепмена на всем графе

$$\Gamma = \bigcup_{(n_1,n_2)\in\mathbf{N}} \Gamma((n_1, n_2)).$$

переходных интенсивностей дискретного марковского процесса $y(t)$. Причем базовый граф $\Gamma(n_1, n_2)$ также удовлетворяет замечанию 1.

Пример 1. Рассмотрим СМО с двумя типами заявок, в которой при некоторых $\lambda_1 = \lambda_2 = \lambda,\ \mu > 0,\ 1 \le k < \infty,\ 0 < K < N \le \infty$, (здесь K, N натуральные числа) выполняются равенства

$$\mu_1(n_1, n_2) = \mu_2(n_1, n_2) = \mu/k,\ 0 \le n_1, n_2 \le K,$$

$$\mu_1(n_1, n_2) = \mu_2(n_1, n_2) = \mu k,\ K < n_1, n_2 \le N,$$

$$\mu_1(n_1, n_2) = \mu/k,\ \mu_2(n_1, n_2) = \mu k,\ 0 \le n_1 \le K < n_2 \le N,$$

$$\mu_1(n_1, n_2) = \mu k,\ \mu_2(n_1, n_2) = \mu/k,\ 0 \le n_2 \le K < n_1 \le N.$$

Эта модель возникает при рассмотрении склада, на который поступают контейнеры, пришедщие морским транспортом для отправки наземным транспортом и контейнеры, пришедшие наземным транспортом для отправки морским транспортом.

Разобъем множество состояний данной СМО $\mathcal{N} = \{0 \le n_1 \le N,\ 0 \le n_2 \le N\} \setminus (N, N)$ на непересекающиеся подмножества

$$\mathcal{N}_1 = \{0 \le n_1 \le K,\ 0 \le n_2 \le K\},$$

$$\mathcal{N}_2 = \{K+1 \le n_1 \le N,\ K+1 \le n_2 \le N\} \setminus (N, N),$$

$$\mathcal{N}_3 = \{0 \le n_1 \le K,\ ; K+1 \le n_2 \le N\},$$

$$\mathcal{N}_4 = \{0 \le n_2 \le K,\ ; K+1 \le n_1 \le N\}$$

и вычислим среднее число S заявок 1-го и 2-го типов в стационарном режиме, полагая $\rho_1 = \lambda/(\mu/k),\ \rho_2 = \lambda/\mu k$:

$$S = C^{-1}D,\ C = \sum_{(n_1,n_2)\in\mathcal{N}_1} \rho_1^{n_1+n_2} + \sum_{(n_1,n_2)\in\mathcal{N}_2} \rho_2^{n_1+n_2} +$$

$$+ \sum_{(n_1,n_2)\in\mathcal{N}_3} \rho_1^{n_1}\rho_2^{n_2} + \sum_{(n_1,n_2)\in\mathcal{N}_4} \rho_1^{n_2}\rho_1^{n_1},$$

$$D = \sum_{(n_1,n_2)\in\mathcal{N}_1} (n_1+n_2)\rho_1^{n_1+n_2} + \sum_{(n_1,n_2)\in\mathcal{N}_2} (n_1+n_2)\rho_2^{n_1+n_2} +$$

$$+ \sum_{(n_1,n_2)\in\mathcal{N}_3} (n_1+n_2)\rho_1^{n_1}\rho_2^{n_2} + \sum_{(n_1,n_2)\in\mathcal{N}_4} (n_1+n_2)\rho_1^{n_2}\rho_2^{n_1},$$

Вычислительный эксперимент показал, что при $N = 100,\ K = 15,\ \rho = 0,8,\ k = 1,3$ получаем $S \approx 7,13896$, но если положить $k = 1$, то получим $S \approx 8$.

Заключение

В данной заметке получено обобщение мультипликативной теоремы для СеМО и для СМО с несколькими потоками заявок. Это обобщение было вызвано с одной стороны попыткой математически корректного использования мультипликативной теоремы для СеМО, возникшей при имитационном моделирования системы обслуживания морского транспорта. С другой стороны такое математическое обобщение потребовало выделения в графе переходных интенсивностей базовых подграфов. Для этих базовых подграфов выбрана форма записи стационарного распределения, при котором стационарные формулы Колмогорова-Чепмена преобразуются в балансовые соотношения для сети массового обслуживания. Постановка задачи о включении в модель зависимости интенсивности обслуживания позволяет учесть влияние состояния различных узлов СеМО друг на друга, что естественным образом возникает в реальных системах и сетях обслуживания. И, наконец, построен численный пример, показывающий, как такая зависимость интенчсивностей обслуживания от числа заявок может быть использования для управления потоками различных заявок.

СПИСОК ЛИТЕРАТУРЫ

1. *Basharin G. P., Gaidamaka Yu. V., Samouylov K. E.* Mathematical Theory of Teletraffic and Its Applications to the Analysis of Multiservice Communication of Next Generation Networks // Autom. Control. Comput. Sci. 2013. Vol. 47, № 5. P. 62–69.
2. *Balsamo S., De Nitto V.* A survey of Product-form Queueing Networks with Blocking and their Equivalences // Annals of Operations Research. 1994. Vol. 48. P. 31–61.
3. *Лемперт А. А., Жарков М. Л., Казаков А. Л., Ву Х. З.* Моделирование морского контейнерного терминала с использованием сети массового обслуживания. // Управление большими системами. 2024. Т. 112. С 310–337.
4. *Ивченко Г. И. Каштанов В. А. Коваленко И. Н.* Теория массового обслуживания. М.: Высшая школа, 1982. 256 с.

Цициашвили Гурами Шалвович — д.ф.-м.н., профессор, г.н.с. ИПМ ДВО РАН. E-mail: *guram@iam.dvo.ru*

ИССЛЕДОВАНИЕ ГИБРИДНОЙ СМО С МГНОВЕННОЙ И ОТСРОЧЕННОЙ ОБРАТНОЙ СВЯЗЬЮ

М. А. Шкленник[1], М. Д. Лавриненко[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Рассматривается гибридная система массового обслуживания, в которой реализуется два типа обратной связи: мгновенная и отсроченная. Мгновенная обратная связь реализуется в блоке с неограниченным числом приборов, а отсроченная - в блоке в виде RQ-системы, в которой заявки поступают на орбиту. Найдено совместное распределение вероятностей числа заявок в блоках системы. **Ключевые слова:** *гибридная система массового обслуживания, мгновенная обратная связь, отсроченная обратная связь, поступление заявок на орбиту*

Введение

Одной из модификаций систем обработки данных являются системы массового обслуживания с обратной связью, в которых заявки после обработки могут с некоторой вероятностью требовать повторного обслуживания. В зависимости от характера повторной обработки заявок различают: а) системы с мгновенной обратной связью, где повторная обработка происходит без задержки (например, при обнаружении ошибки в данных); б) системы с отсроченной обратной связью, где заявки временно уходят на орбиту и требуют повторного обслуживания через случайные промежутки времени [3]. Как правило, такие системы исследуются отдельно, независимо друг от друга [2, 5, 7]. Однако на практике может иметь место гибридная модель, когда в одной системе реализуются оба типа обратной связи – как мгновенная, так и отсроченная. Одним из примеров таких моделей можно считать системы обработки данных с проверкой качества обработки [4].

1. Описание модели и постановка задачи

Рассмотрим систему обработки данных, представленную на рис. 1. В систему поступает простейший поток заявок с параметром λ. Время обслуживания заявок входящего потока является случайной величиной, имеющей экспоненциальное распределение с параметром μ_1. После об-

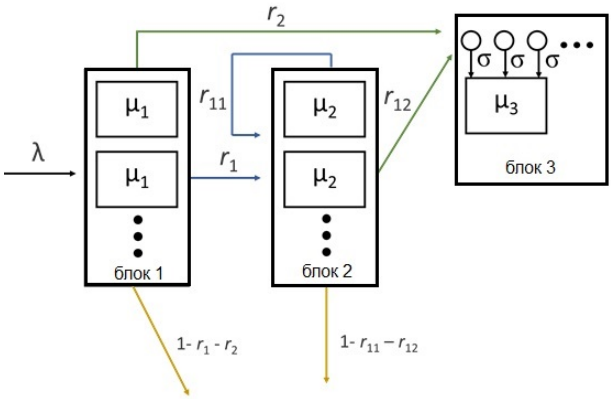


Рис. 1. Гибридная СМО с двумя типами обратной связи

служивания заявок входящего потока выполняется проверка качества обработки на наличие ошибок. Пусть существует два типа ошибок:

- первый тип ошибок устраняется мгновенно, но при этом требуется повторное обслуживание; вероятность возникновения ошибок первого типа после первичного обслуживания равна r_1;
- второй тип ошибок возникает после первичного обслуживания с вероятностью r_2 и для их устранения требуется случайное время, распределенное экспоненциально с параметром σ, после чего заявке также требуется повторное обслуживание, которое осуществляется одним единственным прибором.

Таким образом, каждая заявка входящего потока после завершения первичного обслуживания может с вероятностью r_1 поступить во второй блок, где время обслуживания распределено с параметром μ_2, либо с вероятностью r_2 уйти на орбиту, где осуществляется задержка в течение случайного времени с параметром σ. С орбиты заявка поступает на обслуживающий прибор третьего блока системы и обрабатывается там в течение экспоненциально распределенного времени с параметром μ_3.

После обработки на втором блоке проверка качества обслуживания повторяется, однако вероятность возникновения ошибок первого и второго типа уже равна r_{11} и r_{12} соответственно. Будем считать, что после обработки заявок в третьем блоке (после устранения ошибок второго типа) заявкам больше не требуется обработка и они покидают систему.

Обозначим $i_m(t)$ – число занятых приборов в m-ом блоке системы в момент времени t, $m = 1, 2, 3$. Ставится задача: найти стацио-

нарное распределение вероятностей трехмерного случайного процесса $\{i_1(t), i_2(t), i_3(t)\}$.

2. Система дифференциальных уравнений для частичных производящих функций

Так как процесс $\{i_1(t), i_2(t), i_3(t)\}$ не является марковским, введем случайный процесс $\{k(t)\}$, определяющий состояние прибора третьего блока, обслуживающего заявки после устранения ошибок второго типа, следующим образом

$$k(t) = \begin{cases} 0, \text{ если прибор блока 3 свободен,} \\ 1, \text{ если прибор блока 3 занят.} \end{cases}$$

Найдем стационарное распределение вероятностей четырёхмерного марковского процесса $\{i_1(t), i_2(t), i_3(t), k(t)\}$.

Обозначим вероятность того, что в момент времени t прибор третьего блока находится в состоянии k, в первом блоке обрабатываются i_1 заявок, во втором блоке обрабатываются i_2 заявок, и в третьем блоке находятся i_3 заявок

$$P_k(i_1, i_2, i_3, t) = P\{i_1(t) = i_1, i_2(t) = i_2, i_3(t) = i_3, k(t) = k\}.$$

Тогда распределение вероятностей исследуемого процесса в стационарном режиме можно обозначить как $P_k(i_1, i_2, i_3)$. Введем частичные производящие функции вида

$$G_0(x, y, z) = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} \sum_{i_3=0}^{\infty} x^{i_1} y^{i_2} z^{i_3} P_0(i_1, i_2, i_3),$$
$$G_1(x, y, z) = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} \sum_{i_3=1}^{\infty} x^{i_1} y^{i_2} z^{i_3} P_1(i_1, i_2, i_3).$$

Тогда, записав систему дифференциальных уравнений Колмогорова для стационарного распределения вероятностей процесса $\{i_1(t), i_2(t), i_3(t), k(t)\}$ [6], можем перейти к системе дифференциальных уравнений в частных производных для частичных характеристических функций вида

$$\begin{cases} -\mu_1\left(x-1-r_1(y-1)-r_2(z-1)\right)\dfrac{\partial G_0(x,y,z)}{\partial x}- \\ -\mu_2\left((1-r_{11})(y-1)-r_{12}(z-1)\right)\dfrac{\partial G_0(x,y,z)}{\partial y}- \\ -\sigma z\dfrac{\partial G_0(x,y,z)}{\partial z}+\lambda(x-1)G_0(x,y,z)+\dfrac{\mu_3}{z}G_1(x,y,z)=0, \\ -\mu_1\left(x-1-r_1(y-1)-r_2(z-1)\right)\dfrac{\partial G_1(x,y,z)}{\partial x}- \\ -\mu_2\left((1-r_{11})(y-1)-r_{12}(z-1)\right)\dfrac{\partial G_1(x,y,z)}{\partial y}- \\ \left(\lambda(x-1)-\mu_3\right)G_1(x,y,z)+\sigma z\dfrac{\partial G_0(x,y,z)}{\partial z}=0. \end{cases} \tag{1}$$

Данная система уравнений является системой дифференциальных уравнений в частных производных с переменными коэффициентами. Аналитическое ее решение известными методами представляет собой весьма громоздкую задачу. В связи с этим рассмотрим исследуемую систему как совокупность двух СМО, одна из которых представляет собой двухфазную СМО с неограниченным числом приборов и мгновенной обратной связью – далее СМО 1 , а вторая – RQ-систему, в которой все заявки входящего потока поступают сразу на орбиту, и только с орбиты заявки идут на обслуживание – далее СМО 2. Исследуем процессы, описывающие число заявок в каждой из этих систем, и исходя из полученных результатов, запишем производящие функции одномерных процессов, описывающих число заявок в каждом блоке системы, а также предположим вид производящей функции $G(x,y,z)$ для трехмерного процесса $\{i_1(t),i_2(t),i_3(t)\}$.

3. Производящие функции для маргинальных распределений числа заявок в каждом блоке

Рассмотрим СМО 1, которая представляет собой первые два блока исследуемой системы. Так как из блока 3 (СМО 2) заявки не поступают в СМО 1, то можно предположить, что СМО 1 функционирует независимо от СМО 2. Данная система подробно описана в работе [1], и там же получено выражение для производящей функции стационарного распределения вероятностей двумерного процесса $\{i_1(t),i_2(t)\}$, описывающего число заявок в блоках обслуживания, которое имеет вид

$$F(x,y)=exp\left\{\frac{\lambda}{\mu_1}(x-1)+\frac{\lambda r_1}{\mu_2(1-r_{11})}(y-1)\right\}. \tag{2}$$

По виду производящей функции (2) можно сделать вывод о том, что в стационарном числа заявок в первом и во втором блоках системы не зависят друг от друга и имеют распределение Пуассона. Также несложно показать, что поток заявок, выходящий из СМО 1 в СМО 2 является стационарным Пуассоновским потоком с параметром

$$\Lambda = \lambda \left(\frac{r_1 r_{12}}{1 - r_{11}} + r_2 \right). \tag{3}$$

Рассмотрим СМО 2. Данная система представляет собой RQ-систему, особенностью которой является поступление заявок на орбиту. На вход системы поступает простейший поток заявок с параметром (3). Найдем производящую функцию $G(z)$ стационарного распределения числа заявок в СМО 2 $P_k(i_3) = P\{i_3(t) = i_3, k(t) = k\}$, $k = 0, 1$. Для этого введем частичные производящие функции

$$G_0(z) = \sum_{i_3=0}^{\infty} z^{i_3} P_0(i_3),\ G_1(z) \sum_{i_3=1}^{\infty} z^{i_3} P_1(i_3).$$

Тогда, записав систему дифференциальных уравнений Колмогорова для распределения вероятностей в стационарном режиме [6], можем перейти к системе дифференциальных уравнений для частичных производящих функций в виде

$$\begin{cases} \Lambda(z-1)G_0(z) - \sigma z \dfrac{\partial G_0(z)}{\partial z} + \dfrac{\mu_3}{z} G_1(z) = 0, \\ (\Lambda(z-1) - \mu_3)\, G_1(z) + \sigma z \dfrac{\partial G_1(z)}{\partial z} = 0. \end{cases} \tag{4}$$

Решая систему (4) при условии

$$G_0(1) + G_1(1) = 1,$$

получаем

$$G_0(z) = (1-\rho) e^{\frac{\Lambda}{\sigma}(z-1)} \left(\frac{1-\rho}{1-\rho z} \right)^{\frac{\Lambda}{\sigma}}, \tag{5}$$

$$G_1(z) = \rho z e^{\frac{\Lambda}{\sigma}(z-1)} \left(\frac{1-\rho}{1-\rho z} \right)^{\frac{\Lambda}{\sigma}+1}, \tag{6}$$

где $\rho = \frac{\Lambda}{\mu_3}$.

И тогда, сложив (5) и (6), получисм выражение для производящей функции стационарного распределения вероятностей числа заявок в СМО 2 в виде

$$G(z) = e^{\frac{\Lambda}{\sigma}(z-1)} \left(\frac{1-\rho}{1-\rho z} \right)^{\frac{\Lambda}{\sigma}+1}. \tag{7}$$

4. Производящая функция совместного распределения вероятностей числа заявок в блоках гибридной СМО

Вернемся к системе дифференциальных уравнений (1) для частичных производящих функций совместного стационарного распределения вероятностей числа заявок в блоках исследуемой гибридной СМО. Решение этой системы будем искать в виде

$$\begin{cases} G_0(x,y,z) = F(x,y)G_0(z), \\ G_1(x,y,z) = F(x,y)G_1(z), \end{cases} \tag{8}$$

где $F(x,y)$ определяется выражением (2), а функции $G_0(z)$ и $G_1(z)$ неизвестны. Подставим (8) в систему (1). Получим следующую систему

$$\begin{cases} \lambda\left(r_2 + \dfrac{r_1 r_{12}}{1-r_{11}}\right)(z-1)G_0(z) - \sigma z\dfrac{\partial G_0(z)}{\partial z} + \dfrac{\mu_3}{z}G_1(z) = 0, \\ \left(\lambda\left(r_2 + \dfrac{r_1 r_{12}}{1-r_{11}}\right)(z-1) - \mu_3\right)G_1(z) + \sigma z\dfrac{\partial G_1(z)}{\partial z} = 0. \end{cases} \tag{9}$$

Несложно заметить, что система (9) полностью совпадет с системой (4), если учесть интенсивность входящего потока (3). Следовательно, функции $G_0(z)$ и $G_1(z)$ имеют вид (5) и (6) соответственно. А это значит, что производящая функция совместного распределения вероятностей числа заявок в блоках системы имеет вид

$$G(x,y,z) = \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\Lambda}{\sigma}+1} \cdot \exp\left\{\frac{\lambda(x-1)}{\mu_1} + \frac{\lambda r_1(y-1)}{\mu_2(1-r_{11})} + \frac{\Lambda(z-1)}{\sigma}\right\}, \tag{10}$$

где $\rho = \frac{\Lambda}{\mu_3}$, $\Lambda = \lambda\left(\frac{r_1 r_{12}}{1-r_{11}} + r_2\right)$.

Заключение

Анализируя вид выражения (10), можно сделать следующие выводы:

- случайные процессы $\{i_1(t)\}$, $\{i_2(t)\}$, $\{i_3(t)\}$, описывающие число заявок в каждом блоке системы, при стационарном режиме работы являются независимыми друг от друга;
- производящие функции распределений вероятностей числа заявок в первом и втором блоках системы являются производящими функциями распределений Пуассона;
- производящая функция числа заявок в третьем блоке системы (реализующем отсроченную обратную связь) представляет собой произведение производящих функций распределения Пуассона и отрицательного биномиального распределения.

СПИСОК ЛИТЕРАТУРЫ

1. *Бадмаева С. Б., Шкленник М. А.* О распределении числа занятых приборов в гетерогенных бесконечнолинейных системах массового обслуживания с обратной связью // Системы управления, информационные технологии и математическое моделирование : материалы V Всерос. науч.-практ. конф. с междунар. участием (Омск, 25–26 апр. 2023 г.). Омск: Изд-во ОмГТУ, 2023. С. 308–315.
2. *Данилюк Е. Ю., Моисеева С. П., Фёдорова Е. А.* Асимптотический анализ RQсистемы M|M|1 с конфликтами и нетерпеливыми заявками // Автоматика и телемеханика. 2018. № 12. С. 44–56.
3. *Королюк В. С., Меликов А. З., Пономаренко Л. А., Рустамов А. М.* Методы анализа многоканальной системы обслуживания с мгновенной и отсроченной обратными связями // Кибернетика и системный анализ. 2016. Т. 52, № 1. С. 64–77.
4. *Лавриненко М. Д., Шкленник М. А.* Исследование бесконечнолинейной СМО с проверкой качества обработки при нестационарном режиме работы // Информационные технологии и математическое моделирование (ИТММ-2024) : материалы XXIII Междунар. конф. им. А. Ф. Терпугова, Томск: Изд-во Том. гос. ун-та, 2023. С. 114–119.
5. *Мелошникова Н. П., Назаров А. А., Фёдорова Е. А.* Исследование RQ-системы M/M/1 с катастрофами // Вестник Томского государственного университета. Управление, вычислительная техника и информатика.. 2024. Т. 68, № 68. С. 28–37.
6. *Назаров А. А.* Теория массового обслуживания. Томск: Изд-во НТЛ, 2010. 228 с.
7. *Шкленник М. А., Моисеев А. Н.* Исследование потоков заявок в двухфазной системе массового обслуживания с неограниченным числом приборов и повторными обращениями // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2018. № 45. С. 48–58.

Шкленник Мария Александровна — канд. физ.-мат. наук, доцент кафедры ТВиМС ИПМКН ТГУ, . E-mail: *shklennikm@yandex.ru*

Лавриненко Милена Дмитриевна — аспирант ИПМКН ТГУ. E-mail: *mila-gygy@mail.ru*

ANALYSIS OF A TWO-SERVER QUEUE WITH MAIN SERVER INTERRUPTIONS, VACATION POLICY AND BACKUP SERVER ASSISTANCE

Balaji M. N.[1,3], Dr. Jaya S[2]

[1] *Research Scholar, Department of Mathematics, Maharaja's College, Ernakulam*
[2] *Associate Professor, Department of Mathematics, Maharaja's College, Ernakulam*
[3] *Assistant Professor, Department of Mathematics, KKTM Govt College, Pullut*

This paper analyses a two-server queueing system with prioritized customers, employing a main and standby server. High-priority customers have an infinite queue capacity, while low-priority customers have a finite capacity. The standby server assumes the main server's role during vacations or breakdowns, exclusively serving high-priority customers. The system is depicted as a QBD process and matrix analytic methods are applied to derive performance measures. The findings are numerically substantiated.

Keywords: *Multi-server queuing system, priority queues, matrix analytic methods, vacation, phase type distribution*

Introduction

Queuing theory is a mathematical framework used to analyze systems where entities wait for service, such as in customer service, telecommunications, and manufacturing. Traditionally, it assumes servers are always available to process arrivals, but in real-world scenarios, servers often experience periods of unavailability, such as for maintenance or breaks. To address this, the concept of "vacations"is introduced, modeling these downtime periods and providing a more accurate representation of practical queuing systems.. The concept of vacation was introduced in queuing system first by Y Levy and U Yechiali in 1975 [7] for an M/G/1 queuing system. The survey paper by Doshi (1986) [4] and Tian and Zhang (2006) [9] provide more details about vacation queuing models. The concept of working vacation was first introduced by L. D. Servi and S. G. Finn in [5]. This paper discussed multiple working vacations in the M/M/1 queue. This concept was generalized for the M/G/1 queue with working vacation by Wu and Takagi [10]. Later

Authors would like to thank Dr. A. Krishnamoorthy, Department of Mathematics, Central Univeristy of Kerala, Tejaswani Hills, Kasaragod, Kerala 671325, India, Centre for Research in Mathematics, CMS College, Kottayam, Kerala 686001, India for his valuable suggestions.

Li and Tian in [8] studied GI/M/1 queues with single working vacation in 2011. Queuing models with backup servers and vacation were discussed by Gray, Wang, and Scot in [6]. In [2], Chakravarthy, Shruti, and Kulshrestha propose a model where a backup server assists during the main server's breakdowns or vacations, analysing steady-state performance. In [1], Ayyapan and Karpagam present a queueing model that incorporates breakdowns, optional repairs, a standby server, and multiple vacations. An analysis of a priority queue with phase-type service and failures was done by Dudin, A., and Dudin, S. in [3].
This paper analyzes a two-server queuing model that includes real-world factors like mandatory server vacations, breakdowns, and priority-based customer service.

1. Mathematical formulation

In this paper, we consider an M/PH(1), PH(2)/2 queueing system. The two servers are labeled as the main server and the standby server. Two queues are created based on the priority. High-priority customers are served by the main server, while low-priority customers are served by the standby server. The capacity of the queue for the main server is unlimited, but the capacity of the queue for the standby server is limited, say L. Let λ_1 be the arrival rate for the high- priority customer and λ_2 be the arrival rate of the low- priority customer. The service times of the main server and the stand by server follow PH-distributions with representations (α, T) and (β, U), where $U = \theta T$, $0 < \theta < 1$ respectively. The service rates of the main server and the standby server are respectively represented by $\mu_h = [\alpha(-T)^{-1}e]^{-1}$ and $\mu_l = [\beta(-U)^{-1}e]^{-1}$.We denote T^0 and U^0 as the column vectors satisfying $Te + T^0 = 0$ and $Ue + U^0 = 0$. The main server could experience downtime due to factors like mandatory vacations or a breakdown. The main server takes a vacation after rendering service to K customers at a stretch. During that period, the service of low-priority customers will be terminated and high-priority customers will be served by the standby server. The duration of each vacation is modelled with an exponential distribution with rate η. On the return of the main server, the standby server resumes the service to the interrupted low priority customer. During the absence of main server, the low priority customers are not allowed to join the queue whereas the high priority customers either join the queue with probability p or returns with probability 1-p. Whenever the system becomes empty, the main server goes on vacation. When a high-priority customer arrives, the standby server will serve them. There is a possibility of the main server breaking down. The standby server is shielded from interruption. The standby server will take charge during the repair of the main server, just

like when the main server takes a mandatory vacation. The repair time is exponentially distributed with rate γ. The underlying random variable governing the durations of the inter-arrivals, the services, the vacations and the repairs are all assumed to be mutually independent of each other. The system can be viewed as a QBD process. We denote the variable as follows: $N_1(t)$ and $N_2(t)$ represent the number of high priority customers and the number of low priority customers respectively, $S_1(t)$ and $S_2(t)$ represnt the status of the main server and the standby server respectively, $C_1(t)$ and $C_2(t)$ indicate the phase of the main server and phase of the standby server respectively. Finally $N(t)$ denotes the number of high priority customers that have been already served by the main server.

$$S_1(t) = \begin{cases} 0 & \text{if the main server offers service} \\ 1 & \text{if the main server is on vacation} \\ 2 & \text{if the main server is under repair,} \end{cases}$$

$$S_2(t) = \begin{cases} 0 & \text{if the standby server offers service to low priority customers} \\ 1 & \text{if the standby server offers service to high priority customers} \\ 2 & \text{if the standby server is idle} \end{cases}$$

Then $(N_1(t), N_2(t), S_1(t), S_2(t), C_1(t), C_2(t), N(t))$ is a continuous time Markov chain with state space $\Omega =$

$$\begin{aligned}
&\{(0,0,i,2,j,j,p)\,|\,i \in \{1,2\},\ 1 \le j \le m,\ 0 \le p \le k\} \cup \\
&\{(0,l,i,0,j,j,p)\,|\,i \in \{1,2\},\ 1 \le j \le m,\ 0 \le p \le k,\ 1 \le l \le L\} \cup \\
&\{(q,0,0,2,j,j,p)\,|\,1 \le j \le m,\ q \ge 1\} \cup \\
&\{(q,0,i,1,j,j,p)\,|\,i \in \{1,2\},\ 1 \le j \le m,\ q \ge 1,\ 0 \le p \le k\} \cup \\
&\{(q,l,0,0,j,j,p)\,|\,1 \le l \le L,\ 1 \le j \le m,\ q \ge 1,\ 0 \le p \le k\} \cup \\
&\{(q,l,i,1,j,j,p)\,|\,i \in \{1,2\},\ 1 \le l \le L,\ 1 \le j \le m,\ q \ge 1,\ 0 \le p \le k\}.
\end{aligned}$$

Thus the infinitesimal generator of Ω is of the form

$$\mathcal{Q} = \begin{pmatrix} C_0 & C_1 & & & \\ C_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

Each matrix A_0, A_1 and A_2 are square matrices of order $3Km + 3Km^2L$, C_0 is a square matrix of order $2K + 2KmL$, C_1 is a matrix of order $(2K + 2KmL) \times (3Km + 3m^2KL)$ and C_2 is a matrix of order $(3Km + 3m^2KL) \times (2K + 2KmL)$.

2. Stability condition

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)$ be the steady state probability vector of $A = A_0 + A_1 + A_2$. Then

$$\boldsymbol{\pi} A = \mathbf{0}, \ \boldsymbol{\pi}\mathbf{e} = 1. \tag{2}$$

The queueing system under study with the generator given in (1) is stable if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{3}$$

ie,

$$\boldsymbol{\pi}_0 \cdot B_1 + \boldsymbol{\pi}_1 \cdot B_2 + \boldsymbol{\pi}_2 \cdot (I_{L2} \otimes B_2) + \boldsymbol{\pi}_3 \cdot B_2 < \boldsymbol{\pi}_0 \cdot B_3 + \boldsymbol{\pi}_1 \cdot B_4 + \boldsymbol{\pi}_2 \cdot (I_{L2} \otimes B_4) + \boldsymbol{\pi}_3 \cdot B_4 \tag{4}$$

where

$$A_0 = \begin{pmatrix} B_1 & 0 \\ 0 & I_{L2} \otimes B_2 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} B_3 & 0 \\ 0 & I_{L2} \otimes B_4 \end{pmatrix},$$

B_1 and B_3 are square matrices of order $3mK$, B_2 and B_4 are square matrices of order $3m^2K$ and I_{L2} is an Identity matix of order (L - 2).

3. Steady state probability vector

Let $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ...)$ where $\mathbf{x}_0$ is of dimension $2K + 2KmL$ and $\mathbf{x}_1, \mathbf{x}_2, ...$ are of dimension $3mK + 3m^2KL$ be the steady state probability vector of $\mathcal{Q}$. Then $\mathbf{x}$ must satisfy the set of equations

$$\mathbf{x}\mathcal{Q} = 0, \mathbf{x}\mathbf{e} = 1. \tag{5}$$

Thus the above set of equations reduce to:

$$\begin{aligned} \mathbf{x}_0 C_0 + \mathbf{x}_1 C_2 &= \mathbf{0}, \\ \mathbf{x}_0 C_1 + \mathbf{x}_1 A_1 + \mathbf{x}_2 A_2 &= \mathbf{0}, \\ \mathbf{x}_{n-1} A_0 + \mathbf{x}_n A_1 + \mathbf{x}_{n+1} A_2 &= \mathbf{0}, \quad n \geq 2. \end{aligned} \tag{6}$$

Under the assumption that the stability condition holds, we see that $\mathbf{x}$ is obtained as

$$\mathbf{x}_n = \mathbf{x}_1 R^{n-1}, \quad n \geq 2 \tag{7}$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$R^2 A_2 + R A_1 + A_0 = 0 \tag{8}$$

and the boundary equations are given by

$$\begin{aligned} \mathbf{x}_0 C_0 + \mathbf{x}_1 C_2 &= \mathbf{0}, \\ \mathbf{x}_0 C_1 + \mathbf{x}_1 \left[A_1 + R A_2 \right] &= \mathbf{0}. \end{aligned} \tag{9}$$

4. Some important system performance measures

1) Probability that main server is busy:

$$P_1 = \sum_{i=1}^{\infty}\sum_{r=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{K}\mathbf{x}_i(0,0,2,r,s,u) + \\ + \sum_{i=1}^{\infty}\sum_{j=1}^{L}\sum_{r=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{K}\mathbf{x}_i(j,0,0,r,s,u).$$

2) Probability that the standby server is busy :

$$P_2 = \sum_{i=1}^{\infty}\sum_{j=1}^{L}\sum_{r=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{K}\mathbf{x}_i(j,0,0,r,s,u) + \\ + \sum_{i=1}^{\infty}\sum_{j=0}^{L}\sum_{l=1}^{2}\sum_{r=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{K}\mathbf{x}_i(j,l,1,r,s,u).$$

5. Tables

We present a numerical example to highlight how the system behaves with respect to certain parameters. In Table 1 the probabilities of the main server being busy are calculated for different values of arrival rates of high priority customers and vacation rates. we see that as the arrival rate increases, the probability that the main seerver is busy also increases whereas it decreases with an increase in the vacation rate.

Table 1
Effect of λ_1 and η on P_1: Fix $\mu_h = 2$, $\mu_l = 1$, $\gamma = 0.1$, $K = 3$, $p = 0.8$, $L = 4$, $m = 2$.

λ_1	P_1	η	P_1
0.6	0.5302	0.6	0.9445
0.7	0.6250	0.7	0.9343
0.8	0.7225	0.8	0.9268
0.9	0.8230	0.9	0.9213
1.0	0.9268	1.0	0.9171

6. Conclusion

We analyzed a queuing model, M/PH(1),PH(2)/2, where the main server takes vacations or undergoes repair, and the standby server handles high-priority customers during this time. The matrix analytic method is used to find the stationary probability vector, helping to calculate performance

measures. This model can also be extended to include MAP arrivals with phase-type vacations and repairs.

REFERENCES

1. *Ayyappan G., Karpagam S.* An M[X]/G(a,b)/1 Queueing System with Breakdown and Repair, Stand-By Server, Multiple Vacation and Control Policy on Request for Re-Service // Mathematics. 2018. Vol. 6, № 6.
2. *Chakravarthy, Shrut S. R., Kulshrestha R.* A Queueing Model With Server Breakdowns, Repairs, Vacations And Backup Server // Operations Research Perspectives. 2020. Vol. 7.
3. *Dudin A., Dudin S.* Analysis of a Priority Queue with Phase-Type Service and Failures // International Journal of Stochastic Analysis. 2016. Vol. 216. P. 1–11.
4. *Doshi B. T.* Queueing systems With Vacations: A Survey // Queueing Systems. 1986. Vol. 1. P. 29–66.
5. *Finn S. G., Servi L. D.* M/M/1 Queues With Working Vacations (M/M/1/WV // Performance Evaluation. 2002. Vol. 50, № 1. P. 41–52.
6. *Gray W. J.,Wang P. P.,Scott M.* Queueing Models with Backup Servers and Service Breakdowns // OPSEARCH. 2002. Vol. 39, № 5-6. P. 281–295.
7. *Levy Y., Yechiali U.* Utilization of Idle Time in an M / G /1 Queueing System // Management Science. 1975. Vol. 22, № 2. P. 202–211.
8. *Li J., Tian N.* Performance Analysis of a GI/M/1 Queue With Single Working vacation // Applied Mathematics and Computation. 2011. Vol. 217, № 10. P. 4960–4971.
9. *Tian N., Zhang Z. G.* Vacation Queueing Models Theory and Applications. US: Springer, 2006.
10. *Wu D.A., Takagi H.* M/G/1 Queue With Multiple Working Vacations // Performance Evaluation. 2006. Vol. 63, № 7. P. 654–681.

Balaji M. N. — Research Scholar, Dept. of Mathematics, Maharaja's College, Ernakulam, Kerala–682011, India . E-mail: *balajimn2004@gmail.com*

Dr. Jaya S — Associate Professor, Dept. of Mathematics, Maharaja's College, Ernakulam, Kerala–682011, India. E-mail: *jayasreelakam@gmail.com*

A k-OUT-OF-n:G RETRIAL SYSTEM EXTENDING THE SERVICE TO EXTERNAL CUSTOMERS UNDER MULTIPLE VACATIONS AND N-POLICY

Binumon Joseph[1], K. P. Jose[2]

[1] *Govt. Engineering College Idukki, Kerala, India*
[2] *PG & Research Dept. of Mathematics, St. Peter's College, Kerala, India*

This study examines the reliability of a k-out-of-n:G repairable system with a single server in charge of repairs of faulty components and also serves external customers. The failure times of the system's components follow an exponential distribution, whereas external consumers arrive using the Poisson process. The service durations for both internal components and external customers are exponentially distributed. The system is made up of a finite capacity pool for waiting external customers, and if the pool is full, arriving external consumers enter a retrial orbit with a probability or exit the system with the complementary probability. The N policy governs the servicing of both external customers and internal components. When the system is free of failing external and internal components, the server goes on multiple vacations. We use the Matrix Analytic Method to discuss system stability and steady-state distributions.
Keywords: *Multiple vacation, k-out-of-n system, Matrix-Analytic Method, N-Policy*

Introduction

A common type of redundancy is k-out-of-n reliability systems, which operate if a minimum of k components are functioning. The reliability of a k-out-of-n:G system with multiple repairmen is studied by Yuan [8]. The repairmen are going on multiple vacations, which are independently and identically distributed. A linear consecutive k-out-of-n:F system is investigated by Wang et al. [7] in which the repair facility is subject to breakdown. When the server is busy, an arriving failed component enters a retrial orbit.

Attending more diverse services may improve the server's experience. Krishnamoorthy et al. [5] analysed a k-out-of-n system with a single server offering service to external customers by using the idle time. The system consists of a pool of finite capacity for external customers, and when an arriving external customer finds the pool is fully occupied, the external customer joins a retrial orbit with a probability and leaves the system with a complementary probability. Chakravarthy et al. [1] investigated a k-out-

of-n system with a single reliable server and K spare components, using the Matrix Geometric method. The spare components are used to replace the failed components of the system with a probability that depends on the count of working units.

Dudin et al. [2] studied a k-out-of-n system in which the server's idle periods were used to serve external customers. When the server is occupied, external clients, arriving according to a BMAP, are redirected to an orbit. Hu et al. [3] investigated a repairable k-out-of-n:G system with no waiting space for failed components. The failed components enter into a retrial orbit if the repairman is busy. Joseph and Jose [4] analysed a k-out-of-n system served by a single server, where the server's idle time is utilised to serve external customers. A vacation is triggered for the server whenever the system has no external customers and fewer than N internal component failures.

1. Problem Formulation and Analysis

In this paper, we consider a k-out-of-n:G system, in which the components are subject to failure. When the number of failed internal components reaches the predefined threshold N, the server initiates its repair, handling each component individually in sequence. Once the server has repaired failed components, it begins a vacation, with the vacation time being exponentially distributed with rate ν. However, if the number of failed components reaches N during this vacation, the server interrupts the vacation and immediately resumes servicing the internal components. The idle time of the server is used for the service of customers from outside the system. The lifetime distribution of system components is exponentially distributed with parameter λ_s. The external customers approached the system for service; the inter-arrival time distribution of external customers follows an exponential distribution with parameter λ_e. A newly arrived external customer does not join the system when the server is busy with internal components.

The system consists of a pool of finite capacity K for accommodating the arriving external customers. An arriving external customer gets into the pool if it is vacant; otherwise, they join an orbit of infinite capacity with probability γ. The retrial rate of an orbital customer is exponentially distributed with parameter θ. If there are i external customers in the orbit, then the retrial rate is $i\theta$. If the retrial customer fails to join the pool, then the customer again joins the orbit with probability δ. After completing the vacation, if the pool is free from external customers and the number of failed internal components is less than N, then the server goes for another vacation. When the server is engaged in external service and the count of failed internal components reaches the threshold N, it transitions to servicing in-

ternal failures upon completion of the ongoing external task. The service time distributions of internal failed components and external customers follow exponential distributions with parameters μ_s and μ_e respectively.

Let $N_o(t)$ be the number of external failed customers in the orbit, $N_p(t)$ be the number of external failed customers in the pool, $N_s(t)$ be the number of internal failed components in the system, and $S(t)$ the status of the server at time t, which is 0 when the server is in vacation, 1 if in internal service, and 2 if in external service. Then $\{X(t), t \geq 0\}$ where $X(t) = = (N_o(t), S(t), N_p(t), N_s(t))$ is a continuous time Markov chain with the state space $\Omega = \{(i,0,j,l)/i = 0,1,2,\ldots; j = 0,1,2,\ldots, K; l = 0,1,2,\ldots,N-1\} \cup \{(i,1,j,l)/i = 0,1,2,\ldots,; j = 0,1,2,\ldots,K;\ l = 1,2,\ldots,n-k+1\} \cup \{(i,2,j,l)/i = 0,1,2,\ldots,; j = 1,2,\ldots,K;\ l = 0,1,2,\ldots,\ n-k+1\}$. Let I_n be the n^{th} order identity matrix, E_k be a k^{th} order square matrix in which diagonal elements are -1 and $(i,i+1)^{th}$ elements are 1, $E_k^{'}$ the transpose of E_k, $r_k(i)$ be the 1×k order row matrix with i^{th} element as 1 and all other elements as zeros, $\mathbf{e}$ be a column matrix of appropriate order with all elements 1, $c_k(i)$ be the transpose of $r_k(i)$, and $\otimes$ be the Kronecker product of matrices. Then he infinitesimal generator matrix Q of $\{X(t), t \geq 0\}$ is

$$Q = \begin{pmatrix} A_{(1,0)} & A_0 & & & & & & & \\ A_{(2,1)} & A_{(1,1)} & A_0 & & & & & & \\ & A_{(2,2)} & A_{(1,2)} & A_0 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & A_{(2,r)} & A_{(1,r)} & A_0 & & & \\ & & & & A_{(2,r+1)} & A_{(1,r+1)} & A_0 & & \\ & & & & & \ddots & \ddots & \ddots & \end{pmatrix}.$$

$$A_0 = \begin{pmatrix} A_0^1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & A_0^9 \end{pmatrix},\ A_0^1 = \Big(r_{(K+1)}(K+1) \otimes C_{(K+1)}(K+1)\Big) \otimes \gamma\lambda_e I_N,$$

$A_0^9 = \Big(r_K(K) \otimes C_K(K)\Big) \otimes \gamma\lambda_e I_{(n-k+2)}$. For $i = 1,2,3,\ldots$ $A_{(2,i)} = = \begin{pmatrix} A_{(2,i)}^1 & 0 & 0 \\ 0 & A_{(2,i)}^5 & 0 \\ 0 & 0 & A_{(2,i)}^9 \end{pmatrix}$, where $A_{(2,i)}^1 = (E_{(K+1)} + I_{(K+1)}) \otimes i\theta I_N +$ $+(r_{(K+1)}(K+1) \otimes C_{K+1}(K+1)) \otimes (1-\delta) i\theta I_N$. Also $A_{(2,i)}^5$ and $A_{(2,i)}^5$ have same structure. For $i = 1,2,3,\ldots A_{(1,i)} =$

$$= \begin{pmatrix} A^1_{(1,i)} & A^2_{(1,i)} & A^3_{(1,i)} \\ A^4_{(1,i)} & A^5_{(1,i)} & A^6_{(1,i)} \\ A^7_{(1,i)} & A^8_{(1,i)} & A^9_{(1,i)} \end{pmatrix}, A^1_{(1,i)} = \begin{pmatrix} B^1_{(0,i)} & B^1_{(0,i+1)} & & & & \\ & B^1_{(1,i)} & B^1_{(1,i+1)} & & & \\ & & B^1_{(2,i)} & & & \\ & & & \ddots & & \\ & & & & B^1_{(K-1,i+1)} & \\ & & & & & B^1_{(K,i)} \end{pmatrix},$$

$B^1_{(j,i)} = \lambda_s E_N - (\lambda_e + i\theta) I_N, \quad B^1_{(j,i+1)} = \lambda_e I_N, \quad$ for $j = 0, 1, \ldots, K-1,$

$B^1_{(K,i)} = \lambda_s E_N - (\nu + \gamma\lambda_e + (1-\delta) i\theta)\, I_N.$

$A^2_{(1,i)} = \lambda_s \; I_{K+1} \otimes \Big(r_{n-k+1}(N) \otimes C_N(N) \Big),$

$A^3_{(1,i)} = \begin{pmatrix} 0_{N \times M(n-k+2)} \\ I_K \otimes [\nu I_N \quad 0_{N \times n-k+2-N}] \end{pmatrix},$

$A^4_{(1,i)} = I_{K+1} \otimes \mu_1 \Big(r_N(1) \otimes C_{n-k+1}(1) \Big), \; A^5_{(1,i)} = \begin{pmatrix} A^{(5,1)}_{(1,i)} & 0 \\ 0 & A^{(5,4)}_{(1,i)} \end{pmatrix},$

$A^{(5,1)}_{(1,i)} = I_K \otimes \Big(\lambda_s E_{n-k+1} + \mu_s E'_{n-k+1} + \lambda_s \Big(r_{n-k+1}(n-k+1) \otimes C_{n-k+1}(n-k+1) \Big) - i\theta I_{(n-k+1)} \Big),$

$A^{(5,4)}_{(1,i)} = I_K \otimes \Big(\lambda_s E_{n-k+1} + \mu_s E'_{n-k+1} + \lambda_s \Big(r_{n-k+1}(n-k+1) \otimes C_{n-k+1}(n-k+1) \Big) - (1-\delta) i\theta I_{(n-k+1)} \Big), \quad A^6_{(1,i)} = 0_{(K+1)(n-k+1) \times K(n-k+2)},$

$A^7_{(1,i)} = \begin{pmatrix} \mu_e I_N & 0 \\ 0 & 0 \end{pmatrix}, \; A^8_{(1,i)} = \begin{pmatrix} 0_{N \times (N-1)} & 0 & 0 \\ 0 & \mu_e I_{(n-k+2-N)} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$

$A^9_{(1,i)}$ is a tridiagonal matrix with entries $B^9_{(j,i)} = \mu_e I_{(n-k+2)}$, for $j = 1, 2, \ldots, K$, $B^9_{(j,i+1)} = \lambda_s E_{n-k+2} + \lambda_s \Big(r_{n-k+2}(n-k+2) \otimes C_{n-k+2}(n- -k+2) \Big) - \Big(\lambda_e + \mu_e + i\theta \Big) I_{(n-k+2)}, \quad$ for $i = 0, 1, \ldots, K-1,$

$B^9_{(K,i+1)} = \lambda_s E_{n-k+2} + \lambda_s \Big(r_{n-k+2}(n-k+2) \otimes C_{n-k+2}(n-k+2) \Big) - \Big(\gamma\lambda_e + + \mu_e + (1-\delta) i\theta \Big) I_{(n-k+2)}, \; B^9_{(j,i+2)} = \lambda_e I_{n-k+2}, \quad$ for $i = 0, 1, \ldots, K-1.$

2. Stability Condition

The stability condition is obtained by the Neuts-Rao truncation [6] by assuming $A_{1i} = A_{1M}$ and $A_{2i} = A_{2M}$ for all $i \geq M$. Let $\mathbf{\Pi} = (\mathbf{\Pi_0}, \mathbf{\Pi_1}, \mathbf{\Pi_2})$ be the steady state probability vector of $A = A_{2M} + A_{1M} + A_0$. The stability condition of the system is obtained from $\lim\limits_{M \to \infty} \left(\dfrac{\mathbf{\Pi_M A_0 e}}{\mathbf{\Pi_M A_{2M} e}} \right) < 1.$

3. Steady State Probability Vector

To obtain the steady state probability vector, we use the truncation method by Neuts-Rao [6]. Let the steady state probability vector $\mathbf{x} = (x_0, x_1, x_2, \dots)$ of Q, where, $\mathbf{x}_i = (x_{(i,0,0,1)}, x_{(i,0,0,2)}, \dots, x_{(i,0,0,N)})$, $(x_{(i,0,1,1)}, x_{(i,0,1,2)}, \dots, x_{(i,0,1,N)}), \dots, (x_{(i,0,K,1)}, x_{(i,0,K,2)}, \dots, x_{(i,0,K,N)})$, $(x_{(i,1,0,1)}, x_{(i,1,0,2)}, \dots, x_{(i,1,0,(n-k+1))}), (x_{(i,1,1,1)}, \dots, x_{(i,1,1,(n-k+1))}), \dots,$ $(x_{(i,1,K,1)}, x_{(i,1,K,2)}, x_{(i,1,K,3)}, \dots, x_{(i,1,K,(n-k+1))}), (x_{(i,2,1,0)}, x_{(i,2,1,1)}, \dots,$ $x_{(i,2,1,(n-k+1))}), (x_{(i,2,2,0)}, x_{(i,2,2,1)}, \dots, x_{(i,2,2,(n-k+1))}), \dots, (x_{(i,2,K,0)},$ $x_{(i,2,K,1)}, x_{(i,2,K,2)}, \dots, x_{(i,2,K,(n-k+1))})$.

Then, x_i follows the relation $x_{M+k-1} = x_{M-1}R^k$, $k \geq 1$, where R is the least non-negative solution of the quadratic matrix equation, $R^2A_2 + RA_1 + A_0 = \mathbf{0}$, with $A_1 = A_{1M}, A_2 = A_{2M}$ and $R = \lim_{n\to\infty} R_n$, where $\{R_n\}$ is defined such that $R_{n+1} = -A_0A_1^{-1} - R_nA_2A_1^{-1}; n \geq 0$ and $R_0 = \mathbf{0}$. The vectors $x_0, x_1, \dots, x_{M-1}$ are obtained by solving the equations $x_0A_{10} + x_1A_{21} = 0$, $x_{i-1}A_0 + x_iA_{1i} + x_{i+1}A_{2(i+1)} = 0; (1 \leq i \leq M-2)$, and $x_{M-2}A_0 + x_{M-1}(A_{1(M-1)} + RA_2) = 0$, subject to normalizing condition $[\Sigma_{i=0}^{M-2}x_i + x_{M-1}(1-R)^{-1}]e = 1$.

4. Performance Measures of the System

1) Fraction of time the system was down,
   $P_F = \sum_{i=0}^{\infty}\sum_{j=0}^{K} x_{(i,1,j,(n-k+1))} + \sum_{i=0}^{\infty}\sum_{j=1}^{K} x_{(i,2,j,(n-k+1))}$.
2) The average number of external customers waiting in the pool,
   $N_P = \sum_{i=0}^{\infty}\Big(\sum_{j=0}^{K} j(\, x_{(i,0,j)}\mathbf{e} + x_{(i,1,j)}\mathbf{e}) + \sum_{j=1}^{K}(j-1)\, x_{(i,2,j)}\mathbf{e}\Big)$.
3) The average number of failed system components, $N_{IF} = \sum_{i=0}^{\infty}\Big(\sum_{j=0}^{K}$ $(\sum_{l=0}^{N-1} l\, x_{(i,0,j,l)} + \sum_{l=1}^{n-k+1} l\, x_{(i,1,j,l)}) + \sum_{j=1}^{K}\sum_{l=0}^{n-k+1} l\, x_{(i,2,j,l)}\Big)$.
4) Probability that the server was on vacation, $P_v = \sum_{i=0}^{\infty}\sum_{j=0}^{K} x_{(i,0,j)}\mathbf{e}$.
5) Expected number of external customers lost before entering the orbit,
   $E_{EL1} = \lambda_e(1-\gamma)\sum_{i=0}^{\infty}\Big(x_{(i,0,K)}\mathbf{e} + x_{(i,2,K)}\mathbf{e}\Big)$.
6) Average count of external customer lost due to retrial,
   $E_{EL2} = \sum_{i=0}^{\infty} i\theta(1-\delta)\Big(x_{(i,0,K)}\mathbf{e} + x_{(i,1,K)}\mathbf{e} + x_{(i,2,K)}\mathbf{e}\Big)$.
7) Expected service completion rate of external customers,
   $E_{escr} = \mu_e \sum_{i=0}^{\infty}\sum_{j=1}^{K} x_{(i,2,j)}\mathbf{e}$.

The expected cost/unit time of the system is $ETC = C_1P_F + C_2N_P - C_3E_{escr} + C_4(E_{EL1} + E_{EL2}) + C_5N_{IF} + C_6P_v$, where C_i are the cost associated with the corresponding performance measures.

5. Conclusion

The problem is analysed by a continuous-time Markov chain. The steady-state probability vector and system stability are discussed using the Matrix Analytic method. Several performance measures of the system and a suitable cost function are obtained.

REFERENCES

1. *Chakravarthy S. R.* Analysis of a k-out-of-n system with spares, repairs, and a probabilistic rule // International Journal of Stochastic Analysis. 2006(1). Art.num. 039093.
2. *Dudin A., Krishnamoorthy A., Narayanan V.* Idle time utilization through service to customers in a retrial queue maintaining high system reliability // Journal of Mathematical Sciences. 2013. Vol. 191. P. 506–517.
3. *Hu L., Liu S., Peng R., Liu Z.* Reliability and sensitivity analysis of a repairable k-out-of-n: G system with two failure modes and retrial feature // Communications in Statistics-Theory and Methods. 2022. Vol. 51(9). P. 3043–3064.
4. *Joseph B, Jose K. P.* Reliability Analysis of a k-out-of-n Single Server System Extending Service to External Customers Under N-Policy and Server Vacations // International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2024. P. 242–256.
5. *Krishnamorthy A., Narayanan V. C., Deepak T.* Maximization of reliability of a k-out-of-n system with repair by a facility attending external customers in a retrial queue // Reliability: Theory & Applications. 2007. Vol. 2(2(6)). P. 21–33.
6. *Neuts M.F., Rao B.M.* Numerical investigation of a multiserver retrial model // Queueing systems. 1990. Vol. 7(2). P. 169–189.
7. *Wang Y., Hu L., Yang L., Li J.* Reliability modeling and analysis for linear consecutive-k-out-of-n: F retrial systems with two maintenance activities // Reliability Engineering & System Safety. 2022. Vol. 226. Art.num 108665.
8. *Yuan L.* Reliability analysis for a k-out-of-n: G system with redundant dependency and repairmen having multiple vacations // Applied Mathematics and Computation. 2012. Vol. 218(24). P. 11959–11969.

Binumon Joseph — M.Phil., Assistant Professor, Govt. Engineering College Idukki, Painavu-685603, Kerala, India.. E-mail: *jbinumonjoseph@gmail.com*

K. P. Jose — Ph.D., Associate Professor, Head of the Department, PG & Research Dept. of Mathematics, St. Peter's College, Kolenchery -682311, Kerala, India.. E-mail: *kpjspc@gmail.com*

OPTIMIZATION OF MULTI-DIMENSIONAL INSURANCE NETWORKS WITH INVESTMENT AND REINSURANCE

E. Bulinskaya[1]

[1] *Lomonosov Moscow State University, Moscow, Russian Federation*

The study of insurance networks became popular during the last decade. In this paper we consider two examples of such systems. The first one is devoted to investigation of ruin probability for a network consisting of a main company (reinsurer) and n subsidiary ones. Proportional (or non-proportional) reinsurance is supposed to be applied. The second one treats the Generalized Cramer-Lundberg model. To prove the results we use the integro-differential equations, stochastic calculus and Fourier transforms. **Keywords:** *optimization, insurance network, investment, reinsurance*

Introduction

Risk management or decision making under uncertainty (unfavorable or favorable) is specific for all applications of probability theory, such as insurance, finance, queueing, reliability, population dynamics, medicine, inventory, dams and many others. The main question is how to choose an appropriate mathematical model. The most widely used models are of input-output type described by six-tuple $(T, Y, Z, U, \Psi, \mathcal{L})$, see, for example, [1]. Here

- T – planning horizon;
- $Z = \{Z(t), t \in [0, T]\}$ – input process;
- $Y = \{Y(t), t \in [0, T]\}$ – output process;
- $U = \{U(t), t \in [0, T]\}$ – control;
- Ψ – functional describing the system configuration and its operation mode, thus $X = \Psi(Z, Y, U)$;
- $X = \{X(t), t \in [0, T]\}$ – the system state;
- $\mathcal{L}_T(U) = \mathcal{L}(Z, Y, U, X, T)$ – objective function (target, valuation criterion or risk measure) evaluating the quality of system performance.

At the beginning only limit behavior was studied for dams and queueing systems, that is, no objective function existed. On the other hand, the cost approach was common for inventories and finance, whereas reliability theory and insurance used the so-called reliability approach investigating the ruin

(either survival) probability. Nowadays, objective function (risk measure) is introduced in all the research domains.

Definition 1. Control $U_T^* = \{U^*(t), t \in [0,T]\}$ is optimal if $\mathcal{L}_T(U_T^*) = \inf\limits_{U_T \in \mathcal{U}_T} \mathcal{L}_T(U_T)$,

(or $\mathcal{L}_T(U_T^*) = \sup\limits_{U_T \in \mathcal{U}_T} \mathcal{L}_T(U_T)$),

where $\mathcal{U}_T$ is the class of admissible controls.

Further, $U^* = \{U_T^*, T \geq 0\}$ is called the optimal policy (or strategy).

It is interesting to give the interpretation of the above mentioned parameters for various applications as stated by the Table 1.

Table 1

Interpretation of terms for different applications

Domain of investigation	Input $Z(t)$	Output $Y(t)$	State $X(t)$
Insurance	Premium	Claims	Surplus
Finance	Money inflow	Money outflow	Capital
Inventory	Supply	Demand	Inventory level
Dams	Water inflow	Outflow	Water level
Reliability	New & Repaired elements	Damaged elements	Working elements
Queueing	Clients arrival	Served clients	Queue length
Population growth	Birth & immigration	Death & emigration	Population size

We are interested in the study of insurance models, so we provide the following

Definition 2. Insurance is a contract where one party (the insurer) agrees to compensate another party (the insured) for specified losses, damages, or injuries, in exchange for a premium payment. It's a form of risk management that helps individuals and organizations mitigate potential financial hardship from unforeseen events. All calculations are performed according to the laws of probability theory.

1. Insurance network

Risk networks became popular during the last decade of the 21st century (see [2], [3]). We begin by investigation of ruin probability for a network consisting of a main company (insurer, as well as reinsurer) and n subsidiary ones (insurers). Proportional (or non-proportional) reinsurance is supposed to be applied.

Define the following processes $X_i(t)$, $i = 0, 1, ..., n$

$$X_i(t) = u_i + c_i \cdot t - \alpha_i C_i(t),\ i = 1, .., n \text{ and}$$

$$X_0(t) = u_0 + c_0 \cdot t - \alpha_0 C_0(t) - \textstyle\sum_{i=1}^{n} \beta_i C_i(t),$$

$X_i(t)$ is the capital of the i-th company at time t, $u_i = X_i(0)$ is the initial capital i-th company, c_i is premium inflow to the i-th company, α_i, β_i are respectively fractions of aggregate claims $C_i(t)$ of the i-th company which are paid by insurers and a sole reinsurer. Note that $0 \leq \alpha_i, \beta_i \leq 1$ and $\alpha_i + \beta_i = 1$, $i = 0, 1, ..., n$.

The aggregate claim processes are defined as follows

$$C_i(t) = \sum_{j=1}^{N_{\lambda_i}(t)} C_{i,j}, \quad i = 0, ..., n,$$

where $N_{\lambda_i}(t)$ is a Poisson process with intensity λ_i, $i = 0, 1, ..., n$, and $C_{i,j}$ is the j-th payment of the i-th company. It is supposed that $C_{i,j}$ are independent identically distributed random variables with a density $f_i(x)$, distribution function $F_i(x)$ and finite expectations. Moreover, $C_{i,j}$ and $N_{\lambda_i}(t)$ are independent.

Let $\tau_i(u_i) = \inf\{t \geq 0 : X_i(t) < 0 | X_i(0) = u_i\}$ denote the ruin time of the i-th insurer. We define the system's ruin time $\tau(u_0, ..., u_n)$ as $\min\limits_{i=0,1,...,n} \tau_i(u_i)$.

Our aim is obtaining the ruin probability, that is, $\Psi(u_0, ..., u_n) = = P(\tau(u_0, ..., u_n) < \infty)$.

For investigation we use the Gerber-Shiu function defined as follows:

$$\varphi(u_0, ..., u_n) = E\left[e^{-\delta\tau} w(\textstyle\sum_{i=0}^{n} X_i(\tau^-), \sum_{i=0}^{n} [X_i]_- I\{\tau < \infty\} | \right.$$
$$\left. X_0(0) = u_0, X_1(0) = u_1, ..., X_n(0) = u_n\right],$$

$\delta \in [0; +\infty)$, $w : [0; +\infty) \times (0; +\infty) \to [0; +\infty)$ is a bounded (Borel measurable) function and $[f(x)]_- = \max(0, -f(x))$ denotes the negative part of $f(x)$. Note that putting $\delta = 0$ and $w(x; y) = 1$ we get $\varphi(u_0, ..., u_n)$.

Theorem 1. Gerber-Shiu function $\varphi(u_0, ..., u_n)$ for $(n+1)$-dimensional network satisfies the following integro-differential equation

$$(\delta + \sum_{i=0}^{n} \lambda_i)\varphi(u_0, ..., u_n) = \sum_{i=0}^{n} c_i \frac{\partial}{\partial u_i}\varphi(u_0, ..., u_n) +$$

$$+\lambda_0 \int_0^{u_0} \varphi(u_0 - x, u_1, ..., u_n) f_0(x)dx + \lambda_0 \int_{u_0}^{\infty} w(\sum_{i=0}^{n} u_i, x - u_0) f_0(x)dx +$$

$$+\sum_{i=1}^{n} \left(\lambda_i \int_0^{\min(\frac{u_0}{\beta_i}, \frac{u_i}{\alpha_i})} \varphi(u_0 - \beta_i x, u_1, ..., u_{i-1}, u_i - \alpha_i x, u_{i+1}, ..., u_n) f_i(x)dx + \right.$$

$$\left. + \lambda_i \int_{\min(\frac{u_0}{\beta_i}, \frac{u_i}{\alpha_i})}^{\infty} w(\sum_{j=0}^{n} u_j, (\beta_i x - u_0)^+, (\alpha_i x - u_i)^+) f_i(x)dx \right),$$

where u_i is the initial capital of company i and $f_i(x)$ is the density of its claims, $i = 0, 1, ..., n$.

2. Generalized Cramer-Lundberg model

In order to make the classical model more realistic, we consider n categories of insurance (not a single one), replace the compound Poisson process by a generalized Poisson process and include the investment. Based on these conditions, a risk model is set up to find (by the martingale methods) the formula of ruin probability and its upper bound. Thus, we put

$$U(t) = u_1 + (u_1 - u_2 - u_3)(b - c + d) + S(t),$$

$$S(t) = u_2 r_1 t + u_3(r_2 t + aB(t)) + \sum_{j=1}^{n} \sum_{i=1}^{N_j(t)} X_i^{(j)} - \sum_{j=1}^{n} \sum_{i=1}^{M_j(t)} Y_i^{(j)}.$$

Here $U(t)$ is the capital of insurance company and $S(t)$ is the gain at time t, u_1 is the initial reserve of insurance company, u_2 is invested in a non-risky asset, u_3 is used in venture capital, r_1 is rate of return of u_2, $u_3(r_2 t + aB(t))$ is income from investment of u_3 in a Brownian motion with parameters r_2, a (here r_2 is drift parameter, a is volatility, $B(t)$ is a standard Brownian motion), b is interest rate, c is inflation rate, d is exchange rate.

Clearly, parameters satisfy the following inequalities $t \geq 0, u_1 > 0$, $u_2 > 0$, $u_3 > 0, r_1 > 0, r_2 > 0, a > 0$. Furthermore, $N_j(t)$ and $M_j(t)$ are Generalized Poisson processes providing inhomogeneity of premiums and claims flows, $X_i^{(j)}$ represents the ith premium, while $Y_i^{(j)}$ is the ith claim amount for the jth insurance branch. All r.v.'s possess finite means and variances.

Lemma 1. Surplus process $\{S(t), t \geq 0\}$ has stable independent increments.

Lemma 2. There is a function $h(r)$ satisfying the relation

$$E[exp(-rS(t))] = exp(th(r)).$$

Lemma 3. Equation $h(r) = 0$ has only one positive root R.

Theorem 2. For the Doubly Stochastic Generalized Poisson multiple risk model, the upper bound of ruin is

$$\varphi(u) \leq \exp[-R(u_1 + (u_1 - u_2 - u_3)(b - c + d))].$$

Theorem 3. For the Doubly Stochastic Generalized Poisson multiple risk model, the ruin probability is

$$\varphi(u) = \frac{\exp[-R(u_1 + (u_1 - u_2 - u_3)(b - c + d))]}{E[\exp[-RU(T)]|T < \infty]},$$

here T is the ruin time, that is, $T = \inf\{t \geq 0 : U(t) < 0\}$.

Conclusion

Thus, we investigated two interesting models important for applications: Insurance network with proportional reinsurance and Generalized Cramer-Lundberg model. In addition to equation for the Gerber-Shiu function we obtain some results for its Fourier transform and then calculate the ruin probability. The second model describes a company having several business lines. The model is doubly stochastic and uses investment in risky and non-risky assets. Moreover, the input (premiums flow) and output (claims flow) are generalized Poisson processes (not compound ones). The properties of company surplus are studied. This model illustrates the reliability approach, providing the ruin probability. Next step is investigation of system stability and choice of optimal parameters. The books [4] and [5] and papers [6-8] enable us to carry out the proofs.

Further research directions are

- consideration of more complex claims flows, for example, instead of Poisson process it is possible to introduce renewal either diffusion processes or their combination;
- transfer of the obtained results to the case of Parisian ruin;
- using of a different objective function such as expected discounted dividends paid until the ruin;
- comparison of models by means of stochastic orders of distribution appearing in models description.

REFERENCES

1. *Bulinskaya E.* New research directions in modern actuarial sciences// Modern problems of stochastic analysis and statistics – selected contributions in honor of Valentin Konakov. Berlin: Springer, 2017. P. 349–408.
2. *Avram F., Loke S.-H.* On Central Branch/Reinsurance Risk Networks: Exact Results and Heuristics // Risks. 2018. Vol. 6, № 35. 28 p.
3. *Avram F., Palmowski Z., Pistorius M.* A two-dimensional ruin problem on the positive quadrant // Insurance: Mathematics and Economics. 2008. Vol. 42, № 1. P. 227–234.
4. *Bellman R.* Dynamic programming Princeton New Jersey: Princeton University Press, 1957. 340 p.
5. *Billingsley P.* Convergence of Probability Measures New York: John Wiley and Sons, 1968. 253 p.
6. *Bulinskaya E., Gusak J.* Insurance Models under Incomplete Information // Statistics and Simulation, Springer Proceedings in Mathematics and Statistics, 2018. Vol. 231. P. 171–185.
7. *Bulinskaya E.* Asymptotic analysis and optimization of some insurance models // Applied Stochastic Models in Business and Industry. 2018. Vol. 34, № 6. P. 762–773.
8. *Bulinskaya E.V.* Cost approach versus reliability // Proceedings of International Conference DCCN-2017. Moscow: Technosphera, 2017. P. 382–389.

Bulinskaya Ekaterina — Doctor of Physical and Mathematical Sciences, Professor, Department of Mechanics and Mathematics. E-mail: *ebulinsk@yandex.ru*

MARKOV OPTIMIZATION STRATEGY IN DISTRIBUTED COMPUTER SYSTEMS

A.A. Chudinova[1]

[1] *University ITMO, Saint-Petersburg, Russian Federation*

In cloud-native environments, the issue of rational resource use and intensity distribution is relevant. This research proposes an optimal strategy for traffic intensity measurement under uncertainty, with particular reference to the Kubernetes ecosystem. To improve the effectiveness of the developed mathematical model (MDP), the historical and real-time data are comparable. This model solves the issue with: (1) proactive resource management; (2) service reliability; (3) cost optimization. Based on the metrics of spending (before/after), resource management improvements, effectiveness of CPU/RAM use and ROI, the hypothesis of optimization strategy applicability has been confirmed. This experimental study is innovative in terms of organizing the theoretical analysis as a first part of creating the mathematical model and the experimental study organization applying Kubernetes for collecting the data and running a new algorithm comparing the results to evaluate the effectiveness of the developed Markov optimization strategy as a second part of the research. The hypothesis of Markov decision process implementation effectiveness has been approved because the Markov optimization strategy successfully minimizes the probability of entering undesirable states and resource efficiency increased and overall operating costs decreased. **Keywords:** *optimisation strategy, Markov decision process, distributed computer system, Kubernetes, proactive resource management, service reliability, cost optimization, queuing theory*

Introduction

Contemporary distributed systems have the situations of resource management disbalanced use. Markov optimization strategy can solve this issue. First of all, to set the objectives of constructing a Markov strategy for proactive management, service reliability and system cost means to have the task positioning:

– to assume that actions like a_1, a_2 (if it involves maintenance/inspection) or a_3 (maintenance) are proactive. Assign it appropriate reward (costs)

Supported by V.A. Bogatyrev, Professor, Doctor of Technical Sciences, academic advisor

to encourage its fulfilment if it leads to higher reliability and lower cost in the long run;
- to minimise the probability of reaching a failure state s_6 and maximize the probability of reaching a successful maintenance state s_5;
- to minimise the total expected cost or maximise the net reward.

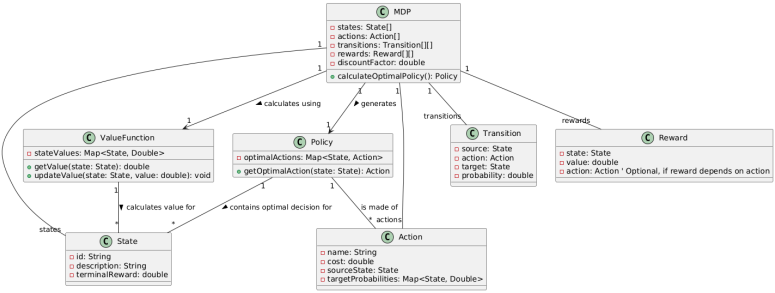


Figure 1. Markov optimization strategy

1. Methods

The Markov Optimization Strategy $\pi^*(s)$ is a function for each state $s \in S$ that defines the optimal action $a \in A(s)$.

- Reward of success $R(s_5) = 100$ (high positive reward)
- Penalty for failure $R(s_6) = -200$ (high negative reward/penalty)

Action costs (negative reward to minimize):

- $R(s_1, b_1) = -1$ (normal work, low cost);
- $R(s_1, a_1) = -2$ (move to management process, medium cost);
- $R(s_1, a_2) = -1.5$ (transition to another substrate, medium cost);
- $R(s_2, c_1) = -1$ (transition back, normal cost);
- $R(s_3, b_3) = -5$ (high cost, alternative way (a fast path) to s_5);
- $R(s_3, a_3) = -1$ (maintain in s_3, low cost);
- $R(s_3, a_4) = -2$ (transition to s_4, proactive verification, medium cost);
- $R(s_4, a_5) = -1$ (completion from s_4, low cost if s_4 is a validation result);
- $R(s_6, a_6) = -1$ (continued failure, very high cost).

Discount factor $\gamma = 0.95$. This value is chosen to emphasize long-term rewards. If γ were too close to 1 (e.g., 0.999), it could lead to an overly "forward-looking" strategy that is slow to adapt and may not be optimal in a rapidly changing environment. Conversely, a γ that is too low (e.g., 0.5) would make the system "myopic," focusing solely on immediate rewards,

which contradicts the goals of proactive management and reliability. A value of 0.95 represents a reasonable compromise, sufficiently discounting very distant future rewards while retaining the importance of medium and long-term consequences.

Applying the algorithm (Value Iteration) iteratively computes $V(s)$ for each state until the values converge [1]:

$$V_{k+1}(s) = \max_{a \in A(S)} \left(R(s,a) + \sum_{S'} P(S'|s,a) V_k(S') \right).$$

Table 1

Value Iteration computational analysis (a few iterations)

Value	**Calculation**
V1(s5)	100
V1(s6)	-10 (from action a6 in s6) / V1(s6) = -200 (if it is an immediate reward) If s6 is a terminal state with a penalty, then $V(s6) = -200$ If in s6 there is action a6 with a reward of -10, then: $V(s_6) = \max_{a_6}(R(s_6,a_6) + \gamma P(s_6|s_6,a_6)V(s_6))$ (2) (2) — Bellman equation applied to a failure state (s6) in the context of a Markov decision process (MDP) [2]. It describes the optimal value (or expected discounted reward) of state s6. $V(s6) = -200$.
V1(s4) (action a5 only)	$V1(s_4) = R(s_4,a_5) + \gamma[P(s_5|s_4,a_5)V0(s_5) + P(s_6|s_4,a_5)V0(s_6)] = 8.5$
*$V0(s) = 0$ for all s because "according to the optimality principle, an optimal strategy guarantees that after the first decision, subsequent decisions are optimal with respect to the new state obtained from the initial decision, regardless of the initial state and the initial decision" [3].	

As $V_k(s_3)$ converges, it will use $V_k(s_4)$, and $V_k(s_1)$ will use $V_k(s_3)$ and $V_k(s_2)$, and so on. The optimal strategy $\pi^*(s)$ will be the action that maximizes the right-hand side of the Bellman equation [4].

After convergence, $\pi^*(s)$ for each state s will look like this:

$\pi^*(s1)$:

- If $V(s3)$ (the value of moving to $s3$ with $a1$) is significantly higher than $V(s1)$ (staying in $s1$ with $b1$) and $V(s2)$ (moving to $s2$ with $a2$), then $\pi^*(s1) = a1$. This would mean that proactive management (moving to phase $s3$) is preferable;
- If $s1$ is a "safe" state with low costs, and $V(s1)$ is high enough, then $\pi^*(s1) = b1$ (maintaining the current state);

$\pi^*(s2) = c1$ (the only path from $s2$, returning to $s1$ for further processing);

$\pi^*(s3)$:

- $\pi^*(s3) = b3$ if the direct path to $s5$ (successful completion) via $b3$ has the highest expected value, despite its high immediate cost. This would be preferred for fast and reliable service.
- $\pi^*(s3) = a4$ if the stochastic path via $s4$ (with a 0.7 chance at $s5$ and 0.3 at $s6$) provides greater value, given proactive diagnostics and possible lower costs along the way. This would be part of proactive management, accepting managed risks;
- $\pi^*(s3) = a3$ if maintaining at $s3$ is most optimal, perhaps to await external conditions or additional processing.

$\pi^*(s4) = a5$ (the only path from $s4$ that leads to the stochastic outcome $s5$ or $s6$);

$\pi^*(s5)$. Terminal state (reward received, system terminated).

$\pi^*(s6) = a6$ (absorbing state with constant penalty if no recovery action is provided).

2. Results

After the implementation of Markov optimization, the Kubernetes system continued to operate as usual, but under the new optimised logic. Continuous data collection continued for a certain, comparable length of time, as presented in Table 2, for all the same metrics as before in Table 1. The observation period takes place one month before integration and one month after. Workload is controllably varying workload for both phases of the experiment.

SLO calculated the ratio of the number of successfully completed requests to the total number of requests for the period. MTTR measures the time it takes for the system to recover from a failure. This was probably the average of all reported failures. Number of critical failures (s6) incidents categorised as "critical failures" were recorded and the number of these per month was calculated. Percentage of proactive inspections/maintenance (actions a1, a4) tracked actions that were performed as proactive maintenance (e.g., preventive checks, updates) and their percentage of the

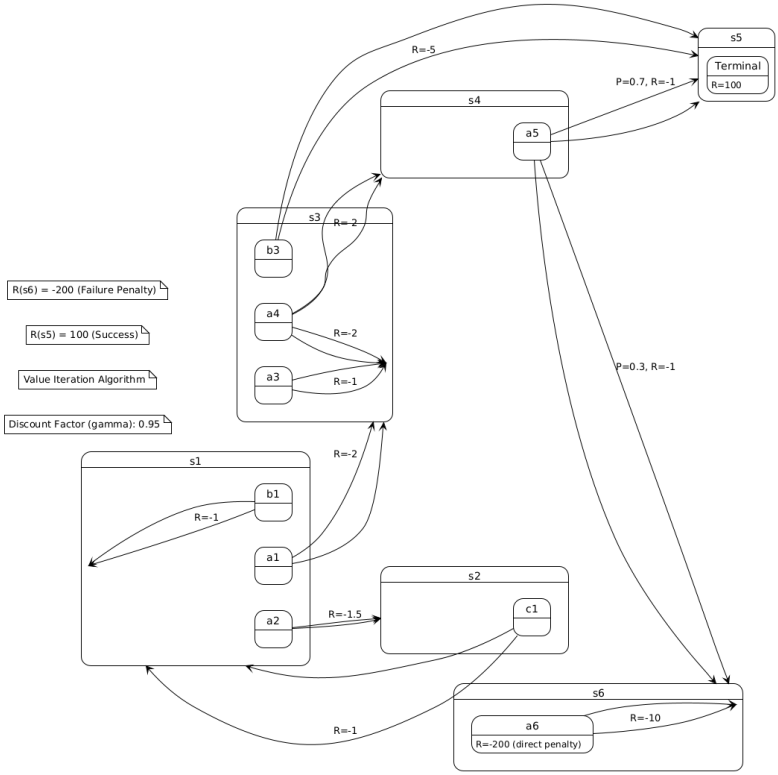


Figure 2. Markov optimization strategy model

total number of all operational actions. The number of alerts that did not correspond to real problems was counted. Resource utilisation across all Kubernetes nodes and containers was measured. The financial costs associated with system operation were estimated, including infrastructure, manual intervention, personnel, etc.

After the implementation of Markov optimization, the Kubernetes system continued to operate as usual, but under the new optimised logic. Continuous data collection continued for a certain, comparable length of time, as presented in Table 2, for all the same metrics as before in Table 1.

Table 2
Real-time data from Kubernetes (before integrating the Markov optimisation strategy)

Metric	**Average Value (before Implementation)**	**Unit**	**Description**
Service Reliability			
Percentage of Successful Requests (SLO)	98.5%	%	Target SLO = 99.9%
Mean Time To Recovery (MTTR)	45	minutes	High recovery time after failures
Number of Critical Failures (s6)	8	times/month	Direct entry into failure state
Average API Response Time	120	ms	Delays affecting UX
Proactive Management			
Percentage of Proactive Checks/Maintenance (actions a1, a4)	15%	% of all actions	Low level of proactive actions
Number of False Alarms	25	times/month	Distracts operators
System Cost			
Average CPU Utilization	75%	%	Inefficient resource usage
Average RAM Utilization	80%	%	Risk of OOM errors
Total Operational Expenses (OpEx)	15000	USD/month	High costs for manual intervention, scaling

Table 3

Optimal data (after integrating the Markov optimization strategy)

Metric	**Average Value (After Implementation)**	**Unit**	**Comments**
Service reliability			
Percentage of Successful Requests (SLO)	99.8%	%	Approaching target SLO
Mean Time To Recovery (MTTR)	10	minutes	Significant reduction in MTTR due to automation
Number of Critical Failures (s6)	2	times/month	Reduction in critical failures
Average API Response Time	80	ms	Performance improvement
Proactive Management			
Percentage of Proactive Checks/Maintenance (actions a1, a4)	60%	% of all actions	Substantial increase in proactive actions
Number of False Alarms	5	times/month	Reduced noise, increased operator efficiency
System Cost			
Average CPU Utilization	60%	%	More efficient scheduling, reduced overprovisioning
Average RAM Utilization	65%	%	Optimized memory consumption
Total Operational Expenses (OpEx)	12000	USD/month	Reduced costs for manual intervention and inefficient scaling

3. Discussion

Service reliability is compared on the **Percentage of successful requests (SR)**, **Mean Time to Recovery (MTTR)**, the **number of critical failures (s6)** and the **average API response time**. SR improved from 98.5% to 99.8%. This demonstrates a significant improvement in availability and quality of service, which is a direct consequence of minimising the probability of reaching a failure state (s6) and maximising the probability of reaching a successful service state (s5). MTTR has been reduced from 45 minutes to 10 minutes. This critical improvement shows that the system is recovering faster from incidents, which may be due to faster response to problems, possibly due to better diagnostics and automated actions provided by the Markov strategy. The number of critical failures (s6) decreased from 8 to 2 times per month. This is a direct confirmation of the effectiveness of the system in avoiding undesirable states. The Markov strategy penalising the s6 state clearly redirects the system along paths with less risk. Average API response time improved from 120 ms to 80 ms. Although not directly stated in the MDP goals, improved reliability and proactive management can lead to a more stable and faster system overall.

Proactive management is compared in the **Percentage of proactive inspections/maintenance (actions a1, a4)** and the **number of false alarms**. The Percentage of proactive inspections/maintenance has increased from 15% to 60%. This is the primary measure of success. The Markovian system, by assigning appropriate rewards/costs to proactive actions, encourages the system to perform diagnostics and maintenance more often, which prevents problems from developing. False positives Decreased from 25 to 5 false positives per month. Improved decision-making algorithms and proactive problem prevention (through more accurate predictions or early remediation) results in less "noise" for operators. **System cost** is measured in **average CPU and RAM utilisation**, and **Total Operating Expenses (OpEx)**. Average CPU usage decreased from 75% to 60%. Average RAM usage has decreased from 80% to 65%. These metrics show that Markov Optimisation has resulted in more efficient use of resources. This may be the result of better planning, avoiding overallocation of resources to reserves, and more accurate diagnostics to avoid sub-optimal configurations. OpEx decreased from 15000 USD/month to 12000 USD/month. This is the cumulative effect of reduced manual intervention, reduced downtime, more efficient resource utilisation and avoidance of costly failures.

4. Conclusion

The implementation of the Markov optimization system significantly improved the key indicators of reliability, proactive management and cost of the Kubernetes system. Markov optimization has been shown to be a powerful tool for building intelligent management strategies for complex systems such as Kubernetes, balancing reliability, proactivity and cost. The results demonstrate that investments in proactive management encouraged by MDP pay off by increasing stability and reducing operational costs.

REFERENCES

1. *Makarov I. M., Vinogradskaya T. M., Rubchinskiy A. A., Sokolov V. B.* Decision making theory. Moscow: Science, 1982. P. 267–273.
2. *Meggendorfer T.* Verification of Discrete-Time Markov Decision Processes. // Doktor der Naturwissenschaften. der Technischen Universität München, 2018. 170 p.
3. *Rachkov M. Yu.* Optimal control in technical systems. 2nd ed., corrected and supplemented. Moscow: Yurait, 2023. 53–59 p.
4. *Semenov V. V., Panteleev A. V., Bortakovsky A. S.* Mathematical Theory of Control in Examples and Problems // Applied Mathematics in Examples and Problems. MAI, 1997. P. 214–216.

Chudinova Alexandra A. — PhD student, University ITMO, The Faculty of Program Engineering and Computer Technologies. E-mail: *alexandra.a.chudinova@mail.ru*

SPATIAL SYNCHRONOUS MARKOVIAN ARRIVAL PROCESS (SPATIAL SYNCHMAP)

Dinata Syalam Ali Wira[1]

[1] *Institute of Applied Mathematics and Computer Science, Tomsk State University, Tomsk, Russia*

This paper introduces and analyzes the Spatial Synchronous Markovian Arrival Process (Spatial SynchMAP), a novel extension of classical Markovian arrival processes that simultaneously captures temporal dynamics through a finite state space $\{1, ..., K\}$ with transition matrix $\mathbf{R}$, state-dependent exponential arrival rates, spatial dynamics via transition displacement kernels, state-specific spatial distributions, modulated transition matrix. The joint characteristic function, combining arrival counts and spatial positions.
Keywords: *Spatial, Synchronous, MAP*

1. Introduction

Markovian Arrival Processes (MAP) [1–3] have significant importance in queuing theory [4, 5] and applied probability. In this work, we analyze the Spatial Synchronous Markovian Arrival Process (Spatial SynchMAP), an extension of synchronous MAP [6, 7] that incorporates both temporal and spatial dimensions. Therefore, we systematically examine its fundamental properties to establish a foundation for subsequent research applications.

2. Spatial Synchronous Markovian Arrival Process (Spatial SynchMAP)

The *Spatial Synchronous Markovian Arrival Process* (Spatial SynchMAP) is an extension of the classical Synchronous MAP that incorporates spatial dynamics by associating each arrival with a spatial location $\mathbf{x} \in \mathbb{R}^d$. The system operates in a finite state space $\{1, 2, \ldots, K\}$, where $k(t)$ denotes the system's state at time t. Transitions between states are governed by an embedded Markov chain with transition probabilities collected in a stochastic matrix $\mathbf{R} = [r_{vk}]$, where r_{vk} is the probability of moving from state v to state k after an arrival.

While in state k, inter-arrival times are exponentially distributed with rate λ_k. After an arrival in state v at location $\mathbf{x}'$, the process transitions to a new state k with probability r_{kv}, and the new arrival location $\mathbf{x}$ is drawn from a spatial transition kernel $g_{kv}(\mathbf{x} - \mathbf{x}')$, which links the current

and previous arrival locations. This kernel introduces spatial correlation between successive arrivals.

Let $P_k(n, \mathbf{x}, t)$ denote the probability density of being in state k at time t, having experienced n arrivals, with the last arrival occurring at location $\mathbf{x}$. With the convention $P_k(n < 0, \mathbf{x}, t) = 0$, the space–time Kolmogorov system governing the evolution of P_k is given by:

$$\frac{\partial}{\partial t} P_k(n, \mathbf{x}, t) = -\lambda_k \, P_k(n, \mathbf{x}, t) + \sum_{v=1}^{K} \lambda_v \int_{\mathbb{R}^d} r_{kv} \, g_{kv}(\mathbf{x} - \mathbf{x}') \, P_v(n-1, \mathbf{x}', t) \, d\mathbf{x}'. \tag{1}$$

The first term on the right-hand side represents the rate of loss of probability mass from configuration $(k, n, \mathbf{x})$ caused by the occurrence of the next arrival while the system is in state k, while the integral term aggregates arrivals into state k from all other states v, weighted by their respective transition probabilities r_{kv} and spatial kernel $g_{kv}(\mathbf{x} - \mathbf{x}')$. The kernel $g_{kv}(\mathbf{x} - \mathbf{x}')$ determines how the new location $\mathbf{x}$ depends on the previous location $\mathbf{x}'$, thereby introducing spatial dependence into the process.

To simplify the analysis, we introduce the joint count–spatial characteristic function, which simultaneously applies the Fourier transform in both the counting index n and the spatial coordinate $\mathbf{x}$:

$$H_k(u, \mathbf{w}, t) = \sum_{n=0}^{\infty} e^{jun} \int_{\mathbb{R}^d} e^{j\mathbf{w}^\top \mathbf{x}} P_k(n, \mathbf{x}, t) \, d\mathbf{x}, \tag{2}$$

where $u \in \mathbb{R}$ is the Fourier variable associated with the counting process, and $\mathbf{w} \in \mathbb{R}^d$ is the Fourier variable corresponding to the spatial coordinate.

Applying the transform (2) to the Kolmogorov equation (1) yields the time–evolution system for $H_k(u, \mathbf{w}, t)$. Using the convolution identity

$$\int_{\mathbb{R}^d} e^{j\mathbf{w}^\top (\mathbf{x} - \mathbf{x}')} g_{kv}(\mathbf{x} - \mathbf{x}') \, d\mathbf{x} = \tilde{g}_{kv}(\mathbf{w}),$$

where $\tilde{g}_{kv}(\mathbf{w})$ denotes the Fourier transform of the spatial kernel $g_{kv}(\mathbf{x})$, the double integral collapses to a product of transforms, and we obtain

$$\frac{\partial}{\partial t} H_k(u, \mathbf{w}, t) = -\lambda_k \, H_k(u, \mathbf{w}, t) + e^{ju} \sum_{v=1}^{K} \lambda_v \, r_{kv} \, \tilde{g}_{kv}(\mathbf{w}) \, H_v(u, \mathbf{w}, t). \tag{3}$$

Equation (3) shows that the temporal evolution of the joint transform H_k depends linearly on the transforms of all other states H_v, weighted by the

transition probabilities r_{kv}, the exponential arrival rate λ_v, and the Fourier-transformed spatial kernel $\tilde{g}_{kv}(\mathbf{w})$. The exponential factor e^{ju} accounts for the increment in the arrival count after each transition.

To separate the convolution integral, we assume that the spatial transition intensity factorizes through a spatial kernel $g_{kv}(\mathbf{x} - \mathbf{x}')$ describing how the new location $\mathbf{x}$ depends on the previous location $\mathbf{x}'$. Applying the Fourier transform with respect to $\mathbf{x}$ and $\mathbf{x}'$ converts the convolution structure into a simple product of Fourier images:

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{j\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')} g_{kv}(\mathbf{x}-\mathbf{x}')\, d\mathbf{x}\, d\mathbf{x}' = \tilde{g}_{kv}(\mathbf{w}), \tag{4}$$

where $\tilde{g}_{kv}(\mathbf{w})$ is the Fourier transform of the spatial kernel $g_{kv}(\mathbf{x})$. This identity ensures that the convolution over space collapses into a simple multiplicative factor in the transform domain.

Substituting (4) into the evolution equation (3), we can express the system compactly in vector form. Let

$$\mathbf{h}(u, \mathbf{w}, t) = [H_1(u, \mathbf{w}, t), \ldots, H_K(u, \mathbf{w}, t)],$$

be the row vector of joint transforms for all states, and define the diagonal rate matrix

$$\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K).$$

We also introduce the spatially modified transition matrix

$$[\mathbf{R}_g(\mathbf{w})]_{kv} = r_{kv}\, \tilde{g}_{kv}(\mathbf{w}), \tag{5}$$

which incorporates both the state transition probabilities and the spatial Fourier structure.

Then the time evolution of the system can be written as a compact row-vector matrix ordinary differential equation:

$$\frac{\partial}{\partial t}\mathbf{h}(u, \mathbf{w}, t) = \mathbf{h}(u, \mathbf{w}, t)\, \mathbf{\Lambda}\, (e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I}). \tag{6}$$

The exponential factor e^{ju} accounts for the increment in the number of arrivals, while $\mathbf{R}_g(\mathbf{w})$ encodes the spatial dependence of transitions. This form unifies the temporal and spatial evolution of the Spatial SynchMAP into a single linear system in the joint Fourier domain.

Since (6) is linear with constant coefficients (for fixed $u, \mathbf{w}$), its solution is given by the matrix exponential [8]. The initial vector $\mathbf{h}(u, \mathbf{w}, 0)$ carries the process's initial conditions (e.g., the initial state distribution and the initial location distribution).

Since (6) is a linear system with constant coefficients (for fixed u and $\mathbf{w}$), its solution is given by the matrix exponential [8]:

$$\mathbf{h}(u, \mathbf{w}, t) = \mathbf{h}(u, \mathbf{w}, 0)\, \exp\left\{t\, \mathbf{\Lambda}\, (e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I})\right\}. \tag{7}$$

The initial vector $\mathbf{h}(u, \mathbf{w}, 0)$ encodes the starting configuration of the process, including both the initial state distribution and the initial spatial distribution of arrivals. If $\boldsymbol{\pi}^\top$ denotes the initial state probabilities and $\tilde{\boldsymbol{\varphi}}(\mathbf{w})$ the Fourier transforms of the initial spatial densities, then

$$\mathbf{h}(u, \mathbf{w}, 0) = \boldsymbol{\pi}^\top \odot \tilde{\boldsymbol{\varphi}}(\mathbf{w}),$$

where $\odot$ denotes elementwise (Hadamard) multiplication.

Finally, the quantity $\mathbf{h}(u, \mathbf{w}, t)$ defined in (2) represents the vector of joint characteristic functions for the pair $(n(t), \mathbf{x}(t))$ conditioned on each state k. Restating its definition for clarity:

$$\mathbf{h}(u, \mathbf{w}, t) = \left[\sum_{n=0}^{\infty} e^{jun} \int_{\mathbb{R}^d} e^{j\mathbf{w}^\top \mathbf{x}} P_k(n, \mathbf{x}, t)\, d\mathbf{x}\right]_{k=1}^{K}, \tag{8}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the spatial Fourier variable. Thus, $\mathbf{h}(u, \mathbf{w}, t)$ serves as the complete spatio-temporal transform of the Spatial SynchMAP, combining the counting dynamics through u and the spatial correlations through $\mathbf{w}$.

We begin with the vector ODE for the joint characteristic function in (6), which can be rewritten in compact operator form as

$$\frac{\partial \mathbf{h}(u, \mathbf{w}, t)}{\partial t} = \mathbf{h}(u, \mathbf{w}, t)\, \mathbf{A}(u, \mathbf{w}), \tag{9}$$

where the system matrix $\mathbf{A}(u, \mathbf{w})$ collects both the counting dynamics and the spatial structure:

$$\mathbf{A}(u, \mathbf{w}) = \mathbf{\Lambda}\left(e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I}\right), \tag{10}$$

and $\mathbf{R}_g(\mathbf{w})$ is the spatially modified transition matrix defined by $[\mathbf{R}_g(\mathbf{w})]_{kv} = r_{kv}\, \tilde{g}_{kv}(\mathbf{w})$. The initial condition is given by $\mathbf{h}(u, \mathbf{w}, 0) = \mathbf{r}$, where $\mathbf{r}$ encodes the initial joint transform determined by the starting state and location distributions.

If $\mathbf{A}(u, \mathbf{w})$ is diagonalizable, its exact time-domain solution is obtained through the matrix exponential in the eigenbasis of $\mathbf{A}$. Let $\mathbf{X}(u, \mathbf{w})$ denote the eigenvector matrix and $\mathbf{\Gamma}(u, \mathbf{w})$ the diagonal matrix of eigenvalues $\gamma_i(u, \mathbf{w})$. Then the solution of (9) takes the closed form:

$$\mathbf{h}(u, \mathbf{w}, t) = \mathbf{r}\, \mathbf{X}(u, \mathbf{w})\, \mathrm{diag}\left(e^{\gamma_i(u, \mathbf{w})t}\right)\, \mathbf{X}(u, \mathbf{w})^{-1}. \tag{11}$$

This representation follows directly from the eigen–decomposition of the system matrix,

$$\exp(t\,\mathbf{A}(u,\mathbf{w})) = \mathbf{X}(u,\mathbf{w})\,\mathrm{diag}(e^{\gamma_i(u,\mathbf{w})t})\,\mathbf{X}(u,\mathbf{w})^{-1},$$

and is consistent with the initial condition $\mathbf{h}(u,\mathbf{w},0) = \mathbf{r}$.

The spectral matrix $\mathbf{A}(u,\mathbf{w})$ thus governs the spatio–temporal propagation of the joint characteristic function. Its eigenvalues $\gamma_i(u,\mathbf{w})$ describe the exponential decay or growth rates of each spatio–temporal mode, while the eigenvectors in $\mathbf{X}(u,\mathbf{w})$ determine how these modes are mixed across states. For $\mathbf{w} = 0$, the operator $\mathbf{A}(u,0) = \mathbf{\Lambda}(e^{ju}\mathbf{R} - \mathbf{I})$ reduces to the standard synchronous MAP generator, ensuring full consistency with the non-spatial formulation.

The scalar (aggregated) joint characteristic function for the total arrival count and spatial locations is obtained by summing over all states. Let $\mathbf{e} = [1, 1, \ldots, 1]^\top$ be the column vector of ones. Then the total characteristic function is

$$h(u,\mathbf{w},t) = \mathbf{h}(u,\mathbf{w},t)\,\mathbf{e} = \mathbf{r}\,\mathbf{X}(u,\mathbf{w})\,\mathrm{diag}\left(e^{\gamma_i(u,\mathbf{w})t}\right)\,\mathbf{X}(u,\mathbf{w})^{-1}\mathbf{e}, \quad (12)$$

where $\gamma_i(u,\mathbf{w})$ denote the eigenvalues of the system matrix $\mathbf{A}(u,\mathbf{w}) = \mathbf{\Lambda}(e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I})$.

The joint probability density $P(n,\mathbf{x},t)$ of observing n arrivals by time t and the last arrival at location $\mathbf{x}$ can be recovered by inverse Fourier transforms in both variables:

$$P(n,\mathbf{x},t) = \frac{1}{(2\pi)^{d+1}} \int_{-\pi}^{\pi}\int_{\mathbb{R}^d} e^{-jun - j\mathbf{w}^\top\mathbf{x}} h(u,\mathbf{w},t)\,d\mathbf{w}\,du. \quad (13)$$

For the marginal distribution of arrival counts, we set $\mathbf{w} = \mathbf{0}$ (ignoring spatial dependence) and invert only with respect to u:

$$P(n,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jun}\,h(u,\mathbf{0},t)\,du. \quad (14)$$

This expression corresponds to the classical inversion formula for the count distribution in the Synchronous MAP, recovered as a special case of the Spatial SynchMAP when $\mathbf{w} = 0$.

Conversely, the spatial probability density is obtained by summing over all arrival counts and inverting with respect to the spatial Fourier variable:

$$p(\mathbf{x},t) = \sum_{n=0}^{\infty} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-j\mathbf{w}^\top\mathbf{x}} h(0,\mathbf{w},t)\,d\mathbf{w}. \quad (15)$$

Equation (15) describes the spatial density of arrivals aggregated over all event counts. In practice, the factor $\tilde{g}_{kv}(\mathbf{w})$ in $\mathbf{R}_g(\mathbf{w})$ acts as a spectral filter that controls the degree of spatial smoothness of the density $p(\mathbf{x}, t)$. For small spatial frequencies $\mathbf{w}$, the expansion $\tilde{g}_{kv}(\mathbf{w}) \approx 1 - \frac{1}{2}\mathbf{w}^\top \Sigma_{kv}\mathbf{w} + \mathcal{O}(\|\mathbf{w}\|^4)$ illustrates a diffusive effect: the process behaves locally like a Gaussian field with spatial covariance matrices Σ_{kv}.

In the high-intensity regime ($\lambda_k \to \infty$), we introduce a common scaling parameter $\lambda \gg 1$ and write

$$\mathbf{\Lambda} = \lambda\,\tilde{\mathbf{\Lambda}}, \quad \tilde{\mathbf{\Lambda}} = \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_K),$$

where the matrix $\tilde{\mathbf{\Lambda}}$ captures the relative state rates and remains $O(1)$ as $\lambda \to \infty$.

At large λ, the generator in (10) is dominated by the term $\lambda e^{ju}\tilde{\mathbf{\Lambda}}\mathbf{R}_g(\mathbf{w})$, and the smaller contribution $-\lambda\tilde{\mathbf{\Lambda}}$ affects only the first-order correction. Thus, the leading-order dynamics of the joint characteristic function are governed by

$$\frac{\partial}{\partial t}\mathbf{h}(u, \mathbf{w}, t) \simeq \lambda\,\mathbf{h}(u, \mathbf{w}, t)\,\tilde{\mathbf{\Lambda}}\,(e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I}),$$

whose spectral modes satisfy

$$\gamma_i(u, \mathbf{w}) \simeq \lambda\,\tilde{\gamma}_i(u, \mathbf{w}), \qquad h(u, \mathbf{w}, t) \sim \exp(\lambda t\,\tilde{\gamma}(u, \mathbf{w})).$$

In this asymptotic regime, the Spatial SynchMAP approaches a *spatial Poisson process* with state-dependent corrections. The high-frequency arrivals make successive transitions nearly independent, and the aggregated process obeys a diffusion-type Central Limit Theorem (CLT). At the macroscopic scale, both the count process $n(t)$ and the spatial component $\mathbf{x}(t)$ converge in distribution to Gaussian laws:

$$n(t) \sim \mathcal{N}\big(\lambda\,\mu_n t,\ \lambda\,\sigma_n^2 t\big)\,, \qquad \mathbf{x}(t) \sim \mathcal{N}(\lambda\,\boldsymbol{\mu}_x t,\ \lambda\,\mathbf{\Sigma}_x t)\,, \tag{16}$$

where μ_n, σ_n^2, $\boldsymbol{\mu}_x$, and $\mathbf{\Sigma}_x$ are derived from the local expansion of $\tilde{\gamma}(u, \mathbf{w})$ near $(u, \mathbf{w}) = (0, \mathbf{0})$.

To formalize the time and intensity normalization, we rescale the governing equation by a large normalization constant $N = \lambda t$ with

$$\frac{1}{N}\frac{\partial \mathbf{h}(u, \mathbf{w}, t)}{\partial t} = \mathbf{h}(u, \mathbf{w}, t)\,\tilde{\mathbf{\Lambda}}\,(e^{ju}\mathbf{R}_g(\mathbf{w}) - \mathbf{I}). \tag{17}$$

Under this scaling, the effective dynamics of the aggregated scalar transform $h(u, \mathbf{w}, t)$ reduce to the leading-order scalar approximation

$$\frac{1}{N}\,\frac{\partial h(u, \mathbf{w}, t)}{\partial t} \approx (e^{ju} - 1)\,\lambda^*(\mathbf{w})\,h(u, \mathbf{w}, t), \tag{18}$$

where the *effective spatial arrival rate* is given by

$$\lambda^*(\mathbf{w}) = \boldsymbol{\pi}^\top \tilde{\boldsymbol{\Lambda}}\, \mathbf{R}_g(\mathbf{w})\, \mathbf{e},$$

and $\boldsymbol{\pi}$ is the stationary distribution of the embedded chain generated by $\mathbf{R} = \mathbf{R}_g(\mathbf{0})$. The effective rate $\lambda^*(\mathbf{w})$ captures both temporal intensity and spatial coherence through $\mathbf{R}_g(\mathbf{w})$. At $\mathbf{w} = \mathbf{0}$, this rate reduces to the classical mean arrival rate of a standard synchronous MAP $\lambda^*(\mathbf{0}) = \boldsymbol{\pi}^\top \tilde{\boldsymbol{\Lambda}}\mathbf{e}$.

However, the proposed asymptotic and spectral approximation introduces several computational challenges especially in numerically evaluating the matrix exponential for large λ and non-commuting $\mathbf{A}(u, \mathbf{w})$ operators. In future research, we plan to extend this framework to queuing systems driven by spatial synchronous MAP arrivals, and to conduct a detailed asymptotic and numerical analysis of the high-intensity limit, following the methods of [10, 11, 12].

3. Conclusion

This paper investigates a spatial synchronous MAP-flow under high-intensity conditions, with an unlimited number of devices and arbitrary service patterns. The analysis is carried out under the assumption of infinitely increasing intensity of the incoming flow. It is demonstrated that, at sufficiently high flow intensities, the multidimensional probability distribution of the number of requests at the system's synchronous phases in steady state can be accurately approximated by a multidimensional normal distribution. The study provides the expectation vector and covariance matrix for this Gaussian approximation. Additionally, numerical experiments using simulation modeling help define the range of applicability for the derived approximation.

REFERENCES

1. *Chakravarthy S. R.* Markovian Arrival Processes. Hoboken: Wiley, 2010. (Encyclopedia of Operations Research and Management Science).
2. *Chakravarthy S. R. et al.* The Batch Markovian Arrival Process: A Review and Future Work // Advances in Probability Theory and Stochastic Processes. 2001. Vol. 1, No. 1. P. 21–49. Publisher: Notable Publications Inc.: Branchburg, NJ, USA.
3. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The Theory of Queuing Systems with Correlated Flows. Cham: Springer, 2020. 410 p.
4. *Bhat U. Narayan* An Introduction to Queueing Theory: Modeling and Analysis in Applications // Springer, 2008. Vol. 36.
5. *Cooper R. B.* Introduction to Queueing Theory. New York: Elsevier North Holland, 1981. 347 p.

6. *Vasilevskaya T. P., Zavgorodnyaya M. E., Shmyrin I. S.* On the relationship between MAP-event flow models and asynchronous, semi-synchronous, and synchronous doubly stochastic event flows // Vestnik of Tomsk State University. 2004. No. S9–2. P. 138–144.
7. *Gortsev A. M., Nezhelskaya L. A.* On the Connection Between MC-Flows and MAP-Flows of Events // Bulletin of Tomsk State University. Control and Computer Science. 2011. No. 1(14). P. 13–21.
8. *Piskunov N. S.* Differential and Integral Calculus for Technical Universities. Vol. 2. Moscow: Nauka, Main Editorial Office of Physical and Mathematical Literature, 1985. 560 p.
9. *Bhatia R.* Matrix Analysis. New York: Springer, 1997. 360 p. (Graduate Texts in Mathematics. Vol. 169).
10. *Moiseeva S. P., Nazarov A. A.* Metody asimptoticheskogo analiza v teorii massovogo obsluzhivaniya [Methods of Asymptotic Analysis in Queueing Theory]. Tomsk: NTL, 2006.
11. *Nazarov A. A. et al.* Study of Queueing Systems and Random Flows by Methods of Asymptotic Analysis // 2013.
12. *Moiseev A. N., Nazarov A. A.* Beskonechnolineynye sistemy i seti massovogo obsluzhivaniya [Queueing Systems and Networks with an Infinite Number of Servers]. Tomsk: NTL, 2015.

Dinata Syalam Ali Wira — Post-graduate Student, Candidate of Physical and Mathematical Sciences, National Research Tomsk State University, Tomsk, Russian Federation. E-mail: *syalam_ali_wira_dinata@lecturer.itk.ac.id*

THE MULTI-SERVER RETRIAL QUEUEING SYSTEM WITH DIVERSE SERVERS AND BATCH MARKOV ARRIVAL PROCESS

A. Dudin[1], M. Liu[1]

[1] *Belarusian State University, Minsk, Belarus*

A multi-server queueing system with jobs retrials and diverse servers is studied. The primary jobs arrival is described by a batch Markovian arrival process. The service times follow the exponential distribution with the parameter that depends on the server number. The available servers having the minimal number are always assigned for jobs service. If a server completes service of a job, it immediately starts service of another job which is processed by the busy server with the maximal number. When the number of idle servers at a batch arrival moment is insufficient for admission of a whole batch, the number of jobs corresponding to the number of available servers start service. The other job attend the virtual place called orbit to retry to reach the servers in exponentially distributed periods of time. Upon arrival or retrial from the orbit, a job occupies the server having the minimal number among all idle servers, if any. The system behavior is described by a Markov chain in several dimensions having the special block upper-Hessenberg structure of the infinitesimal generator. The explicit expression for the blocks of the generator are presented. Ergodicity condition of the system is given. The expressions for computation of the key performance characteristics are given. **Keywords:** *Batch Markovian arrival process, retrials, heterogeneous servers, stationary distribution, asymptotically quasi-Toeplitz Markov chains*

1. Introduction

Operation of versatile contact centers, communications and communication networks, including mobile cellular networks, and other real-world systems can all be described using retrial queues. Compared to examining queues with job loss or waiting in a buffer when servers are busy, analyzing such queues is much more complicated. This explains why the retrial queues have received significantly less attention. Refer to the surveys in [1, 2] for additional details regarding the retrial waits. Due of the increased size of the state space of the random process characterizing the system behavior, multi-server retrial queues are researched to a far lesser amount than single-server queues. The monograph [1] presents the background results achieved

for the multi-server retrial queues of the $M/M/N$ type. In these queues service times have an exponential distribution and jobs arrive in the stationary Poisson process. However, the $BMAP$ (Batch Markovian Arrival Process), see, for example, [4, 10, 11] is essentially a more appropriate model of arrivals in the modern systems.

The quite general $BMAP/PH/N$ type retrial queue was studied in [6] and [7]. It is suggested there that the servers of the system are identical. Here, we omit this assumption.

The following is a quick outlook of the manuscript's content. Section 2 provides an explanation of the mathematical model being studied. Section 3 introduces the random process that describes the dynamics of the system under consideration. It presents the explicit expression for the chain's generator. In this section, the difficulty of computing the stationary distribution is touched upon and the sufficient condition for the ergodicity of this chain is stated. Formulas for calculating the values of the system's primary performance metrics are provided in Section 4. The conclusion is in Section 5.

2. Mathematical model

We analyse a queueing system having N diverse servers. A job service time in the nth server is exponentially distributed with parameter μ_n, $n = \overline{1, N}$.

The jobs arrive to the system in a $BMAP$. Let $\nu_t,\ t \geq 0$, be the directing process of the $BMAP$. The transition intensities of this process are defined as the entries of the square matrices $D_k,\ k \geq 0$, of size W. The matrix $D = \sum_{k=0}^{\infty} D_k$ is a generator of the process ν_t. The row vector $\boldsymbol{\theta}$, which satisfies equations $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$ is the invariant probability vector of the process ν_t. From this point on, $\mathbf{e}$ represents a column vector of suitable size made up of 1s, and $\mathbf{0}$ represents a row vector of suitable size made up of zeroes. The mean arrival rate of the $BMAP$ is defined as $\lambda = \boldsymbol{\theta} \sum_{k=1}^{\infty} k D_k \mathbf{e}$.

If the arriving batch of jobs does not meet any server being idle, then all arrived jobs transit to the virtual place that is called orbit and try to enter the service individually later on. Capacity of the orbit is unlimited. When the number of jobs on the orbit is equal to i, the rate of jobs retrials is $\alpha_i \ \ i > 0$. We suppose that $\lim_{i \to \infty} \alpha_i = \infty$.

If the arriving batch meets the number of available servers sufficient for service of all jobs from the batch, these jobs enter the service. If the arriving batch meets the number of idle servers insufficient for service of all jobs from the batch, the corresponding part of jobs enters the service while the rest

of a batch moves to the orbit. They will try to enter service later on, as it was described above.

We assume that the jobs in a batch are enumerated in such a way that if the batch consists of n jobs then an arbitrary job receives the number k with the probability $\frac{1}{n}$, for $k = \overline{1,n}$. Jobs arriving in the batch occupy the idle servers according to their numbers. Jobs having minimal number enter the service first.

We assume the following discipline of servers occupation / releasing. Let the number of busy servers be $n,\ n = \overline{0,N-1}$. Then the admitted job will be processed by the $(n+1)$th server. When the number of busy servers is $n,\ n = \overline{1,N}$, and service in one of these servers is completed, the released server immediately starts service of the job that was serviced by the nth server. This server becomes idle.

3. Process describing the dynamics of the system and its analysis

Let i_t be the number of jobs on the orbit, $i_t \geq 0$; n_t be the number of busy servers, $n = \overline{0,N}$; ν_t be the state of the directing process of the arrivals, $\nu_t = \overline{1,W}$, at the moment $t,\ t > 0$.

Let us analyze the process $\xi_t = \{i_t, n_t, \nu_t\},\ t \geq 0$. It is easy to verify that this process is an irreducible continuous-time Markov chain. We arrange the states of Markov chain ξ_t in the direct lexicographic order and We call the set of the states $((i,n,1),(i,n,2),\ldots,(i,n,W))$ of the chain as sub-level (i,n). The level i the set of the sub-levels $(i,0),(i,1),\ldots,(i,N))$. Let the matrix $Q_{i,j}^{(n,n')}$ consists of the rates of transition from the states belonging to sub-level (i,n) to the states belonging to sub-level (j,n') and denote the matrix consisting of the blocks $Q_{i,j}^{(n,n')},\ n,n' = \overline{0,N}$, as $Q_{i,j}$.

Let us denote $\tilde{\mu}_0 = 0,\ \tilde{\mu}_n = \sum\limits_{k=1}^{n} \mu_k,\quad n = \overline{1,N},\ M_n = \tilde{\mu}_n I_W,\ n = \overline{0,N},\ A_i = \alpha_i I_W,\ i \geq 0$, where I_W is the identity matrix of size W.

Lemma 1. The matrix Q has the following structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \cdots \\ O & O & Q_{3,2} & Q_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where the blocks $Q_{i,j}$ of size $W \times (N+1)$ have the following form:

$$Q_{i,i-1} = \begin{pmatrix} O & A_i & O & \dots & O \\ O & O & A_i & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & A_i \\ O & O & O & \dots & O \end{pmatrix},$$

$$Q_{i,i+k} = \begin{pmatrix} O & \dots & O & D_{k+N} \\ O & \dots & O & D_{k+N-1} \\ O & \dots & O & D_{k+N-2} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & D_k \end{pmatrix}, k \geq 1,$$

$$Q_{i,i} = -J_i + \begin{pmatrix} -M_0 + D_0 & D_1 & D_2 & \dots & D_N \\ M_1 & -M_1 + D_0 & D_1 & \dots & D_{N-1} \\ O & M_2 & -M_2 + D_0 & \dots & D_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & -M_N + D_0 \end{pmatrix},$$

where $J_i,\ i \geq 0$, is the block diagonal matrix of block size $N+1$ with all diagonal entries equal to the matrix A_i except the last zero block.

Proof evidently follows from analysis of transitions between the states of the chain ξ_t.

Lemma 2. The Markov chain ξ_t is the member of the class of asymptotically quasi-Toeplitz Markov chains ($AQTMC$), see [8].

Proof is performed via the validation that the generator Q has the form that fits the definition of $AQTMC$ given in [8].

Theorem 1. The Markov chain ξ_t is ergodic if the following inequality holds true:

$$\lambda < \sum_{k=1}^{N} \mu_k.$$

Proof is implemented via numerous algebraic transformation of ergodicity condition given in [8] in matrix form to the scalar form.

Let the ergodicity condition be satisfied. Then there exist the invariant probabilities of the chain ξ_t defined by the limits

$$\pi(i,n,\nu) = \lim_{t\to\infty} P\{i_t = i, n_t = n, \nu_t = \nu\},\ i \geq 0,\ n = \overline{0,N},\ \nu = \overline{1,W}.$$

The invariant probabilities of the states from sub-level (i,n) can be assembled into the row vectors $\boldsymbol{\pi}(i,n),\ i \geq 0,\ n = \overline{0,N}$, and the invariant probabilities of the states from level i can be assembled into the row vectors $\boldsymbol{\pi}_i,\ i \geq 0$.

The vectors $\boldsymbol{\pi}_i,\ i \geq 0$, can be calculated as the solution of equations

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)Q = \mathbf{0},\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)\mathbf{e} = 1.$$

When the matrix Q does not possess quasi-Toeplitz property, the task of this system solution is quite difficult. However, it can be successfully solved using the numerically stable algorithms developed for the $AQTMC$s, see, e.g., [8].

4. The key performance indicators

The mean quantity L_{orbit} of jobs in the orbit is computed by

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

The probability $P_{empty-orbit}$ that the orbit is empty is computed by

$$P_{empty-orbit} = \boldsymbol{\pi}_0\mathbf{e}.$$

The mean quantity N_{busy} of occupied servers is computed as

$$N_{busy} = \sum_{i=0}^{\infty}\sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability $P_{busy}^{(n)}$ that the nth server is occupied at an arbitrary epoch is computed by

$$P_{busy}^{(n)} = \sum_{i=0}^{\infty}\sum_{k=n}^{N} \boldsymbol{\pi}(i,n)\mathbf{e},\ n = \overline{1,N}.$$

The probability P_{imm} that an arbitrary job succeeds to enter service at an arrival moment is computed by

$$P_{imm} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\left[\sum_{k=1}^{N-n} kD_k + \sum_{k=N-n+1}^{\infty}(k-N+n)D_k\right]\mathbf{e}.$$

The rate $P_{interruption}^{(n)}$ of service interruptions in the nth server is

$$P_{interruption}^{(n)} = \sum_{i=0}^{\infty}\sum_{k=1}^{n-1} \mu_k\boldsymbol{\pi}(i,n)\mathbf{e},\ n = \overline{2,N}.$$

5. Conclusion

The generator of the multidimensional Markov chain that describes the behavior of the system being studied is given in explicit form. This enables us to calculate the stationary distribution of the system states and other performance metrics of the system, as well as to show the transparent ergodicity condition.

REFERENCES

1. *Falin G.I., Templeton J.G.C.* Retrial Queues. London: Chapman & Hall. 1997.
2. *Artalejo J.R., Gomez-Corral A.* Retrial Queueing Systems. Springer: Berlin/Heidelberg. 2008.
3. *Lucantoni D.* New results on the single server queue with a batch Markovian arrival process // Communication in Statistics-Stochastic Models. 1991. Vol. 7. P. 1–46.
4. *Chakravarthy S. R.* The batch Markovian arrival process: A review and future work. // Adv. Probab. Theory Stoch. Process. 2001. Vol. 1. P. 21–49.
5. *Dudin A.N., Klimenok V.I., Vishnevsky V.M.*: The theory of queuing systems with correlated flows. Cham: Springer Nature. 2020.
6. *He Q. M., Li H., Zhao Y. Q.* Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times // Queueing Systems. 2000. Vol. 35. P. 323-347.
7. *Breuer L., Dudin A., Klimenok V.* A retrial $BMAP/PH/N$ system // Queueing Systems. 2002. Vol. 40, №. 4. P. 433–457.
8. *Klimenok V. I., Dudin A. N.* Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory // Queueing Systems. 2006. Vol. 54. P. 245–259.

Dudin Alexander — Doctor of Sciences, Professor, Head, laboratory of applied probabilistic analysis. E-mail: *dudin@bsu.by*

Liu Mei — Researcher, laboratory of applied probabilistic analysis. E-mail: *liumei19910101@126.com*

RETRIAL QUEUEING SYSTEM WITH A FINITE NUMBER OF SOURCES AND ADDITIONAL OPPORTUNISTIC FLOW OF CUSTOMERS

O. Dudina[1], S. Dudin[1], A. Dudin[1]

[1] *Belarusian State University, Minsk, Belarus*

The article studies a multi-server queueing system with a finite number of sources of repeated requests. Each source, independently of other sources in the system, generates a request for service. If all servers are busy at the moment of generating a request from a source, the latter makes repeated attempts to enter the service until this request finds a free server. An additional Markov arrival process of requests arrives to the system. These requests are accepted for service opportunistically. The request is accepted only if the number of busy servers is less than some preassigned threshold. Otherwise, the request permanently leaves the system (is rejected). The time of request service by a server has a phase-type distribution. Operation of this system is described by the multi-dimensional Markov chain. The explicit form of the blocks of the generator of this Markov chain as the block structured matrices is obtained.
Keywords: *retrial queue with a finite number of sources, phase-type distribution, Markov arrival process, opportunistic service of additional flow*

1. Introduction

Queueing systems in which an arriving request that meets all servers busy has an opportunity to make repeated attempts (retrials) to receive service adequately describe many important real-world systems. This explains the great interest in their study. For a survey of the state of the art in the investigation of retrial queues, see, e.g., [1–3]. The overwhelming majority of research is done for the systems where the number of requests, which make retrials, is not limited. But there are a lot of real-world systems where the number of requests that can make the retrials is finite. It is assumed that in a system with a finite number of sources of repeated requests, each source can stay in one of three states: the source is free, the source generates repeated requests, or the source is busy. If a source is free, it can generate a primary request. If this request finds the server busy, the corresponding source starts the generation of repeated attempts to get

service. If a request coming from a source finds the server free, it begins to be serviced, and the source goes into the busy state.

The corresponding queueing systems were considered in the literature; see, e.g., in [4–8], mainly in single-server settings, under various assumptions about the availability of the servicing server and various types of distributions describing the processes of servicing and generating initial and repeated requests.

Most of the existing papers have assumed that the request's service times are distributed according to an exponential law, which is a quite restrictive assumption when studying real systems. In this paper, it is assumed that the request service time has a much more general phase-type distribution that can adequately fit the real service time distribution in the system, in particular, accounting not only for the mean service time but also its variance. It is also assumed in our paper that the source may not wait for the end of servicing the request it has generated, but generate the next request.

In addition, the possibility of existence of an additional request flow is considered. Request from this flow is accepted for service only if the number of busy servers upon arrival epoch is less than some preassigned threshold. Service of additional requests allows to increase the throughput of the system and service provider's revenue. The optimal choice of the threshold allows receiving the maximum revenue. In modeling cognitive radio systems, the additional requests correspond to the cognitive requests.

2. Mathematical model

Consider a multi-server system with N identical servers providing service to requests and a finite number R of sources. Each free source, independently of other sources in the system, generates a service request after an exponentially distributed time with intensity α, $\alpha > 0$. If there is a free server at the moment of request generation, then this request starts servicing on any of the idle servers.

The service time of a request has a phase-type distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. Such distribution is defined by a Markov chain m_t, $t \geq 0$, with a state space $\{1, 2, \ldots, M\}$ of transient states and an absorbing state $M+1$. A stochastic row vector $\boldsymbol{\beta}$ of dimension M determines the initial state of the chain at the moment of service beginning. The square matrix S of dimension M is a sub-generator that determines the intensities of transitions of the chain m_t within the set $\{1, 2, \ldots, M\}$. The transition to the absorbing state means the end of service. The intensities of transitions to absorbing states are determined by the column vector $\mathbf{S}_0 = -S\mathbf{e}$. Here

and below, the vector $\mathbf{e}$ has all entries equal to 1. Also, we will use the row vector $\mathbf{0}$ that has all entries equal to 0.

If all servers are busy at the moment the source generates a request, then this source decides to make repeated attempts to service the generated request. The time between repeated attempts to send a request for service has an exponential distribution with the parameter γ, $\gamma > 0$. Attempts will be made until the request is received for service.

In contrast to the usual assumption in the literature that a source can generate the next primary request only when its previous request ends service, we assume that the source can generate the next primary request when its previous request starts service.

An additional flow of requests, which is described as a Markov arrival process (MAP), enters the system. This input flow is determined by the control process v_t, $t \geq 0$, which is an irreducible Markov chain with continuous time. This chain has a finite state space $\{1, 2, \ldots, V\}$ and a generator H of size V, which can be represented as the sum of two matrices, H_0 and H_1. The probabilistic meaning of the matrices H_0 and H_1 and more information about the MAP can be found in [9–11].

If at the moment of arrival of an additional request the number of occupied servers is greater than or equal to K, $K = \overline{0, N}$, then this additional request is rejected; otherwise, it occupies a free server and begins service.

3. Process describing behavior of the system

Let, at the moment t, $t \geq 0$,

i_t, $i_t = \overline{0, R}$, be the number of free sources;

n_t, $n_t = \overline{0, N}$, be the number of busy servers;

v_t, $v_t = \overline{1, V}$, be the state of the control process of an additional MAP flow;

$s_t^{(l)}$, $l = \overline{1, M}$, be the number of servers in which service is provided at the l-th phase, $s_t^{(l)} = \overline{0, n_t}$, $\sum\limits_{l=1}^{M} s_t^{(l)} = n_t$.

It can be verified that the random process $\xi_t = \{i_t, n_t, v_t, s_t^{(1)}, \ldots, s_t^{(M)}\}$, $t \geq 0$, is an irreducible continuous-time Markov chain.

To facilitate work with the multidimensional Markov chain ξ_t, $t \geq 0$, we enumerate its states in the lexicographic order of components n_t, v_t and inverse lexicographic order of components $s_t^{(l)}$, $l = \overline{1, M}$. We call the corresponding set of the states of Markov chain ξ_t having the value i of the component i_t as level i.

Let the matrix $Q_{i,j}$ consists of transition intensities from level i to level j, $i, j \geq 0$, $|i - j| \leq 0$. The following statement is true.

Theorem 1. The generator Q of the Markov chain ξ_t has the block tridiagonal structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & O & O & O \\ O & Q_{2,1} & Q_{2,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Q_{R-1,R-2} & Q_{R-1,R-1} & Q_{R-1,R} \\ O & O & O & \dots & O & Q_{R,R-1} & Q_{R,R} \end{pmatrix}.$$

The diagonal blocks $Q_{i,i}$, $i = \overline{0,R}$, have a block tridiagonal structure:

$$Q_{i,i} = \begin{pmatrix} Q_{i,i}^{0,0} & Q_{i,i}^{0,1} & O & \dots & O & O & O \\ Q_{i,i}^{1,0} & Q_{i,i}^{1,1} & Q_{i,i}^{1,2} & \dots & O & O & O \\ O & Q_{i,i}^{2,1} & Q_{i,i}^{2,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Q_{i,i}^{N-1,N-2} & Q_{i,i}^{N-1,N-1} & Q_{i,i}^{N-1,N} \\ O & O & O & \dots & O & Q_{i,i}^{N,N-1} & Q_{i,i}^{N,N} \end{pmatrix},$$

where the diagonal blocks $Q_{i,i}^{n,n}$, $n = \overline{0,N}$, are determined by

$$Q_{i,i}^{0,0} = \delta_{K=0} H_0 - [i\alpha + (R-i)\gamma] I_V,\ n = \overline{0,N},$$

$$Q_{i,i}^{n,n} = H_0 \otimes I_{T_n} + I_V \otimes (A_n + \Delta_n) -$$

$$-[i\alpha + \delta_{n<N}(R-i)\gamma] I_{VT_n} + \delta_{n\geq K} H_1 \otimes I_{T_n},\ n = \overline{1,N},$$

the overdiagonal blocks $Q_{i,i}^{n,n+1}$, $n = \overline{0,N-1}$, are given by the formula

$$Q_{i,i}^{n,n+1} = i\alpha I_V \otimes P_n(\boldsymbol{\beta}) + \delta_{n<K} H_1 \otimes P_n(\boldsymbol{\beta}),\ n = \overline{0,N-1},$$

and the sub-diagonal blocks $Q_{i,i}^{n,n-1}$, $n = \overline{1,N}$, have the form

$$Q_{i,i}^{n,n-1} = I_V \otimes L_n,\ n = \overline{1,N}.$$

The overdiagonal blocks $Q_{i,i+1}$, $i = \overline{0,R-1}$, have the form

$$Q_{i,i+1} = \begin{pmatrix} O & Q_{i,i+1}^{0,1} & O & \dots & O & O & O \\ O & O & Q_{i,i+1}^{1,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & O & Q_{i,i+1}^{N-2,N-1} & O \\ O & O & O & \dots & O & O & Q_{i,i+1}^{N-1,N} \\ O & O & O & \dots & O & O & O \end{pmatrix},$$

where the overdiagonal blocks $Q_{i,i+1}^{n,n+1}$ are given by

$$Q_{i,i+1}^{n,n+1} = (R-i)\gamma I_V \otimes P_n(\boldsymbol{\beta}),\ n = \overline{0, N-1},\ i = \overline{0, R-1}.$$

The subdiagonal blocks $Q_{i,i-1}$, $i = \overline{1, R}$, have the form

$$Q_{i,i-1} = \begin{pmatrix} O_{V\sum_{n=0}^{N-1} T_n} & O_{V\sum_{n=0}^{N-1} T_n \times VT_N} \\ O_{VT_N \times V\sum_{n=0}^{N-1} T_n} & Q_{i,i-1}^{N,N} \end{pmatrix},$$

where the diagonal block $Q_{i,i-1}^{N,N}$ is defined by the formula

$$Q_{i,i-1}^{N,N} = i\alpha I_{VT_N},\ i = \overline{1, R}.$$

Here

I is the identity matrix, O is the zero matrix;

$\otimes$ is the matrix Kronecker product symbol;

$\delta_a = \begin{cases} 1, & \text{if the condition } a \text{ is satisfied;} \\ 0, & \text{otherwise;} \end{cases}$

$T_n = C_{n+M-1}^{M-1}$, $n = \overline{1, N}$.

Algorithms for finding matrices L_n, $n = \overline{1, N}$, Δ_n, A_n, $n = \overline{0, N}$, and $P_n(\boldsymbol{\beta})$, $n = \overline{0, N-1}$, are given in [12].

4. Conclusion

The explicit form of the generator of the multidimensional Markov chain describing the behavior of the system under study is presented. This allows us to compute the stationary distribution of the system states and various performance characteristics of the system. The outcome of such computation as well as the results of investigation of the impact of the input parameter on the performance characteristics of the system will be presented. The problem of the choice of optimal value of the threshold K will be solved. All these results are planned to be presented in the extended version of this paper.

REFERENCES

1. *Falin G. I., Templeton J. G. C.* Retrial Queues. Chapman & Hall: London, UK, 1997.
2. *Artalejo J. R., Gomez-Corral A.* Retrial Queueing Systems. Springer: Berlin, Heidelberg, Germany, 2008.
3. *Falin G.* A survey of retrial queues // Queueing Systems. 1990. Vol. 7. P. 127–167.

4. *Falin G. I., Artalejo J. R.* A finite source retrial queue // European Journal of Operational Research. 1998. Vol. 108, № 2. P. 409–424.
5. *Zhang F., Wang J.* Performance analysis of the retrial queues with finite number of sources and service interruptions // Journal of the Korean Statistical Society. 2013. Vol. 42. P. 117–131.
6. *Harikrishnan T. et al.* A finite source retrial queueing inventory system with stock dependent arrival and heterogeneous servers // Scientific Reports. 2024. Vol. 14, № 1. P. 30588.
7. *Sztrik J. et al.* The effect of operation time of the server on the performance of finite-source retrial queues with two-way communications to the orbit // Journal of Mathematical Sciences. 2022. Vol. 267, № 2. P. 196.
8. *D'Apice C. et al.* Analysis of a multi-server retrial queue with a varying finite number of sources // AIMS Mathematics. 2024. Vol. 9, № 12. P. 33365-33385.
9. *Chakravarthy S. R.* Introduction to matrix-analytic methods in queues 1: Analytical and simulation approach-queues and simulation. John Wiley & Sons, 2022.
10. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The theory of queuing systems with correlated flows. Cham: Springer, 2020.
11. *Lucantoni D. M.* New results on the single server queue with a batch Markovian arrival process // Communications in Statistics. Stochastic Models. 1991. Vol. 7, № 1. P. 1–46.
12. *Kim C., Dudin A., Dudin S., Dudina O.* Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users // IEEE Access. 2021. Vol. 9. P. 106933–106946.

Dudina Olga — Candidate of Sciences, Leading Researcher, laboratory of applied probabilistic analysis. E-mail: *dudina@bsu.by*

Dudin Sergei — Candidate of Sciences, Leading Researcher, laboratory of applied probabilistic analysis. E-mail: *dudins@bsu.by*

Dudin Alexander — Doctor of Sciences, Professor, Head, laboratory of applied probabilistic analysis. E-mail: *dudin@bsu.by*

PERFORMANCE ANALYSIS OF A FAULT TOLERENCE QUEUEING SYSTEM WITH INTERRUPTION AND SELF HEALING MECHANISM IN RANDOM ENVIRONMENT

Jaya S.[1]

[1] *Maharaja's College, Ernakulam, Kerala, India.*

In this paper, we investigate the role of various environmental factors in causing service interruptions in queueing models and analyze their consequent effects on system performance. Specifically, we consider a single-server queueing system disrupted by a finite number of environmental influences. Some of these interruptions are mild, allowing service to continue at a reduced rate, with the possibility of self-correction, after which the server resumes normal operation. The duration of service under mild interruptions is tracked using an interruption clock, which, upon expiration, leads to server repair. On the epoch of repair the interrupted customer may leave the system without completing service. Depending on the situation, interrupted service may either resume or restart after the completion of repair. Severe interruptions, caused by the remaining environmental factors, necessitate immediate server replacement before service can continue. We examine system stability, determine the steady-state probability vector using the matrix-analytic method, and present numerical results to illustrate key performance measures. **Keywords:** *environmental factor, self correction, protected service, interruption clock, ignored interruption*

Introduction

Interruptions are intrinsic to numerous real-world systems and play a critical role in determining their operational efficiency. In queueing systems, service disruptions can markedly affect performance by altering queue dynamics and compromising overall system functionality. Characterized as temporary pauses in service, these interruptions have been extensively analyzed in queueing theory to elucidate their impact. Foundational work by White and Christie (1958) [7] established the basis for research in this area, leading to extensive studies on queueing models incorporating service interruptions. For a comprehensive overview of such models, we refer readers to the work of A. Krishnamoorthy, P. K. Pramod, and S. R. Chakravarthy (2012) [4]. In certain scenarios, to mitigate customer dissatisfaction, systems

may attempt to maintain service continuity during interruptions whenever feasible. In practice, servers often continue operating during disruptions, albeit at reduced efficiency.

Postponement of interruption refers to delaying the interruption until the current service finishes. Kalidass and Kasturi (2012) [3], and Deepa, Kalidass and Vijayalakshmi(2021) [1] considered a breakdown policy called working break down where customers receive service at a reduced rate when the system experiences partial failure. To ensure stability, it is essential to identify their root causes and address them promptly. In [5] the interruption causing environmental factors and duration of interruption are the deciding factors of whether to repeat, resume the service after interruption, or replace the server. In such cases, the likelihood of server damage and self-correction of the interruption is equal.

Modern automated systems are increasingly designed to facilitate self-correction as exemplified by the recent work by Hafaiedh and Slimane(2022) [2] on designing autonomous systems that can identify and fix problems themselves. These autonomic computing systems, inspired by the human body's self-regulating nervous system, strive for a high degree of self-management. In [6], Jaya(2019) explores the concept of partially ignored interruptions with the possibility of self-correction. Queueing models are fundamentally designed to achieve a balance between uninterrupted service, cost-effectiveness, and high levels of customer satisfaction.

This work focuses on a single-server queueing system characterized by Poisson arrivals and Erlang-distributed service times. The system is prone to service interruptions, and we analyze its performance in response to such disruptions, with the duration of each interruption assumed to follow an exponential distribution.

1. Model description

We consider a single-server queueing system. Customers arrive according to a Poisson process with arrival rate λ. The customers are served in FIFO service order. Service times follow an Erlang distribution with shape parameter k and scale parameter μ. During service, interruptions can occur due to n environmental factors.These factors are ranked from 1 to n based on the severity of the interruption they cause. Interruption stream follow a Poisson process with parameter ϑ. Interruption due to $i^{th}, (1 \leq i \leq n)$ factor occurs with probability p_i. If the interruptions originate from one of the first m factors $(m < n)$, it is considered mild. Then the actual service phase does not change and the service continues at a reduced rate $\mu_i (i = 1, 2, \ldots, m)$. An interruption clock start ticking at the onset of interruption. This clock is exponentially distributed with parameter $\delta_i, (i = 1, 2, \ldots, m)$. During this

interrupted service period, there is a chance for self-correction, which is exponentially distributed with parameter $\tau_i, i = 1, 2, \ldots, m$. If self-correction occurs, the service rate returns to μ. If the customer finishes their service before a pre-defined repair window opens (triggered by a random clock), the system moves on to the next customer(assuming it is still operational). On the realisation of interruption clock the server directly goes for repair. In such situation the customer in service either leaves the system without completing service with probability p or waits for the return of server from repair with probability $1-p$. Interruptions caused by the remaining $n-m$ factors are considered severe and the faulty server is replaced with a new one. The repair time depends on the specific environmental factor that caused the interruption. Each factor has its own associated repair time distribution, denoted by the parameter $\eta_i, i = 1, 2, \ldots, m$. Once the repair is complete, the interrupted service is resumed from the halted phase with probability q or restarted with probabilty $1-q$ depending on the circumstances of interruption. Service protection, ensuring no further interruptions to the remaining service, is provided to the customer whose service was disrupted by the i^{th} factor, where $i = m+1, \ldots, n$, starting from the moment service resumes following repair.

2. Mathematical Description

The behavior of the queueing system described above can be analyzed using a Markov chain. Let $\mathbf{X} = \{X(t), t \geq 0\} = \{(N(t), S(t), I_1(t), I_2(t)), t \geq 0\}$ where $N(t)$ is the number of customers in the system, $S(t)$ is the status of the server, $I_1(t)$ is the environmental factor causing interruption and $I_2(t)$ is the phase of service:

$$S(t) = \begin{cases} 0, & \text{servier is idle ;} \\ 1, & \text{service without any interruption ;} \\ 2, & \text{if interrupted service going on;} \\ 3, & \text{if server under repair;} \\ 4, & \text{if protected service is going on.} \end{cases}$$

The state space of the process is
$\{(0,0) \cup (0,3,j) \cup (r,1,i) \cup (r,2,j,i) \cup (r,3,j,i) \cup (r,4,l,i); r = 1, \ldots, \infty; i = 1, \ldots, k; j = 1, \ldots, m; l = m+1, \ldots, n\}$.
The infinitesimal generator matrix of the process is given by

$$Q = \begin{bmatrix} B_0 & B_1 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots \end{bmatrix}$$

3. Analysis of service process

The service time follows PH distribution with representation $(\boldsymbol{\alpha}, S)$ where $\boldsymbol{\alpha} = (1, 0, \ldots, 0)_{1\times(m+n+1)k}$ and $S = \begin{bmatrix} D_0' & D_1 & 0 & D_2 \\ D_3 & D_4' & D_5 & 0 \\ D_6 & 0 & D_7' & 0 \\ D_8 & 0 & 0 & D_9' \end{bmatrix}$.

$D_0' = D_0 + \lambda I$, $D_4' = D_4 + \lambda I$, $D_7' = D_7 + \lambda I$ and $D_9' = D_9 + \lambda I$. The absorbing state is represented by

$$S^0(i,j) = \begin{cases} \mu, & \text{for } i = k, \text{and } i = (m+n+1)k + m - n + h, \\ & \text{h} = 1, \ldots, (n-m) \text{and } j = 1; \\ \mu_r, & \text{for } i = ((k-1)m + k + r, j = 1, r = 1, \ldots, m; \\ (1-p)\delta_l & \text{for } i = k + mr + l, j = 1, l = 1, \ldots, m, r = 0, \ldots, (k-1); \\ 0; & \text{otherwise.} \end{cases}$$

which is a column matrix.

- The response time of the service process, $E(S) = -\boldsymbol{\alpha} S^{-1} e$.
- Hence the expected service rate $\mu_s = \frac{1}{E(S)}$.

Theorem: *The queueing system is stable when* $\lambda < \mu_s$.

4. Stationary Distribution

The stationary distribution, under the condition of stability, $\lambda < \mu_s$ of the model, has Matrix Geometric solution. Let χ=$(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ...)$ be the steady state probability vector of the Markov chain $\{V(t), t \geq 0\}$. Each $\mathbf{x}_i, i > 0$ are vectors with $(1+m+n)K$ elements. We assume that $\mathbf{x}_2 = \mathbf{x}_1.R$, and $\mathbf{x}_i = \mathbf{x}_1.R^{i-1}, i \geq 2$, where R is the minimal non- negative solution to the matrix quadratic equation

$R^2 A_2 + RA_1 + A_0$=0.

From χQ=0 we get

$$\mathbf{x}_0 B_0 + \mathbf{x}_1 B_2 = 0,$$

$$\mathbf{x}_0 B_1 + \mathbf{x}_1 (A_1 + RA_2) = 0.$$

Solving the above two equations we get $\mathbf{x}_0$ and $\mathbf{x}_1$ subject to the normalizing condition $\mathbf{x}_0 e + \mathbf{x}_1 (I - R)^{-1} e$=1. The stationary distribution, under the condition of stability, $\lambda < \mu_s$ of the model, has Matrix Geometric solution.

5. Important performance measures

- Probability that the system is idle, $P(I) = \mathbf{x}_{00}$.
- Fraction of time the system is under repair $P(R) = \sum_{i=1}^{\infty} \sum_{l=1}^{m} \sum_{j=1}^{k} \mathbf{x}_{i3lj}$.

– Expected number of customers in the system, $E(C) = \sum_{i=1}^{\infty} i\mathbf{x}_i\boldsymbol{e}$.

– Effective interruption rate, $E_{int} = \vartheta(\sum_{i=1}^{\infty}\sum_{l=1}^{m}\sum_{j=1}^{k}\mathbf{x}_{i2lj} + \sum_{i=1}^{\infty}\sum_{l=m+1}^{n}\sum_{j=1}^{k}\mathbf{x}_{i4lj})$.

6. Numerical Illustrations

In this section assuming arbitrary values for the parameters, subject to stability, we obtained the numerical values for important performance measures. Let $n = 5, m = 3, \eta_1 = 4, \eta_2 = 3, \eta_3 = 2; \tau_1 = 1, \tau_2 = 2;\ \delta_1 = 2,$ $\delta_2 = 1, \delta_3 = 0.5; p_1 = p_2 = p_3 = p_4 = p_5 = 0.2$.
The conclusion drawn are purely based on the values of input parameters.

Table 1

Effect of μ on various performance measures

μ	$E(S)$	$E(C)$	$P(I)$	$P(R)$	E_{int}
8	2.8208	6.9016	0.0618	0.8279	0.8726
9	2.9444	4.8883	0.0906	0.5891	0.6120
10	3.0704	3.9329	0.1160	0.4717	0.4852
11	3.1985	3.3722	0.1385	0.4001	0.4087
12	3.3281	3.0017	0.1586	0.3509	0.3566
13	3.4590	2.7375	0.1767	0.3145	0.3183
14	3.5910	2.5390	0.1929	0.2861	0.2888
15	3.7239	2.3839	0.2077	0.2632	0.2651

7. Conclusion

In this study, we have explored the impact of environmental factors on service interruptions in a single-server queueing system. By distinguishing between mild and severe types of interruptions and modeling their respective effects on service continuity, recovery, and customer behavior, we provide a comprehensive framework that reflects more realistic operational scenarios. The incorporation of an interruption clock and the possibility of self-correction add depth to the modeling of mild interruptions, while the necessity of server replacement in severe cases highlights the critical role of system resilience. Using the matrix-analytic method, we derived the steady-state probability vector and evaluated performance measures This work lays

the foundation for future studies on sustainable and resilient queueing systems under uncertainty and dynamic environments.

REFERENCES

1. *Deepa B., Kalidass K., Vijayalakshmi V.* An M/M/1/N Queue with Working Breakdowns and Vacations-An Approach to Cost Analysis // Journal of Physics: Conference Series. 2021. Vol. 1850, № 1.
2. *Hafaiedh I. B., Slimane M. B.* A distributed formal-based model for self-healing behaviors in autonomous systems: from failure detection to self-recovery. // Journal of Supercomputing. 2022. Vol. 678. P. 18725–18753.
3. *Kalidass K., Kasturi R.* A queue with working breakdowns. // Computers and Industrial Engineering.2012.Vol. 63,№. 4.P. 779–783.
4. *Krishnamoorthy A., Pramod P. K., Chakravarthy S. R.* Queues with Interruptions: A Survey. // TOP. 2014. Vol. 22. P. 290-320.
5. *Krishnamoorthy A., Jaya S.* Queue with Partially Ignored Interruption in Markovian Environment. // Vishnevskiy V., Samouylov K., Kozyrev D. (eds) Distributed Computer and Communication Networks.DCCN., Lecture Notes in Computer Science. 2019. Vol. 11965. P. 289–301.
6. *Jaya S.* Queue with ignored interruption and self correction // Bulletin of Kerala Mathematics Association.2019. Vol. 16, №. 1. P. 55-64.
7. *White H., Christie L. S.*Queueing with preemptive priorities or with breakdown //Operations Research. 1958. Vol. 6, № 1. P. 79–95.

Dr. Jaya S. — Ph.D., Associate Professor and Head of Department of Mathematics, Maharaja's College, Ernakulam, Kerala, India. E-mail: *sreelakamjaya2024@gmail.com*

MODELING THE CPU PERFORMANCE WITH NON-POISSON WORKLOAD

I. Lapatin[1], A. Nazarov[1], S. Paul[1], O. Lizyura[1], A. Ivanova[1]

[1] *Tomsk State University, Tomsk, Russian Federation*

The paper proposes to study a mathematical model of a computing node in the form of an infinite queueing system with an input Markov modulated Poisson process. This correlated flow can model complex stochastic processes in a node, in which the rate of occurrence of arrivals depends on the state of the underlying process, which makes it useful taking into account the correlation between the moments of occurrence of tasks on the node. The service rate for each request depends on the number of requests in the system. The purpose of the study is to apply the asymptotic analysis method and obtain the Gaussian approximation of the probability distribution of the number of requests in the system. **Keywords:** *infinite-server queueing system, Markov modulated Poisson process, asymptotic analysis method*

Introduction

The application of the queuing theory apparatus in modeling computing systems allows us to predict and calculate their performance metrics, such as average latency or the probability that the service level agreement (SLA) will exceed, the number of requests waiting to be serviced, the average number of servers in use, average waiting time, server load, throughput, buffer overflow probability, waiting time distribution, and server activity period.

On the one hand, when several tasks arrive at a computing node, the data transfer rate may change depending on external factors and will switch between different levels. On the other hand, the deterioration of service quality in computing nodes is associated with the problem of parallelism (contention for resources). When tasks are executed on a node, this can slow down its operation, even if the node has enough resources to service them. Modeling the service rate in such systems should take into account the dependence on the number of tasks executed in the node [1–3].

Queueing models with an infinite number of devices, correlated incoming flows and the rate of task execution depending on the total number of tasks on the node are suitable for modeling such computing systems, since they can take into account the variable nature of the system parameters. To

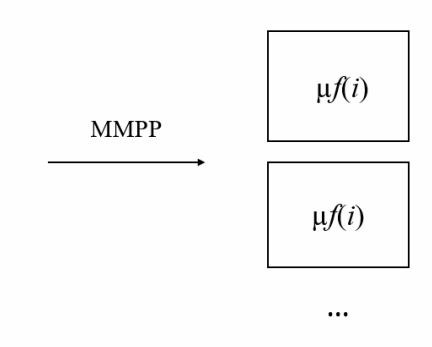


Figure 1. Mathematical Model

model the processor operation, we will use a queuing system with an unlimited number of devices. The incoming Markov modulated Poisson flow [4, 5] describes the flow of tasks that arrive at the processor. Each busy device will be identified with the task being executed. The more devices are busy, the more tasks are simultaneously executed on the node. To model the decrease in service rate, we will use a degradation function that changes the service intensity depending on the number of requests in the system. Thus, the randomness of the request execution time is generated by the different volume of incoming tasks and the variable speed of their execution. The study will be carried out using the method of asymptotic analysis.

1. Mathematical Model

Let us consider an infinite queueing model (Fig. 1) The input to the system is a Markovian modulated Poisson process of requests (MMPP) specified by the generator $\mathbf{Q}_{dop} = [q_{\nu m}]$ of infinitesimal characteristics of the Markov chain $m(t) = 0, 1, ..., M$, diagonal matrix $\mathbf{\Lambda}_{dop}$ conditional rates λ_m occurrence of arrivals in the m-th state of the MMPP, when process $m(t)$ takes the value $m = 0, ..., M$. The process $m(t)$ is called the Markov control chain for the MMPP considered. The states of the control process $m(t)$ will also be called MMPP states. Each request of the incoming flow immediately begins servicing. The servicing time is distributed with the parameter depends on the number of requests in the system. The service intensity can be written as $\mu f(i)$, where μ is the service rate when the only one request in the system (without competition), and $f(i)$ is a degradation function whose values are a dimensionless coefficient of service rate reduction depending on the number of requests i in the system. The more requests are simultaneously in the system, the lower the intensity of their service.

Let us denote: $i(t)$ is the number of requests in the system at time t. For the probability distribution

$$P(m,i,t) = P\{m(t) = m, i(t) = i\} \tag{1}$$

of the two-dimensional Markov process $\{m(t), i(t)\}$ we will compose a Kolmogorov system in stationary mode. Let is denote vector

$$\mathbf{P}(i) = \{P(1,i), P(2,i), ..., P(M,i)\}$$

and identity matrix $\mathbf{I}$. We have

$$\mathbf{P}(i)\left(\mathbf{Q}_{dop} - \mathbf{\Lambda}_{dop} - i\mu f(i)\mathbf{I}\right) + \mathbf{P}(i-1)\mathbf{\Lambda}_{dop} + \\ + \mathbf{P}(i+1)(i+1)\mu f(i+1) = 0. \tag{2}$$

We obtain the solution to system (2) using the asymptotic analysis method under the limit condition of a high intensity incoming flow, which is provided by the parameter T and the conditions $\mathbf{\Lambda}_{dop} = \mathbf{\Lambda}T$, $\mathbf{Q}_{dop} = \mathbf{Q}T$, $f(i) = \overline{f}\left(\frac{i}{T}\right)$:

$$\mathbf{P}(i)\left(\mathbf{Q} - \Lambda - \frac{i}{T}\mu\overline{f}\left(\frac{i}{T}\right)\mathbf{I}\right) + \\ + \mathbf{P}(i-1)\Lambda + \mathbf{P}(i+1)\frac{i+1}{T}\mu\overline{f}\left(\frac{i+1}{T}\right) = 0. \tag{3}$$

Let us write down the consistent equation, which we obtain by multiplying equation (3) by the unit column vector $\mathbf{e}$:

$$-\mathbf{P}(i)\left(\Lambda + \frac{i}{T}\mu\overline{f}\left(\frac{i}{T}\right)\mathbf{I}\right)\mathbf{e} + \\ + \mathbf{P}(i-1)\Lambda\mathbf{e} + \mathbf{P}(i+1)\frac{i+1}{T}\mu\overline{f}\left(\frac{i+1}{T}\right)\mathbf{e} = 0. \tag{4}$$

2. Asymptotic Analysis Method

We will study the system (3)–(6) under the limit condition of a high-intensity incoming flow $T \to \infty$. By the method of asymptotic analysis, we obtain the limiting Gaussian probability distribution density $i(t)$ of the number of servers in the system under study, which is determined by the equality

$$p(x) = \frac{1}{\sqrt{2\pi\varkappa_2 T}} e^{-\frac{(x-\varkappa_1 T)^2}{2\varkappa_2 T}}, \tag{5}$$

where $\varkappa_1$ is the solution to the equation

$$\mathbf{r\Lambda e} - \varkappa_1 \overline{f}(\varkappa_1) = 0. \tag{6}$$

$\varkappa_2$ is defined by equality

$$\varkappa_2 = \frac{\mathbf{r\Lambda e} - \boldsymbol{\varphi\Lambda e}}{a}, \tag{7}$$

$$a = \mu \left. \frac{d\left(x\overline{f}(x)\right)}{dx} \right|_{x=\varkappa_1},$$

the vector $\boldsymbol{\varphi}$ is the solution of the following system

$$\boldsymbol{\varphi}\mathbf{Q} = \mathbf{r}\left(\mathbf{\Lambda} - \varkappa_1 \mu \overline{f}(\varkappa_1)\mathbf{I}\right),$$

$$\boldsymbol{\varphi}\mathbf{e} = 0.$$

To obtain a discrete probability distribution of the process $i(t)$, we will use the following formula

$$p_{discret}(i) = \frac{p(i)}{\sum\limits_{i=0}^{\infty} p(i)}, \quad i \geq 0. \tag{8}$$

The Gaussian probability distribution $p(i)$ is determined by the average number $\varkappa_1 T$ and the variance $\varkappa_2 T$ of service requests in the system.

3. Numerical Example

Let us consider the following values of the model parameters

$$\mathbf{\Lambda} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} -3 & 2.4 & 0.6 \\ 0.1 & -0.2 & 0.1 \\ 0.3 & 0.7 & -1 \end{bmatrix}.$$

The average number of tasks sent to the processor per unit of time is $\mathbf{r\Lambda e} = 1$, i.e. the intensity of task arrival; $\mu = 2$ is the value inverse to the average time of request execution in the absence of other tasks on the processor; the parameter of high intensity of the incoming flow $T = 5$; we define the service rate degradation function as a non-negative decreasing function

$$f(i) = \frac{1}{\sqrt{1+i}}.$$

Fig. 2 shows a Gaussian approximation of the probability distribution of the number of service requests in the system, taking into account the degradation function.

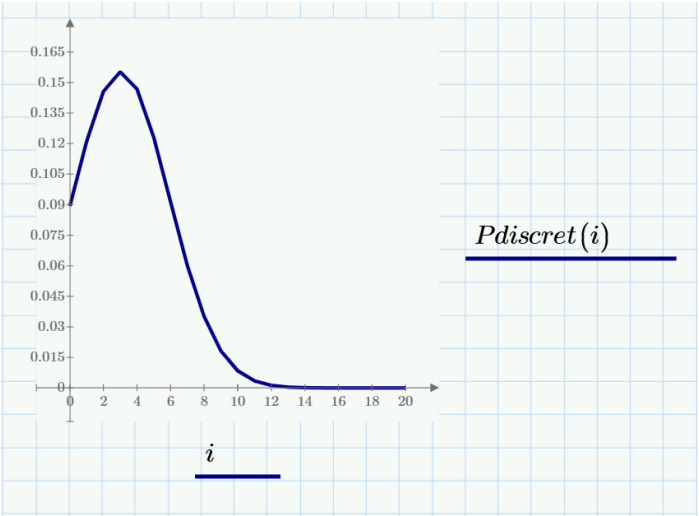


Figure 2. Gaussian approximation of the probability distribution of the number of service requests in the system

4. Conclusion

The proposed paper considers a mathematical model of a computing node in the form of an infinite queueing system taking into account the dependence of the rate of servicing requests on the number of busy servers, at the input of which a Markov modulated Poisson process of arrives for servicing, with which it is possible to take into account the correlation of intervals between the moments of receipt of requests in the system.

REFERENCES

1. *Liu X., Li S., Tong W.* A queuing model considering resources sharing for cloud service performance // The Journal of Supercomputing. 2015. Vol. 71, № 11. P. 4042—4055.
2. *Choudhary A., Chakravarthy S. R., Sharma D. C.* Analysis of MAP/PH/1 Queueing System with Degrading Service Rate and Phase Type Vacation // Mathematics. 2021. Vol. 9, №. 19. P. 2387.
3. *Morozov E.* A general multiserver state-dependent queueing system // Centre de Recerca Matematica. 2010. Vol. 927. 17 p.
4. *Nazarov A., Paul S., Phung-Duc T., Morozova M.* Analysis of Tandem Retrial with Common Orbit and MMPP In coming Flow // Lecture Notes in Computer Science, 2022. Vol. 13766. P. 270–283.

5. *Nazarov A., Paul S., Phung-Duc T., Lizyura O.* Two-Way Communication Retrial Queue with Markov Modulated Poisson Input and Multiple Types of Outgoing Calls // Communications in Computer and Information Science, 2022. Vol. 1748. P. 370–381.

Lapatin Ivan — candidate of physics and mathematics, associate professor of applied informatics department. E-mail: *ilapatin@mail.ru*

Nazarov Anatoly — doctor of technical sciences, professor, full professor of probability theory and statistics department. E-mail: *nazarov.tsu@gmail.com*

Paul Svetlana — doctor of physics and mathematics, associate professor, full professor of probability theory and statistics department. E-mail: *paulsv82@mail.ru*

Lizyura Olga — seniour lecturer of probability theory and statistics department. E-mail: *oliztsu@mail.ru*

Ivanova Alena — bachelor student of the third year, institute of applied mathematics and computer science.

A PRODUCTION INVENTORY SYSTEM WITH DETERIORATING ITEMS AND MAP ARRIVALS: PH-TYPE SERVICE, AND CUSTOMER RETRIALS

A. Melikov[1], K. P. Jose[2], Bobina J. Mattam[2]

[1] *Dept. of Mathematics, Baku Engineering University, Baku, Azerbaijan*
[2] *P. G. & Research Dept. of Mathematics, St. Peter's College, Kerala, India*

Abstract This article deals with a production inventory system with perishable items where the customers arrive according to the Markovian Arrival Process. The time interval between the production of each item is exponentially distributed. It is assumed that the service time follows Phase-type distribution. The server is considered unreliable and upon arrival, a customer receives service if the server is available and the inventory is non-empty. If the server is busy or experiencing a breakdown, the arriving customer either joins an infinite waiting space. The Matrix Analytic Method is used to solve the system. A profit function is constructed based on the system parameters to analyze the efficiency of the system model.
Keywords: *Perishable Inventory, Markovian Arrival Process, Unreliable Server, Retrial, Production*

Introduction

Services may be frequently halted for several reasons, including wear and tear, exhaustion, and other innate problems. The impact of such failures has to be studied to anticipate the overall compatibility of the system. This paper considers a MAP/PH perishable inventory system with an unreliable server and a production unit. A real life example of such an inventory model is a pharmaceutical manufacturing and distribution center.

A single server queueing approach with consumers arriving with MAP complying with the FCFS regulation was investigated by Chakravarthy [2]. Krishnamoorthy et al. [4] study looked at a single server queue that included multiple services. Melikov and Shahmaliyev [5] constructed a queueing system model with a single server, perishable inventory, and repeated customers forming an infinite orbit. Salini and Jose [7]incorporated varying service rates to a production inventory model. Reshmi and Jose [3] modelled production inventory with detriorating items. Bobina and Jose [1] investigated a system with deteriorating items under the MAP/PH frame, where the server is unreliable and the system is equipped with a production unit.

1. Mathematical Model

Consider a single server production inventory system of perishable items governed by a continuous review (s, S) policy. The following are the assumptions of the model:

- production is turned On at inventory level s. Production time follows an exponential distribution with rate β. Production stops at maximum capacity S.
- Arrivals follow the Markovian arrival process having representation $MAP(D^0, D^1)$ with m phases;
- service time is phase type distributed with representation $PH(\eta, T)$ having n phases;
- a customer who arrives when the server is busy or when inventory is zero may wait in the orbit with probability γ or leave the system with complementary probability;
- retrying customer, on finding zero inventory or busy server waits in the orbit with probability δ or exits the system with probability $1 - \delta$;
- the lifetime of an item and inter-retrial times of consecutive retrials are exponentially distributed with linear rates $j\omega$ and $i\theta$, respectively;
- the server may breakdown while in service and it follows Poisson process with parameter δ_1. The repair times of the server are exponentially distributed with parameter δ_2.

We use the notations, $N(t)$ for the number of customers in the orbit at time t. $C(t)$ takes the values 0, 1 or 2 whenever the server is idle, busy or breakdown respectively. The production status, $P(t)$ takes values 0 or 1 if the production is off or on respectively. $I(t)$, represents inventory level at time t. $F_1(t)$ and $F_2(t)$ represents phases of the arrival and service at time t respectively.

Then $\{X(t); t \geq 0\}$ is a level dependent quasi-birth-death process on the state space E, where $X(t) = (N(t), C(t), P(t), I(t), F_1(t), F_2(t))$ and $E = H_1 \cup H_2 \cup H_3 \cup H_4 \cup H_5$ with

$$H_1 = \underline{(i,0,0,j)} = \{(i,0,0,j,k) : s+1 \leq j \leq S, i \geq 0, 1 \leq k \leq m\},$$
$$H_2 = \underline{(i,0,1,j)} = \{(i,0,1,j,k) : i \geq 0, 0 \leq j \leq S-1, i \geq 0, 1 \leq k \leq m\},$$
$$H_3 = \underline{(i,1,0,j)} = \{(i,1,0,j,k,l) : s+1 \leq j \leq S, i \geq 0, 1 \leq k \leq m, 1 \leq l \leq n\},$$
$$H_4 = \underline{(i,1,1,j)} = \{(i,1,1,j,k,l) : 1 \leq j \leq S-1, i \geq 0, 1 \leq k \leq m, 1 \leq l \leq n\},$$
$$H_5 = \underline{(i,2,p,j)} = \{(i,2,p,j,k) : p = 0,1; 1 \leq j \leq S-1, i \geq 0, 1 \leq k \leq m\}.$$

The infinitesimal generator of the process is

$$\mathcal{Q} = \begin{bmatrix} A_{10} & A_0 & & & \\ A_{21} & A_{11} & A_0 & & \\ & A_{22} & A_{12} & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

Order of $A_0, A_{1i}(i \geq 0)$ and $A_{2i}(i \geq 1)$ are $(2S-s)m + (2S-s-1)mn$. The transitions of A_0, A_{1i} and A_{2i} are given in Table 1.

2. System Stability

For obtaining the system stability, we apply the Neuts-Rao truncation, [6] by assuming $A_{1i} = A_{1N}$ and $A_{2i} = A_{2N}$ for all $i \geq N$, where N is sufficiently large. In this scenario, the change in the steady state probability vector is negligible if the number of customers in the orbit is limited to a suitably chosen N.

The system under consideration is stable if and only if,

$$\lim_{N\to\infty} \left(\frac{\pi_N A_0 e}{\pi_N A_{2N} e}\right) < 1.$$

3. System Performance Measures

Let $\mathbf{x} = (x_0, x_1, x_2, ...)$ be the stationary probability vector of $\mathcal{Q}$, where each $x_i(i \geq 0)$ has $(4S-2s-1)m + (2S-s-1)mn$ elements where

$$\begin{aligned} x_i = (&z_{i,0,0,s+1,1}, \ldots, z_{i,0,0,S,m}, z_{i,0,1,0,1}, \ldots, z_{i,0,1,S-1,m}, \\ &z_{i,1,0,s+1,1}, \ldots, z_{i,1,0,S,m}, z_{i,1,1,1,1}, \ldots, z_{i,1,1,S-1,m}, \\ &z_{i,2,0,s+1,1}, \ldots, z_{i,2,0,S,m}, z_{i,2,1,1,1}, \ldots, z_{i,2,1,S-1,m}); (i \geq 0). \end{aligned}$$

in which

$$\begin{aligned} z_{i,1,0,j,k} &= (z_{i,1,0,j,k,1}, ..., z_{i,1,0,j,k,n}), s+1 \leq j \leq S, 1 \leq k \leq m, \\ z_{i,1,1,j,k} &= (z_{i,1,1,j,k,1}, ..., z_{i,1,1,j,k,n}), 0 \leq j \leq S-1, 1 \leq k \leq m. \end{aligned}$$

The performance measures of the system in terms of stationary probability vectors are:

a) expected Inventory level

$$\begin{aligned} EI = &\sum_{i=0}^{\infty} \sum_{j=s+1}^{S} \sum_{k=1}^{m} j \left(z_{i,0,0,j,k} + \sum_{l=1}^{n} z_{i,1,0,j,k,l} + z_{i,2,0,j,k} \right) + \\ &+ \sum_{i=0}^{\infty} \sum_{j=1}^{S-1} \sum_{k=1}^{m} j \left(z_{i,0,1,j,k} + \sum_{l=1}^{n} z_{i,1,1,j,k,l} + z_{i,2,1,j,k} \right); \end{aligned}$$

Table 1

Transitions

From	To	Rate	Description
		Transitions of A_0	
$(i,0,1,0)$	$(i+1,0,1,0)$	γD^1	
$(i,1,0,j)$	$(i+1,1,0,j)$	$\gamma D^1 \otimes I_n$	$s+1 \le j \le S$
$(i,1,1,j)$	$(i+1,1,1,j)$	$\gamma D^1 \otimes I_n$	$1 \le j \le S-1$
$(i,2,0,j)$	$(i+1,2,0,j)$	γD^1	$s+1 \le j \le S$
$(i,2,1,j)$	$(i+1,2,1,j)$	γD^1	$1 \le j \le S-1$
		Transitions of A_{2i}	
$(i,0,0,j)$	$(i-1,1,0,j)$	$i\theta I_m \otimes \eta$	$s+1 \le j \le S$
$(i,0,1,0)$	$(i-1,0,1,0)$	$i\theta(1-\delta)I_m$	
$(i,0,1,j)$	$(i-1,1,1,j)$	$i\theta I_m \otimes \eta$	$1 \le j \le S$
$(i,1,0,j)$	$(i-1,1,0,j)$	$i\theta(1-\delta)I_{mn}$	$s+1 \le j \le S$
$(i,1,1,j)$	$(i-1,1,1,j)$	$i\theta(1-\delta)I_{mn}$	$1 \le j \le S-1$
$(i,2,0,j)$	$(i-1,2,0,j)$	$i\theta(1-\delta)I_{mn}$	$s+1 \le j \le S$
$(i,2,1,j)$	$(i-1,2,1,j)$	$i\theta(1-\delta)I_{mn}$	$1 \le j \le S-1$
		Transitions of A_{1i}	
$(i,0,0,j)$	$(i,0,0,j)$	$D^0 - i\theta I_m - j\omega I_m$	$s+1 \le j \le S$
$(i,0,0,s+1)$	$(i,0,1,s)$	$(s+1)\omega I_m$	
$(i,0,0,j)$	$(i,0,0,j-1)$	$j\omega I_m$	$s+2 \le j \le S$
$(i,0,0,j)$	$(i,1,0,j)$	$D^1 \otimes \eta$	$s+1 \le j \le S$
$(i,0,1,S-1)$	$(i,0,0,S)$	βI_m	
$(i,0,1,j)$	$(i,0,1,j)$	$D^0 + (1-\gamma)D^1 - i\theta(1-\delta)I_m - \beta I_m$	$j=0$
$(i,0,1,j)$	$(i,0,1,j)$	$D^0 - i\theta I_m - \beta I_m - j\omega I_m$	$1 \le j \le S-1$
$(i,0,1,j)$	$(i,0,1,j+1)$	βI_m	$0 \le j \le S-2$
$(i,0,1,j)$	$(i,0,1,j-1)$	$j\omega I_m$	$1 \le j \le S-1$
$(i,0,1,j)$	$(i,1,1,j)$	$D^1 \otimes \eta$	$1 \le j \le S-1$
$(i,1,0,j)$	$(i,0,0,j-1)$	$I_m \otimes T^0$	$s+2 \le j \le S$
$(i,1,0,s+1)$	$(i,0,1,s)$	$I_m \otimes T^0$	
$(i,1,0,j)$	$(i,1,0,j)$	$(D^0 + (1-\gamma)D^1) \oplus T - i\theta(1-\delta)I_{mn} - \delta_1 I_m \otimes c_n$	$s+1 \le j \le S$
$(i,1,0,j)$	$(i,1,0,j-1)$	$j\omega I_{mn}$	$s+2 \le j \le S$
$(i,1,0,s+1)$	$(i,1,1,s)$	$(s+1)\omega I_{mn}$	
$(i,1,0,j)$	$(i,2,0,j)$	$\delta_1 I_m \otimes c_n$	$s+2 \le j \le S$
$(i,1,1,j)$	$(i,0,1,j-1)$	$I_m \otimes T^0$	$1 \le j \le S-1$
$(i,1,1,S-1)$	$(i,1,0,S)$	βI_{mn}	
$(i,1,1,j)$	$(i,1,1,j)$	$(D^0 + (1-\gamma)D^1) \oplus T - i\theta(1-\delta)I_{mn} - \delta_1 I_m \otimes c_n - \beta I_{mn}$	$1 \le j \le S-1$
$(i,1,1,j)$	$(i,1,1,j-1)$	$j\omega I_{mn}$	$2 \le j \le S-1$
$(i,1,1,j)$	$(i,1,1,j+1)$	βI_{mn}	$1 \le j \le S-2$
$(i,1,1,j)$	$(i,2,1,j)$	$\delta_1 I_m \otimes c_n$	$1 \le j \le S-1$
$(i,2,0,j)$	$(i,1,0,j)$	$\delta_2 I_m \otimes \eta$	$s+1 \le j \le S$
$(i,2,0,j)$	$(i,2,0,j)$	$D^0 + (1-\gamma)D^1 - i\theta(1-\delta)I_m - \delta_2 I_m - j\omega I_m$	$s+1 \le j \le S$
$(i,2,0,j)$	$(i,2,0,j-1)$	$j\omega I_m$	$s+2 \le j \le S$
$(i,2,0,s+1)$	$(i,2,1,s)$	$(s+1)\omega I_m$	
$(i,2,1,j)$	$(i,1,1,j)$	$\delta_2 I_m \otimes \eta$	$1 \le j \le S-1$
$(i,2,1,j)$	$(i,2,0,j+1)$	βI_m	$j=S-1$
$(i,2,1,j)$	$(i,2,1,j)$	$D^0 + (1-\gamma)D^1 - i\theta(1-\delta)I_m - \delta_2 I_m - \beta I_m - j\omega I_m$	$1 \le j \le S-1$
$(i,2,1,j)$	$(i,2,1,j-1)$	$j\omega I_m$	$2 \le j \le S-1$
$(i,2,1,j)$	$(i,2,1,j+1)$	βI_m	$1 \le j \le S-2$

b) expected perishable rate

$$E_p = \sum_{i=0}^{\infty} \sum_{j=s+1}^{S} \sum_{k=1}^{m} j\omega \left(z_{i,0,0,j,k} + \sum_{l=1}^{n} z_{i,1,0,j,k,l} + z_{i,2,0,j,k} \right) +$$

$$+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m} j\omega z_{i,0,1,j,k}+\sum_{i=0}^{\infty}\sum_{j=2}^{S-1}\sum_{k=1}^{m} j\omega\left(\sum_{l=1}^{n} z_{i,1,1,j,k,l}+z_{i,2,1,j,k}\right);$$

c) fraction of time the production is on

$$FP_{On}=\sum_{i=0}^{\infty}\sum_{j=0}^{S-1}\sum_{k=1}^{m} z_{i,0,1,j,k}+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m}\sum_{l=1}^{n} z_{i,1,1,j,k,l}+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m} z_{i,2,1,j,k};$$

d) expected number of customers in the orbit

$$EC=\sum_{i=0}^{\infty}\sum_{j=s+1}^{S}\sum_{k=1}^{m} i\left(z_{i,0,0,j,k}+\sum_{l=1}^{n} z_{i,1,0,j,k,l}+z_{i,2,0,j,k}\right)+$$
$$+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m} i\left(z_{i,0,1,j,k}+\sum_{l=1}^{n} z_{i,1,1,j,k,l}+z_{i,2,1,j,k}\right);$$

e) expected departures after service

$$E_{ds}=\sum_{i=0}^{\infty}\sum_{j=s+1}^{S}\sum_{k=1}^{m}\sum_{l=1}^{n} z_{i,1,0,j,k,l}T_l^0+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m}\sum_{l=1}^{n} z_{i,1,1,j,k,l}T_l^0;$$

f) expected rate of breakdown

$$E_{br}=\delta_1\left(\sum_{i=0}^{\infty}\sum_{j=s+1}^{S}\sum_{k=1}^{m}\sum_{l=1}^{n} z_{i,1,0,j,k,l}+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m}\sum_{l=1}^{n} z_{i,1,1,j,k,l}\right);$$

g) expected number of primary loss

$$E_{l1}=(1-\gamma)\sum_{i=0}^{\infty}\left(\sum_{k=1}^{m}\sum_{k_1=1}^{m} z_{i,0,1,0,k}D^1_{kk_1}+\sum_{j=s+1}^{S}\sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{k_1=1}^{m} z_{i,1,0,j,k,l}D^1_{kk_1}+\right.$$
$$\left.+\sum_{j=1}^{S-1}\left(\sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{k_1=1}^{m} z_{i,1,1,j,k,l}D^1_{kk_1}+\sum_{k=1}^{m}\sum_{k_1=1}^{m}\left(z_{i,2,1,j,k}+z_{i,2,0,j,k}\right)D^1_{kk_1}\right)\right);$$

h) expected rate of repair

$$E_{rr}=\delta_2\left(\sum_{i=0}^{\infty}\sum_{j=s+1}^{S}\sum_{k=1}^{m} z_{i,2,0,j,k}+\sum_{i=0}^{\infty}\sum_{j=1}^{S-1}\sum_{k=1}^{m} z_{i,2,1,j,k}\right);$$

i) expected number of retrial customer loss

$$E_{l2}=\theta\sum_{i=1}^{\infty} i\left(\sum_{k=1}^{m}(1-\delta)z_{i,0,1,0,k}+\sum_{j=s+1}^{S}\sum_{k=1}^{m}\sum_{l=1}^{n}(1-\delta)z_{i,1,0,j,k,l}+\right.$$
$$+\sum_{j=1}^{S-1}\sum_{k=1}^{m}\sum_{l=1}^{n}(1-\delta)z_{i,1,1,j,k,l}+\sum_{j=s+1}^{S}\sum_{k=1}^{m}(1-\delta)z_{i,2,0,j,k}+$$
$$\left.+\sum_{j=1}^{S-1}\sum_{k=1}^{m}(1-\delta)z_{i,2,1,j,k}\right).$$

The construction of the total profit per unit time is as follows:

$$E_{profit}=c_1E_{ds}-c_2EI-c_3EC-c_4(E_{l1}+E_{l2})-c_5E_p-c_6FP_{ON}-c_7E_{br}-c_8E_{rr},$$

where c_1 =revenue from selling unit item per unit time and c_2, c_3, c_4, c_5 and c_6 are the inventory holding cost, waiting space maintenance cost, penalty due to customer loss, disposing cost of decayed items and production running cost, per unit per unit time respectively. c_7 is penalty due to breakdown and c_8 is the repair cost.

4. Conclusion

The paper examines a perishable inventory system with an unreliable server under a MAP/PH framework, incorporating (s,S) production policy and retrial mechanisms. Using the matrix-geometric method, it analyzes system performance and profit. Future extension of this work may done by replacing MAP with BMAP.

REFERENCES

1. *Bobina J. M., Jose K. P.* Stochastic Analysis of a Production Inventory System with Deteriorating Items, Unreliable Server and Retrial of Customers // ITMM. Queueing Theory and Applications. 2024. Vol. 2163. P. 91–105.
2. *Chakravarthy S. R.* The batch Markovian arrival process: A review and future work // Advances in Probability Theory and Stochastic Processes. 2001. P. 21–39.
3. *Jose K.P., Reshmi P.S.* A Production Inventory Model with Deteriorating Items and Retrial Demands // Opsearch(Springer). 2020. Vol. 58. P. 71–82.
4. *Krishnamoorthy A., Vishnevsky V.,Dhanya S.* Single server queues with several services // Automatic control theory and applications. 2017. Vol. 12.
5. *Melikov A. Z., Shahmaliyev M. O.* Queueing system m/m/1/∞ with perishable inventory and repeated customers. // Automation and Remote Control. 2019. Vol. 38, № 80. P. 53–65.
6. *Neuts M.F., Rao B.* Numerical investigation of a multiserver retrial model. // Queueing systems. 1990. Vol. 7, № 2. P. 160–190.
7. *Salini S. N., Jose K. P.* A MAP/PH/1 Production Inventory Model with Varying Service Rates. // International Journal of Pure and Applied Mathematics. 2017. Vol. 117, № 12. P. 373–381.

Agassi Melikov — Ph.D., Vice- Rector of Scientific Affairs, Dept. of Mathematics, Baku Engineering University,Hasan Aliyev Str. 120, Baku AZ 0101, Azerbaijan. E-mail: *amelikov@beu.edu.az*

K. P. Jose — Ph.D., Associate Professor, P. G. & Research Dept. of Mathematics, St. Peter's College, Kolenchery, Kerala, India. E-mail: *kpjspc@gmail.com*

Bobina J. Mattam — M.Sc., Research Scholar, P. G. & Research Dept. of Mathematics, St. Peter's College, Kolenchery, Kerala, India. E-mail: *itsboby7@gmail.com*

BUSY PERIOD ANALYSIS OF M/G/1-TYPE RETRIAL QUEUE

E. V. Morozov[1], K. A. Zhukova[1]

[1] *Institute of Applied Mathematical Research, Karelian Research Center of RAS, Petrozavodsk, Russia*

We study the busy period of a stationary single-server retrial system with constant retrial rate. The analysis is based on the comparing the considered retrial system with a buffered $M/G/1$-type model with the first exceptional service. It is shown that the distribution of the busy period satisfies a functional equation which is similar to the well-known equation for the conventional $M/G/1$ system. The main result is expressed in the terms of the Laplace-Stieltjes transforms (LST) and an explicit example is given as well.
Keywords: *busy period, retrial queue, Laplace-Stieltjes transform, constant retrial rate*

Introduction

In this paper, we perform the busy period analysis of a single-server single class $M/G/1$-type system with constant retrial rate with Poisson input process and general service time distribution. The analysis is based on the reduction of the original retrial system to a buffered system with the 1st exceptional service (of the customer starting each busy period) [2]. That is, the customer beginning each busy period has in general another distribution than that other customers arriving during busy period.

We show that the target Laplace-Stieltjes transform (LST) satisfies a functional equation which is similar to the corresponding equation in the stationary classic $M/G/1$ queueing system, and the further analysis is based on the application of the Last-In-First-Out (LIFO) service discipline, see for instance, [1]. Under this discipline each new customer arriving during the 1st service generates a new busy period and, as a result, an integral equation containing series of convolutions of the service time distribution is obtained. Then the application of the LST gives the mentioned functional equation.

The structure of the paper is as follows. Section 1 contains a detailed description of the constant retrial rate system under consideration. In Section 2, using the approach similar to that is applied in the conventional buffered $M/G/1$ system, we derive the functional equation for the LST of the busy period in the original retrial system. Also we give an explicit example when all governing distributions are exponential.

1. Model Description

We consider a single-server retrial system fed by Poisson input with rate λ. Denote by $\{t_n,\, n \geq 1\}$ the arrival times of the customers to the system. Let τ be the generic exponential interarrival time and the independent identically distributed (iid) service times $\{S_n\}$ be generally distributed with function F_S and the mean $\mathsf{E}S < \infty$. When the new customer enters the system, there are two possible cases. If server is idly upon an arrival, he/she occupies it. Otherwise, the arrival joins a virtual orbit working as First-In-First-Out queue, where the top customer makes attempts to get service after the exponential time ξ with parameter γ. This type of retrial discipline is called constant retrial rate policy. The stability criterion of such a system is [4]

$$\rho := \lambda \mathsf{E}S < \frac{\gamma}{\lambda + \gamma}. \tag{1}$$

The processes describing the dynamics of the system are all regenerative, and we show it for the process $N(t)$ which is the number of customers in the system at instant t, $t \geq 0$. The process $\{N(t)\}$ is regenerative, and the regeneration instants $\{\theta_n\}$ in which the process stochastically restarts are defined recursively as

$$\theta_{n+1} = \inf(k > \theta_n : N_k = 0),\, n \geq 0 \ (\theta_0 := 0), \tag{2}$$

where $N(t_n^-) = N_n$, $n \geq 1$. The existence of such epochs follows because the so-called regeneration condition $\mathsf{P}(\tau > S_1) > 0$ holds, see [4]. It means that each regeneration period starts by a customer which meets the system fully empty. This period is thus composed by a busy period (when there are customers in the system) and the following idle exponential period τ. During busy period, after each departure of the served customer from the system, there exists an idle period, which is the minimum of retrial time and interarrival time $\zeta := \min(\xi, \tau)$. This period ζ is exponentially distributed, and it can be considered as an addition to the following 'service time'. It allows us to treat the retrial system (from the viewpoint of busy period analysis) as a buffered queueing system with an exceptional 1st service because the service time of the 1st customer beginning busy period is simply its service time S without the additional term ζ. Other customers have the service time $\zeta + S$ where ζ and S are independent random variables. We apply this idea to the busy period analysis of the retrial queuing system under consideration.

2. Functional equation for the LST of busy period

The well-known analysis of the busy period in the conventional $M/G/1$ queue [1] is directly extended to the system with the exceptional 1st service in which we denote busy period by B_1 and its distribution function (d.f.) by G_1. An important observation is that busy periods generated by the customers arrived during the 1st service are distributed as a random variable B with d.f. G in the standard $M/G/1$ system. Then, adapting the approach from [1], we obtain

$$G_1(x) = \mathsf{P}(B_1 \le x) = \int_0^x \sum_{n=0}^{\infty} e^{-\lambda u} \frac{[\lambda u]^n}{n!} G^{(n)}(x-u) dF_{S_1}(u) \ , \quad x \ge 0 \ , \quad (3)$$

implying the functional equation for the corresponding LST's:

$$g_1(z) = \varphi_1(z + \lambda - \lambda g(z)), \quad (4)$$

where $g_1(z) = \mathsf{E}e^{-zB_1}$, $\varphi_1(z) = \mathsf{E}e^{-zS_1}$ and $G^{(n)}$ denotes n-convolution of the d.f. G with itself.

Now we specify this analysis to the system under consideration and recall that, because in our setting $S_1 = S$ while other service times are distributed as $S + \zeta$, then (4) becomes

$$g_1(z) = \varphi(z + \lambda - \lambda \widehat{g}(z)), \quad (5)$$

where $\widehat{g}(z) = \mathsf{E}e^{-z\widehat{B}}$, $\varphi(z) = \mathsf{E}e^{-zS}$ and $\widehat{B}$ denotes a (continuous-time) busy period in $M/G/1$ system (denoted by $\widehat{\Sigma}$) in which each service time is distributed as $S + \zeta$. Hence, $\widehat{g}(z)$ satisfies functional equation

$$\widehat{g}(z) = \widehat{\varphi}(z + \lambda - \lambda \widehat{g}(z)), \quad (6)$$

with $\widehat{\varphi}(z) = \mathsf{E}e^{-z(S+\zeta)} = \varphi(z)\mathsf{E}e^{-z\zeta}$. In particular, the discrete-time busy period in the system $\widehat{\Sigma}$, denoted by $\widehat{\theta}$, has the mean

$$\mathsf{E}\widehat{\theta} = -\widehat{g}'(0) = \frac{1}{1 - \lambda \mathsf{E}(S + \zeta)}.$$

(We recall that $\lambda \mathsf{E}(S + \zeta) < 1$ by (1).) In the basic system we have by (5)

$$\mathsf{E}B_1 = -g_1'(0) = \frac{\mathsf{E}S}{1 - \lambda \mathsf{E}(S + \zeta)}. \quad (7)$$

Example. In $M/M/1$-type retrial system, equation (6) can be solved in an explicit form. Let service time S be exponential with parameter μ. In this case

$$\widehat{\varphi}(z) = \frac{\mu(\lambda + \gamma)}{(\mu + z)(\lambda + \gamma + z)},$$

and (6) can be written as

$$\widehat{g}(z) = \frac{\mu(\lambda+\gamma)}{(\mu+\lambda+z-\lambda\widehat{g}(z))(2\lambda+\gamma+z-\lambda\widehat{g}(z))}. \tag{8}$$

The third degree equation (8) (with respect to $\widehat{g}(z)$) can be solved explicitly with given parameters λ, μ, γ (satisfying stability condition (1)). As a result, $g_1(z)$ also can be derived from equation (4). In particular, by (7),

$$\mathsf{E}B_1 = \frac{\lambda+\mu+\gamma}{\mu(\mu+\gamma)}.$$

More numerical examples are assumed to be considered.

3. Conclusion

The busy period analysis of a stationary single-server retrial system with constant retrial rate is performed. The basic idea of the analysis is to compare the original retrial system with a buffered $M/G/1$-type model with the 1st exceptional service. It is shown that the distribution of the busy period satisfies a functional equation which is similar to the well-known equation for the conventional $M/G/1$ system. The case of the exponential service time is considered as a numerical example.

REFERENCES

1. *Gnedenko B., Kovalenko I. N.* An Introduction to Queueing Theory. Birkhäuser, Basel, 1989.
2. *Welch P. D.* On a generalized M/G/1 queueing process in which the first customer of each busy period receives exceptional service // Operations Research. 1964. Vol. 12. P. 736–752.
3. *Asmussen S.* Applied probability and queues. New York: Springer, 2003.
4. *Morozov E. V., Steyaert B.* Stability Analysis of Regenerative Queueing Models. Springer, Cham, 2021.

Morozov Evsey Viktorovich — Doctor of Physical and Mathematical Sciences, Professor, Learding researcher, Institute of Applied Mathematical Research, KarRC of RAS. E-mail: *emorozov@karelia.ru*

Zhukova Ksenia Alekseevna — Candidate of of Physical and Mathematical Sciences, researcher, Institute of Applied Mathematical Research, KarRC of RAS. E-mail: *kalininaksenia90@gmail.com*

A QUEUING INVENTORY MODEL WITH BATCH DEMANDS AND FINITE BUFFER

Samaja S Naboothri[1], Jaya S[1]

[1] *Department of Mathematics, Maharaja's College, Ernakulam, India*

This paper considers an inventory queuing model with batch demands where customers arrive according to a Markovian Arrival Process and service times is phase type distributed. An arriving customer can request between 1 and N items, where N denotes the maximum allowable number of items per demand. If the demanded unit of item is unavailable, customers can leave the system or can enter into service with suitable probabilities. Further, if they enter the service, they can purchase the available items or can wait by moving to a buffer of finite capacity. An (s, S) policy is used for inventory replenishment and customers in the buffer are instantaneously provided with items when replenishment occurs. Analysis of this model was done by classical matrix analytic method and some important performance measures are derived. A suitable cost function is also constructed. **Keywords:** *queuing inventory, batch demands, finite buffer*

Introduction

Traditional queuing inventory models with positive service time often assume that customer demands arrive individually, one unit at a time and many scholars have undertaken detailed analyses in this domain. The earliest contributions to this field were made by Sigman and Simchi-Levi [8] and Melikov and Molchanov [7] followed by Berman et al. [2], Berman and Kim [3] and so on. But there are many real world scenarios where customer demands often occur in batches, meaning that a single arriving customer may request multiple items from the inventory. Yue et al. [9] analyze an M/M/1 queueing inventory system with batch demands modelled by a compound Poisson process and geometric distribution, deriving stationary distributions under two loss scenarios. Later, Chakravarthy [4] investigates two batch-demand queueing inventory models with positive service times and customer loss scenarios, analyzing them in steady state using matrix analytic methods. Chakravarthy and Rumyantsev [5] extend this work by

Authors thank Prof. (Dr.) A. Krishnamoorthy, Director of Centre for Research in Mathematics, CMS College, Kottayam, Professor of Department of Mathematics, Central University of Kerala, India.

incorporating MAP arrivals and phase-type service times, analyzing both single-server and multi-server cases. Some other contributions to this area had been done by Akash Verma et al. [1] and Linhong et al. [6] and so on.

This paper introduces an inventory queuing model with batch demands for each customer but with a limit of N units of items. If the demanded units is unavailable, customer can either leave without buying any item or can proceed to service where they are free to choose between the options

- leave by purchasing available items from the inventory or
- move to a finite buffer and can wait.

This scenario is common in everyday life, where customers may find items out of stock. Providing them the option to wait for restocking can help buisnesses retain customers and build strong relationships.

1. Model Formulation

We consider a single server single inventory system with infinite queue where services are offered in FCFS manner. The customer arrival pattern is modelled as a Markovian Arrival Process with representation (D_0, D_1) of order m_1 and the service time is assumed to be phase type distributed with representation (η, T) of order m_2. Assume (s, S) replenishment policy is used for replenishment. It is further assumed that each customer announces their required item quantity at the arrival epoch and each customer has the option to request a batch of items, labeled as Z, from the inventory. Each customer can place a demand for up to N items where $N < s$. The probability of demanding i items is given by

$$P(Z = i) = p_i,\ 1 \leq i \leq N \text{ with } \sum_{i=1}^{N} p_i = 1.$$

If the requested quantity of items is unavailable, customers are free to

- leave the system without taking any item from the inventory with probability $(1 - p)$ or
- can enter into the service with probability p.

Additionally, when a customer enters the service, they can either leave the system by purchasing all the available items from the inventory with probability $(1 - q)$ or after completing the service, the customer may move to a buffer of finite capacity, (say K), without taking any item from the inventory and can wait there with probability q.

This system employs an (s, S) replenishment policy in such a way that if the inventory level drops to s or below, a replenishment order is triggered. The order quantity is calculated to bring back inventory level to S together

with an additional amount of inventory to meet the demands of customer in buffer.

That is, when the replenishment occurs, customers in the buffer are instantaneously provided with items. When the total demands in the buffer reaches its maximum and the requested number of items is unavailable, then the arriving customers are lost.

2. Mathematical Analysis of Model

The following notations are required for the mathematical analysis of the model. Let $N_1(t), I(t), N_2(t), J_1(t)$ and $J_2(t)$ represents the number of customers in the system, inventory level, number of customers in the buffer, phase of service process and phase of arrival process at time t, respectively. Then the process $X(t) = (N_1(t), I(t), N_2(t), J_1(t), J_2(t))$ constitutes a CTMC possessing the state space $\{(0, i, n_2, \star, j_2)|0 \leq i \leq N-1, 0 \leq n_2 \leq K, 1 \leq j_2 \leq m_1\}\cup$
$\{(0, i, 0, \star, j_2)|N \leq i \leq S, 1 \leq j_2 \leq m_1\}\cup$
$\{(n_1, i, 0, j_1, j_2)|n_1 \geq 1, 0 \leq i \leq N-1, 0 \leq n_2 \leq K, 1 \leq j_1 \leq m_2,$
$1 \leq j_2 \leq m_1\}\cup$
$\{(n_1, i, 0, j_1, j_2)|n_1 \geq 1, N \leq i \leq S, 1 \leq j_1 \leq m_2, 1 \leq j_2 \leq m_1\}$.

To effectively represent the infinitesimal generator matrix of the Markov Process, we need to introduce some notations beforehand.

- Define $f = (K+1)N + S - (N-1), n = K+1, m = s-(N-1)$, $b = S - s$, $a = Nn + m$.
- With the probabilities $\{p_i|1 \leq i \leq N\}$ representing the demand of i items, we proceed to define the probabilities corresponding to demands greater than i as, $\tilde{p}_i = \sum_{k=i+1}^{N} p_k, 1 \leq i \leq N-1$, and demands less than or equal to i as, $\hat{p}_i = \sum_{k=1}^{i} p_k, 2 \leq i \leq N$.
- We use notations I and **e** for the identity matrix and column matrix of ones of appropriate order.
- We use the notations $\bigotimes$ and $\bigoplus$ to denote Kronecker product and Kronecker sum respectively.

Then the infinitesimal generator matrix Q of the process is given by:

$$\begin{bmatrix} B_0 & B_1 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

3. Stability Condition

Let π denote the steady state probability vector of the generator $A = A_0 + A_1 + A_2$. Then we have $\pi A = 0$ and $\pi \mathbf{e} = 1$. And the system is stable if and only if $\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e}$.

On solving we get

$$(\pi_0 \mathbf{e} - \pi_{0,K})(I_{m_2} \bigotimes pD_1) + \sum_{i=1}^{N-1} \pi_i (I_{m_2} \bigotimes (\hat{p}_i + p.\tilde{p_i} D1)) +$$

$$+ \sum_{i=N}^{S} \pi_i (I_{m_2} \bigotimes D_1) < T^0 \bigotimes \eta \bigotimes I_{m_1}.$$

Stationary Distribution

Assuming the stability condition is satisfied, the steady state probability vector of the system is obtained by solving the equations $xQ = 0$ and $x\mathbf{e} = 1$, where $x = (x_0, x_1, \cdots)$ denote the steady state probability vector of Q.

From $xQ = 0$ we get

$$x_0 B_0 + x_1 B_2 = 0, \tag{1}$$

$$x_0 B_1 + x_1 A_1 + x_2 A_2 = 0, \tag{2}$$

$$x_{n-1} A_0 + x_n A_1 + x_{n+1} A_2 = 0. n \geq 2 \tag{3}$$

Now x is obtained as

$$x_n = x_1 R^{n-1}, \tag{4}$$

where R is the minimal non negative solution to the matrix quadratic equation $R^2 A_2 + R A_1 + A_0 = 0$, subject to the normalizing condition

$$x_0 \mathbf{e} + x_1 (I - R)^{-1} \mathbf{e} = 1. \tag{5}$$

4. System Performance Measures

Some of the important performance measures are given by

– Probability that system is idle

$$p_{idle} = x_0 \mathbf{e}.$$

– Expected number of customers in the system

$$E_{cus} = \sum_{n_1=1}^{\infty} n_1.x_{n_1}\mathbf{e}.$$

– Mean inventory level

$$M_{Inv} = \sum_{n_1=0}^{\infty}\sum_{i=1}^{S} i.x_{n_1,i}\mathbf{e}.$$

– Expected number of customers in the buffer

$$E_{Buff} = \sum_{n_1=0}^{\infty}\sum_{i=0}^{N-1}\sum_{n_2=0}^{K} n_2.x_{n_1,i,n_2}\mathbf{e}.$$

– Expected rate of customer loss

$$E_{Loss} = (1-p)\lambda \sum_{n_1=1}^{\infty}\sum_{i=0}^{N-1}\sum_{n_2=0}^{K-1} \tilde{p}_i.n_1.x_{n_1,i,n_2}\mathbf{e}+$$

$$+\lambda \sum_{n_1=1}^{\infty}\sum_{i=0}^{N-1} \tilde{p}_i.n_1.x_{n_1,i,K}\mathbf{e}.$$

– Expected number of partially lost customers,

$$E_{P_{loss}} = (1-q)\lambda \sum_{n_1=1}^{\infty}\sum_{i=0}^{N-1} \tilde{p}_i.n_1.x_{n_1,i}\mathbf{e}.$$

Cost Function

Expected Total Cost is given by $ETC = c_1.M_{Inv} + c_2 + c_3E_{Buff} + c_4.E_{Loss} + c_5.E_{P_{loss}}$ where c_1 is the holding cost of inventory per unit time, c_2 is the fixed cost for buffer, c_3 is buffer maintenance cost, c_4 is cost for customer loss and c_5 is cost for partially lost customers.

We take

$$D_0 = \begin{pmatrix} -3.0067 & 3.0067 & 0 \\ 0 & -3.0067 & 0 \\ 0 & 0 & -677.2500 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.0301 & 0 & 2.9766 \\ 670.4775 & 0 & 6.7725 \end{pmatrix},$$

$$T = \begin{pmatrix} -10 & 4 \\ 8 & -9 \end{pmatrix},\ T^0 = \begin{pmatrix} 6 \\ 1 \end{pmatrix},\ \eta = \begin{pmatrix} 0.3 & 0.7 \end{pmatrix}.$$

5. Numerical Illustration

In this section we provide numerical illustration of some of the system performance measures with varying parameters. Here, (see Table 1) as the buffer size increases, expected number of customers in the system as well as expected number of customers in the buffer also increases.

Table 1

Effect of E_{Buff} and E_{Cus} on K: Fix S=20, s=8, K=7, N=3, p=0.8, 1-p=0.2, q=0.6, v=0.4, γ=0.7

K	E_{Buff}	E_{Cus}
3	0.4358	3.0891
4	0.5210	3.2903
5	0.5884	3.4579
6	0.6424	3.5967
7	0.6860	3.7098

6. Conclusion

This paper studied a queuing inventory model with batch demands and finite buffer, where each arriving customer can demand between 1 and N units of item and if the demanded unit is unavailable, customers have the provision to wait. We analyzed the system through matrix analytic method and some performance measures are derived. A suitable cost function is also constructed on the basis of system performance measures.

REFERENCES

1. *Akash Verma, ujit Kumar Samanta, Sapna Isotupa K. P.* A Batch Demands Queueing-inventory Model with Positive Replenishment Time // Queuing Models and service Mangement. 2025. Vol. 8, № 1. P. 1–20.
2. *Berman O., Kaplan E. H., Shimshak D. G.* Deterministic approximations for inventory management at service facilities // IIE Transaction. 1993. Vol. 25, № 5. P. 98–104.
3. *Berman O., Kim E.* Stochastic models for inventory managements at service facilities // Communications in Statistics Stochastic Models. 1999. Vol. 15, № 4. P. 695–718.
4. *Chakravarthy S. R.* Queueing-Inventory Models with Batch Demands and Positive Service // Times. Autom Remote Control. 2020. Vol. 81. P. 713–730.
5. *Chakravarthy S. R.,Rumyantsev A.* Analytical and simulation studies of queueing-inventory models with MAP demands in batches and positive phase

type services // Simulation Modelling Practice and Theory. 2020. Vol. 103. P. 102092.

6. *Linhong Li, Liwei Liu, Wei Xu, Zhen Wang* Stochastic decomposition in a queueing-inventory system with batch demands, randomized order policy and multiple vacations // Communications in Statistics-Theory and Methods. 2024. Vol. 53, № 12. P. 4407–4427.
7. *Melikov A. Z., Molchanov A. A.* Stock optimization in transportation/storage systems // Cybernetics and Systems Analysis. 1992. Vol. 28, № 3. P. 484–487.
8. *Sigman K., Simchi-Levi D.* Light traffic heuristic for an M/G/1 queue with limited inventory // Annals of Operation Research. 1992. Vol. 40. P. 371–380.
9. *Yue D.,Zhao G.,Qin Y.* An M/M/1 Queueing-Inventory System with Geometric Batch Demands and Lost Sales // J Syst Sci Complex. 2018. Vol. 31. P. 1024–1041.

Samaja S Naboothri — Research Scholar, Department of Mathematics, Maharaja's College, Ernakulam, Kerala, India. E-mail: *samajas1995@gmail.com*

Jaya S — Research Guide, Department of Mathematics, Maharaja's College, Ernakulam, Kerala, India. E-mail: *sreelakamjaya2024@gmail.com*

A MAP/PH/1 PRODUCTION INVENTORY SYSTEM WITH SERVICE INTERRUPTION AND DIFFERENT RATES OF PRODUCTION

Susmi Skaria[1], Salini S. Nair[2], K.P. Jose[2]

[1],*Dept. of Mathematics, Union Christian College, Kerala, India*
[2]*Dept. of Mathematics, St. Peter's College, Kerala, India*

Absract: This paper studies a production inventory model under (s,S) policy incorporating customer retrials and service interruptions. Customer arrivals follow a Markovian Arrival Process(MAP) ,and service times follow a Phase Type (PH) distribution. When the inventory falls to s, production starts at an increased rate until the inventory reaches the level s+1. Customers encountering a busy server or stockout enter an orbit of infinite capacity and retry after an exponentially distributed time. Service interruptions also redirect customers to the orbit. Customers who experience excessive delays may leave the system permanently. Key performance measures and the relevant cost function are constructed to optimize the total expected cost. **Keywords:** *retrial inventory, (s,S) policy, Markovian arrival process, phase type distribution, service interruption, cost function*

Introduction

Efficient inventory control is crucial in uncertain markets to ensure optimal resource use and customer satisfaction. Modern research focuses on stochastic inventory models incorporating random arrivals, service interruptions, retrials and system failures.

Krishnamoorthy et al.[3] analyzed an (s,S) production inventory system with Poisson arrivals and Erlang service times. They introduced a novel method to obtain an explicit expression for the stability of the system and several system performance measures were derived. Salini and Jose [4] were studied the production inventory system with retrial and varying service rates by assuming MAP arrivals with Phase Type service times. They defined some important system performance measures and constructed a suitable cost function. Nisha Mathew et al. [2] studied models with server breakdowns, customer impatience and multi-stage production using MAP

The authors acknowledge the financial support provided by the FIST Program, Department of Science and Technology, Govt.of India, to the PG and Research Department of Mathematics through SR/FST/College-2018-XA 276(C)

and Phase Type distributions. A production inventory system with different rates of production and retrials was studied by Jose and Salini [1]. They used the Matrix analytic method to find an algorithmic solution. Serife Ozkar [5] studied a production inventory model with phase type service times where customers join the system according to a Markovian arrival process. The steady-state analysis was done by using the Matrix Geometric Method.

This paper analyzes a production inventory model running with different production rates, MAP arrivals and with the concept of interruption to the service facility.

1. Mathematical Modeling and Analysis of the Problem

Consider a production inventory system operating under the (s, S) policy, which allows customer retrials and service interruptions. Customer arrivals follow a Markovian arrival process with representation (D_0, D_1) of order m_1 and service time follows a Phase type distribution with representation (η, T) of order m_2. The item in the inventory is served through a single server counter. Production starts when the inventory falls to s and continues the production till the inventory reaches the level S. The production rate is $\alpha\beta$ where $\alpha \in [1, c]$ and c is a finite number greater than 1, when production starts and the rate falls to β when it exceeds s. Customer service can be interrupted at any time, but only during active service. Inter occurrence time of interruption is exponentially distributed with parameter δ_1 and time to resume the service is also exponentially distributed with parameter δ_2 . An arriving customer facing a busy server, an interruption or zero inventory enters with probability γ, or is lost forever with probability $1 - \gamma$. A retrial customer encountering the same condition re-enters the orbit with probability δ, or is lost forever with probability $1 - \delta$. The inter retrial time is exponentially distributed with linear rate $i\theta$, when there are i customers in the orbit. No inventory or customers are lost due to server interruptions.

The following are the notations used in this model.

$N(t)$: Number of customers in the orbit at time t.

$I(t)$: Inventory level at time t.

$J(t)$: The production status,

$$J(t) = \begin{cases} 0, \text{if the production is in OFF mode,} \\ 1, \text{if the production is in ON mode.} \end{cases}$$

$S(t)$: The server status

$$S(t) = \begin{cases} 0, \text{if the server is idle,} \\ 1, \text{if the server is busy,} \\ 2, \text{if the sever is in interruption.} \end{cases}$$

$F_1(t)$: Phase of the arrival process at time t.
$F_2(t)$: Phase of the service process at time t.
Now $X(t) = \{(N(t), S(t), J(t), I(t), F_1(t), F_2(t)) : t \geq 0.\}$ is a level dependent quasi birth death process on the state space

$$\{(i,0,0,k,l) : i \geq 0; k = s+1, ..., S; l = 1, ...m_1\} \bigcup$$

$$\{(i,j,0,k,l,p) : i \geq 0; j = 1,2; k = s+1, ..., S; l = 1, ...m_1; p = 1, ...m_2\} \bigcup$$

$$\{(i,0,1,k,l) : i \geq 0; k = 0, ..., S-1; l = 1, ...m_1\} \bigcup$$

$$\{(i,j,1,k,l,p) : i \geq 0; j = 1,2; k = 1, ..., S-1; l = 1, ...m_1; p = 1, ...m_2\}.$$

The infinitesimal generator Q of the process is a block tri-diagonal matrix and has the form:

$$Q = \begin{bmatrix} A_{1,0} & A_0 & & & & \\ A_{2,1} & A_{1,1} & A_0 & & & \\ & A_{2,2} & A_{1,2} & A_0 & & \\ & & A_{2,3} & A_{1,3} & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix},$$

where the block matrices $A_0, A_{1,i}; (i \geq 0)$ and $A_{2.i}; (i \geq 1.)$ are square matrices of order $(4S - 2s - 2)m_1m_2 + (2S - s)m_1$ and are given by

$$A_0 = \begin{matrix} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & B_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & B_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_5 \end{bmatrix}, \text{ where } (B_1)_{mn} = \begin{cases} \gamma D_1; m = n = 1 \\ 0; \text{otherwise} \end{cases},$$

$$(B_2)_{mn} = \begin{cases} \gamma D_1 \otimes I_{m_2}; m = 1, ...S - s, n = m \\ 0; \text{otherwise} \end{cases},$$

$$(B_3)_{mn} = \begin{cases} \gamma D_1 \otimes I_{m_2}; m = 1, ...S - 1, n = m \\ 0; \text{otherwise} \end{cases},$$

$$(B_4)_{mn} = \begin{cases} \gamma D_1 \otimes I_{m_2}; m = 1, ...S - s, n = m \\ 0; \text{otherwise} \end{cases},$$

$$(B_5)_{mn} = \begin{cases} \gamma D_1 \otimes I_{m_2}; m = 1, ...S - 1, n = m \\ 0; \text{otherwise} \end{cases}.$$

$$A_{2,i} = \begin{array}{c} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{array} \begin{bmatrix} 0 & 0 & C_1 & 0 & 0 & 0 \\ 0 & C_2 & 0 & C_3 & 0 & 0 \\ 0 & 0 & C_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & C_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & C_5 \end{bmatrix},$$

$$A_{1,i} = \begin{array}{c} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{array} \begin{bmatrix} E_1 & 0 & E_2 & 0 & 0 & 0 \\ E_3 & E_4 & 0 & E_5 & 0 & 0 \\ E_6 & E_7 & E_8 & 0 & E_9 & 0 \\ 0 & E_{10} & E_{11} & E_{12} & 0 & E_{13} \\ 0 & 0 & E_{14} & 0 & E_{15} & 0 \\ 0 & 0 & 0 & E_{16} & 0 & E_{17} \end{bmatrix},$$

$$(C_1)_{mn} = \begin{cases} (i\theta) I_{m_1} \otimes \eta; m = 1, ...S - s, n = m \\ 0; \text{otherwise} \end{cases},$$

$$(C_2)_{mn} = \begin{cases} (i\theta)(1 - \delta) I - m_1; \ m = n = 1 \\ 0; \ \text{otherwise} \end{cases},$$

$$(C_3)_{mn} = \begin{cases} (i\theta) I_{m_1} \otimes \eta; \ m = 2, ...S, n = m - 1 \\ 0; \ \text{otherwise} \end{cases},$$

$$(C_4)_{mn} = \begin{cases} (i\theta)(1 - \delta) I_{m_1 m_2}; \ m = 1, ...S - s, n = m \\ 0; \ \text{otherwise} \end{cases}$$

$$(C_5)_{mn} = \begin{cases} (i\theta)(1 - \delta) I_{m_1 m_2}; \ m = 1, ...S - 1, n = m \\ 0; \ \text{otherwise} \end{cases},$$

$$(E_1)_{mn} = \begin{cases} \Delta; \ m = 1, ...S - s, n = m \\ 0; \ \text{otherwise} \end{cases},$$

$$(E_2)_{mn} = \begin{cases} D_1 \otimes \eta; \ m = 1, ...S - s, n = m \\ 0; \ \text{otherwise} \end{cases},$$

$$(E_3)_{mn} = \begin{cases} \beta I_{m_1}; \ m = n = S \\ 0; \ \text{otherwise} \end{cases},$$

$$(E_4)_{mn} = \begin{cases} \Gamma_0;\ m = n = 1 \\ \Gamma_1;\ m = 2, ...s+1, n = m \\ \Gamma_2;\ m = s+2, ...S, n = m \\ \alpha\beta I_{m_1};\ m = 1, ...s+1, n = m+1 \\ \beta I_{m_1};\ m = s+2, ...S-1, n = m+1 \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_5)_{mn} = \begin{cases} D_1 \otimes \eta;\ \text{m=2,...S,n=m-1} \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_6)_{mn} = \begin{cases} I_{m_1} \otimes T^0;\ \text{m=2,...S-s,n=m-1} \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_7)_{mn} = \begin{cases} I_{m_1} \otimes T^0;\ m = 1, n = s+1 \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_8)_{mn} = \begin{cases} \Delta_1;\ m = 1, ...S-s, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_9)_{mn} = \begin{cases} \delta_1 I_{m_1 m_2};\ m = 1, ...S-s, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{10})_{mn} = \begin{cases} I_{m_1} \otimes T^0;\ m = 1, ...S-1, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{11})_{mn} = \begin{cases} \beta I_{m_1 m_2};\ m = S-1, n = S-s \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{12})_{mn} = \begin{cases} \Delta_2;\ m = 1, ...s, n = m \\ \Delta_3;\ m = s+1, ...S-1, n = m \\ \alpha\beta I_{m_1 m_2};\ m = 1, ...s, n = m+1 \\ \beta I_{m_1 m_2};\ m = s+1, ...S-2, n = m+1 \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{13})_{mn} = \begin{cases} \delta_1 I_{m_1 m_2};\ m = 1, ...S-1, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{14})_{mn} = \begin{cases} \delta_2 I_{m_1 m_2};\ m = 1, ...S-s, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{15})_{mn} = \begin{cases} \Delta_4;\ m = 1, ...S-s, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{16})_{mn} = \begin{cases} \delta_2 I_{m_1 m_2};\ m = 1, ...S-1, n = m \\ 0;\ \text{otherwise} \end{cases},$$

$$(E_{17})_{mn} = \begin{cases} \Delta_5;\ m = 1, ...s, n = m \\ \Delta_6;\ m = s+1, ...S-1, n = m \\ \alpha\beta I_{m_1 m_2};\ m = 1, ...s, n = m+1 \\ \beta I_{m_1 m_2};\ m = s+1, ...S-2, n = m+1 \\ 0;\ \text{otherwise.} \end{cases}$$

2. System Stability

Take the Lyapunov test function $\varphi(r)$ as, $\varphi(r) = i$, if r is in level i. Then the mean drift y_r for any r belonging to the level $i \geq 1$, is given by $y_r = \sum_{p \neq r} q_{rp}(\varphi(p) - \varphi(r))$, ie,

$$y_r = \begin{cases} -i\theta;\ \text{if the server is idle and inventory level is positive,} \\ -i\theta(1-\delta) + \gamma(D_1 e)_l;\ \text{if the server is idle} \\ \text{and the inventory level is zero,} \\ -i\theta(1-\delta) + (\gamma(D_1 e) \otimes e_{m_2})_{(l-1)m_1+n};\ \text{otherwise.} \end{cases}$$

Since $(1-\delta) > 0$, for any $\varepsilon > 0$, we can find N large enough that $y_r < -\varepsilon$ for any r belonging to the level $i \geq N$. By Tweedi's [6]result, the system under consideration is stable.

3. System Performance Measures

Let the steady state probability vector $\mathbf{x}$ of Q be
$\mathbf{x} = (x(0), x(1), ..., x(n-1), x(n), ...)$ Then $x(i)$'s are given by

$$x(i) = (y_{i,0,0,s+1}, ..., y_{i,0,0,S}, y_{i,0,1,0}, ..., y_{i,0,1,S-1}, y_{i,1,0,s+1}, ..., y_{i,1,0,S},$$

$$y_{i,1,1,1}, ..., y_{i,1,1,S-1}, y_{i,2,0,s+1}, ..., y_{i,2,0,S}, y_{i,2,1,1}, ..., y_{i,2,1,S-1}).$$

Some important performance measures are

(i) Expected inventory level in the system

$$EIL = \sum_{i=0}^{\infty} \sum_{k=s+1}^{S} \sum_{l=1}^{m_1} k y_{i,0,0,k,l} + \sum_{i=0}^{\infty} \sum_{j=1}^{2} \sum_{k=s+1}^{S} \sum_{l=1}^{m_1} \sum_{p=1}^{m_2} k y_{i,j,0,k,l,p} +$$

$$\sum_{i=0}^{\infty} \sum_{k=0}^{S-1} \sum_{l=1}^{m_1} k y_{i,0,1,k,l} + \sum_{i=0}^{\infty} \sum_{j=1}^{2} \sum_{k=1}^{S-1} \sum_{l=1}^{m_1} \sum_{p=1}^{m_2} k y_{i,j,1,k,l,p}.$$

(ii) Expected Number of customers in the orbit

$$ENC = \left(\sum_{i=1}^{\infty} i x_i \right) e = \left(\sum_{i=1}^{N-1} i x_i + x_N (N(I-R)^{-1} + R(I-R)^{-2}) \right) e.$$

(iii) Expected interruption rate

$$EIR = \delta_1 \sum_{i=0}^{\infty} \left(\sum_{k=s+1}^{S} y_{i,1,0,k} e_{m_2} + \sum_{k=1}^{S-1} y_{i,1,1,k} e_{m_2} \right).$$

(iv) Expected repair rate of the server

$$ERR = \delta_2 \sum_{i=0}^{\infty} \left(\sum_{k=s+1}^{S} y_{i,2,0,k} e_{m_2} + \sum_{k=1}^{S-1} y_{i,2,1,k} e_{m_2} \right).$$

(v) Expected number of external customers lost, before entering the orbit is

$$ECL_P = \left((\sum_{i=0}^{\infty} y_{i,0,1,0})(1-\gamma) D_1 e \right) +$$

$$+((1-\gamma) D_1 \otimes e_{m_2}) \left(\sum_{i=0}^{\infty} \sum_{j=1}^{2} \sum_{k=s+1}^{S} y_{i,j,0,k} + \sum_{i=0}^{\infty} \sum_{j=1}^{2} \sum_{k=1}^{S-1} y_{i,j,1,k} \right).$$

(vi) Expected number of departures after completing service is

$$END = \left(\sum_{i=0}^{\infty} \sum_{k=s+1}^{S} y_{i,1,0,k} (e_{m_1} \otimes T^0) + \sum_{i=0}^{\infty} \sum_{k=1}^{S-1} y_{i,1,1,k} (e_{m_1} \otimes T^0) + \right).$$

(vii) Expected number of customers lost due to retrials

$$ECL_R = \theta(1-\delta) \left(\sum_{i=1}^{\infty} i \left(y_{i,0,1,0} e_{m_1} + \right. \right.$$

$$\left. \left. + \sum_{j=1}^{2} \sum_{k=s+1}^{S} y_{i,j,0,k} e_{m_1 m_2} + \sum_{j=1}^{2} \sum_{k=1}^{S-1} y_{i,j,1,k} e_{m_1 m_2} \right) \right).$$

(viii) Expected switching rate

$$ESR = \sum_{i=0}^{\infty} y_{i,1,0,s+1} (e_{m_1} \otimes T^0).$$

Cost Function: The Expected total Cost is defined as $ETC = c_1 ESR + c_2 EIL + c_3 ENC + c_4 ECL_P + c_5 ECL_R + c_6 END + c_7 ERR$, where c_1 =switching cost for production, c_2 = holding cost of inventory /unit/unit time, c_3 =holding cost of the customers in the orbit/unit/unit time, c_4 =cost due to loss of primary customers/unit/unit time, c_5 =cost due to loss of retrial customers/unit/unit time, c_6 =cost due to service/unit/unit time, c_7 = repair cost for the server.

4. Conclusion

This paper proposed a production inventory system incorporating varying production rates, customer retrials and server interruptions. Assuming that customer arrivals follow a Markovian Arrival Process (MAP), suitable production strategies and a cost function were developed. This study can be further extended by considering a Batch Markovian Arrival process.

REFERENCES

1. *Jose K. P., Salini S. Nair* A production inventory system with different rates of production and retrials, International Journal of Mathematical Archive, 2018, Vol. 9 (12). P. 30–40.
2. *Nisha Mathew, V. C. Joshu, A. Krishnamoorthy, A. Melikov, Ambily P. Mathew* A production inventory model with server breakdown and customer impatience, Annals of Operations Research, 2023. Vol. 331. P. 1269–1304.
3. *Krishnamoorthy A., Sajeev S. Nair, Viswanath C. Narayanan* Production Inventory with service time and interruptions, International Journal of System Science, 2015. Vol. 46. P. 1800–1816.
4. *Salini S. Nair, Jose K. P.* A MAP/PH/1 Production inventory model with varying service rates, International Journal of Pure and Applied Mathematics 2017. Vol. 117(12). P. 373–380.
5. *Serife Ozkar* MAP/PH/1 Production Inventory Model Uludag University of The Faculty of Engineering, 2023. Vol. 28, P. 89–160.
6. *Tweedie, Richard L.* Sufficient conditions for regularity, recurrence and ergodicity of Markov processes // Mathematical Proceedings of the Cambridge Philosophical Society, 1975, P. 125–136.

Susmi Skaria — Assistant Professor of Mathematics, Union Christian College, Aluva-683102, Kerala,India. E-mail: *susmiarun.arun@gmail.com*

Salini S. Nair — Ph.D.,Associate Professor of Mathematics,Saint Peter's College, Kolenchery -682311, Kerala, India. E-mail: *saliniajithk@gmail.com*

K.P. Jose — Ph.D., Associate Professor & Head, Post Graduate and Research Dept. of Mathematics, Saint Peter's College, Kolenchery -682311, Kerala, India. E-mail: *kpjspc@gmail.com*

SIMULATION OF A RETRIAL QUEUEING SYSTEM WITH TWO-WAY COMMUNICATION, AN UNRELIABLE SERVER AND IMPATIENT CUSTOMERS ACROSS VARIOUS SCENARIOS

Á. Tóth[1], J. Sztrik[1]

[1] *University of Debrecen, Debrecen, Hungary*

The paper examines a retrial queueing model featuring a finite source, two-way communication, and a server susceptible to breakdowns. Primary customers, coming from a limited pool, arrive according to an exponential distribution. Should the server be available, the service initiates instantaneously; in the alternative, customers are routed to the orbit and attempt reconnection subsequent to an exponentially distributed delay. A salient attribute of the model is that, in instances of server idleness, outbound calls—designated as secondary customers—are directed toward both the orbit and the source, with each pathway controlled by distinct parameters. As a consequence of the server's inherent unreliability, the provision of customer service may be subject to disruptions, wherein such interruptions adhere to an exponential distribution. Additionally, the system accounts for customer impatience, where customers may abandon the queue if delays exceed their tolerance. The study conducts a sensitivity analysis by testing different service time distributions to evaluate their effects on key performance metrics in two distinct scenarios. Results and comparisons are presented through graphical illustrations. **Keywords:** *Simulation, queueing system, finite-source model, sensitivity analysis, two-Way communication, unreliable operation, impatience*

Introduction

Two-way communication is a widely studied concept due to its relevance in modeling retrial queueing systems across numerous real-world applications. A key example is call centers, where agents, during idle periods, perform tasks like sales, marketing, or product promotion while handling incoming customer calls. A critical factor in these systems is utilization, with ongoing efforts to enhance the efficiency of service units or agents, as

J. Sztrik was supported by the Thematic Excellence Programme (TKP2021-NKTA-34) of the Ministry for Innovation and Technology in Hungary.

discussed in studies such as [1].The defining feature of two-way communication resides in the capability to originate calls both internally and externally to the system during intervals of server idleness. In the context of our model, the server possesses the ability to dispatch outbound calls directed toward either the source or the orbit. Extant literature, encompassing the work of Nazarov et al. [6], has delved into retrial queueing systems with infinite sources that incorporate two-way communication, whereas Dragieva and Phung-Duc [3] scrutinized a setup wherein a secondary outbound call reverts to the source upon service completion.

Building on [7], this study introduces other aspect like impatience where the service unit is prone to periodic breakdowns following an exponential distribution.

Waiting is a common experience across many areas of life, frequently causing frustration due to prolonged time spent in queues. This frustration may lead to requests leaving the system prematurely without receiving service, a phenomenon known as impatience. This behavior is evident in various fields, such as healthcare systems, call centers, and telecommunication networks. The impatience mechanism plays a vital role in the model, as it impacts system performance by potentially decreasing the number of customers remaining in the queue and altering service dynamics. Research exploring these behaviors includes [5].

The novelty of this research lies in conducting a sensitivity analysis to assess how various service time distributions affect key performance metrics. The results are generated using our stochastic simulation program, developed based on SimPack ([4]). This toolkit includes C/C++ libraries and executable programs tailored for computer simulation, supporting discrete event, continuous, and hybrid (multi-model) simulations. It provides the flexibility to model any queueing system or simulation type, calculating performance metrics with customizable random number generators for any specified random variable. The paper presents a table of input parameters and graphical representations illustrating operational modes and the impacts of different distributions.

1. System model

The present section delineates a retrial queueing model characterized by a finite source and a solitary server. The source encompasses N requests, wherein each entity is capable of originating a primary inbound call directed toward the server, with inter-arrival intervals conforming to an exponential distribution parameterized by λ. Upon server availability, the servicing of an arriving customer commences without delay, with service durations adhering to gamma, hyper-exponential, Pareto, or lognormal distributions,

each exhibiting distinct parameters yet sharing an identical mean value. Upon the conclusion of service, customers revert to the originating source. If an incoming customer finds the server occupied, they are redirected to the orbit instead of being lost. These become secondary incoming jobs that may retry connecting to the server after a random delay, governed by an exponential distribution with parameter σ.

Furthermore, in periods of server idleness, outbound calls may be initiated toward both the source and the orbit. Two distinct categories of outbound calls are delineated:

- The server is capable of soliciting a job from the source for servicing—termed a primary outbound call–subsequent to an exponentially distributed duration parameterized by λ_2.
- The server is capable of engaging a job from the orbit–designated as a secondary outbound call–subsequent to an exponentially distributed duration parameterized by σ_2.

The duration of service for outbound customers conforms to an exponential distribution parameterized by μ. Two distinct scenarios are outlined for outbound calls emanating from the orbit:

- Case 1: The call involves an unserved incoming request, and after the outgoing service, it returns to the orbit to await service for its incoming request.
- Case 2: The call includes an unserved incoming request, but after the outgoing service, the server immediately processes the incoming request. This results in a two-phase service process – first the outgoing call, then the incoming one – after which the call returns to the source upon completion of both phases.

The server is susceptible to breakdowns, which adhere to an exponential distribution parameterized by γ. In the event of a failure, the restoration process initiates instantaneously, with repair durations conforming to an exponential distribution specified by parameter γ_2.

Every primary customer in the system possesses an impatience attribute, determining their likelihood of exiting the system if service is not provided within a specific timeframe. This decision to depart transpires subsequent to a stochastic interval, which adheres to an exponential distribution characterized by the rate parameter τ.

It is posited that the arrivals of primary inbound calls, the retrial periods for secondary inbound calls, the service durations for both inbound and outbound calls, the failure and repair intervals, the impatience durations,

and the intervals preceding the initiation of outbound calls exhibit mutual independence.

2. Simulation results

We utilized a statistical module equipped with an advanced analytical toolkit to quantitatively evaluate the mean and variance of the observed variables through the application of the batch means method. This methodology involves partitioning n successive observations derived from a steady-state simulation into batches, thereby yielding a sequence of approximately independent sample means. The batch means method is esteemed for its efficacy in deriving confidence intervals pertaining to the steady-state mean of a stochastic process. Attaining sample means that are sufficiently independent necessitates the selection of adequately large batch sizes. An exhaustive exposition of the batch means method is available in [2]. The simulations were executed at a 99.9% confidence level, terminating upon attainment of a relative half-width for the confidence interval of 0.00001.

Table 1

Numerical values of model parameters

$\mathbf{N}$	$\boldsymbol{\tau}$	$\boldsymbol{\mu}$	$\boldsymbol{\lambda_2}$	$\boldsymbol{\sigma_2}$	$\boldsymbol{\gamma_2}$	$\boldsymbol{\sigma}$	$\boldsymbol{\gamma}$
100	0.001	1	0.02	0.02	1	0.01	0.1

In the present section, our objective is to specify the parameters for service times across each distribution, thereby ensuring equivalence in their mean values and variances. Four distinct distributions are scrutinized to evaluate their impacts on key performance metrics. The hyper-exponential distribution is selected to ensure a squared coefficient of variation exceeding one. The input parameters for these distributions are delineated in Table 2, whereas Table 1 enumerates the values of other pertinent parameters.

Table 2

Parameters of service time of primary customers

Distribution	**Gamma**	**Hyper-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 0.011$ $\beta = 0.011$	$p = 0.494$ $\lambda_1 = 0.989$ $\lambda_2 = 1.011$	$\alpha = 2.005$ $k = 0.501$	$m = -2.257$ $\sigma = 2.125$
Mean	1			
Variance	90.25			
Squared coefficient of variation	90.25			

Figure 1 depicts the correlation between the mean response time of customers who are successfully served and the arrival intensity. Successfully

served customers refer to those who remain in the system and do not leave prematurely due to impatience. The Pareto distribution yields the highest mean response time, while the distinctions among the other distributions become more evident. In particular, the gamma distribution produces the lowest mean response time.

A notable finding is that as arrival intensity grows, the mean response time initially increases but then declines after surpassing a certain threshold. This behavior is typical of finite-source retrial queueing systems and emerges under specific parameter settings. In the expanded version of our paper, we intend to include results based on an alternative parameter setup with a squared coefficient of variation below one and placing greater emphasis on how customer impatience influences performance metrics.

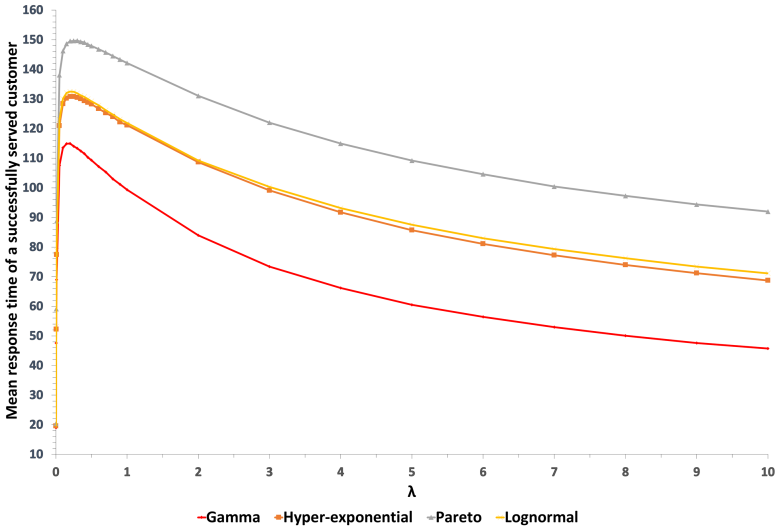


Figure 1. Mean response time vs. arrival intensity

3. Conclusion

The present study proposes a retrial queueing system characterized by a finite source and a two-way communication system, which encompasses an unreliable server, impatient customers, and a range of service time distributions. Different scenarios were analyzed using distinct parameters for a sensitivity analysis focused on the average customer response time and the service unit's utilization. Results were derived from our simulation program,

accompanied by graphical representations that illustrate how different service time distributions impact system performance metrics. These visuals reveal slight variations in performance measures when the squared coefficient of variation exceeds one, emphasizing the importance of distribution selection. In prospective endeavors, we aim to further examine additional finite-source retrial queueing systems incorporating two-way communication or to integrate a secondary service unit functioning as a backup mechanism.

REFERENCES

1. *Artalejo J., Gomez-Corral A.* Retrial Queueing Systems: A Computational Approach Springer, 2008.
2. *Chen J. E., Kelton D. W.* A Procedure for Generating Batch-Means Confidence Intervals for Simulation: Checking Independence and Normality // SIMULATION. 2007. Vol. 83, № 10. P. 683–694.
3. *Dragieva V., Phung-Duc T.* Two-way communication M/M/1//N retrial queue // In: Thomas, N., Forshaw, M. (eds.) ASMTA 2017. LNCS. 2017. Vol. 10378. P. 81–94.
4. *Fishwick P. A.* SimPack: Getting Started With Simulation Programming In C And C++ // In 1992 Winter Simulation Conference. P. 154–162.
5. *Kumar R., Jain N., Mokdad L., Som B.* Optimization of an $M/M/1/N$ Feedback Queue with Retention of Reneged Customers // Operations Research and Decisions. 2014. Vol. 24. P. 45–58.
6. *Nazarov A., Phung-Duc T., Paul S.* Heavy outgoing call asymptotics for MMPP/M/1/1 retrial queue with two-way communication // in: A. Dudin, A. Nazarov, A. Kirpichnikov (Eds.), Information Technologies and Mathematical Modelling. Queueing Theory and Applications. Vol. 800. Springer International Publishing, Cham, 2017. P. 28–41.
7. *Sztrik J., Tóth Á.* Analysis of retrial queueing system with two-way communication in different scenarios using simulation // in: A. Dudin, A. Nazarov, A. Moiseev (Eds.), Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Springer Nature Switzerland, Cham, 2023. P. 138–148.

Ádám Tóth — PhD, assistant professor, Department of Informatics Systems and Networks. E-mail: *toth.adam@inf.unideb.hu*

János Sztrik — Doctor of the Hungarian Academy of Sciences, full professor, Department of Informatics Systems and Networks. E-mail: *sztrik.janos@inf.unideb.hu*

Интеллектуальные системы, информационные технологии и программная инженерия

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНЫХ АДАПТИВНЫХ ПРОГРАММ В ПОЛИКУЛЬТУРНОЙ ИНФОРМАЦИОННО-ОБРАЗОВАТЕЛЬНОЙ СРЕДЕ

О. В. Блинова[1], Ю. В. Таратухина[1]

[1] *Институт проблем управления им. В. А. Трапезникова РАН, г. Москва, Россия*

Активное внедрение и использование систем искусственного интеллекта (ИИ) и информационно-коммуникационных технологий (ИКТ) в образовательный процесс требует осмыслить проблему этической и коммуникативной релевантности его работы в процессе управления обучением в информационных образовательных средах (ИОС). Соблюдение этических, культурных и нравственных норм особенно важно в рамках образовательной деятельности в поликультурной информационно-образовательной сфере. Вопрос обеспечения безопасной и комфортной среды достаточно остро стоит на повестке дня и требует междисциплинарного комплексного подхода. Необходимо обучить языковые модели учитывать индивидуально-психологические и культурно-когнитивные особенности обучающихся для оптимизации управления учебным процессом в поликультурных ИОС. В статье рассматриваются шаги для адаптации существующих программ под различные культурные профили. **Ключевые слова:** *этика искуственного интеллекта, большие языковые модели, культурный профиль искуственного интеллекта,локальная адаптация интеллектуальных систем*

Введение

Одна из основных задач разработчиков и пользователей систем искусственного интеллекта – научить их вести диалог с учетом общечеловеческих и культурно-специфичных принципов морали и этики для предотвращения серьезных негативных последствий. Существует набор этических принципов: приватность, прозрачность, надёжность, подотчетность, безопасность, контролируемость, и в идеале системы, исполь-

Исследование выполнено в рамках научной программы Национального центра физики и математики, направление № 9 «Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах

зующие ИИ, должны их обеспечивать. Важно уделять внимание предотвращению проблем на стадии проектирования информационного сервиса, использующего ИИ.

При подготовке интеллектуальных образовательных сервисов этический профиль людей, участвующих в разработке, будет влиять на конечный продукт. Возможность гибкой адаптации под особенности каждого конкретного человека — это уникальная возможность, открывающая большие перспективы для более комфортного и качественного обучения и в целом обеспечения более высокого уровня жизни пользователей. Научить ИИ этическому общению в рамках этнокультурных и религиозных особенностей возможно, задав ему основные алгоритмы и скрипты коммуникации, характерные для разных культур: базовые ценности, специфика дискурса, табу, религиозные табу, стереотипы, предрассудки и т. п. [1, 2].

1. Предварительные этапы создания интеллектуальных образовательных систем с учетом культурного профиля

Говоря об этичном ИИ, мы говорим о наличии в нем дополнительных сервисов пред- и постмодерации, предотвращающих вывод потенциально опасных результатов. Система человеческих ценностей сложна и часто может быть интерпретирована в зависимости от различных социальных контекстов, и здесь немалую роль играет человеческий фактор [3, 4]. Решение вопроса этичности и культурности ИИ, видимо, должно состоять из следующих шагов:

1. Обучение ИИ общечеловеческим культурным нормам и правилам. Здесь два основных пути: попытка выработать свод правил и формальных признаков или итерационный путь обучения. На выходе языковые модели должны быть обучены так, чтобы они могли коммуницировать с учетом культурного контекста по методу code switching (переключение кода). Когнитивное расширение может позволить, с одной стороны, действовать в рамках принятых для данной культуры норм и моделей коммуникации, а с другой стороны, оперативно пополнять пул стратегий коммуникации.

2. Способы выявления и учета индивидуальных особенностей по типу составления индивидуального профиля каждого пользователя. Здесь тоже могут быть вопросы как технического, так и этического или законодательного характера (хранение по сути «досье», характеристики личности). Нельзя допустить дискриминации по гендерному, возрастному и расовому признакам.

3. Обучение и выработка способов общения, основанного на понимании существования различных групп людей. Здесь может пригодиться опыт работы психотерапевтов, которые основываются на безусловном уважении личности человека независимо от его убеждений. В отличие от первого пункта здесь нужно формирование понимания осторожных, болезненных тем и особой осторожности в высказываниях, а не только запрета определенных идей [5–7].

Однако в кросс-культурных коммуникациях для достижения уровня этического общения требуются дополнительные компетенции разработчиков, учет различных культурных профилей, где помимо общей безопасности систем ИИ с точки зрения потенциально возможного вреда для человека и универсальных этических норм будут учитываться еще и культурные параметры. Можно выделить ряд параметров, которые лежат в основе анализа культурной специфики поведения, мышления и деятельности и которые будут обуславливать собственно специфику культурно-когнитивного профиля личности: восприятие времени, пространства и, как следствие, специфика деятельности; представления информации, мышления и внимания; социальных коммуникаций; доминирующие ценности. В любой коммуникации можно условно выделить когнитивный, эмоциональный и операционный компоненты. Когнитивный компонент будет включать в себя особенности когнитивного стиля и способов работы с информацией. Операционный компонент – отношение ко времени, к инструкциям, к целеполаганию, место в социуме, отношение к внешней среде, к командной работе. Также будет значимым эмоциональный компонент, как контекст, детерминирующий когнитивный и операционный компоненты (контекстуальные параметры).

2. Учет культурных профилей в образовательных программах

С появлением в учебных заведениях информационно-образовательной среды появилась техническая возможность персонифицированного подхода к образовательной траектории обучающегося и отслеживанию его прогресса. Дизайн и архитектура поликультурной ИОС позволяют адаптировать контент и интерфейс под культурно-когнитивные особенности обучающихся и формировать его динамическую компетентностную модель. Технологии ИИ способствуют инклюзии лиц с ограниченными возможностями в учебный процесс, а также данные технологии способствуют быстрой адаптации учащихся из других регионов, стран и культур. Более того, сбор и анализ данных, персонализированные треки и электронное портфолио, рекомендательные сервисы и т. п. [8, 9]. В

данном контексте речь идет о возможности технической адаптации под индивидуальные особенности интерфейса, контента и способа реализации учебного материала. При проектировании поликультурной ИОС ориентируемся на следующие три блока различий: специфика эргономического дизайна, специфика организации и структурирования контента, специфика образовательной коммуникации. Предполагается, что данная ИОС должна обладать «культурным интеллектом», то есть иметь способность отображать различные интерфейсы и стратегии подачи материала в зависимости от культурных особенностей обучаемого.

Система должна уметь интерпретировать поведение учащегося в соответствии с его культурной группой и в результате «предлагать» ему ту или иную программу обучения и соответствующий интерфейс [10]. В первую очередь осуществляется интерпретация культурных профилей на основе информации из базы знаний о типах мышления и деятельности в разных культурах и консультаций с профильными специалистами, вырабатываются соответствующие учебные профили. Далее полученные данные передаются в модуль выбора учебной программы и сверяются с данными, получаемыми от студента при заполнении анкет и выполнении заданий. Получаемые от студента данные постоянно поступают обратно в систему для обработки и дополнения как самих культурных профилей, так и для проверки правильности отнесения обучающегося к той или иной группе. На первых итерациях на каждом шаге требуется модерация человеком и проверка получаемых результатов. Предполагается, что за несколько итераций интеллектуальная система сможет подбирать культурный профиль для студента с минимальным контролем со стороны человека (рис. 1).

Заключение

Таким образом, ИИ в поликультурных ИОС может выполнять роль ассистента и ментора в управлении учебным процессом, то есть брать на себя задачи по адаптации учащегося в учебный процесс, отслеживании его состояния и навыков, обращении к педагогическому персоналу в случае возникновения сложно интерпретируемых проблем. Предлагаемый алгоритм не является примером конкретного решения, а является общим планом шагов для учета культурных профилей. Выбор конкретных интрументов ИИ, методов и критериев оценки результатов, способов обработки обратной связи и т. д. должны быть привязаны к конкретной программе, платформе обучения, квалификации сотрудников и доступности необходимых специалистов и других факторов. Предложенный авторами алгоритм является общей схемой, руководствуясь которой специалисты могут выбирать методики для реализации кон-

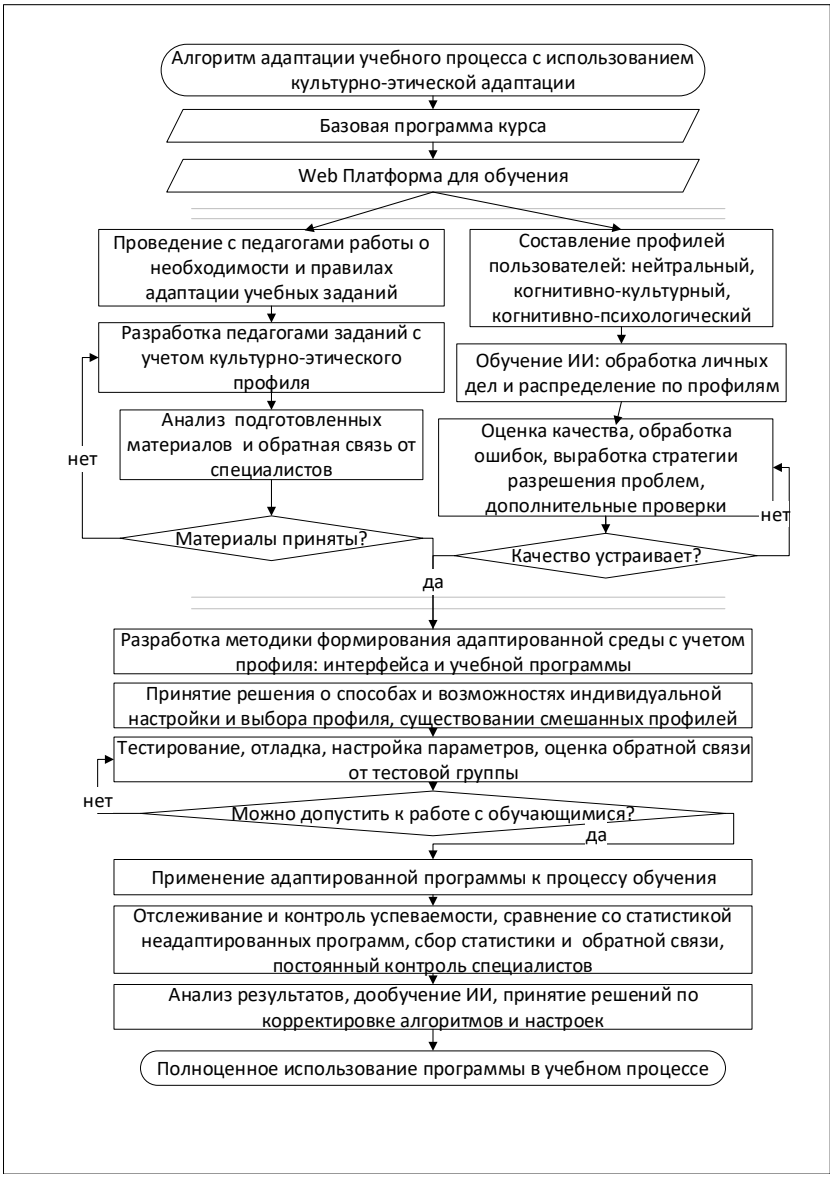


Рис. 1. Алгоритм управления обучением в поликультурной ИОС

кретных шагов. Необходимо продолжать работу в данном направлении, подключать адаптацию под особенности физического и психического характера, собирать обратную связь, корректировать алгоритмы и методы. Только с помощью итерационной работы и объединения специалистов разных специальностей можно научить ИИ этичному общению и «развить культурный интеллект». Необходимо привлекать к данной работе специалистов по нейронным сетям, ИИ, адаптивным интерфейсам, педагогов, методистов, этнопсихологов, психологов, инженеров, тестировщиков и разработчиков.

СПИСОК ЛИТЕРАТУРЫ

1. *Блинова О. В., Таратухина Ю. В.* Вопросы соблюдения законодательных и культурно-этических норм системами искусственного интеллекта // Материалы 6-й Всероссийской конференции «Системы управления, информационные технологии и математическое моделирование» (СУИТММ'2024, Омск). Омск: Колос-с, 2024. Т. 5. С. 296–300.
2. *Блинова О. В., Таратухина Ю. В.* Проблема нравственный и этических аспектов искусственного интеллекта // Новые информационные технологии в исследовании сложных структур : материалы Пятнадцатой Международной конференции, 16–20 сентября 2024 г. – Томск : Издательство Томского государственного университета, 2024. 94 с.
3. *Bostrom N.* Strategic implications of openness in AI development //Artificial intelligence safety and security. Chapman and Hall/CRC, 2018. P. 145–164.
4. *Nguyen A., Ngo H. N., Hong Y., Dang B., Nguyen B.P.* Ethical principles for artificial intelligence in education. Education and information technologies, Vol. 28(4). P. 4221–4241.
5. *Разин А. В.* Этические проблемы искусственного интеллекта и цифровых технологий // Философская аналитика цифровой эпохи: сборник научных статей / отв. ред. Л.В. Шиповалова, С.И. Дудник. СПб.: Изд-во СПб. ун-та, 2020. С. 154–163.
6. *Былевский П. Г.* Социально-культурные риски мультимодальных больших генеративных моделей «искусственного интеллекта»(GenAI) // Культура и искусство. 2024. № 6. С. 213–224.
7. *Пушкарев А. В.* Мировоззренческий и социокультурный аспект идеи искусственного интеллекта // Исторические, философские, политические и юридические науки, культурология и искусствоведение. Вопросы теории и практики. 2014. № 8-1. С. 136–138.
8. *Holmes W., Porayska-Pomsta, K., Holstein K., Sutherland E., Baker T., Shum S. B., Koedinger K. R.* Ethics of AI in education: Towards a community-wide framework // International Journal of Artificial Intelligence in Education, 2022. P. 1–23.

9. *Кладько С. С.* Уроки кросс-культурного менеджмента в преодолении предкризисной ситуации в этике в сфере искусственного интеллекта // Информационное общество: образование, наука, культура и технологии будущего. 2024. Т. 1, № 8. С. 67–79.

10. *Ряпина Н. Е., Трубина И. И.* Интерпретация студентами этических аспектов искусственного интеллекта в образовании // Отечественная и зарубежная педагогика. 2025. Т. 1, № 1. С. 101–111.

Блинова Ольга Викторовна — научный сотрудник, лаб № 17. E-mail: *blinova_olga_v@mail.ru*

Таратухина Юлия Валерьевна — к.ф.н., старший научный сотрудник, лаб № 17. E-mail: *jvt6@ipu.ru*

ЭФФЕКТИВНОСТЬ АДАПТИРОВАННЫХ НЕЙРОСЕТЕЙ В ДИАГНОСТИКЕ ПЕРЕЛОМОВ

Е. Е. Голденок[1,2], Н. А. Лукьянова[1,2], И. С. Лихачев[1]

[1] *Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого, г. Красноярск, Россия*
[2] *Сибирский федеральный университет, г. Красноярск, Россия*

Проведено сравнительное исследование эффективности адаптированных предобученных сверточных нейронных сетей для автоматизированной диагностики переломов костей конечностей по рентгеновским снимкам. В качестве материала исследования использована объединенная база рентгеновских снимков, содержащая 5720 изображений из открытого репозитория Kaggle. Для решения проблемы недостаточности данных применены методы аугментации, позволившие увеличить набор данных до 21 173 изображений (10 523 без патологии и 10 650 с переломом). Разработаны и реализованы модификации четырех архитектур сверточных сетей: AlexNet, MobileNetV2, EfficientNet-B0 и ResNet-18, адаптированные для задачи бинарной классификации переломов. Сравнительный анализ обученных моделей по показателю accuracy показал, что все архитектуры достигают точности свыше 96%, при этом наилучшие результаты продемонстрировали ResNet-18 и EfficientNet-B0. Полученные результаты подтверждают перспективность использования глубокого обучения для автоматизации первичного скрининга рентгеновских снимков при диагностике переломов.

Ключевые слова: *сверточные нейронные сети, распознавание переломов костей, классификация, трансферное обучение*

Введение

Переломы костей представляют собой серьезную медико-социальную проблему, являясь одной из ведущих причин временной нетрудоспособности и инвалидности. По данным ВОЗ ежегодно регистрируется около 5 миллионов летальных исходов, связанных с травмами, что составляет почти 9% от от общей смертности, травмы являются основной причиной смерти в возрастной группе от 1 до 40 лет [1].

Традиционные методы диагностики, основанные на визуальном анализе рентгенограмм, КТ и МРТ, обладают рядом ограничений, включая субъективность оценки, зависимость от квалификации специалиста и высокую вероятность ошибок при сложных переломах.

Применение технологий искусственного интеллекта, в частности сверточных нейронных сетей (CNN, Convolutional Neural Network) в сфере медицины, открывает новые возможности для автоматизации диагностики. Современные исследования показывают, что медицинские технологии активно используют методы машинного обучения для анализа данных пациентов, включая медицинские изображения, биометрические показатели [2]. Глубокое обучение способно выполнять сложную интерпретацию на уровне специалистов здравоохранения, однако их клиническое применение сталкивается с рядом проблем, включая недостаток размеченных данных, высокие требования к вычислительным ресурсам и сложности интеграции с медицинскими информационными системами [3]. Это имеет особую значимость в контексте распознавания переломов, так как подобные повреждения могут принимать различные формы и локализации, что затрудняет их идентификацию на рентгеновских снимках. Несмотря на то, что задача детекции переломов является активно изучаемой, проведение сравнительного анализа эффективности и практической применимости адаптированных версий современных архитектур в условиях ограниченных вычислительных ресурсов сохраняет свою актуальность для потенциального внедрения в клиническую практику.

Цель исследования: сравнительный анализ эффективности адаптированных предобученных нейронных сетей для распознавания наличия или отсутствия переломов на рентгеновских снимках.

1. Материалы и методы исследования

Материалом для исследования послужили рентгеновские снимки из двух открытых наборов данных репозитория Kaggle [4, 5], содержащие изображения костей верхних и нижних конечностей. Объединенный датасет включает 5 720 аннотированных изображений (2 667 без патологий и 3 053 с переломами).

Выполнены предобработка (препроцессинг) и аугментация данных. К этапу предобработки относились: конвертация в градации серого, центральное кадрирование, нормализация значений пикселей. Аугментация данных включала аффинные преобразования (повороты на $\pm 20°$, горизонтальные отражения), коррекцию яркости и контрастности, что позволило решить проблему недостаточности данных и повысить устойчивость моделей к вариациям входных данных. Техническая реализация этих преобразований осуществлялась с использованием библиотек PyTorch и torchvision. После аугментации объем выборки увеличен до 21 173 изображений (10 523 без переломов и 10 650 с переломами), классы сбалансированные.

Для решения задачи бинарной классификации использованы четыре предобученные архитектуры CNN: AlexNet, MobileNetV2, EfficientNet-B0 и ResNet-18. Все архитектуры были модифицированы с учетом особенностей решаемой задачи, включая преобразование входного слоя для обработки одноканальных изображений, коррекцию классификатора и добавление механизмов регуляризации.

По оригинальной архитектуре AlexNet [6] выполнены следующие модификации:

- адаптация входа: был изменен первый сверточный слой для обработки одноканальных изображений ($1 \times 400 \times 400$), сохранена оригинальная стратегия инициализации весов;
- улучшения нормализации: были добавлены слои Batch Normalization после каждого сверточного слоя, удалена Local Response Normalization как устаревшая;
- модификация классификатора: уменьшена размерность полносвязных слоев, увеличен dropout (0,5) для лучшей регуляризации, выход адаптирован для бинарной классификации.

По оригинальной архитектуре MobileNetV2 [7] выполнены следующие модификации:

- кастомная головка классификатора: добавлен дополнительный полносвязный слой (128 нейронов), dropout 0,5, на выходе используется сигмоидная активация;
- стратегия обучения: полная заморозка предобученных слоев, использование предустановок ImageNet;
- аугментация данных: увеличенный диапазон поворотов ($\pm 20°$), коррекция яркости (диапазон $0{,}8 - 1{,}2$), горизонтальные отражения;
- оптимизация: размер входа 224×224, специфичная нормализация.

Оригинальная архитектура EfficientNet-B0 [8] основана на улучшенных MBConv-блоках с depthwise-свёртками и расширением каналов. В качестве активации применяется функция Swish, предложенная Google, которая повышает качество модели по сравнению с ReLU. В работе выполнены модификации:

- добавлен дополнительный скрытый слой, содержащий 256 нейронов, Batch Normalization, увеличенный dropout (0,3);
- стратегия обучения: частичное размораживание (после 5-го уровня), использование AdamW оптимизатора [6], L2 регуляризация (weight_decay=1e−4);
- балансировка классов: применение WeightedRandomSampler, вычисление весов классов;
- гиперпараметры: learning rate равен 0,0005, batch size = 32.

По оригинальной архитектуре ResNet-18 [9] выполнены следующие модификации:

- адаптация входа: сделана модификация первого слоя для grayscale, инициализация весов средним по каналам;
- стратегия обучения: частичное размораживание (последние 5 слоев), механизм ранней остановки (patience= 3);
- классификатор: добавлен скрытый слой (256 нейронов), dropout 0,3, ReLU-активация;
- оптимизация: взвешенная функция потерь, L2 регуляризация.

Анализ эффективности моделей проводился на основе ключевых метрик: точности (Accuracy), F1-меры, доли ложноположительных (FPR) и ложноотрицательных (FNR) распознований переломов, а также времени обучения и вычислительных требований.

2. Результаты и обсуждения

В табл. 1 приведены результаты сравнения эффективности четырех модифицированных архитектур CNN для выявления переломов костей по рентгенографическим изображениям.

Таблица 1

Сравнение эффективности моделей машинного обучения

Показатели	AlexNet	MobileNetV2	EfficientNet-B0	ResNet-18
Accuracy, %	99,25	96,01	99,37	99,76
F1-score, %	99,25	96,00	99,35	99,80
FNR, %	0,60	2,89	0,40	0,30
FPR, %	0,90	5,02	0,84	0,19
Эпохи, шт	15	25	8	15
Время,мин	780	90	300	480

Модель ResNet-18 показала наивысшую точность (99,76%) и минимальные значения доли ложноположительных детекций FPR= 0,19% и доли ложноотрицательных FNR= 0,3% (пропущенные переломы), что делает её наиболее подходящей для клинического применения, где критически важна минимизация диагностических ошибок. EfficientNet-B0 при значительно меньшем времени обучения по сравнению с ResNet-18 и AlexNet (5 часов против 8 и 13 часов соответственно) продемонстрировала близкую к ResNet-18 точность (99,37%), что делает её предпочтительной для внедрения в медицинских учреждениях с ограниченными вычислительными ресурсами. AlexNet обеспечила высокую точность (99,25%), однако требовала 13 часов обучения, что снижает её

практическую привлекательность по сравнению с более современными архитектурами. MobileNetV2 показала более низкие метрики точности (96,01%) при наименьших вычислительных затратах — обучение завершилось за 90 минут (25 эпох), что указывает на её потенциальную применимость в системах с ограниченными вычислительными ресурсами, однако относительно низкая специфичность (94,98%) может привести к увеличению числа исследований, требующих верификации врачом, что следует учитывать при внедрении.

Модифицированные нейросети демонстрируют особую эффективность при анализе сложных случаев (рис. 1), таких как, микротрещины и скрытые переломы, переломы с артефактами и низкой контрастностью, атипичные локализации повреждения.



(а) (б) (в)

Рис. 1. Примеры сложных случаев переломов: (а) — микротрещина диафиза, (б) — перелом с артефактами, (в) — атипичная локализация повреждения

Все модели, особенно ResNet-18 и EfficientNet-B0, показали высокую чувствительность к минимальным повреждениям, которые трудно обнаружить визуально (рис. 1а). Архитектуры с глубокими слоями (например, AlexNet) успешно фильтруют шумы и выделяют значимые паттерны (рис. 1б) на снимках с артефактами. Благодаря автоматическому извлечению признаков, нейросети выявляют переломы в нетипичных областях, где ручные методы часто ошибаются (рис. 1в).

3. Заключение

В ходе работы были модифицированы и обучены четыре модели сверточных нейронных сетей для автоматизированного распознавания переломов на рентгеновских снимках. Проведенные эксперименты показали, что модели ResNet-18 и EfficientNet-B0 демонстрируют наилучшие результаты, достигая точности свыше 99%, и особенно хорошо справляются со сложными случаями диагностики, включая микротрещины и изображения с артефактами. Полученные результаты подтверждают

возможность практического применения данных архитектур в клинической практике в качестве вспомогательного инструмента для первичного скрининга рентгеновских снимков.

СПИСОК ЛИТЕРАТУРЫ

1. Тлевцежев З. Х. Травматология Майкоп: Изд-во МГТУ, 2015. 92 с.
2. Переломова П. А. Анализ алгоритмов ML для выявления патологий на основе медицинских данных // Вестник науки. 2025. Т. 1, № 3(84). С. 526–531.
3. Smith J., Brown R. Deep Learning in Fracture Detection: A Systematic Review // Journal of Medical AI. 2023. Vol. 10. P. 112–125.
4. https://www.kaggle.com/datasets/vuppalaadithyasairam/bone-fracture-detection-using-xrays — Открытый репозиторий Kaggle
5. https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data — Открытый репозиторий Kaggle
6. Krizhevsky A., Sutskever I., Hinton G. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems (NIPS). 2012. P. 1097–1105.
7. Sandler M., Howard A., Zhu M. [et al.] MobileNetV2: Inverted Residuals and Linear Bottlenecks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 4510–4520.
8. Tan M., Le Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of the 36th International Conference on Machine Learning (ICML). 2019. P. 6105–6114.
9. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778.

Голденок Елена Евгеньевна — к.ф.-м.н., доцент, доцент кафедры медицинской кибернетики и информатики, КрасГМУ, доцент кафедры высшей и прикладной математики, СФУ. E-mail: *egoldenok@gmail.com*

Лукьянова Наталья Александровна — к.ф.-м.н., доцент кафедры медицинской кибернетики и информатики, КрасГМУ, доцент кафедры высшей математики 2, СФУ. E-mail: *nalukyanovakrsk@gmail.com*

Лихачев Иван Сергеевич — студент 6 курса специальности 30.05.03 Медицинская кибернетика, КрасГМУ. E-mail: *ivanlihachev321@gmail.com*

ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ АЛГОРИТМОВ РЕШЕНИЯ ЗАДАЧИ КОРРЕЛЯЦИОННОЙ КЛАСТЕРИЗАЦИИ

Э. И. Ибрагимова[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

В работе рассматривается ранее предложенная техника построения алгоритмов решения задачи корреляционной кластеризации, основанных на структуре графа с точки зрения вычислительной сложности. Техника позволяет конструировать новые алгоритмы такого типа из модулей, имитирующих поведение известных алгоритмов. В работе проанализирована вычислительная сложность модулей и их комбинации. **Ключевые слова:** *задача корреляционной кластеризации, вычислительная сложность*

Введение

Задача корреляционной кластеризации (ССР) заключается в поиске оптимальной кластеризации на заданном знаковом графе. При этом количество кластеров заранее не задано, а функционал ошибки чаще всего представляется в виде суммы межкластерной и внутрикластерной ошибок. Рассматриваемая задача является NP-трудной [4].

Алгоритмы решения ССР разделяют на два подхода [9]. В первом задача рассматривается в терминах математического программирования, а во втором в теоретико-графовой постановке. Второй подход включает в себя алгоритмы, использующие матрицу графа и алгоритмы, основывающиеся на структуре графа (NS-алгоритмы).

Данная работа продолжает развитие предложенного системного подхода к построению и анализу NS-алгоритмов [1, 2]. Подход позволяет комбинировать различные модули и получать новые NS-алгоритмы, а также облегчает доказательство вычислительной сложности таких алгоритмов. Далее будет представлен анализ вычислительной сложности алгоритмов, сконструированный с помощью рассматриваемай техники.

1. Постановка задачи

В работе исследуются знаковые графы вида $\Sigma = (G, \sigma) = (V, E, \sigma)$, где $G = (V, E)$ — неориентированный, невзвешенный граф без кратных ребёр и петель с множеством вершин V, $|V| = n \geq 2$, и множеством рёбер E, $|E| = m \geq 1$.

Обозначим множество вершин смежных с v как $\Gamma(v) = \{u\colon (v,u) \in E\}$. Под степенью вершины v традиционно понимается число рёбер, инцидентных ей. Очевидно, что степень вершины v можно представить в виде $\delta(v) = |\Gamma(v)|$. Тогда степенью графа будет $\Delta = \max\limits_{v \in V} \delta(v)$. На ребрах $(u,v) \in E$ графа G задана функция знака $\sigma : E \to \{+,-\}$, которая порождает разбиение множества рёбер графа $E = E^+ \cup E^-$, где E^+ — множество положительных, E^- множество отрицательных рёбер. Тогда $\Sigma^+ = (V, E^+, \sigma)$ и $\Sigma^- = (V, E^-, \sigma)$ будут знаковыми графами с только положительными и только отрицательными ребрами соответственно.

Под положительной или межкластерной ошибкой $P(\mathcal{C})$ разбиения $\mathcal{C}$ будем понимать число положительных ребер между кластерами.

Под отрицательной или внутрикластерной ошибкой $N(\mathcal{C})$ будем понимать число отрицательных ребер внутри кластеров. В [7] авторы предлагают представлять суммарную ошибку в виде выпуклой комбинации положительной и отрицательной ошибок, зависящей от параметра $\alpha \in [0,1]$:

$$Q_\alpha(\mathcal{C}) = \alpha N(\mathcal{C}) + (1-\alpha)P(\mathcal{C}). \tag{1}$$

Заметим, что функционал ошибки (1) всегда удовлетворяет неравенству

$$0 \le Q_\alpha \le \alpha|E^-| + (1-\alpha)|E^+|.$$

В работе задача кластеризации знакового графа рассматривается в следующей постановке [6].

Correlation Clustering problem (CCP)

Условие: задан знаковый граф $\Sigma = (G, \sigma)$, где $G = (V, E)$ — неориентированный граф, $n = |V| \ge 2$, $m = |E| \ge 1$.

Вопрос: для заданного $\alpha \in [0,1]$ требуется найти разбиение $\mathcal{C}$ множества вершин V знакового графа Σ с минимальной суммарной ошибкой $Q_\alpha(\mathcal{C})$.

2. Вычислительная сложность сконструированных алгоритмов

Схема построения NS-алгоритмов 1 состоит из шести блоков, которые выполняются последовательно пока не будет выполнено условие останова.

В работе [2] с помощью данной схемы были проанализированы следующие NS-алгоритмы решения CCP: *relocation heuristic* (RH) [6], *tabu search* (TS) [5], *variable neighborhood search* (VNS) [5], $KwikCluster$ [3], *iterated local search* (ILS) [8], $SGClust_\alpha$ [2] и $CarVeR$ [2]. Для каждого

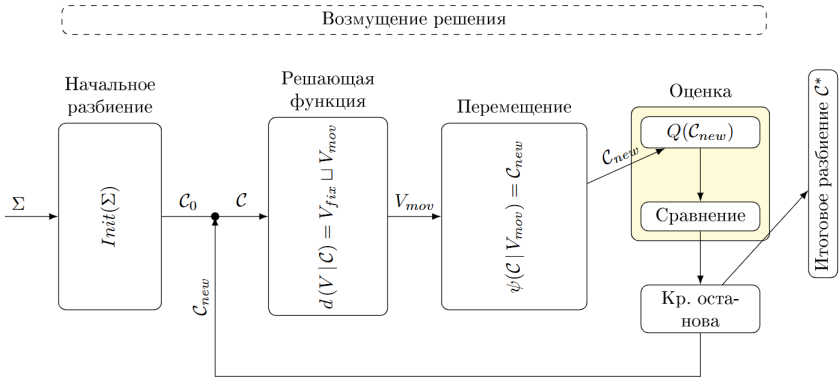


Рис. 1. Общая схема алгоритмов, основанных на структуре графа

блока были выделены модули, имитирующие поведение проанализированных NS-алгоритмов. В таблицах 1 и 2 представлена вычислительная сложность однократного повторения рассмотренных модулей.

Итоговая вычислительная сложность сконструированного алгоритма зависит от вычислительной сложности каждого использованного модуля, количества повторений каждого блока и заданных параметров (если такие задаются для модулей).

Пусть вычислительная сложность выбранного модуля для блока «Начальное разбиение» будет C_{init}, для блока «Решающая функция» — $C_{decision}$, для блока «Перемещение» — $C_{relocation}$, для блока «Оценка» — C_{eval}, для блока «Возмущение решения» — C_{pert}. Если в алгоритме отсутствует какой-либо блок, то можно считать, что его вычислительная сложность равна 0.

Вычислительная сложность одной итерации сконструированного алгоритма $\mathbf{C}_{iter}$ будет суммой сложностей его модулей:

$$\mathbf{C}_{iter} = C_{decision} + C_{relocation} + C_{eval} + C_{pert}.$$

Если критерием останова алгоритма являются время или количество итераций, то общая сложность $C_{general}$ будет:

$$C_{general} = C_{init} + x \cdot C_{iter},$$

где x — либо заданное количество итераций, либо произведенное за заданное время. Если критерием останова является посещение всех вершин, то общая сложность будет:

$$C_{general} = C_{init} + n \cdot C_{iter}.$$

Таблица 1

Вычислительная сложность модулей алгоритмов. Часть 1

Блок	**Модуль**	**Сложность**	**Источник**
Начальное разбиение	Тривиальное — случайное разбиение	$O(n)$	*RH*
	Готовое разбиенние — результат работы другого алгоритма	$O(1)$	*TS, VNS*
	Компоненты связности — компоненты в Σ^+	$O(n+m)$	$SGClust_\alpha$, *CarVeR*
	Ранжирование и срез	$O(n^2 \log(n)m)$	*ILS*
	KwikCluster	$O(nm)$	*KwikCluster*
Решающая функция	Тривиальная функция — возвращает все вершины	$O(1)$	*RH, VNS, ILS*
	Табу — возвращает вершины, которые не находятся в списке табу, в который помещаются затронутые вершины на заданное количество итераций	$O(n)$	*TS*
	Отрицательная ошибка — возвращает все вершины, которые вносят вклад в отрицательную ошибку	$O(nm)$	$SGClust_\alpha$
	Потенциальная функция — возвращает вершины, для которых значение некоторой потенциальной функции удовлетворяет заданному условию	$O(PC \cdot n)$	*CarVeR*

В случае многократного перезапуска алгоритма, $C_{general}$ будет домножаться на количество запусков. Также подразумевается, что каждый блок реализуется в алгоритме единожды.

Таблица 2

Вычислительная сложность модулей алгоритмов. Часть 2

Блок	**Модуль**	**Сложность**	**Источник**
Перемещение	По вкладу — вершина, которая вносит наибольший вклад в отрицательную ошибку	$O(n^2m)$	$SGClust_\alpha$, $CarVeR$
	По целевой функции — вершина, при перемещении которой, целевая функция минимальна	$O(n^2m)$	*RH, TS, VNS*
	Перемещение r вершин — $1 \leq r \leq r_{max}$ вершин, при одновременном перемещении которых, целевая функция минимальна	$O(n^{2r}m)$	*ILS*
Оценка	Оценка всех вершин	$O(m)$	-
	Оценка в окрестности вершины	$O(\|\Delta\|)$	-
Критерий останова	Время — остановка по истечению заданного количества времени	$O(1)$	*RH, TS, VNS, ILS*
	Итерации — остановка после заданного количества итераций	$O(1)$	*ILS*
	Посещение вершин — остановка после посещения всех вершин	$O(1)$	$KwikCluster$, $SGClust_\alpha$, $CarVeR$
Возмущение решения	Перемешивание вершин — заданное число вершин перемещается случайным образом	$O(n)$	*VNS, ILS*

Заключение

Проведен анализ вычислительной сложности NS-алгоритмов, который показывает, что сложность конкретного NS-алгоритма может быть легко получена из сложностей использованных модулей. Одним из направлений развития исследования является дополнение списка модулей.

Также рассматриваемая техника была реализована в виде программного комплекса. Комплекс состоит из описанных модулей, алгоритмов генерации знаковых графов и содержит функционал для анализа результатов. Он позволяет конструировать NS-алгоритмы, тестировать на сгенерированных тестовых графах и сравнивать результаты работы алгоритмов.

СПИСОК ЛИТЕРАТУРЫ

1. *Ibragimova E. I., Semenova D. V., Soldatenko A. A.* A Technique of Algorithms Construction for Correlation Clustering Problem // Discrete and Continuous Models and Applied Computational Science. In press.
2. *Солдатенко А. А., Семенова Д. В., Ибрагимова Э. И.* Подход к анализу и построению алгоритмов решения одной задачи кластеризации на знаковых графах // Прикладная дискретная математика. 2024. № 64. С. 112–127.
3. *Ailon N., Charikar M., Newman A.* Aggregating inconsistent information: Ranking and clustering // Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing, 2008, P. 684–693.
4. *Bansal N., Blum A., Chawla S.* Correlation Clustering // Machine Learning, 2002, №. 56. P. 89–113.
5. *Brusco M. J., Doreian P.* Partitioning signed networks using relocation heuristics, tabu search, and variable neighborhood search // Soc. Networks, 2019, №. 56. P. 70–80.
6. *Doreian P., Mrvar A.* A partitioning approach to structural balance // Social Networks, 1996, №. 18. P. 149–168.
7. *Doreian P., Mrvar A.* Structural balance and partitioning signed graphs // In A. Ferligoj & A. Kramberger (Eds.), Developments in data analysis. Metodoloskizvezki 12. Ljubljana, Slovenia: FDV, 1996, P. 195–208.
8. *Levorato, M., Figueiredo, R., Frota, Yu., Drummond, L.* Evaluating balancing on social networks through the efficient solution of correlation clustering problems // EURO Journal on Computational Optimization, 2017. Vol. 4, №. 5. P. 467–498.
9. *Wahid D. F., Hassini E.* A Literature Review on Correlation Clustering: Cross-disciplinary Taxonomy with Bibliometric Analysis // Oper. Res. Forum. 2022. Vol. 47, №. 3.

Ибрагимова Эллада Ибрагимовна — аспирант, Сибирский федеральный университет. E-mail: *ibragimovaei@mail.ru*

ОБРАБОТКА ГАММА-СПЕКТРОМЕТРИЧЕСКИХ ПОЛЕВЫХ ИЗМЕРЕНИЙ В ГЕОФИЗИКЕ С ПРИМЕНЕНИЕМ ПОНЯТИЯ ИНТЕРВАЛОВ СТАЦИОНАРНОСТИ

С. Г. Катаев[1], С. С. Катаева[2], А. А. Миллер[3], Е. Ю. Ермолин[4]

[1] *Томский государственный педагогический университет, г. Томск, Россия*
[2] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*
[3] *АО НПП ВНИИГИС, г. Санкт-Петербург, Россия*
[4] *ООО «ДЖИ М Сервис», г. Санкт-Петербург, Россия*

Рассматриваются особенности обработки гамма-спектроскопических данных алгоритмами метода выявления структур в различных геофизических исследованиях. Представленные в работе результаты демонстрируют широкие возможности метода при анализе данных. **Ключевые слова:** *гамма-спектрометр, гамма-каротаж, графы, поток случайных событий, метод выявления структур*

Введение

В настоящее время при геофизических исследованиях широко применяются многоканальные спектрометрические методы. Наибольшее распространение в скважинных измерениях получили такие спектрометрические методы, как спектрометрический гамма-каротаж и спектрометрический импульсный нейтронный гамма-каротаж. Анализ потока гамма квантов как при погружении датчика в скважину, так и при перемещении его по поверхности, позволяет выделять литологические границы в разрезе с привязкой их по глубине, либо определять пространственные границы различных пород. Главным требованием к обработке и интерпретации данных спектрометрических методов каротажа является точность определения концентраций элементов породы, устойчивость алгоритмов и отсутствие систематических погрешностей.

Все геофизические методы имеют статистический разброс при измерениях, но именно для радиоактивных измерений из-за специфики регистрации сигнала статистическая ошибка намного больше других составляющих результирующей ошибки измерения. Основная трудность при обработке спектрометрических данных заключается в необходимости накопления необходимого числа импульсов для получения заданной

погрешности измерения. Так, прямые эксперименты показали, что локальная аномалия не проявляется при непрерывном движении прибора, но если осуществлять съемку квазипоточечно, останавливаясь каждые полметра на 10 секунд, то аномалия видна.

Это обстоятельство, плюс требование оперативности интерпретации, что немаловажно для промышленного использования, приводят к необходимости разработки новых подходов, позволяющих осуществлять экспресс анализ данных в автоматическом режиме.

1. Подход к обработке гамма-спектроскопических данных

Можно попытаться улучшить результаты обработки данных измерений, если задаться априорным предположением, что существуют интервалы (пространственные или временные), в пределах которых параметры гамма-поля не меняются. Геологически это будет соответствовать пластам определенных пород. Назовем такие интервалы **интервалами стационарности**, то есть интервалами, в пределах которых случайный процесс поступления гамма-квантов в детектор можно считать стационарным. Но статистическая ошибка измерений при таком подходе должна быть меньше, чем разница между отдельными пластами, границы которых мы пытаемся зафиксировать. Введение понятия ИС оправдано, учитывая, что в каждой точке измерения поток радиоактивных частиц можно считать пуассоновским, исходя из условий его формирования.

Такой подход был нами применен при анализе скважинных измерений [1], а также при наземных исследованиях [2] с использованием одного из алгоритмов метода выявления структур (МВС) [3]. А именно алгоритм APFLOW [4], который ранее нами применялся для определения параметров случайного потока событий, являющегося стандартным объектом в теории массового обслуживания. Он позволяет выявлять структуру потока, т. е. представлять исходный массив данных в виде последовательности ИС, со своим набором параметров.

Измеренные количества гамма-квантов в секунду при погружении датчика в скважину либо при перемещении его по поверхности представляют собой ряд случайных чисел $n_1, n_2, \ldots, n_m$. В каротажных исследованиях в каждой точке датчик принимает поток частиц, идущий из горизонтального слоя, и этот поток можно считать пуассоновским, исходя из условий его формирования. Если проводить измерения в точке определенное (не малое) время, то разности времен срабатывания датчика будут подчиняться пуассоновскому распределению. Это утверждение справедливо для любой точки, находящейся на произвольной глубине. Причем, число срабатываний датчика за единицу времени в

точках одного слоя будут примерно одинаковыми, т. к. мы фиксируем интенсивности $n_1, n_2, \ldots, n_m$, фактически, одного и того же пуассоновского потока, полученные для разных его участков.

Таким образом, задача определения границ слоев при каротажных измерениях, может быть сформулирована следующим образом. Найти номера отсчетов $m_j, j = 1, \ldots, k$, где k — количество слоев, разделяющие различные ИС, где ИС — последовательность идущих подряд номеров ряда, интенсивности при которых примерно одинаковы, а суммарная дисперсия выделенных таким образом ИС — минимальна. То же и в случае перемещения датчика по поверхности.

2. Метод нахождения интервалов стационарности

Для нахождения ИС мы используем МВС. Для этого необходимо и достаточно построить матрицу отношений между объектами. В качестве объектов в задаче выступают номера отсчетов $i = 1, \ldots, m$, а в качестве признаков — накопленные значения интенсивности t_i, $i = 1, \ldots, m$. Для того чтобы поток был упорядоченным, значения t_i строились рекуррентно, то есть

$$t_i = t_{i-1} + n_i \tag{1}$$

При этом, как видно из (1) разность между двумя соседними значениями равна просто n_i. В качестве характеристики отношения между объектами i и j удобно выбрать «расстояние» между ними, т.е. величину d_{ij} определяемую по (2).

$$d_{ij} = \begin{cases} \frac{t_{j+1} - t_{i-1}}{j-i+1} & : \ i \le j; \ i, j = 1, \cdots, N \\ 0 & : \ i > j; \ i, j = 1, \cdots, N \end{cases} \tag{2}$$

Определенные таким образом элементы d_{ij} образуют верхнетреугольную матрицу $D = [d_{ij}]_{m \times m}$. Тогда формализация МВС для данной задачи может быть изложена в следующей редакции. Сопоставим каждому i-му событию потока, $i = 1, ..., m$, вершину с тем же номером i некоторого графа $G(V, E)$, где V – множество вершин графа мощности N, а E – множество пар вершин, т.е. множество ребер графа. Стоимость каждого ребра графа, связывающего i-ю и j-ю ($i < j$) вершины – величинf d_{ij}, соответствующий элемент матрицы D. Мощность множества E составит $m(m-1)/2$. Поскольку МВС отыскивает на графе заранее заданные структуры, надо определить, что в данной задаче является структурой. Определение структуры можно сформулировать, исходя из понятия ИС. А именно, если группа событий с номерами $i_1^p, i_2^p, \ldots, i_k^p$ составляет участок стационарности с интенсивностью $\lambda = \lambda_p$, то это озна-

чает, что: во-первых, среднее значение стоимостей ребер, связывающих между собой соответствующие вершины, примерно равно интенсивности λ_p:

$$\frac{2}{k(k-1))} \sum_{(i,j)\in I} d_{ij} \cong \lambda_p, \tag{3}$$

где I – множество всевозможных упорядоченных по номерам пар из $\{i_1^p, i_2^p, \ldots, i_k^p\}$; во-вторых, номера i_j^p, идут подряд; в-третьих все $d_{ij}, i, j \in I$ принадлежат, в основном, некоторому интервалу $(\lambda_p \pm \Delta\lambda)$.

Оценка интенсивности, полученная по всей группе событий, т.е. (i_1^p, i_k^p)-я компонента матрицы D, должна давать наиболее точную оценку λ_p. Эти соображения определяют идею алгоритма выделения участков стационарности потока. При выборе удачного интервала $(\lambda_p \pm \Delta\lambda)$ и удалении из графа G всех ребер, стоимость которых не попадает в указанный интервал, в графе выделятся компоненты связности $G_k = (V_k, E_k)$, соответствующие участкам стационарности, относящимся к p-му состоянию МС-потока.

3. Пример

На территории Сибири была произведена пешеходная съемка с использованием гамма-спектрометра GM-GAMMA-1000. Всего было пройдено около 60 трасс с общим количеством точек порядка150000. В идеале, в каждой пространственной точке результатом измерения гамма-спектрометром является энергетический спектр, который показывает, сколько гамма-квантов было зафиксировано в определенном интервале энергии за единицу времени. Основные источники радиации — это изотопы атомов K, U и Th. В реальности, при больших объемах проводимых исследований получаемый спектр в отдельной пространственной точке не дает возможности определять аномальные содержания указанных радиоактивных элементов в силу небольшого числа поступивших импульсов в единицу времени. По этой причине, как и в проделанных нами ранее исследованиях, для последующего анализа исходные данные группировались в пять интегральных каналов: $k1$, $k2$, $k3$, $k4$, $k5$. Количество импульсов в каждом интегральном канале получалось суммированием по идущим подряд аппаратурным каналам (от 1 до 4084). Причем, границы каналов $k2$, $k3$, $k4$, $k5$ пподбирались таким образом, чтобы выделенные области энергии включали в себя пики, соответствующие гамма излучению атомов Cs, K, U и Th, соответственно.

Таким образом, каждой пространственной точке (x, y) исследуемых профилей присваивался упорядоченный набор из 5 чисел. Учитывая,

что более объективным показателем распределения числа импульсов по интегральным каналам $k1 - k5$ является не абсолютное значение – количество импульсов в данном интегральном канале, а его относительная величина – доля этих импульсов относительно полного числа импульсов, то дальнейшее исследование проводилось именно для долей каналов калия ($k3$) , урана ($k4$) и тория ($k5$) относительно полного числа импульсов.

$$dol_i(x,y) = \frac{Imp_{k_i}(x,y)}{\sum_{j=1}^{5} Imp_{k_j}(x,y)}, i = 3,4,5, \qquad (4)$$

где $Imp_{k_i}(x,y)$ – суммарное число импульсов канала k_i.

С целью получения представления о распределении на исследуемой территории естественных изотопов – гамма-излучателей и нахождения возможных аномальных областей была проведена классификация по сформированной системе признаков всех профилей отдельно для каждого признака dol_i. Классификация по каналу $k4$ выявила две хорошо локализованные области с повышенным содержанием урана. Для тория подобные области не наблюдаются.

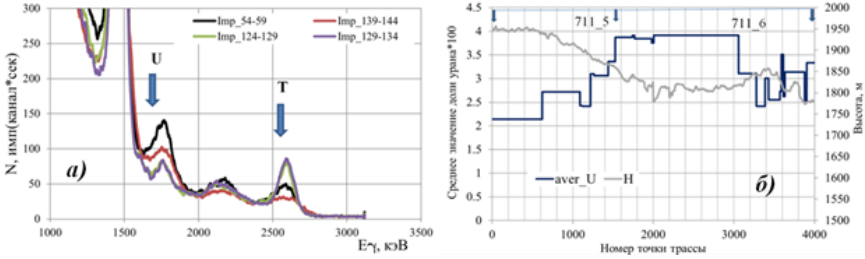


Рис. 1. а) Спектры территории с урановой аномалией (Imp_54_59) в сравнении со спектрами фоновых ареалов (без урановой аномалии). Стрелки указывают на положения основных пиков; урановый при E =*1,76 Мэв* и ториевый при E =*2,61 Мэв* . б) Положения ИС на этом участке

Расчет суммарного спектра точек, располагающихся в этих областях, также показывают наличие урановой аномалии. На рис. 1а показан спектр, построенный по точкам, входящих в область 54-59. Видно, что суммарное количество импульсов в области E =*1760 Кэв* для точек ареала 54-59 практически в два раза превышает фоновые значения. Методом APFLOW были найдены ИС для ареала 54-59, положения которых на исследуемом профиле изображены на рис. 1б.

Заключение

Предложенный новый подход к обработке гамма-спектроскопических данных показал свою работоспособность. Используемый алгоритм метода выявления структур дает возможность представить каждую трассу исследуемой территории в виде дискретной последовательности ИС. Распределение ИС на небольшое число состояний с близкими средними значениями числа импульсов позволяет получить пространственное положение всех ИС. Вместе с обработкой спектров по долям радиоактивных элементов, особенно урана, это позволяет лучше выявлять геологические границы по сравнению с традиционным вычислением процентного содержания урана.

СПИСОК ЛИТЕРАТУРЫ

1. *Kataev S., Kataeva S., Miller A.* On a New Method of Detecting Layer Boundaries at the Borehole Measurements by Gamma Logging. // Near Surface Geoscience Conference & Exhibition. The Hague, Netherlands. September 2019. P. 8–12. DOI:10.3997/2214-4609.202152102
2. *Катаев С. Г., Катаева С. С., Миллер А. А., Ермолин ,Е. Ю.* Применение статистических методов обработки спектра естественного гамма-излучения на примере съемок на полигоне твердых бытовых отходов в Смоленске. // Сборник материалов 20-ой научно-практической конференции и выставки "Инженерная и рудная геофизика". Казань. 2024. С. 279–283
3. *Катаев С. Г.* Алгоритмические и программные средства аналитического и неразрушающего контроля параметров природной среды на основе метода выделения структур: дис. ... д-ра техн. наук. Томск: 2013. 355 с.
4. *Bekkerman E. N., Kataev S. G., Kataeva S. S.* Heuristic approximation method for a random flow of events by an mc-flow with arbitrary number of states. // Automation and Remote Control. 2013. Vol. 74, № 9. P. 1449–1459.

Катаев Сергей Григорьевич — д.т.н., профессор, Томский государственный педагогический университет. E-mail: *sgkataev2010@yandex.ru*

Катаева София Семёновна — к.т.н., доцент, Национальный исследовательский Томский государственный университет. E-mail: *sskataeva@yandex.ru*

Миллер Андрей Аскольдович — к.т.н., АО НПП ВНИИГИС, Санкт-Петербург. E-mail: *andreymiller@yandex.ru*

Ермолин Евгений Юрьевич — к.г-м.н., генеральный директор ООО «ДЖИ М Сервис», Санкт-Петербург.

ОПТИМИЗАЦИЯ БЫСТРОДЕЙСТВИЯ ПРОЦЕССОРНОГО КОНВЕЙЕРА

А. В. Корсун[1], С. П. Сущенко[1,2], А. С. Шкуркин[1]

[1] *Томский государственный университет, Томск, Россия*
[2] *Томский филиал Федерального исследовательского центра информационных и вычислительных технологий, Томск, Россия*

Рассмотрена конвейерная модель процесса выполнения неоднородного потока команд однородным процессором. Получена функциональная зависимость оптимальной длины конвейера от времени выполнения отдельных команд приложения, размера приложения, накладных расходов на запись и чтение промежуточных результатов на каждом этапе конвейера и неоднородных вероятностей перезагрузки конвейера при промахах предсказания ветвлений на различных участках программного кода. Проведен анализ эффекта от оптимальной конвейеризации. Найдены условия ускорения конвейеризированных вычислений.

Ключевые слова: *процессорный конвейер, длина конвейера, оптимизация*

Введение

Важнейшим параметром, определяющим быстродействие вычислителя, является количество фаз процессорного конвейера. Классический конвейер процессора [1] имеет пять фаз – чтение команды, декодирование команды, чтение операндов, выполнение команды и запись результата. Для увеличения быстродействия каждый из этапов разбивается на отдельные фазы, что дает возможность повысить рабочую частоту процессора и как следствие – быстродействие. Фазы чтения команд и операндов, а также фазу записи результатов ускоряют за счет опережающей выборки команд, организации многоуровневого ассоциативного кэширования адресуемых объектов с высоким коэффициентом неблокируемости иерархической памяти, обеспечивающим параллельное выполнение большого числа неблокирующих транзакций доступа к многоуровневой подсистеме памяти, и возможностью обратной записи [1]. Быстродействие фазы декодирования команд ускоряют за счет организации параллельной работы нескольких конвейеризированных дешифрирующих устройств различной сложности [1]. Высокая скорость фазы выполнения команд при выполнении предусловий обработки достига-

ется с помощью неупорядоченного и спекулятивного выполнения команд несколькими независимыми исполнительными устройствами различных подмножеств системы команд процессора. Также ускорение перечисленных фаз достигается за счет увеличения объема регистровой памяти для преодоления проблем «ложных» взаимозависимостей с помощью архитектурного приема «переименование регистров», беспечивающего отображение ограниченного количества архитектурных регистров общего назначения на существенно большее число физических регистров [1]. Кроме того, ускорение исполнительного этапа при «истинных» взаимозависимостях последовательных команд приложений достигается с помощью архитектурного приема «обход и продвижение данных», снижающего накладные расходы на запись результатов предыдущей команды и чтение операндов для следующей. Вопросы оптимизации процессорного конвейера по критерию минимума времени исполнения приложений заданного объема рассматривались в работах [2–4]. Однако в них получены результаты только для однородного потока обрабатываемых команд приложения и равновероятной перезагрузки конвейера при неверном предсказании всех условных программных переходов приложения. Кроме того, в указанных публикациях не содержится исследование эффекта от оптимизации операционных характеристик процесса конвейерной обработки.

1. Модель центрального процессора

Центральный процессор [1] моделируется конвейером с количеством фаз $D \geq 1$, обрабатывающим поток команд приложения различной трудоемкости с одинаковым быстродействием на различных фазах конвейера. Полагаем, что на вычислительной системы выполняется неоднородное приложение из $N \geq 1$ команд. Считаем, что общее время выполнения команды приложения в виде одной фазы $T_n, n = \overline{1, N}$. Обычно этапы конвейера реализуются как независимые исполнительные устройства [1]. Тогда при выполнении команды на отдельных этапах промежуточные результаты сохраняются в буферных полях конвейера, а перед обработкой на следующей фазе – считываются из буфера (рис. 1).
Считаем, что время чтения-записи промежуточных результатов на каждом этапе конвейера $t_n, n = \overline{1, N}$ пропорционально длительности выполнения команды. Тогда время выполнения n-й команды на отдельной фазе конвейера с учетом накладных расходов на чтение-запись промежуточных результатов составит $b_n = T_n/D + t_n$. Процессор с неупорядоченным исполнением команд видит исполняемое приложение через скользящее вдоль программного кода окно команд, реализует прогнозирование переходов при ветвлении программы, выбирает команды с

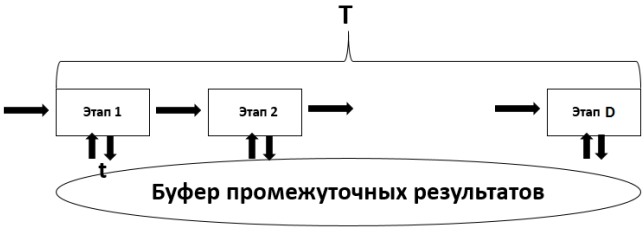


Рис. 1. Структурная модель процессорного конвейера

предсказанного адреса перехода до того, как он узнает, верным ли был его прогноз и спекулятивно их выполняет. При неверном предсказании ветвления программы требуется перезагрузка конвейера. Полагаем, что вероятности неверного предсказания переходов $p_n, n = \overline{1,N}$ зависят от типа команды, после которой выполняется перезагрузка конвейера. Для перезагрузки после выполнения n-й команды приложения потребуется дополнительное время, пропорциональное длине конвейера без одной фазы $(D-1)\tau_n$ [5], где $\tau_n, n = \overline{1,N}$ – средняя задержка на отдельной фазе конвейера при выполнении n-й команды приложения и последующей перезагрузке.

2. Время выполнения неоднородного приложения

Важнейшей операционной характеристикой конвейерной обработки приложения является время его выполнения конвейеризированным процессором. При выполнении неоднородного приложения на конвейере длины $D > 1$ среднее время, затрачиваемое на перезагрузки, составит $(D-1)\sum_{n=2}^{N} p_n\tau_n$. Средняя задержка $\tau_n, n = \overline{2,N}$ определяется взвешенной по вероятностям суммой длительностей отдельных фаз для всех вариантов перезагрузок, предшествующих перезагрузке после n-й команды: $\tau_n = b(1,n)\prod_{i=2}^{n-1}(1-p_i)+\sum_{i=2}^{n-1} p_i b(i,n)\prod_{j=i+1}^{n-1}(1-p_j), n = \overline{2,N}$, где $b(i,n)$ – длительность отдельной конвейерной фазы самой трудоемкой команды между последовательными перезагрузками после i-й и n-й командами программного кода [5] $b(i,n) = \max\limits_{j=\overline{i+1,n}} b_j, i = \overline{1,n-1}, n = \overline{2,N}$. Среднее время выполнения неоднородного по составу команд приложения процессором с динамическим исполнением примет вид

$$T(D,N) = (D-1)\left(\frac{T_m}{D} + t_m + \sum_{n=2}^{N} p_n\left(\frac{\bar{T}_n}{D} + \bar{t}_n\right)\right) + \sum_{n=1}^{N}\left(\frac{T_n}{D} + t_n\right),$$

$$T_m = \max_{n=\overline{1,N}} T_n, t_m = \max_{n=\overline{1,N}} t_n.$$

$$\bar{T}_n = \max_{j=\overline{1,n}} T_j \prod_{l=2}^{n-1}(1-p_l) + \sum_{i=2}^{n-1} \max_{j=\overline{i+1,n}} T_j p_i \prod_{l=i}^{n-1}(1-p_l),$$

$$\bar{t}_n = \max_{j=\overline{1,n}} t_j \prod_{l=2}^{n-1}(1-p_l) + \sum_{i=2}^{n-1} \max_{j=\overline{i+1,n}} t_j p_{i-1} \prod_{l=i}^{n-1}(1-p_l). \qquad (1)$$

3. Оптимизация длины конвейера

Экстремальная зависимость времени выполнения приложения от длины конвейера обусловлена двумя факторами. С одной стороны, рост количества конвейерных фаз приводит к снижению времени обработки программного кода за счет параллельной обработки на различных этапах конвейера различных команд. С другой стороны, увеличение числа этапов приводит к пропорциональному нарастанию накладных расходов на запись/чтение промежуточных результатов каждого этапа и росту времени перезагрузки после ошибок предсказания переходов. Определим из (1) выигрыш от конвейерного исполнения приложения:

$$\Delta(D,N) = T(1,N) - T(D,N) = \\ = \frac{D-1}{D}\left[\sum_{n=1,n\neq m}^{N} T_n - Dt_m - \sum_{n=2}^{N} p_n\left(\bar{T}_n + D\bar{t}_n\right)\right]. \qquad (2)$$

Условие $\Delta(D,N) > 0$ определяет рациональное множество длин конвейера. Выигрыш положителен при интервальных ограничениях на длину конвейера $1 < D < \left(\sum_{n=1,n\neq m}^{N} T_n - \sum_{n=2}^{N} p_n\bar{T}_n\right) \Big/ \left(t_m + \sum_{n=2}^{N} p_n\bar{t}_n\right)$. Зависимость выигрыша при выполнении интервальных условий положительности является унимодальной функцией от длины конвейера (рис. 2). Для однородного потока команд и равновероятной перезагрузки конвейера после каждой команды выигрыш (2) преобразуется к $\Delta(D,N) = \frac{D-1}{D}\left[T(N-1)(1-p) - Dt\left(1+p(N-1)\right)\right]$. Условие положительности выигрыша принимает вид $1 < D < \frac{T(N-1)(1-p)}{t(1+p(N-1))}$. Кроме того, положительный выигрыш достигается для вероятности перезагрузки $p < \frac{T(N-1)-Dt}{(N-1)(T+Dt)}$ и если отношение полного времени выполнения команды T к накладным расходам t удовлетворяет ограничению снизу $\frac{T}{t} > \frac{D(1+p(N-1))}{(N-1)(1-p)}$.

Из (2) получаем оптимальную длину конвейера

$$D_o = \sqrt{\left(\sum_{n=1,n\neq m}^{N} T_n - \sum_{n=2}^{N} p_n\bar{T}_n\right) \Big/ \left(t_m + \sum_{n=2}^{N} p_n\bar{t}_n\right)}. \qquad (3)$$

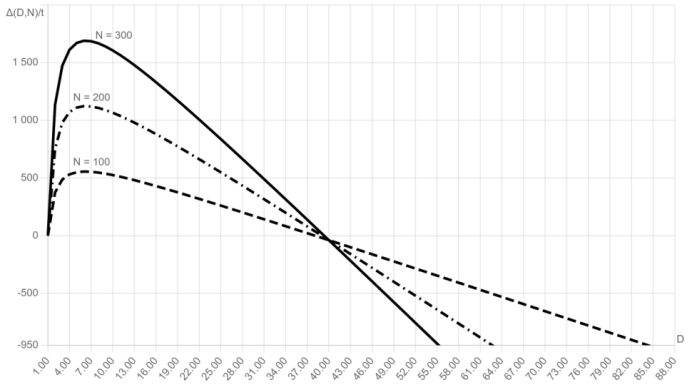


Рис. 2. Зависимость относительного выигрыша от длины конвейера D при $p = 0{,}2;\ T/t = 10$

Однородный поток команд и равновероятная перезагрузка конвейера в любой точке приложения приводит выражение (3) к соотношению для оптимальной длины вида $D_o = \sqrt{\frac{T(N-1)(1-p)}{t(1+p(N-1))}}$. Потенциальный положительный эффект от оптимальной конвейеризации определяется оптимальным выигрышем

$$\Delta(D_o, N) = 2\left[\frac{\sum_{n=1,n\neq m}^{N} T_n - \sum_{n=2}^{N} p_n\bar{T}_n + t_m + \sum_{n=2}^{N} p_n\bar{t}_n}{2} - \right.$$
$$\left. - \sqrt{\left(\sum_{n=1,n\neq m}^{N} T_n - \sum_{n=2}^{N} p_n\bar{T}_n\right)\left(t_m + \sum_{n=2}^{N} p_n\bar{t}_n\right)}\right].$$

Для однородного потока команд и равновероятной перезагрузке процессорного конвейера оптимальный выигрыш принимает вид

$$\Delta(D_o, N) = 2\left[\frac{T(N-1)(1-p)+t(1+p(N-1))}{2} - \right.$$
$$\left. - \sqrt{T(N-1)(1-p)t\,(1+p(N-1))}\right].$$

Отсюда следует, что вероятность перезагрузки должна удовлетворять ограничению $p < \frac{T(N-1)-t}{(N-1)(T+t)}$. Из приведенных распределений по длинам приложений видно, что оптимальное количество фаз процессорного конвейера в значительной мере определяется вероятностью перезагрузки.

Заключение

Предложена модель процесса выполнения неоднородного потока команд однородным процессорым конвейером. В аналитическом виде найдена оптимальная длина конвейера как функция монолитного времени выполнения отдельных команд одностадийным процессором, издержек на сохранение и чтение промежуточных результатов обработки на независимых фазах конвейера, размера приложения и вероятности перезагрузки конвейера при промахах предсказания ветвлений программных алгоритмов. Выполнен анализ потенциальных возможностей ускорения вычислительного процесса. Показана преемственность к ранее полученным результатам других авторов. На основе анализа функции выигрыша получены условия целесообразности конвейерной обработки в виде неравенств.

СПИСОК ЛИТЕРАТУРЫ

1. *Танненбаум Э., Остин Т.* Архитектура компьютера. 6-е изд. СПб.: Питер, 2021. 816 с.
2. *Беляев А.А.* Влияние программных переходов и зависимостей по данным в исполняемом программном коде на производительность конвейера DSP-ядра // Известия вузов. Электроника. 2009. № 3 (77). С. 75–80.
3. *Беляев А.А.* Оптимальная по производительности глубина программного конвейера для приложений с программными переходами и зависимостью по данным // Известия вузов. Электроника. 2010. № 2 (82). С. 48–51.
4. *Хусаинов А.А., Титова Е.А.* Оптимальная глубина вычислительного конвейера при заданном объеме входных данных // Вычислительные технологии. 2018. Т. 23, № 1. С. 96–104.
5. *Сущенко С.П.* Анализ сквозной задержки сообщения в многозвенном виртуальном канале // Автоматика и вычислительная техника. 1989. № 3. С. 52–64.

Корсун Александр Викторович — ассистент кафедры прикладной информатики. E-mail: *aleksnfsl5@gmail.com*

Сущенко Сергей Петрович — д-р техн. наук, профессор, заведующий кафедрой прикладной информатики. E-mail: *ssp.inf.tsu@gmail.com*

Шкуркин Алексей Сергеевич — канд. техн. наук, доцент, доцент кафедры прикладной информатики. E-mail: *shkurkin@mail.tsu.ru*

СПОСОБЫ ПРИМЕНЕНИЯ БАЙЕСОВСКОЙ СЕТИ НА ОСНОВЕ СТАНДАРТА OMG ESSENCE ДЛЯ АВТОМАТИЧЕСКОГО СОСТАВЛЕНИЯ ПЛАНА ИТЕРАЦИИ

О. В. Лазарев[1], Д. О. Змеев[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

Значительное количество принятых в программной инженерии методов разработки имеют итеративно-инкрементальный характер, то есть подразумевают выполнение работы небольшими фрагментами, каждый из которых занимает определённое время — итерацию. Важной задачей руководителя проекта и проектной команды при использовании таких методов является планирование итераций, то есть определение объёма работ на итерацию. В настоящей работе рассмотрены способы автоматизации этого процесса, основанные на использовании байесовской сети, построенной на основе стандарта OMG Essence. Приводится сравнение различных алгоритмов составления плана итерации. **Ключевые слова:** *OMG Essence, управление проектами, байесовская сеть, итерация, план итерации*

Введение

Значительное количество методов разработки программного обеспечения, принятых в программной инженерии, имеют итеративно-инкрементальный характер [1], что подразумевает, что реализация проекта разбивается на небольшие временные отрезки (именуемые итерациями), в рамках каждой из которых к разрабатываемому программному обеспечению добавляется некоторая ценность (именуемая инкрементом). Итерации обычно имеют равную или примерно равную продолжительность в рамках одного проекта, при этом планирование итераций обычно осуществляется не заранее, в начале проекта, а непосредственно по окончании одной итерации и перед началом следующей.

Одним из самых известных процессов такого типа является SCRUM, в терминологии которого итерации именуются спринтами [2]. В данном процессе принято, чтобы продолжительность итерации составляла от 1 недели до 1 месяца (в качестве стандартного решения предлагаются 2 недели), и в начале каждой итерации проводилось планирование, на

котором определялся бы инкремент, который планируется создать в течение итерации, то есть составлялся бы план итерации. Очевидно, что от качества планирования итераций во многом зависит успех проекта.

В связи с высокой распространённостью итеративно-инкрементальных методов разработки и важностью процесса планирования итерации приобретает актуальность задача автоматизации этого процесса. Для решения этой задачи используются различные методы, включая генетические алгоритмы [3], экономические модели [4] и графовые представления процессов разработки [5].

В настоящей модели рассматривается автоматическое составление плана итерации с использованием байесовской сети на основе стандарта OMG Essence, выполняемое в рамках системы поддержки принятия решений, предложенной в [6].

1. Формулировка задачи

Стандарт OMG Essence выделяет семь основных аспектов проекта по разработке программного обеспечения, именуя их альфами и предлагая использовать их для отслеживания прогресса проекта. Для этого в каждой альфе выделяются состояния, обычно последовательно достигаемые альфой в ходе проекта. Каждое состояние характеризуется набором чекпоинтов. Считается, что если каждый чекпоинт состояния альфы выполнен, то альфа достигла этого состояния. Таким образом, в момент инициации проекта ни один из чекпоинтов не выполнен, а при его завершении — выполнены все чекпоинты. Отметим, что стандарт OMG Essence охватывает весь жизненный цикл программного обеспечения, поэтому под завершением проекта понимается вывод системы из эксплуатации, хотя он обычно не осуществляется в рамках проекта по разработке программного обеспечения.

В рамках диссертационного исследования [6] предложен алгоритм оценки степени уверенности в достижении чекпоинтов, т. е. вероятностей их достижения, основанный на прямом вероятностном выводе в байесовской сети, где в качестве свидетельств используется мнение менеджера проекта о достижении или недостижении чекпоинтов. В [7] алгоритм дополнен учётом объективных свидетельств — состояний суб-альф и уровней детализации рабочих продуктов. При этом учтено, что суб-альфы и рабочие продукты, в отличие от чекпоинтов альф, могут вносить не только положительный, но и отрицательный вклад в прогресс проекта, а также менять «знак» своего вклада в зависимости от степени детализации.

Поскольку прогресс проекта равнозначен выполнению чекпоинтов, он может быть численно представлен как сумма вероятностей достиже-

ния всех чекпоинтов. Таким образом, задача оптимального планирования итерации сводится к определению такого множества состояний суб-альф, уровней детализации рабочих продуктов и чекпоинтов (далее — элементов), что суммарные трудозатраты на их исполнение не превышают продолжительности итерации (оба параметра должны измеряться в единицах размерности человек·время; в рамках работы будут использоваться человеко-часы), а их реализация приведёт к наибольшему увеличению суммы вероятностей достижения чекпоинтов среди всех таких множеств.

2. Алгоритмы решения задачи

В указанной выше постановке задача автоматического составления плана итерации становится аналогична одной из классических задач комбинаторной оптимизации — задаче о рюкзаке. Действительно, имеется множество «предметов» (элементов), для каждого из которых определены «вес» (оценка трудозатрат в человеко-часах) и «стоимость» (прирост суммы вероятностей достижения чекпоинтов). Поскольку чекпоинты, состояния суб-альфы и уровни детализации рабочего продукта не могут быть достигнуты дважды, задачу можно рассматривать как задачу о «рюкзаке 0-1» [8].

Однако, существует важное различие между традиционной формулировкой задачи о рюкзаке и решаемой задаче автоматического составления плана итерации. Задачей о рюкзаке обычно предполагается, что стоимость предметов не меняется в зависимости от того, какие из предметов помещены в рюкзак. Это, однако, не так в случае рассматриваемого подхода к автоматическому составлению плана итерации. При добавлении любого элемента необходимо выполнить перерасчёт прироста вероятностей достижения чекпоинтов.

Рассмотрим в качестве примера два чекпоинта:

(1) Stakeholder groups identified («Группы стейкхолдеров идентифицированы»)

(2) Requirement's format agreed («Формат требований согласован»)

В присутствии (1) ценность (2) существенно возрастает, так как без идентификации групп стейкхолдеров вообще непонятно, с кем был согласован формат требований. Следовательно, отметка (2) достигнутым без отметки (1) достигнутым не приведёт к существенному росту суммы вероятностей достижения чекпоинтов, так как вероятность достижения (2) будет оценена низко вопреки мнению менеджера. Однако, при наличии отмеченного выполненным (1) отметка (2) выполненным приведёт к существенному росту суммы вероятностей достижения чекпоинтов.

Указанное различие делает невозможным использование классического подхода к решению задачи о рюкзаке — динамического программирования, поскольку не выполняется требование о наличии оптимальной подструктуры, обязательное в динамическом программировании. Рассмотрим возможные подходы к решению задачи с учётом данного ограничения.

Полный перебор. Данный способ гарантированно позволяет получить оптимальное решение, однако является неприемлемым по трудоёмкости. Поскольку каждый из элементов может быть либо включён, либо не включён в решение, общее количество возможных решений составляет 2^n, где n — общее количество элементов.

Поскольку алгоритм прямого вероятностного вывода для всего множества элементов имеет трудоёмкость $O(n \cdot 2^p)$, где n — среднее количество родителей у элемента, трудоёмкость полного перебора составляет $O(2^n \cdot n \cdot 2^p)$, что является неприемлемым даже при небольшом n.

Метод ветвей и границ. Представим полный перебор вариантов может быть в виде дерева, где каждый узел представляет возможное решение (набор элементов), а его потомками являются наборы, отличающиеся от него добавлением ровно одного элемента. Введём на множестве элементов какой-либо строгий порядок и для каждой вершины добавлять в качестве потомков только решения, отличающиеся добавлением элемента, «большего», чем добавленный в текущей вершине (это позволит избежать идентичных поддеревьев, отличающихся лишь порядком добавления вершин).

Выполняя поиск в таком дереве оптимального решения, будем отсекать поддеревья, вершина которых нарушает ограничение на максимальную трудоёмкость итерации (т. к. добавление элемента в решение не может снизить общую трудоёмкость, в этом поддереве не будет допустимых решений).

В наихудшем случае метод ветвей и границ имеет ту же трудоёмкость, что и полный перебор, однако на практике итерация обычно содержит лишь небольшое количество элементов, поэтому отсечение поддеревьев существенно сократит перебор.

Жадный алгоритм. Поскольку множество элементов едва ли является матроидом, жадный алгоритм является лишь приближённым методом решения рассматриваемой задачи. Однако, рассмотрим его в качестве приближённого метода. На каждом шаге будем рассчитывать сумму вероятностей для каждого из ещё не отмеченных элементов, после чего выбирать оптимальный, отмечать его и повторять эту процеду-

ру, пока суммарная трудоёмкость итерации не превысит установленное ограничение.

В качестве критерия оптимальности можно использовать максимизацию прироста суммы вероятностей достижения чекпоинтов или максимизацию её же прироста в единицу времени. Второй подход позволит отдать приоритет небольшим задачам, позволяющим за короткое время добиться существенного прогресса на проекте, в то время как первый отдаёт приоритет крупным задачам. Если же трудоёмкости элементов близки друг к другу, разница между подходами не является существенной.

Поскольку расчёт суммы вероятностей для всех элементов имеет трудоёмкость $O(n^2 \cdot 2^p)$, трудоёмкость жадного алгоритма в худшем случае (итерация, включающая весь проект) составляет $O(n^3 \cdot 2^p)$. Если же принять среднее количество элементов в итерации равным m ($m < n$), трудоёмкость в среднем составит $O(m \cdot n^2 \cdot 2^p)$.

Наивное решение. В качестве другого приближённого метода можно проигнорировать отсутствие оптимальной подструктуры у решаемой задачи и использовать стандартный алгоритм решения задачи о рюкзаке.

Трудоёмкость такого алгоритма составляет $O(nw)$, где w — предельная трудоёмкость итерации («грузоподъёмность» рюкзака). (При достаточно малых w трудоёмкость составляет $O(n^2 \cdot 2^p)$, так как основным элементом алгоритма становится изначальный расчёт прироста суммы вероятностей для каждого чекпоинта.)

В связи с тем, что решение полным перебором обладает неприемлемо высокой трудоёмкостью, оптимальным будет использование метода ветвей и границ или одного из предложенных приближённых алгоритмов. Для выбора алгоритма необходимо реализовать их провести сравнительный анализ по времени исполнения и качеству предлагаемых решений. Важной подзадачей при этом также является определений значений трудоёмкостей для элементов, поскольку они не вводятся стандартом OMG Essence.

Заключение

В настоящей работе задача автоматического составления плана итерации сформулирована для системы поддержки принятия решений на основе байесовской сети и стандарта OMG Essence, предложенной в [6] Задача сведена к классической задаче о рюкзаке, однако, с нарушением требования о наличии оптимальной подструктуры, что делает невозможным использование динамического программирования для точного решения задачи. Рассмотрены варианты алгоритма решения этой за-

дачи. Определены дальнейшие направления исследований — практическая реализация рассмотренных алгоритмов и их сравнительный анализ, а также определение трудоёмкостей элементов.

СПИСОК ЛИТЕРАТУРЫ

1. *Larman C., Basili Vol. R.* Iterative and Incremental Development: A Brief History // Computer. 2003. № 36 (6). P. 47–56.
2. `https://www.scrum.org/resources/what-is-a-sprint-in-scrum` — What is a Sprint | Scrum.org. 2025.
3. *Ancveire I., Polaka I.* Application of Genetic Algorithms for Decision-Making in Project Management: A Literature Review // Information Technology and Management Science. 2019. Vol. 22. P. 22–31.
4. *Soltan S., Ashrafi M.* Predicting project duration and cost, and selecting the best action plan using statistical methods for earned value management // Journal of Project Management. 2020. № 5. P. 157–166.
5. *Ishaque M., Zaidi A. K., Levis A. H.* Project management using point graphs // Systems Engineering. 2009. Vol. 12, № 1. P. 36–54.
6. *Змеев Д. О.* Прототип системы поддержки принятия решений для управления проектами на основе стандарта OMG Essence и байесовских сетей: дис. ... канд. техн. наук. Томск, 2022. 152 с.
7. *Иванова Л. С.* Практики OMG Essence как источник изменяющихся объективных свидетельств в байесовской сети для систем поддержки принятия решений: дис. ... канд. техн. наук. Томск, 2023. 148 с.
8. *Martello S., Toth P.* Knapsack problems: Algorithms and computer implementations Chichester: John Wiley & Sons Ltd., 1990. 296 p.

Лазарев Олег Вадимович — аспирант, факультет инновационных технологий. E-mail: *laoltomsk@gmail.com*

Змеев Денис Олегович — к. т. н., доцент НОЦ «Высшая ИТ-школа». E-mail: *denis.zmeev@accounts.tsu.ru*

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ВЫЖИВАЕМОСТИ ПРИ РАКЕ МОЛОЧНОЙ ЖЕЛЕЗЫ

Н. А. Лукьянова[1,2], Е. Е. Голденок[1,2], Е. Ю. Зарубин[1]

[1] *Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого, г. Красноярск, Россия*
[2] *Сибирский федеральный университет, г. Красноярск, Россия*

Цель работы: оценка прогнозирования выживаемости пациентов с раком молочной железы. Анализ проводился на основе обезличенной базы данных пациентов Красноярского края (2014–2024 гг.), включающей 18 228 случаев рака молочной железы как у мужчин, так и у женщин. Реализована серия вычислительных экспериментов с пятью методами машинного обучения, адаптированными для анализа выживаемости. Модели обучались на совокупности отобранных значимых предикторов в регрессионной модели случайного леса. Все модели выживаемости показали хорошие прогностические результаты с коэффициентом соответствия C–индекс$>$0,9.

Ключевые слова: *правостороннее цензурирование, машинное обучение, прогнозирование, выживаемость*

Введение

Рак молочной железы (РМЖ) продолжает оставаться одним из ведущих онкологических заболеваний [1]. На территории Российской Федерации в 2023 г. в общей структуре заболеваемости населения злокачественными заболеваниями (ЗНО) РМЖ занимал 1-е место среди женского населения, составляя 22,5% [2]. Наблюдается тенденция к увеличению заболеваемости РМЖ на 18–22%, что связано как с совершенствованием методов ранней диагностики, так и с ростом влияния факторов окружающей среды, гормональными и генетическими предрасположенностями [3]. Важным показателем, характеризующим эффективность противоопухолевой помощи, является уровень своевременной диагностики. По данным [2] доля пациентов с IV стадией заболевания при постановке диагноза составляет около 30% среди пациентов с РМЖ. В

Авторы благодарят заместителя главного врача по организационно-методической работе Красноярского краевого клинического онкологического диспансера имени А. И. Крыжановского, к.м.н. Сафонцева Ивана Петровича за предоставление данных и возможность проведения исследования

Российской Федерации в 2023 г. в структуре смертности от ЗНО удельный вес РМЖ составил 7,2% [2]. О необходимости разработки и внедрения более эффективных скрининговых программ, направленных на выявление этих заболеваний на ранних этапах отмечено в работе [4]. Особую значимость приобретает анализ выживаемости, прогнозирование течения заболевания и выявление прогностически значимых факторов, которые могут повлиять на исход лечения.

Цель исследования: оценка выживаемости больных со злокачественной патологией молочной железы в Красноярском крае.

1. Материалы и методы исследования

Материалом исследования является обезличенная база данных (БД) о пациентах с диагнозом С50 из информационной системы «Канцер-регистр», предоставленная Красноярским краевым клиническим онкологическим диспансером имени А. И. Крыжановского за период 2014–2024 гг. После предобработки в БД находятся данные о 18 228 пациентах, из них 73 мужчины и 18 155 женщин, и о 167 признаках, включающих демографические (возраст, пол), клинические (стадия, морфологический тип опухоли, TNM-классификация) и терапевтические (химиотерапия, лучевая терапия, хирургические вмешательства) характеристики. Особое внимание уделено временным параметрам: дате рождения, дате установления диагноза и дате смерти или последнего наблюдения для цензурированных случаев. Дата завершения исследования – дата выгрузки из «Канцер-регистр» (01.04.2025). На основе временных переменных определен возраст пациента (в годах) на момент установления диагноза. При создании признака *Событие* заполнение данных происходило по принципу, если *Дата смерти* заполнена в БД, то присваивался статус 1 (полное событие), если же дата не введена, то присваивался статус 0 (цензурированное событие). Цензурирование правостороннее: дата начала наблюдения каждого пациента – дата установления диагноза. Период наблюдения (в днях) вычислялся, как разность между датой клинического исхода (смерти) и датой установления диагноза.

Для анализа выживаемости использовали кривые Каплана–Мейера [5], а также была реализована серия вычислительных экспериментов (рис. 1) с применением алгоритмов машинного обучения [6], адаптированных для работы с цензурированными данными и обученных на совокупности признаков, предварительно отобранных регрессионной моделью случайного леса (Random Forest Regressor):

1) случайный лес выживания (Random Survival Forests);

2) компонентный градиентный бустинг анализа выживаемости (Component wise Gradient Boosting Survival Analysis);
3) градиентный бустинг анализа выживаемости (Gradient Boosting Survival Analysis);
4) машина опорных векторов выживания (Survival Support Vector Machine);
5) дополнительные деревья выживания (Extra Survival Trees).

Для измерения точности прогностических моделей используется индекс согласованности (C-индекс).

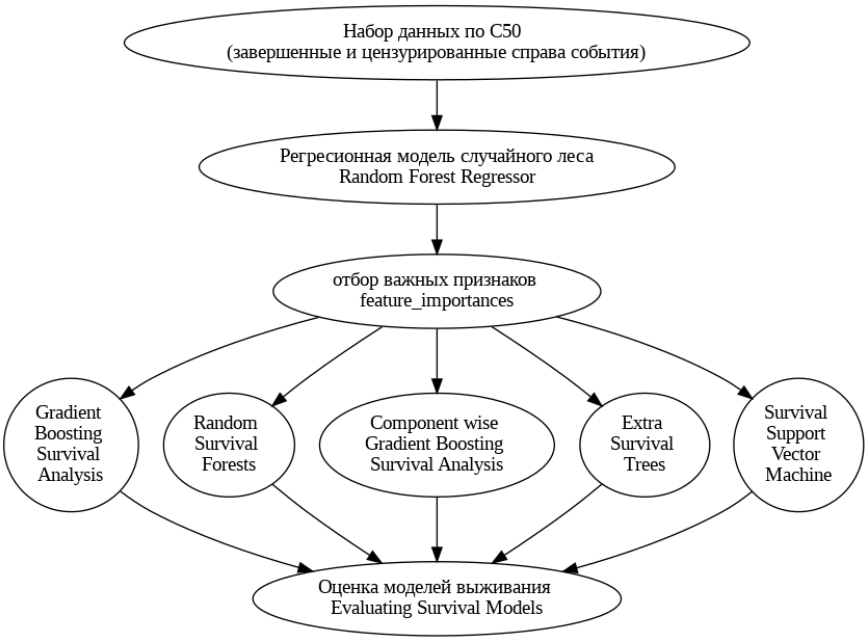


Рис. 1. Схема проведения вычислительных экспериментов

Обработка и анализ данных проведены в рабочей среде Google Colaboratory на языке программирования Python версия 3.11.

2. Результаты и обсуждения

Возраст мужчин варьируется от 35 до 91 года, у женщин от 22 до 98 лет. Медиана возраста ($\mathrm{Me}_{\mathrm{age}}$ [Q1, Q3]) на момент постановки диагноза у мужчин составила 65 [58; 71] лет, женщин: $\mathrm{Me}_{\mathrm{age}} = 61$ [50; 69] год.

Доля цензурированных событий больше доли полных:

- полные – 4850 (27%) пациентов, $\mathrm{Me}_{\mathrm{age}} = 64$ [54; 74] года, минимальное время наблюдения 1 день, максимальное – 3993 дня;
- цензурированные – 13378 (73%) пациентов, $\mathrm{Me}_{\mathrm{age}} = 59$ [48; 67] лет, минимальное время наблюдения – 95 дней, максимальное – 4108.

По всем пациентам БД медиану выживаемости ($\mathrm{Me}_{\mathrm{time}}$) определить невозможно, событие не наступило у половины пациентов вообще, до 1617 дня дожили три четверти всех наблюдаемых пациентов (75%), что говорит о высокой выживаемости пациентов. По функции выживаемости $S(t)$ вероятность (P, %):

- 1-летней выживаемости: $\mathrm{P}_1 = 93{,}18\%$;
- 5-летней выживаемости: $\mathrm{P}_5 = 72{,}70\%$;
- 10-летней выживаемости: $\mathrm{P}_{10} = 59{,}29\%$.

Высокая 1-летняя выживаемость говорит о том, что большинство пациентов успешно переживают первый год после постановки диагноза и начала лечения. Далее, с увеличением периода времени идет постепенное снижение вероятности выживания.

В группе женщин $\mathrm{Me}_{\mathrm{time}}$ определить невозможно, событие не наступило у половины испытуемых вообще, до 1623 дней дожили 75% наблюдаемых пациентов, что говорит о высокой выживаемости пациенток. В группе мужчин $\mathrm{Me}_{\mathrm{time}} = 2252$ дня, интерквартильный размах не имеет верхней границы. Вероятность 5-летней выживаемости в группе женщин 72,75%; в группе мужчин – 60,06%. Различия в функциях выживаемости между полами являются статистически значимыми (Log-Rank test: $p = 0{,}01$).

Проанализированы распределения пациентов по стадиям и функции выживаемости, n – количество пациентов:

- I стадия – $n = 5284$ (29%), $\mathrm{Me}_{\mathrm{time}}$ найти невозможно, $\mathrm{P}_5 = 90{,}10\%$, $\mathrm{P}_{10} = 79{,}63\%$;
- II стадия – $n = 7885$ (43%); $\mathrm{Me}_{\mathrm{time}}$ найти невозможно, 75% пациентов дожили до 2276 дней (больше 6 лет), $\mathrm{P}_5 = 79{,}32\%$, $\mathrm{P}_{10} = 65{,}57\%$;
- III стадия – $n = 3817$ (21%), $\mathrm{Me}_{\mathrm{time}} = 2182$ дня, интерквартильный размах не имеет верхней границы, нижняя граница 930 дней, $\mathrm{P}_5 = 54{,}97\%$, $\mathrm{P}_{10} = 37{,}12\%$;
- IV стадия – $n = 1242$ (7%), $\mathrm{Me}_{\mathrm{time}} = 413$ [78; 1106] дней, $\mathrm{P}_1 = 52{,}39\%$, $\mathrm{P}_5 = 16{,}02\%$, $\mathrm{P}_{10} = 10{,}72\%$.

Различия в функциях выживаемости по стадиями при РМЖ являются статистически значимыми (Log-Rank test: $p < 0{,}005$).

Согласно схеме (рис. 1) проведения вычислительных экспериментов с использованием методов машинного обучения необходимо выделить важные предикторы. Для этого использовали RandomForestRegressor. Целевая переменная *Время* – время жизни после установления диагноза (в днях). Регрессионная модель случайного леса показала коэффициент детерминации $R^2 = 0{,}96$. Из 167 признаков, вошедших в модель, 45 признаков имеют ненулевой вклад в модель. В топ-10 вошли: 1) морфологический тип, 2) событие (полное/цензура), 3) возраст, 4) топография РМЖ, 5) метастазирование в регионарные лимфатические узлы в соответствии с классификацией TNM, 6) количество курсов лекарственного лечения (химиотерапия или гормоноиммунотерапия), 7) количество операций, 8) объем первичной опухоли в соответствии с классификацией TNM, 9) вид лучевого лечения, 10) мастэктомия по Маддену.

В методах машинного обучения анализа выживаемости целевую переменную формируют, учитывая переменные *Событие* и *Время*. На тренировочную и тестовую выборку набор по РМЖ делится в соотношении 7 : 3 соответственно. Обучаем модели и делаем оценку прогнозирования. Результаты вычислительных экспериментов прогнозирования:

1) модель RandomSurvivalForests на тестовой выборке дает значение C-индекс= 0,94, на тренировочной – C-индекс= 0,95;
2) модель ComponentwiseGradientBoostingSurvivalAnalysis на тестовой выборке имеет значение C-индекс= 0,92, на тренировочной выборке – C-индекс= 0,92;
3) модель GradientBoostingSurvivalAnalysis на тестовой выборке имеет C-индекс= 0,94, на тренировочной – C-индекс= 0,94;
4) модель Survival Support Vector Machine на тестовой выборке имеет C-индекс= 0,93, на тренировочной – C-индекс= 0,93;
5) модель Extra Survival Trees на тестовой выборке имеет значение C-индекс= 0,94, на тренировочной – C-индекс= 0,95.

Для всех моделей выживания значение коэффициента C-индекс$>$0,9, что говорит о высокой предсказательной точности прогнозирования выживаемости при РМЖ с учётом цензурирования. Классические методы анализа выживаемости (модель Кокса, модель Вейбулла) в сравнении с методами машинного обучения показали хуже результаты, на тестовой выборке обе модели имели значения C-индекс= 0,71. Тот факт, что наиболее значимыми в моделях оказались именно те параметры, которые уже используются в клинической практике, свидетельствует о легитимности полученных результатов и их практической применимости.

3. Заключение

Рак молочной железы остается наиболее распространенным онкологическим заболеванием среди женщин, а заболеваемость им растет, особенно в Красноярском крае. Наибольшая заболеваемость РМЖ приходится на возрастные группы старше 60 лет. Все модели машинного обучения при анализе выживаемости показали хорошие прогностические результаты с C-индексом выше 0,9 на основе отобранных значимых 45 предикторов в регрессионной модели случайного леса.

СПИСОК ЛИТЕРАТУРЫ

1. World Cancer Report 2020 International Agency for Research on Cancer (IARC). Geneva: World Health Organization. 2020. 568 p.
2. Злокачественные новообразования в России в 2023 году (заболеваемость и смертность) / под ред. А. Д. Каприна, В. В. Старинского, Г. В. Петровой. М.: Московский научно-исследовательский онкологический институт имени П. А. Герцена – филиал ФГБУ «НМИЦ радиологии» Минздрава России. 2024. 276 с.
3. Ефремова Т. В., Киселёва Н. С., Боровкова Е. И. Анализ заболеваемости раком молочной железы в РФ за последние 10 лет // Российский онкологический журнал. 2022. № 2. С. 45–50.
4. Шахзадова, А. О., Старинский В. В., Лисичникова И. В. Методологические подходы к оценке показателей заболеваемости, выживаемости и смертности при раке молочной железы в Российской Федерации // Онкология. Журнал им. П. А. Герцена. 2024. № 1. С. 36–43.
5. Kaplan, E. L., Meier P. Nonparametric estimation from incomplete observations // J. Am. Stat. Assoc. 1958. Vol. 53. P. 457–481.
6. Davidson-Pilon, C. Lifelines: Survival Analysis in Python // Journal of Open Source Software. 2019. Vol. 4, № 40. P. 1317. doi: https://doi.org/10.21105/joss.01317.

Лукьянова Наталья Александровна — к.ф.-м.н., доцент кафедры медицинской кибернетики и информатики, КрасГМУ, доцент кафедры высшей математики 2, СФУ. E-mail: *nalukyanovakrsk@gmail.com*

Голденок Елена Евгеньевна — к.ф.-м.н., доцент, доцент кафедры медицинской кибернетики и информатики, КрасГМУ, доцент кафедры высшей и прикладной математики, СФУ. E-mail: *egoldenok@gmail.com*

Зарубин Евгений Юрьевич — студент 6 курса специальности 30.05.03 Медицинская кибернетика, КрасГМУ. E-mail: *zaruba1263@gmail.com*

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ НЕЙРОСЕТИ НА КАЧЕСТВО ВИДЕО, СОЗДАВАЕМОГО НА ОСНОВЕ НАБОРА ВИДЕОЗАПИСЕЙ

Р. Д. Маскалёв[1], И. В. Баранова[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

Целью данной работы является исследование влияния параметров генеративно-состязательной нейросети, генерирующей видео по заданной выборке, на качество получаемого видео. Задача генерации заключается в том, чтобы создать видео, максимально похожее на видеоматериалы из обучающей выборки. В работе предлагается нейронная сеть генерации видео на основе архитектуры DCGAN. Проводятся вычислительные эксперименты по генерации видео с помощью разработанной нейросети с заданием различных значений параметров нейросети. Исследуются следующие параметры нейросетей: шум и архитектурные особенности. Выполняется анализ качества сгенерированных видеоматериалов. **Ключевые слова:** *нейронная сеть, генеративно-состязательная нейросеть, генерация, видео, шум*

Введение

Генерация видео является одним из наиболее востребованных и актуальных направлений в области анализа данных, искусственного интеллекта и машинного обучения. Поставленная задача является практически значимой и имеет широкий спектр применения в различных областях деятельности: в образовании, науке, маркетинге, рекламе, производстве, искусстве и т.д. В работе решается следующая разновидность задачи генерации: генерация чёрно-белых видео с движущимися геометрическими объектами. Имеется исходная выборка чёрно-белых видео с движущимися геометрическими объектами, на основе которой выполняется обучение нейронной сети, способной генерировать новые видео с использованием шума. Под шумом понимается множество случайных чисел, которые используются при генерации пикселей кадра видео. Получаемые видео должны быть похожи на видео из исходной выборки (в смысле близости относительно выбранной метрики).

Целью исследования является оценка влияния, оказываемого различными параметрами нейронных сетей на генерацию видео, и подбор таких параметров, чтобы обучение было максимально эффективным. Для этого в данной работе выполняется статистический анализ оценок качества генерации полученных видео.

1. Этапы решения задачи генерации

Решение задачи распознавания объекта на изображениях выполняется в несколько этапов:

1) Предобработка данных: осуществляется загрузка и предварительная обработка данных (изменение размера видео, нормализация и аугментация данных).
2) Обучение модели, генерация и распознавание: выполняется обучение модели генерации видео и модели распознавания, затем выполняется генерация видео.
3) Оценка качества сгенерированных видео: оцениваются результаты генерации видео с использованием следующих метрик качества: *Average Content Distance* (*ACD*), *CLIP Tem-Con*, *Mean Squared Error* (*MSE*) и *Mean Absolute Error* (*MAE*) [1].

2. Математическая постановка задачи

Пусть имеется множество объектов (в данном случае видео) $X = x_1, \ldots, x_n$, где $x_i \in X$ – объект генерации. Пусть F – пространство признаков, тогда функция $f(x) : X \to F$ ставит в соответствие каждому объекту x точку $f(x)$. Вектор $f(x)$ является образом объекта x [2–3].

Требуется построить такое генерирующее правило $g(x)$ на основании доступной информации о функциях $f_i = f(x_i)$, чтобы новый сгенерированный объект $x \in X$, причем полученный объект был близок к объектам исходного множества относительно выбранной метрики ρ.

3. Вычислительные эксперименты по исследованию влияния параметров нейросети на результат генерации видео

Генеративно-состязательная нейросеть (GAN) – алгоритм машинного обучения без учителя, построенный на комбинации из двух нейронных сетей, одна из которых генерирует образцы, а вторая старается отличить сгенерированные объекты от настоящих [4].

В данной работе была предложена нейронная сеть генерации, представляющая собой модификацию нейросети DCGAN [5–6] и использующая трехмерные сверточные слои вместо двумерных. При обучении

в качестве функции потерь использовалась функция бинарной кросс-энтропии [7].

Были проведены вычислительные эксперименты по генерации видео с помощью разработанной нейросети с заданием различных значений параметров нейросети. В качестве исследуемых параметров в работе рассматривались: дополнительные слои исключения нейронов (dropout), различные виды случайных шумов (шум с нормальным распределением, с равномерным распределением, шум Перлина), размер ядер сверточных слоев.

Шум Перлина – математический алгоритм по генерированию процедурной текстуры псевдослучайным методом. Используется в компьютерной графике для увеличения реализма (или графической сложности) поверхности геометрических объектов, для генерации эффекта дыма, тумана и т. д.

Как уже было сказано ранее, в работе решается следующая разновидность задачи генерации: генерация чёрно-белых видео с движущимися геометрическими объектами. В качестве исходного набора видеозаписей в работе использовался датасет Moving MNIST, содержащий 10 000 видеозаписей. Каждое видео состоит из 20 кадров. Размеры кадра – 64 × 64 пикселя. В каждой видеопоследовательности две цифры независимо перемещаются по кадру, пересекаются друг с другом и отскакивают от краев кадра. Каждой разновидностью нейросети (с различным набором параметров) было сгенерировано 50 видео.

На рис. 1 слева приведено несколько примеров с кадрами черно-белых видео из обучающей выборки, справа – кадры видео, сгенерированные нейросетью (с различными значения параметров).

Для набора видео, полученного каждой разновидностью нейросети, были найдены оценки качества сгенерированных видео. После этого было выполнено сравнение полученных значений метрик между собой. Оценивание качества генерации выполнялось с использованием следующих метрик: Average Content Distance (ACD), CLIP Tem-Con, Mean Squared Error (MSE) и Mean Absolute Error (MAE) [1].

Метрика ACD характеризует скорость изменения содержания видео, она находится следующим образом:

$$ACD = \frac{1}{T-1}\sum_{t=1}^{T-1}|V_{t+1} - V_t|,$$

где T – общее количество кадров в видео, $P_t(i,j)$ – значение пикселя в позиции (i,j) кадра t (1 – если пиксель белый, -1 – если чёрный),

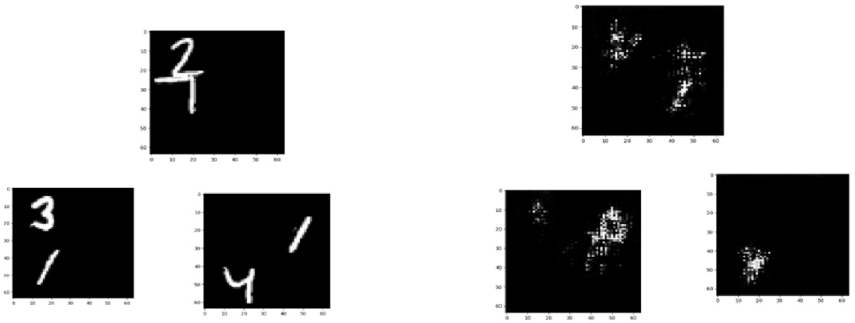


Рис. 1. Слева приведены кадры видео из обучающей выборки, справа – полученные видео, сгенерированные нейросетью

$V_t = \sum_{i=1}^{M} \sum_{j=1}^{N} P_t(i,j)$ – сумма всех пикселей кадра t, M – высота кадра в пикселях, N – ширина кадра в пикселях.

Вычисление CLIP Tem-Con выполняется так: два кадра рассматриваются как матрицы. Берётся каждая строка матрицы, к ней присоединяется следующая, потом к получившейся строке следующая и т.д. Так матрицы превращаются в вектора. После этого находится косинусное расстояние двух кадров как скалярное произведение полученных векторов, поделенное на произведение их норм. Такое косинусное расстояние вычисляется для всех соседних кадров, а затем находится среднее арифметическое между полученными значениями.

Метрики MSE и MAE определяют реалистичность видео по мнению дискриминатора. Метрика MSE вычисляется как среднее квадратичное ошибок дискриминатора:

$$MSE = \frac{1}{n}(y^i{}_{\text{true}} - y^i{}_{\text{pred}})^2.$$

MAE вычисляется как сумма модулей ошибок дискриминатора, делённых на их количество:

$$MAE = \frac{1}{n}|y^i{}_{\text{true}} - y^i{}_{\text{pred}}|,$$

где $y^i{}_{\text{true}}$ – истинное значение для данного видео, а $y^i{}_{\text{pred}}$ – вывод дискриминатора.

В табл. 1 представлены результаты сравнения качества реальных и полученных видео, сгенерированных нейросетью с различными видами шумов, дополнительными слоями исключения нейронов (dropout) и

различным размером ядер сверточных слоев. Оценивание качества генерации выполнялось с использованием вышеописанных метрик ACD, CLIP Tem-Con, MSE и MAE.

Таблица 1

Результаты вычислительных экспериментов по сравнению качества реальных и генерируемых видео

	ACD	**CLIP Tem-Con**	**MSE**	**MAE**
Реальные видео	0,002	0,9	–	–
Шум с нормальным распределением	0,04	0,94	0,5	0,5
Шум с равномерным распределением	0,02	0,94	0,5	0,5
Шум Перлина	**0,01**	**0,97**	**0,5**	**0,5**
Шум Перлина + дропаут	0,01	0,95	0,35	0,39
Шум Перлина + уменьшение ядра	0,1	0,28	0,5	0,5

Из табл. 1 видно, что нейросеть, генерирующая видео с помощью шума Перлина, показывает результаты наиболее близкие к реальным видео относительно метрики ACD. Данная метрика характеризует скорость изменения содержания видео. Нейросети с двумя другими шумами показали существенное отклонение от реальных видео по этой метрике. Добавление дополнительного слоя исключения нейронов (дропаут) к нейросети с шумом Перлина не оказало существенного влияния на значение метрики ACD, а уменьшение ядра сверточного слоя для этой нейросети только ухудшило ситуацию. По метрике CLIP Tem-Con, показывающей временную согласованность, можно увидеть, что нейросети генерируют более плавные видео, чем исходные, и наиболее плавными получаются видео, сгенерированные из шума Перлина. Применение дропаута к сети с шумом Перлина сделало видео менее плавным, но, в целом, временная согласованность в этом случае была всё еще выше, чем для нейросетей с шумами с нормальным или равномерным распределением. Уменьшение ядра сверточного слоя для нейросети с шумом Перлина оказало достаточно сильный отрицательный эффект и привело к существенной временной несогласованности полученных видео.

Как уже было сказано ранее, метрики MSE и MAE определяют реалистичность видео по мнению дискриминатора. Как следует из табл. 1, значения этих метрик практически для всех исследуемых типов нейросетей равны 0,5, что свидетельствуют об обученности дискриминатора у нейросетей, кроме нейросети с дропаутом, она обучается значительно медленнее (поскольку ее значение равно 0,35).

Таким образом, вычислительные эксперименты показали, что наилучшей нейросетью, генерирующей самые качественные видео, т.е. самые близкие к реальным видео по скорости изменения содержания видео, временной согласованности и реалистичности, является нейронная сеть, генерирующая видео на основе шума Перлина. Также проведенные эксперименты выявили, что наибольшее влияние на качество создаваемого видео оказывает шум, а уменьшение ядра свёртки всегда ухудшает результат генерации.

Заключение

В работе была предложена нейронная сеть генерации чёрно-белых видео с движущимися геометрическими объектами на основе исходной выборки. Разработано программное обеспечение на языке программирования Python, реализующее работу разновидностей нейросети генерации видео на основе исходной выборки.

Было проведено исследование влияния, оказываемого различными параметрами нейронной сети на генерацию видео: различными видами шумов, дополнительными слоями исключения нейронов (dropout) и размером ядер сверточных слоев. Проведенный анализ позволил выделить значительный вклад, оказываемый шумом на качество генерируемого видео. Также было обнаружено, что существенное влияние на качество генерации оказывает процесс обучения нейросети.

СПИСОК ЛИТЕРАТУРЫ

1. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2017. 402 с.
2. *Баранова И. В., Гилин С. В.* Гибридный алгоритм распознавания строений на спутниковых снимках на основе метода жука и алгоритма исключения областей // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2024. Т. 13, № 2. С. 56–76.
3. *Лепский А. Е., Броневич А. Е.* Математические методы распознавания образов. Таганрог: Изд-во ТТИ ЮФУ, 2009. 155 с.
4. *Гафаров Ф. М, Галимянов А. Ф.* Искусственные нейронные сети и приложения. Казань: Изд-во Казан. университета, 2018. 121 с.

5. `www.tensorflow.org/tutorials/generative/dcgan?hl=ru` — Глубокая свёрточная генеративно-состязательная сеть Tensorflow. 2025.
6. *Аггарвал Ч.* Нейронные сети и глубокое обучение: учебный курс. М.: Вильямс, 2021. 752 с.
7. *Хайкин С.* Нейронные сети: полный курс. М.: Вильямс, 2020. 1104 с.

Маскалёв Роман Дмитриевич — студент, Институт математики и фундаментальной информатики Сибирского федерального университета. E-mail: *romandivnogorsk999@gmail.com*

Баранова Ирина Владимировна — кандидат физико–математических наук, доцент, доцент, базовая кафедра вычислительных и информационных технологий Института математики и фундаментальной информатики Сибирского федерального университета. E-mail: *irinabar@yandex.ru*

ОБ АЛГОРИТМАХ ГЕНЕРАЦИИ СЕТЕВОГО ГРАФИКА ПРОЕКТОВ ДЛЯ ЗАДАЧИ RCPSP

А. В. Пастушенко[1], Д. В. Семенова[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

В работе рассматриваются две модификации алгоритма, предназначенного для генерации инвестиционных проектов для задачи RCPSP с максимизацией NPV – чистой приведённой стоимости. Эти модификации позволяют создавать проекты с различными характеристиками сети: в первом случает сохраняется способ распределения вершин по уровням, но количество дуг задается случайным образом; во втором случае вершины равномерно распределены по уровням, а вероятность связи с предыдущим уровнем задается как параметр. Полученные экземпляры проектов могут быть использованы для оценки и сравнения оптимизационных алгоритмов решения задачи.

Ключевые слова: *задача календарного планирования, инвестиционный проект, сетевой график проекта*

Введение

Задача календарного планирования проекта (RCPSP – Resource-Constrained Project Scheduling Problem) с критерием максимизации чистой приведенной прибыли (NPV) с ресурсами в денежной форме впервые была рассмотрена в работе Рассела [1] в 1970 г.. С тех пор в литературе активно исследуются различные постановки задачи, многие из которых относятся к труднорешаемым [2–5]. В наших исследованиях основной акцент сделан на задачу календарного планирования инвестиционных проектов с ограниченными ресурсами в денежной форме [6, 4]. Критерием оптимального расписания начала для каждой из работ проекта выступает максимум чистой приведённой стоимости (NPV-критерий), при котором выполняются ограничения на достаточность средств и учитываются технологические взаимосвязи между работами. Данная задача является NP-трудной в сильном смысле [4].

Отдельной темой для исследования в этой области стала разработка генераторов сетей проектов и создания библиотек тестовых задач анализа эффективности методов и алгоритмов решения. Одним из первых случайных генераторов, был генератор для сети типа AoA (Activity-On-Arrow), а также набор Паттерсона из 120 проектов [7] и генератор DAGEN [8]. В 1950-х гг. все ранние проекты использовали этот тип сети.

Только в 1960-х гг. появилась сеть типа AoN (Activity-On-Node). Для решения задач с таким типом сети применяется генератор ProGen, его модификации ProGen/Max, RanGen1, RanGen2 и, основанная на них, открытая библиотека PSLIB эталонных задач [9]. Они используются для тестирования большей части исследований RCPSP. Но стоит отметить, что они предназначены только для задач ресурсно-календарного планирования проекта с ограниченными ресурсами по критерию минимизации длины расписания выполнения всех работ проекта.

Целью работы является разработка и анализ алгоритмов генераторов экземпляров задач календарного планирования проекта с критерием максимизации чистой приведенной прибыли с ресурсами в денежной форме, удовлетворяющих заданным требованиям.

1. Алгоритмы для генерации сети проектов типа AoN

В [10, 11] были предложены алгоритм для генерации сети проектов типа AoN и его модификации, которые рассматривают случаи с постоянной вероятностью появления дуги к вершинам с предыдущего уровня и «затухающей» вероятностью появления дуги к вершинам из всех предыдущих уровней. В данной работе предложены две новых модификации этих алгоритмов – алгоритмы A_1 и A_2.

Первый A_1 – с аналогичным [10] распределением вершин по уровням, но со случайным числом дуг.

Алгоритм A_1 работает следующим образом:

1) На вход подаются общее количество проектов n и количество уровней k. Создается начальный граф с одной вершиной (источник) и инициализируется список вершин предыдущего уровня.
2) Определение размера уровня: для каждого уровня i количество вершин X_i выбирается равномерно из целых чисел от 1 до m_i [10], где

$$m_i = \left\lfloor \frac{n - \sum\limits_{l=1}^{i-1} X_l}{k+1-i} \right\rfloor .$$

   Заметим, что математическое ожидание и дисперсия для числа вершин на каждом уровне имеют вид: $\mathbb{E}[X_i] = (m_i+1)/2$, $\mathbb{D}[X_i] = (m_i^2-1)/12$.
3) Добавление вершин: вычисленное количество вершин добавляется в граф и в список вершин текущего уровня.
4) Создание связей: для каждой вершины текущего уровня случайным образом выбирается количество входящих дуг в диапазоне от 1 до

m_i – количества вершин предыдущего уровня. Затем выбираются случайные вершины для соединения.

5) Шаги 2–4 повторяются для каждого уровня, пока не будет достигнут последний уровень. На последнем уровне добавляется оставшееся количество вершин, необходимое для достижения общего числа проектов n.
6) Добавление стока: добавляется конечная вершина (сток). Если у какой-либо вершины нет исходящих дуг, она соединяется со стоком.
7) Вычисление кратчайшего пути и позиционирование: вычисляется длина кратчайшего пути от источника. На основе этой информации, задается координатное положение уровней и вершин в графе.

На рис. 1 представлены все возможные условные распределения количества вершин X_i на i-м слое ($i = \overline{1,n}$), которые можно получить на втором и третьем шагах алгоритма A_1 при заданном числе вершин $n = 7$ и количестве слоёв $k = 3$. Заметим, что распределения на каждом уровне вычисляются рекурсивно, а при заданных параметрах возможны пять различных условных распределений $\mathbf{P}(X_i|X_{i-1})$. Благодаря такому дереву можно наглядно показать, какие существуют варианты распределения вершин, сколько их всего и найти вероятность осуществления каждого из вариантов.

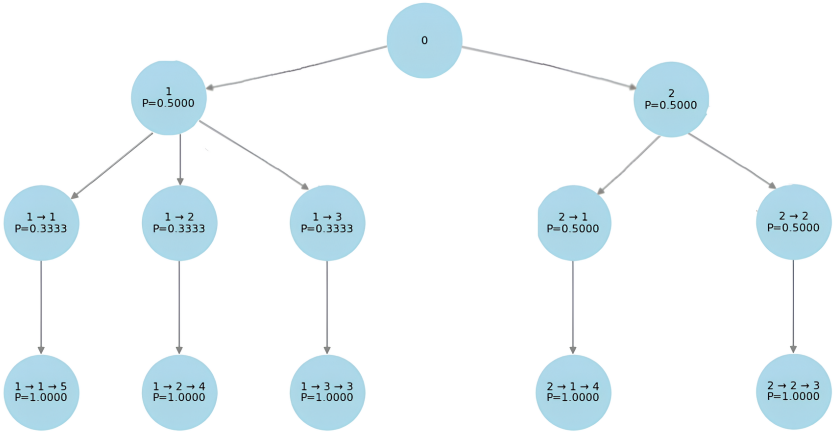


Рис. 1. Дерево вариантов при $n = 7$, $k = 3$

Второй случай – с равномерным распределением вершин по уровням и с заданной вероятностью появления дуги к вершинам с предыдущего уровня реализован в алгоритме A_2.

На вход алгоритм A_2 получает: n – общее количество вершин, представляющих проекты; k – количество уровней в графе; p – вероятность создания ребра между вершинами соседних уровней.

Работает алгоритм следующим образом:

1) На вход подаются n, k и p. Создается начальный граф с источником и инициализируется словарь уровней и список вершин.
2) Определение размера уровня. Вычисляется количество вершин на каждом уровне: $L = \max(1, \lfloor \frac{n}{k} \rfloor)$. Это гарантирует, что на каждом уровне будет хотя бы одна вершина, но не больше, чем пропорционально общему числу вершин и уровней.
3) Добавление вершин. Вычисленное количество первых L вершин добавляется в граф и в список вершин текущего уровня, затем удаляются из общего списка.
4) Создание связей. С вероятностью p появляется дуга между вершинами с предыдущего уровня и с текущего уровня. Если вдруг вершина текущего уровня не связана ни с одной вершиной с предыдущего, добавляется дуга от случайной вершины предыдущего уровня.
5) Шаги 2–4 повторяется для всех уровней, если в списке остались вершины, они уходят в последний слой.
6) Добавление стока. Добавляется конечная вершина (сток). Если у какой-либо вершины нет исходящих дуг, она соединяется со стоком.
7) Позиционирование. Используя словарь уровней, задается координатное положение уровней и вершин в графе.

Оба алгоритма создают сети с $n + 2$ вершинами (n промежуточных проектов, источник и сток), которые упорядочены по уровням. Максимальная глубина сети (количество уровней) не превышает $k + 2$. Важно отметить, что n больше k, и в сетях нет лишних связей (избыточных дуг).

Характеристики сетей, полученных алгоритмом A_2 представлены в табл. 1. Для оценки сетей использовались популярные метрики из работы [12]. В таблице используются следующие обозначения: CNC – коэффициент сложности сети; CNC' – коэффициент Кайманна; $T_density$ – общая плотность активности; ADD – средняя плотность активности; C – сложность сети; A' – количество избыточных дуг.

Заключение

Обшей целью исследования является разработка генераторов сетей проектов для RCPSP с критерием максимизации чистой приведённой стоимости (NPV) и создания библиотеки задач анализа эффективности методов и алгоритмов решения. В работе сделан акцент на алгорит-

Таблица 1
Характеристики сетей, полученных алгоритмом A_2, с числом вершин $n = 13$

p_{const}	k	CNC	CNC'	C	$T_{density}$	ADD	A'
0,25	3	1,53	35,26	0	10	0,66	0
0,50	3	1,6	38,4	0	10	0,66	0
0,75	3	2,15	60,26	0	12	0,92	0
0,25	5	1,30	22,23	0	6	0,46	0
0,50	5	1,42	28,57	0	7	0,5	0
0,75	5	1,58	30,08	0	9	0,75	0

мы генерации сети проектов типа AoN и предложены две модификации алгоритма. Оценки сложности сетей, полученных алгоритмом A_2, проиллюстрированы на примере сети малой размерности, однако аналогичная методика применима для генерации сетей с большим числом проектов. Перспективным направлением исследований является анализ влияния параметров алгоритма на характеристики сети, а также модификация алгоритмов генерации сетей с учётом длительности проектных работ (на данном этапе длительность работ равна и период считается максимальным).

СПИСОК ЛИТЕРАТУРЫ

1. *Russell A. H.* Cash Cows in networks // Management Science. 1970. P. 357–373.
2. *Гимади Э. Х., Залюбовский В. В., Севастьянов С. В.* Полиномиальная разрешимость задач календарного планирования со складируемыми ресурсами и директивными сроками // Дискретный анализ и исследование операций. Сер. 2. 2000. Т. 7. № 1. С. 9–34.
3. *Leyman P. and Vanhoucke M.* Payment models and net present value optimization for resource constrained project scheduling // Computers& Industrial Engineering. 2016. Vol. 91. P. 139–153.
4. *Казаковцева Е. А., Сервах В. В.* Сложность задачи календарного планирования с кредитованием // Дискретный анализ и исследование операций. 2015. Т. 22, № 4. С. 35–49.
5. *Гимади Э. Х., Гончаров Е. Н., Штепа А. А.* Быстрый алгоритм вычисления нижней оценки для решения задачи ресурсно-календарного планирования с тестированием на примерах библиотеки PSPLIB // Тр. ИММ УрО РАН. 2021. Т. 27, № 1. С. 22–36.

6. *Булавчук А. М., Семенова Д. В.* Применение методов идемпотентной алгебры в генетическом алгоритме для решения задачи календарного планирования // Прикладная дискретная математика. 2022. № 58. С. 112–124.
7. *Patterson J. H.* A comparison of exact approaches for solving the multiple constrained resource, project scheduling problem // Management Science. 1984. Vol. 30, № 7. P. 854–867.
8. *Agrawal M. K., Elmaghraby S. E., Herroelen W. S.* DAGEN: a generator of test sets for project activity nets // European Journal of Operational Research. 1996. Vol. 90, № 2. P. 376–382.
9. *Kolisch R., Hartmann S.* Heuristic algorithms for solving the resource-constrained project scheduling problem – Classification and computational analysis // Project scheduling: Recent models, algorithms and applications Berlin: Springer, 1999. P. 147–178.
10. *Булавчук А. М., Семенова Д. В., Солдатенко А. А.* Об одном генераторе проектов для задачи RCPSP // ИТ. НАУКА. КРЕАТИВ. Т. 5. Системы управления, информационные технологии и математическое моделирование: материалы I Междунар. форума (Омск, 14–16 мая 2024 г.) / науч. ред. П. С. Ложников, отв. ред. И. Г. Ольгина. Омск: Изд-во ОмГТУ, 2024. С. 70.
11. *Булавчук А. М., Семенова Д. В., Солдатенко А. А.* О модификациях алгоритма генерации сетевого графика проектов для задачи RCPSP // Информационные технологии и математическое моделирование (ИТММ-2024), Томск: Информационные технологии и математическое моделирование (ИТММ-2024), 2024. С. 287–292.
12. *Browning T. R., Yassine A. A.* A random generator of resource-constrained multi-project network problems // J Sched. 2010. Vol. 13. P. 143–161.

Пастушенко Анастасия Витальевна — студент Института математики и фундаментальной информатики СФУ.. E-mail: *apastushenko-im22@stud.sfu-kras.ru*

Семенова Дарья Владиславовна — кандидат физ.-мат. наук, доцент, доцент кафедры высшей и прикладной математики СФУ. E-mail: *DVSemenova@sfu-kras.ru*

МЕТОД ВЫБОРА МОДЕЛИ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ СО СТЕПЕННЫМИ ПАРАМЕТРАМИ ДЛЯ ОБРАБОТКИ СИГНАЛОВ

Е. А. Пчелинцев[1], З. М. Курдошев[1,2]

[1] *Томский государственный университет, г. Томск, Россия*
[2] *Самаркандский государственный университет, г. Самарканд, Узбекистан*

В работе развивается подход на основе нейросетевого моделирования для реконструкции неизвестного сигнала по зашумленным наблюдениям, описываемым стохастическим дифференциальным уравнением с малым шумом. В гибридную нейронную архитектуру введен новый слой Model Selection Method (MSM), включающий сверточные и статистические компоненты. Для контролируемого обучения ИНС создан синтетический набор данных из «чистых» и зашумленных сигналов. Предложена процедура выбора модели на основе взвешенных оценок МНК со степенными весами. **Ключевые слова:** *статистическая обработка сигналов, нейронные сети, гибридные модели*

Введение

Рассматривается проблема оценивания неизвестного детерминированного сигнала S по дискретным наблюдениям процесса y_t, который является решением стохастического дифференциального уравнения

$$dy_t = S_t dt + \varepsilon dw_t, \quad 0 \leqslant t \leqslant 1, \tag{1}$$

где S_t – неизвестный квадратично интегрируемый сигнал, шум $(w_t)_{t \geqslant 0}$ – стандартный винеровский процесс, $\varepsilon > 0$ – интенсивность шума. Процесс доступен наблюдению лишь в точках $t_l = l/n$, $l = \overline{0,n}$ где $n \geq 3$ – объем выборки. Эта модель имеет важное значение в изучении различных стохастических процессов. Они стали ключевым инструментом для моделирования и анализа важных процессов в медицине [1, 2] финансах [3], статистике [4, 5], метрологии [6], обработке сигналов [7], биологии [8] и многих других областях. Реконструкция функции $S(t)$ в этой модели имеет принципиальное значение для принятия решений и прогнозирования, а ее наиболее точная оценка очень важна с точки зрения статистического анализа и практической применимости. Традиционные параметрические и непараметрические методики оценивания

Работа поддержана РНФ, проект № 24-11-00191.

зависят от параметров функционального класса, которому принадлежит $S(\cdot)$. Важной практической задачей является выбор параметров, минимизирующих ошибку оценивания. Для выбора наилучших параметров широко применяются методы, основанные на пенализационных критериях Акаике, Шварца, кросс-валидации и оракульных неравенствах [9–12]. Цель данной работы – автоматизировать процесс выбора модели, основанного на точных оракульных неравенствах, применяя методы машинного обучения.

1. Построение нейросетевой процедуры выбора модели

Построение нейросетевого алгоритма выбора модели основано на статистическом непараметрическом методе обработки сигналов из [5]. Пусть неизвестный сигнал S в (1) принадлежит Соболевскому шару в $\mathcal{L}_2[0,1]$:

$$\Theta_{k,r} = \{S \in \mathcal{L}_2[0,1] : \sum_{j=1}^{\infty} j^{2k}\theta_j^2 \leqslant r/\pi^{2k}\}. \tag{2}$$

Здесь $k \geq 1$ и $r > 0$, $\theta_j = (S, \varphi_j)$, $j \geqslant 1$, – коэффициенты Фурье для функции S относительно базиса $(\varphi_j)_{j\geqslant 1}$ в пространстве $\mathcal{L}_2[0,1]$, определенного ниже. Для оценивания сигнала $S \in \Theta$ в (1) используем следующие взвешенные оценки наименьших квадратов

$$\hat{S}_{\lambda,\varepsilon}(t) = \sum_{j=1}^{n_*} \lambda_j \hat{\theta}_j \varphi_j(t). \tag{3}$$

Здесь весовые коэффициенты

$$\lambda_j = \lambda_j(k,r) = 1 - \left(\frac{j}{n_*}\right)^k, \tag{4}$$

где

$$n_* = \max\left\{ l \geqslant 1 \,:\, l^k \sum_{j=1}^{l} j^k - \sum_{j=1}^{l} j^{2k} \leqslant \varepsilon^{-2} r/\pi^{2k} \right\}. \tag{5}$$

Оценки МНК для коэффициентов Фурье:

$$\hat{\theta}_j = \frac{1}{n}\sum_{l=1}^{n} \varphi_j(t_l)\Delta y_l, \quad \Delta y_l = y_l - y_{l-1}, \tag{6}$$

и $(\varphi_j)_{j\geqslant 1}$ – тригонометрический базис в $\mathcal{L}_2[0,1]$: $\varphi_1 = 1$, а для $j \geqslant 2$

$$\varphi_j(t) = \begin{cases} \sqrt{2}\sin(2\pi[\frac{j}{2}]t), & \text{для нечетных } j, \\ \sqrt{2}\cos(2\pi[\frac{j}{2}]t), & \text{для четных } j, \end{cases} \tag{7}$$

где $[\cdot]$ обозначает целую часть.

Для измерения качества оценки $\hat{S}$ будем использовать квадратичный риск

$$R(S;\hat{S}) = \mathbf{E}_S||S-\hat{S}||^2, \quad ||f||= \int_0^1 f^2(t)dt, \tag{8}$$

где $\mathbf{E_S}$ обозначает ожидание относительно распределения $\mathbf{P_S}$ наблюдений $(y_t)_{0\leqslant t\leqslant 1}$ в (1) при заданной S.

Согласно подходу из [12] введем следующее дискретное параметрическое пространство:

$$\mathcal{A}_\uparrow = \{(1,\ldots,k_d)\times(r_1,...,r_d)\}, \tag{9}$$

где

$$k_d = \left[k_0 + \sqrt{\ln\left(1+\frac{1}{\varepsilon^2}\right)}\right], \quad k_0 \geq 0, \tag{10}$$

$$r_i = \frac{i}{\ln\left(1+\frac{1}{\varepsilon^2}\right)}, \quad i=1,\ldots,d, \quad d=d_\varepsilon = \left[\frac{1}{\varepsilon^2}\right]. \tag{11}$$

В [5] установлено, что оценки (3) являются асимптотически эффективными при $\varepsilon\to 0$ для любых значений параметров $k \geqslant 1$ и $r > 0$. Заметим, что эффективная скорость сходимости и константа Пинскера (определяющая минимальное значение риска) для модели (1) зависят от выбранных параметров и определяются соответственно как $v_\varepsilon = \varepsilon^{-2k/(2k+1)}$ и

$$l(k,r) = \left(\frac{(2k+1)r}{\pi^{2k}}\right)^{1/(2k+1)} \left(\frac{2k}{2k+2}\right)^{2k/(2k+1)}.$$

Таким образом, параметры напрямую влияют на точность оценки $\hat{S}(t)$ через n_* (максимальное количество базисных функций, используемых при оценке), определенного в (5) и оптимальный выбор k и r определяет сходимость $\hat{S}(t)$ к $S(t)$. При определении оптимальных параметров в связи с вышеперечисленными зависимостями возникают следующие проблемы:

- неопределенное влияние параметров, т.е. каждая комбинация k и r по-разному влияет на точность $\hat{S}(t)$;
- расчет n_* для каждой пары параметров требует итерационного процесса, что затрудняет обработку в реальном времени.

Поэтому актуальным является разработка автоматического механизма выбора оптимальных значений k и r на основе ИНС, которые позволяют осуществить быстрый поиск в пространстве параметров при минимизации ошибок. После определения параметров производится реконструкция сигналов с использованием всех найденных комбинаций по формулам (3)–(7). Комбинация параметров k и r, используемая для достижения оптимальных результатов при реконструкции сигнала, зависит от характеристик каждого сигнала. Механизм стробирования, основанный на принципе «Смесь экспертов» [14, 15], используется для выбора наиболее подходящей комбинации для каждого сигнала из определенных параметров. Механизм состоит из трех слоев:

$$h^{(1)} = \hat{S}_{k,r_d} \cdot W^{(1)} + b^{(1)}, \quad W^{(1)} \in \mathbb{R}^{d \times 64}, \tag{12}$$

$$h^{(2)} = \mathrm{ReLU}(h^{(1)}), \tag{13}$$

$$z = h^{(2)} \cdot W^{(2)} + b^{(2)}, \quad W^{(2)} \in \mathbb{R}^{64 \times d}. \tag{14}$$

где $\hat{S}_{k,r}$ – кандидат восстановленного сигнала. $W^{(1)}, W^{(2)}$ – веса. В процессе обучения выполняются следующие шаги:

1) из всех $\hat{S}_{k,r}$ выбирается модель с наибольшим весом:

$$(k^*, r^*) = \underset{(k,r) \in A}{arg\,max}\, w_{k,r}; \tag{15}$$

2) выбранные параметры k^* и r^* считаются оптимальными для данного типа сигналов.

Рис. 1 иллюстрирует архитектуру ИНС. Для предлагаемой обучения

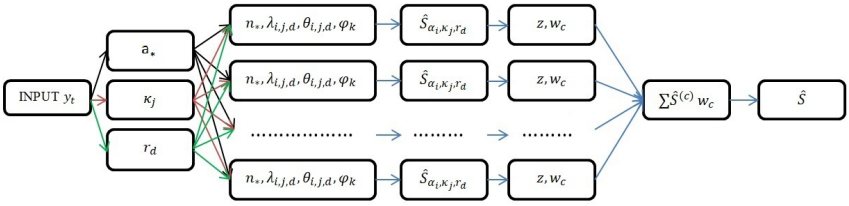


Рис. 1. Архитектура ИНС

ИНС используется 5 различных синтетических сигналов со случайными параметрами, перечисленные в табл. 1.

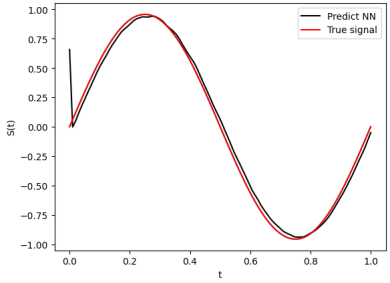
2. Результаты

После обучения ИНС тестировали для разных сигналов с разными свойствами. Например, сигналы, которых не было в обучающей базе данных. Результаты тестирования показаны в табл. 1 и на рис. 2.

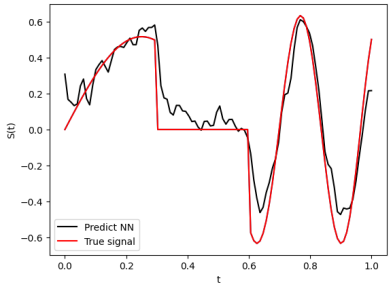
Таблица 1

Эмпирические квадратические риски при обработке сигналов для различных ε

Тип сигнала	$\varepsilon = 0,01$	$\varepsilon = 0,05$	$\varepsilon = 0,1$
Синусоидальных сигналов	0,0093	0,1041	0,2927
Электрические импульс	0,0279	0,0705	0,2389
Линейная частотная модуляция	0,0322	0,1346	0,3316
Сумма сигналов	0,0105	0,0869	0,2692
Комбинация сигналов	0,0368	0,1822	0,2612



(a) Синусоида при $\varepsilon = 0,01$, $S(t) = 0,98\sin(2\pi t)$



(b) Комбинация сигналов при $\varepsilon = 0,1$: $S(t) = 0,54\sin(2\pi t)_{(0<t<0,3)} + 0,47 \cdot \mathbb{1}_{(t_{0,1},\ t_{0,1}+0,29)}(t)_{(0,3<t<0,6)} + 0,63\sin(2\pi(0,3t + 5t^2/2))_{(0,6<t<1)}$

Рис. 2. Средние значения предсказания НС и вид исходного сигнала

Полученные результаты показывают, что предложенный гибридный алгоритм MSM достаточно эффективен, особенно с очень высокой точ-

ностью восстанавливает синусоидальные сигналы, а также сигналы, которые представляют собой различные комбинации базовых сигналов.

Заключение

Основной вклад работы включает разработку механизма стробирования для динамического выбора модели и интеграцию гибридного статистического и нейросетевого подходов, который использует преимущества обеих методологий. MSM продемонстрировал надежную производительность на 5 типах базовых сигналов и их комбинациях, дает надежные оценки даже при высоких уровнях шума, при этом представленный метод является вычислительно эффективным и действенным для приложений в реальном времени.

СПИСОК ЛИТЕРАТУРЫ

1. *Hamzi B., Owhadi H., Paillet L.* A note on microlocal kernel design for some slow–fast stochastic differential equations with critical transitions and application to EEG signals // Physica A: Statistical Mechanics and its Applications. 2023. Vol. 616. P. 128583.
2. *Tajmirriahi M., Amini Z.* Modeling of seizure and seizure-free EEG signals based on stochastic differential equations // Chaos, Solitons & Fractals. 2021. Vol. 150. P. 111104.
3. *Platen E., Bruti-Liberati N.* Numerical solution of stochastic differential equations with jumps in finance // Springer Science & Business Media, 2010. Vol. 64.
4. *Ruttor A., Batz P., Opper M.* Approximate Gaussian process inference for the drift function in stochastic differential equations // Advances in Neural Information Processing Systems. 2013. Vol. 26.
5. *Pchelintsev E., Pergamenshchikov S., Povzun M.* Efficient estimation methods for non-Gaussian regression models in continuous time // Annals of the Institute of Statistical Mathematics. 2022. P. 1–30.
6. *Eichstadt S., Ruhm K., Shestakov A.* Dynamic measurement and its relation to metrology, mathematical theory and signal processing: A review // Journal of Physics: Conference Series. IOP Publishing, 2018. Vol. 1065, № 21. P. 212018.
7. *Seghouane A. K., Shah A., Ting C. M.* fMRI hemodynamic response function estimation in autoregressive noise by avoiding the drift // Digital Signal Processing. 2017. Vol. 66. P. 29–41.
8. *Braumann C. A.* Introduction to stochastic differential equations with applications to modelling in biology and finance. // John Wiley & Sons, 2019.
9. *Akaike H.* A new look at the statistical model identification // IEEE Transactions On Automatic Control. 1974. Vol. 19. P. 716–723.

10. *Schwarz G.* Estimating the dimension of a model // The Annals Of Statistics. 1978. P. 461–464.

11. *Browne M. W.* Cross-validation methods // Journal of mathematical psychology. 2000. Vol. 44, № 1. P. 108–132.

12. *Pchelintsev E. A., Pergamenshchikov S. M.* Oracle inequalities for the stochastic differential equations // Statistical inference for stochastic processes. 2018. Vol. 21. № 2. P. 469–483.

13. *Pinsker M. S.* Optimal filtering of square-integrable signals in Gaussian noise // Problemy Peredachi Informatsii. 1980. Vol. 16. № 2. P. 52–68.

14. *Masoudnia S., Ebrahimpour R.* Mixture of experts: a literature survey // Artificial Intelligence Review. 2014. Vol. 42. P. 275–293.

15. *Yuksel S. E., Wilson J. N., Gader P. D.* Twenty years of mixture of experts // IEEE transactions on neural networks and learning systems. 2012. Vol. 23. № 8. P. 1177–1193.

Пчелинцев Евгений Анатольевич — доктор физико-математических наук, доцент, заведующий кафедрой математического анализа и теории функций, механико-математический факультет, Томский государственный университет. E-mail: *evgen-pch@yandex.ru*

Курдошев Зарифжон Мансур угли — аспирант, механико-математический факультет, Томский государственный университет, ассистент, кафедра математического моделирования, математический факультет, Самаркандский государственный университет. E-mail: *zarifbek001@gmail.com*

РАЗРАБОТКА ВЕБ-САЙТА ДЛЯ ПОИСКА ИНФОРМАЦИИ О НАУЧНЫХ ИЗДАНИЯХ

А. А. Солдатенко[1]

[1] *Сибирский федеральный университет, г. Красноярск, Россия*

В работе представлен процесс создания веб-сайта для поиска журналов, входящих в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук. Основной целью разработки было быстрое получение информации о принадлежности журнала конкретной научной специальности, а также получение информации о включении соответствующего журнала в белый список. **Ключевые слова:** *веб-разработка, рецензируемые издания, поисковые системы, базы данных*

Введение

В настоящее время, подбор научных журналов для публикации своих исследований соискателями ограничен перечнем рецензируемых научных изданий, который представлен Высшей аттестационной комиссией при Министерстве науки и высшего образования Российской Федерации [1]. Данный перечень представлен в виде PDF файла, содержащего 3206 научных изданий (на момент 19.05.2025). Подбор журнала по желаемой научной специальности в данной ситуации возможен только ручным поиском по файлу.

Также Министерством науки и высшего образования Российской Федерации организован «Белый список» – список научных журналов, созданный в целях обеспечения мониторинга и оценки публикационной активности. Российский центр научной информации представил вебсайт с актуальной информацией по научным изданиям, входящим в «белый список», а также по другим наукометрическим базам данных [2]. Однако, подбор научных изданий по научной специальности также не предусмотрен.

Целью данной работы является восполнение указанного недостатка, в виде создания веб-сайта, который будет содержать информацию о рецензируемом перечне научных изданий, вхождении в «белый список», и предоставлять возможность быстрого поиска по заданной научной специальности и отрасли науки. В работе будут описаны шаги, которые были проделаны для достижения данной цели.

1. Получение информации из перечня рецензируемых изданий

Как было сказано ранее, перечень представлен на официальном вебсайте ВАК в виде PDF файла. Файл от 19 мая 2025 г. содержит 1281 страницу, на которых представлены 3206 научных издания. Для каждого издания указан уникальный номер ISSN, перечень научных специальностей и соответствующих отраслей науки, а также дата включения издания в перечень. Как известно, создание PDF может быть различным, существенно зависит от входных данных, алгоритма его создания и множества других переменных [3]. Поэтому автоматическое выделение информации требует индивидуального подхода к каждому файлу.

В работе для этих целей использовался язык Python с пакетами Camelot, Pandas и NumPy. Перечень представляет собой многостраничную таблицу, каждый фрагмент которой можно извлечь пакетом Camelot с каждой страницы. Для общего хранения всего массива данных используется пакет Pandas. Поскольку в документе встречаются научные журналы, информация о которых представлена на нескольких страницах, поэтому во время извлечения таблицы требуется хранить предыдущую запись, чтобы с помощью пакета Pandas их можно было соединить.

При обработке перечня были выявлены следующие особенности, которые затрудняют автоматическую обработку:

- опечатки в наименовании специальностей, научных направлений и дат принятия;
- отсутствие единообразия названия в некоторых научных направлениях;
- многочисленные пробелы до, в и после некоторых строк.

Первые две особенности легко отслеживаются при проверки уникальных значений, однако последняя особенность доставляет серьезные неудобства, поскольку итоговая строка с пробелами может выступать за границы страницы, вследствие чего пакет Camelot и другие, неправильно распознают строку и могут поместить ее в другую ячейку таблицы. В большинстве случаев данная особенность встречается в наименовании научных специальностей.

В ходе извлечения данных из перечня, для каждого издания была получена информация о сайте издания, если она представлена на портале регистрации ISSN [4]. Стоит отметить, что только для 870 изданий при регистрации ISSN были указаны веб-адреса, остальные, в данный момент, указаны как “None”. Также из-за неидеальности выделения информации из файла, у некоторых научных специальностей



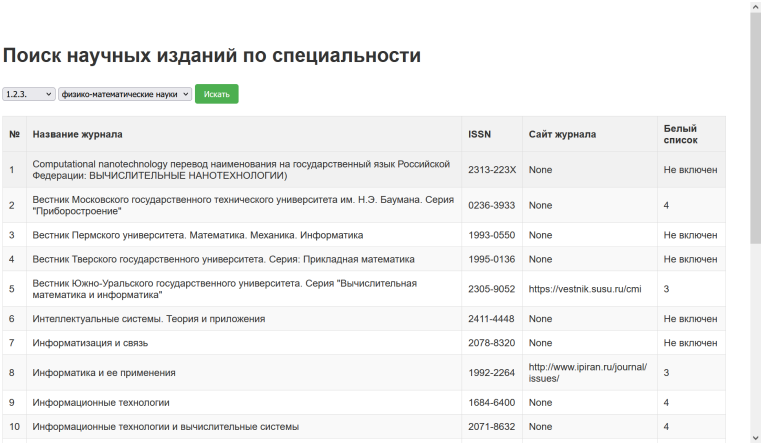
Рис. 1. ER-диаграмма базы данных рецензируемых научных изданий

была утеряна отрасль науки, поэтому в базе данных она будет представлена как «-».

2. Организация базы данных

Для хранения полученной информации используется база данных PostgreSQL. Схема базы данных представлена в виде ER-диаграммы на рис. 1.

Основным ключом и уникальным идентификатором журнала является ISSN. За каждым научным журналом закреплен список научных специальностей и отраслей, которые соответствуют тематике журнала, что отражено в таблицах `journal_spec_table` и `spec_table`. Дополнительно хранится булева переменная `new_spec` для отражения, новой или старой номенклатуре соответствует научная специальность. В таблице `date_table` отражено, когда научному журналу была присвоена научная специальность. Поскольку журналы могут иметь несколько ISSN (для печатной и online версий), то дополнительный номер хранится в таблице `issn_table`. Таблица `white_table` хранит в себе уровень журнала в «белом списке», и по существу является кэширующей, чтобы снизить число обращений к API серверам «белого списка». Для обновления информации о положении журнала в «белом списке», используется переменная `stale_date`, которая отмечает время устаревания информации и равна одной неделе от последнего обращения к «белому списку».



№	Название журнала	ISSN	Сайт журнала	Белый список
1	Computational nanotechnology перевод наименования на государственный язык Российской Федерации: ВЫЧИСЛИТЕЛЬНЫЕ НАНОТЕХНОЛОГИИ)	2313-223X	None	Не включен
2	Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия "Приборостроение"	0236-3933	None	4
3	Вестник Пермского университета. Математика. Механика. Информатика	1993-0550	None	Не включен
4	Вестник Тверского государственного университета. Серия: Прикладная математика	1995-0136	None	Не включен
5	Вестник Южно-Уральского государственного университета. Серия "Вычислительная математика и информатика"	2305-9052	https://vestnik.susu.ru/cmi	3
6	Интеллектуальные системы. Теория и приложения	2411-4448	None	Не включен
7	Информатизация и связь	2078-8320	None	Не включен
8	Информатика и ее применения	1992-2264	http://www.ipiran.ru/journal/issues/	3
9	Информационные технологии	1684-6400	None	4
10	Информационные технологии и вычислительные системы	2071-8632	None	4

Рис. 2. Внешний вид сайта с запросом по специальности 1.2.3. и отрасли физико-математические науки

3. Создание веб-сайта

Сайт разработан с использованием стандартной тройки PHP, HTML и JavaScript, без привлечения веб-фреймворков. Основной функционал сайта состоит в следующем:

– Выбор научной специальности и отрасли науки для поиска;
– Получение информации об уровне журнала в «белом списке».

Вид сайта представлен на рис. 2.

Дополнительную информацию об издании можно получить путем нажатия на соответствующую строку таблицы. Внешний вид информационного блока представлен на рис. 3.



Рис. 3. Подробная информация о журнале, представляемая при нажатии на строку таблицы

Заключение

В рамках работы представлен процесс создания веб-сайта для поиска научных изданий, подходящих для публикации научных работ по определенным специальностям, а также получению дополнительных сведений о журнале. В дальнейшем предполагается развитие функционала сайта для более гибкой фильтрации научных изданий, добавление информации о вхождении в другие наукометрических базы данных, добавление описания научного издания, его требования к оформлению.

СПИСОК ЛИТЕРАТУРЫ

1. `https://vak.minobrnauki.gov.ru/` — Высшая аттестационная комиссия при Министерстве науки и высшего образования Российской Федерации. 2025.
2. `https://journalrank.rcsi.science/ru/` — «БЕЛЫЙ СПИСОК» научных журналов РЦНИ. 2025.
3. `https://opensource.adobe.com/dc-acrobat-sdk-docs/pdfstandards/PDF32000_2008.pdf` — Document management – Portable document format – Part 1: PDF 1.7. 2008.
4. `https://portal.issn.org/` — The ISSN Portal. 2025.

Солдатенко Александр Александрович — кандидат физико-математических наук, доцент, Институт высшей и прикладной математики. E-mail: *ASoldatenko@sfu-kras.ru*

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ЦИФРОВОГО ПРЕДЫСКАЖЕНИЯ СИГНАЛА С ИСПОЛЬЗОВАНИЕМ АРХИТЕКТУРЫ КОСВЕННОГО ОБУЧЕНИЯ И РАЗЛИЧНЫХ АДАПТИВНЫХ АЛГОРИТМОВ

В. Т. Хоанг[1]

[1] *Московский физико-технический институт Московский физико-технический институт,141701, Московская область, г. Долгопрудный, Институтский переулок, д.9*

В данной статье представлен сравнительный анализ алгоритмов цифрового предыскажения, реализованных в архитектуре косвенного обучения и предназначенных для линеаризации усилителя мощности. Рассмотрены алгоритмы LMS, RLS, RPEM и RLS-Калмана. Моделирование показало, что использование данных методов позволяет существенно снизить нелинейные искажения, улучшить спектральные характеристики усилителя и повысить его энергетическую эффективность. Полученные результаты подтверждают, что при применении адаптивного цифрового предыскажения позволять повысить производительность системы передачи данных. **Ключевые слова:** *DPD,цифровое предыскажение, усилитель мощности,LMS, RLS, RPEM, RLS-Калмана*

Введение

В условиях стремительного развития стандартов беспроводной связи нового поколения (DVB-T2, LTE, 5G и др.) [1, 2] возрастает потребность в обеспечении высокой линейности и энергетической эффективности передающих трактов. Усилитель мощности (УМ) является критически важным элементом, однако при работе вблизи насыщения возникают нелинейные искажения и эффекты памяти, ухудшающие качество сигнала. Цифровое предыскажение (DPD) [3] позволяет компенсировать эти искажения. Среди различных архитектур особый интерес представляет косвенное обучение (ILA) [4–7], отличающееся низкой вычислительной сложностью и практической реализации. Однако эффективность ILA в значительной степени определяется качеством применяемых адаптивных алгоритмов, которые непосредственно влияют на итоговую производительность системы. В работе [4] рассматривались

только RLS, фильтр Калмана и RPEM, без учёта LMS в отдельный метод. Кроме того, не проводился анализ вычислительной сложности, что ограничивает практическую оценку, Дополнительно используется модель усилителя Винера–Салеха, позволяющая точнее учитывать как AM/AM, так и AM/PM нелинейности при наличии эффектов памяти.Настоящая работа посвящена сравнительному анализу эффективности четырёх адаптивных алгоритмов — LMS (алгоритм наименьших среднеквадратичных ошибок), RLS (рекурсивный метод наименьших квадратов), RPEM (рекурсивный метод минимизации ошибки предсказания) и RLS-Калмана (модифицированный RLS с фильтрацией Калмана) — для линеаризации усилителя мощности.

1. Модель цифрового предыскажения

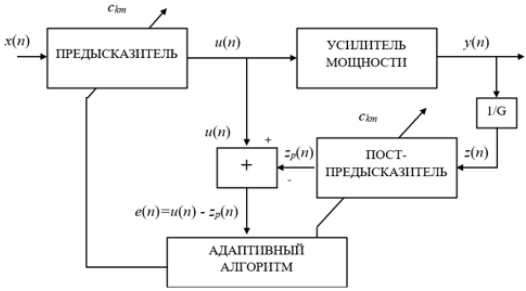


Рис. 1. Архитектура косвенного обучения

Рис. 1 показывает блок-схему системы цифрового предыскажения [7], реализованной по архитектуре косвенного обучения (ILA). Взаимосвязь между входным и выходным сигналом модели предыскажителя математически описывается следующим образом:

$$u(n) = \sum_{k=1}^{K} \sum_{m=0}^{M} c_{km}\, x(n-m)\, |x(n-m)|^{k-1}, \tag{1}$$

$$z_p(n) = \sum_{k=1}^{K} \sum_{m=0}^{M} c_{km}\, z(n-m)\, |z(n-m)|^{k-1}, \tag{2}$$

$$y(n) = \sum_{k=1}^{K} \sum_{m=0}^{M} a_{km}\, u(n-m)\, |u(n-m)|^{k-1}, \tag{3}$$

где a_{km} характеризуют параметры модели УМ.

Алгоритм 1. LMS-DPD

Выход: $c(n)$
Вход: $n = 0$, $c(0)$, μ
for $n = 1$ to L
1. $e(n) = u(n) - z_p(n)$
2. $c(n) = c(n-1) + \mu \cdot e(n) \cdot z^T(n)$

end

2. Адаптивные алгоритмы

2.1. Алгоритм LMS

Алгоритм LMS выводится на основе минимизации среднеквадратичной ошибки (MSE) $E\left\{e^2(n,c)\right\}$, обновление коэффициентов $c(n)$ осуществляется с использованием правила стохастического градиентного спуска с постоянным шагом обучения μ. Преимущества LMS заключаются в его простоте и низкой вычислительной сложности, что делает алгоритм особенно подходящим для реализации в условиях ограниченных вычислительных ресурсов и в системах реального времени. Однако при высокой нелинейности усилителя мощности и наличии эффектов памяти сходимость может быть медленной и чувствительной к выбору μ [5]. Обновление коэффициентов по правилу стохастического градиентного спуска требует порядка $\mathcal{O}(P)$ операций умножения/сложения на один отсчёт (P = K(M+1)).

2.2. Алгоритм RLS

Алгоритм RLS применяется для адаптивной настройки коэффициентов предыскажателя на основе минимизации экспоненциально взвешенной суммы квадратов ошибок. Ключевым элементом является коэффициент забывания $\lambda \in (0,1)$, который регулирует влияние предыдущих наблюдений. Чем меньше значение λ, тем быстрее информация о прошлых данных теряет значение. Таким образом, выбор λ определяет способность алгоритма отслеживать изменения параметров усилителя мощности во времени. Однако недостатками являются чувствительность к шуму при малом λ и высокая вычислительная сложность, связанная с обновлением ковариационной матрицы [4]. Рекурсивное обновление ковариационной матрицы размера $P \times P$ и вектора коэффициентов даёт сложность $\mathcal{O}(P^2)$.

2.3. Алгоритм RPEM

Алгоритм RPEM используется для оценки коэффициентов предыскажателя с целью повышения устойчивости адаптации и снижения колебаний при изменении характеристик усилителя мощности. В отли-

Алгоритм 2. RLS-DPD

Выход: $c(n)$
Вход: $n = 0$, λ, $P(0) = \rho I$, $c(0)$
for $n = 1$ to L
1. $e(n) = u(n) - z_p(n)$
2. $k(n) = \frac{P(n-1)z(n)}{\lambda + z^T(n)P(n-1)z(n)}$
3. $P(n) = \frac{1}{\lambda}[P(n-1) - k(n)z^T(n)P(n-1)]$
4. $c(n) = c(n-1) + k(n)e(n)$

end

Алгоритм 3. RPEM-DPD

Выход: $c(n)$
Вход: $n = 0$, $\lambda(0)$, λ_0, $P(0) = \rho I$, $c(0)$
for $n = 1$ to L
1. $e(n) = u(n) - z_p(n)$
2. $\lambda(n) = \lambda_0\lambda(n-1) + (1 - \lambda_0)$
3. $k(n) = \frac{P(n-1)z(n)}{\lambda(n) + z^T(n)P(n-1)z(n)}$
4. $P(n) = \frac{1}{\lambda(n)}[P(n-1) - k(n)z^T(n)P(n-1)]$
5. $c(n) = c(n-1) + k(n)e(n)$

end

чие от RLS, в RPEM применяется переменный коэффициент забывания $\lambda(n)$, который экспоненциально стремится к 1 при $n \to \infty$, обеспечивая тем самым более стабильную оценку в установившемся режиме и улучшенное поведение в переходных процессах. Несмотря на преимущества, RPEM требует более сложной настройки, а его эффективность чувствительна к выбору начальных параметров и функции изменения $\lambda(n)$. По структуре близок к RLS и также имеет сложность $\mathcal{O}(P^2)$, но с переменным коэффициентом забывания $\lambda(n)$. Дополнительные вычисления для обновления $\lambda(n)$ имеют порядок $\mathcal{O}(1)$.

2.4. Алгоритм RLS-Калман

Алгоритм RLS-Kalman представляет собой модифицированную версию RLS, в которой вектор коэффициентов $\mathbf{c}(n)$ рассматривается как случайный процесс в модели состояний. Это позволяет эффективно отслеживать временные изменения параметров усилителя мощности за счёт ковариационной матрицы шума R_1, которая регулирует скорость адаптации. Однако метод требует точной настройки параметра R_1 и имеет повышенную вычислительную сложность по сравнению с классическим RLS. Теоретическая сложность также $\mathcal{O}(P^2)$, однако метод

Алгоритм 4. RLS-Kalman-DPD

Выход: $c(n)$
Вход: $n = 0$, $P(0) = \rho I$, $c(0)$, R_1
for $n = 1$ to L
1. $e(n) = u(n) - z_p(n)$
2. $k(n) = \frac{P(n-1)z(n)}{1+z^T(n)P(n-1)z(n)}$
3. $P(n) = P(n-1) - k(n)z^T(n)P(n-1) + R_1$
4. $c(n+1) = c(n) + k(n)e(n)$

end

требует хранения и обработки дополнительной ковариационной матрицы шума состояния R_1 размером $P \times P$, а также выполнения операций сложения/вычитания матриц $P \times P$ на каждом шаге. Это приводит к удвоению числа операций сложения по сравнению с классическим RLS и увеличивает объём памяти.

3. Результат моделирования

В работе используется модель усилителя мощности, основанная на структуре Винера. Входной сигнал нелинейного блока задаётся как: $v(n) = h_0u(n) + h_1u(n-1) + h_2u(n-2), \quad h = [0,7692,\ 0,06538,\ 0,00769]$, Выходной сигнал определяется моделью Saleh [6]:

$$y(n) = \frac{\alpha_a|v(n)|}{1+\beta_a|v(n)|^2} \cdot \exp\left(j \cdot \arg\left[v(n) + \frac{\alpha_\varphi|v(n)|^2}{1+\beta_\varphi|v(n)|^2}\right]\right). \qquad (4)$$

Таблица 1

Сравнение NMSE [dB], ACPR [dB]

Метод	NMSE [dB]	ACPR [dB]
LMS	-37,1777	-44,7208
RLS	-42,4215	-46,0795
RPEM	-43,4844	-48,3331
RLS-KALMAN	-43,1012	-47,9234

Как видно из табл. 1 и рис. 2, использование цифрового предыскажения (DPD) значительно снижает мощность утечек в соседние каналы. Наилучшие характеристики достигаются при применении алгоритмов RPEM и RLS-Калмана. Однако высокая эффективность этих методов сопровождается существенными вычислительными затратами.

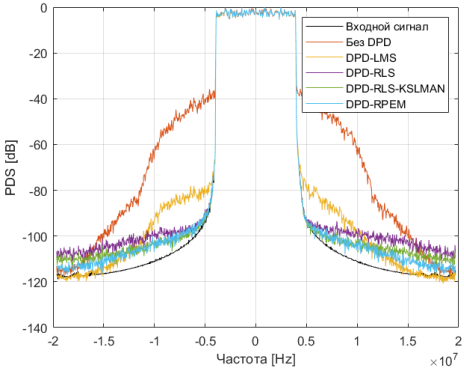


Рис. 2. Спектральная мощность (PSD) на выходе УМ без и с применением DPD.

Заключение

В данной работе проведён сравнительный анализ четырёх адаптивных алгоритмов DPD: LMS, RLS, RPEM и RLS-Калмана. Показано, что алгоритм LMS неэффективен для сильно нелинейных систем, а RLS чувствителен к шуму. RPEM и RLS-Калмана демонстрируют высокую точность, но требуют значительных ресурсов.С целью повышения эффективности простых алгоритмов при сохранении приемлемой сложности вычислений целесообразно использовать комбинированный подход: предварительная offline-настройка коэффициентов (например, методом Левенберга–Марквардта [7]) с последующей online-адаптацией (LMS,RLS). Это улучшает сходимость и снижает требования к вычислительным ресурсам.

СПИСОК ЛИТЕРАТУРЫ

1. *ETSI* Digital Video Broadcasting (DVB); Frame structure, channel coding and modulation for a second generation digital terrestrial television broadcasting system (DVB-T2) // ETSI EN 302 755 Vol. 1.3.1, Apr. 2012.
2. *Дворкович В. П., Дворкович А. В.* Цифровые видеоинформационные системы (теория и практика). М.: Техносфера, 2012. 1008 с.
3. *Kwon J., Eun C.* Digital feedforward compensation scheme for the non-linear power amplifier with memory // IJISTA. 2010. Vol. 9. С. 326–334.
4. *Vu V. K., Tran V. N.* Adaptive digital predistortion of RF power amplifiers based on memory polynomial model and indirect learning architecture // Proc. 2020 Int. Conf. Engineering and Telecommunication (En&T), IEEE, 2020. С. 1–5.

5. *Lv J., Xu G., Zhang C., Zhang H., Liu T.* Full-Wave Digital Predistortion Linearization Based on Memory Polynomials for HF Power Amplifiers // IEEE Access. 2024. Vol. 12. С. 29362–29372.
6. *Le D. H. et al* Linearization of RF power amplifiers in wideband communication systems by adaptive indirect learning using RPEM algorithm // Mobile Networks and Applications. 2020. Vol. 25. С. 1988–1997.
7. *Хоанг Ван Туан, Дворкович А. В.* Исследование эффективности цифрового предыскажения на основе полиномиальной модели памяти и алгоритма Левенберга–Марквардта для усилителей мощности // Цифровая обработка сигналов. 2025. № 1. С. 10–16.

Хоанг Ван Туан — Аспирант кафедры мультимедийных технологий и телекоммуникаций МФТИ. E-mail: *khoang.vt@phystech.edu*

МЕТОДЫ ДЕТЕКТИРОВАНИЯ ВИРТУАЛЬНЫХ МАШИН НА ОСНОВЕ АНАЛИЗА CPUID, DMI И СИСТЕМНЫХ АРТЕФАКТОВ

Г. В. Чалых[1], А. С. Твардовский[1]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия*

В целях защиты корпоративного программного обеспечения от несанкционированного копирования в виртуализированных средах предложены методы детектирования, опирающиеся на особенности работы процессора, данные Desktop Management Interface (DMI) и системные артефакты гипервизора. Предлагаемая реализация включает в себя получение инструкции CPUID для проверки 31-го бита ECX, устанавливающегося внутри виртуальной машины (ВМ). Анализ полей DMI позволяет выявить строки, связанные с виртуализацией. Измерение RDTSC обнаруживает характерные для виртуализации задержки и кэш-промахи. Также детектирование дополняют такие методы, как проверка первых трёх октетов MAC-адреса, наблюдение за динамикой температуры и анализ размера виртуального диска. Методы на основе CPUID, DMI и системных артефактов выявляют исполнение кода в виртуальной среде, что позволяет блокировать запуск программы для предотвращения нелегального копирования лицензии. **Ключевые слова:** *Виртуальные машины, защита от тиражирования, гипервизор, детектирование виртуализации, CPUID, Desktop Management Interface*

Введение

Активное развитие технологий виртуализации [1] позволило им перерасти из идеи организации совместного доступа к файловым системам в уровень целостной эмуляции операционных систем (ОС). Текущие решения в области виртуализации включают в себя поддержку многозадачности, производят распределение ресурсов в динамичном режиме и достигают высокой степени масштабируемости. Виртуальные среды эмулируют некоторые системные датчики, поддерживают низкоуровневые аппаратные функции различных процессорных архитектур и различные регистры процессора. Однако клонирование виртуальных машин создает реальную угрозу незаконного распространения лицензий, что требует надежных средств защиты.

Гипервизор [2] представляет из себя программу или отдельную операционную систему и позволяет одновременно запускать несколько виртуальных ОС на одном и том же хост-компьютере. Гипервизор обеспечивает изоляцию ОС друг от друга, защищает хост-машину от модификаций, разделяет ресурсы между различными запущенными ОС и управляет всеми вычислительными ресурсами, предоставленными гостевой машине. Гипервизоры первого типа не требуют наличия хост-машины с определенной операционной системой и, не являясь приложением, представляют отдельную операционную систему. Гипервизоры второго типа – это особые приложения, которые работают на базе хостовой операционной системы и способны внутри себя в отдельном окне запускать виртуальную систему.

В данной работе рассматриваются методы обнаружения исполнения кода в виртуальных средах, позволяющие при выявлении виртуализации блокировать лицензию. Для детектирования используется инструкция CPUID для проверки 31-го бита регистра ECX [3], активируемого внутри виртуальной машины. Анализируются DMI-поля BIOS и системная информация для обнаружения характерных цифровых меток гипервизора. Проводится измерение времени выполнения инструкции RDTSC [4] с последующим анализом задержек и кэш-промахов. Проверяются первые три октета MAC-адреса на соответствие шаблонам виртуальных сетевых адаптеров. Дополнительными признаками является наличие и динамика температурных датчиков под нагрузкой, общий размер виртуального диска, а также наличие вспомогательных файлов и некоторые поля в системном реестре.

1. Процессорные методы дектирования

Определение факта выполнения программного кода в виртуальной среде возможно посредством анализа особенностей работы процессора и отклонений в его поведении. Классическим примером выступает инструкция CPUID, позволяющая извлекать информацию о характеристиках процессора, включая признаки наличия гипервизора. Таким образом, работа инструкции CPUID в контексте виртуализированной среды основывается на механизме передачи управления через гипервизор. При исполнении CPUID запрос сначала направляется от пользовательского приложения в операционную систему, а затем передаётся на уровень гипервизора. В случае физического выполнения процессор возвращает значения, характерные для оборудования. Однако при запуске внутри ВМ гипервизор может перехватывать и изменять результат, подставляя собственную строку идентификации, например, с указанием названия модели гипервизора. Такой подход позволяет скрывать

или наоборот явно сообщать о наличии виртуализации. Визуально этот процесс представлен как последовательный путь запроса от приложения через ядро, далее – в гипервизор, который эмулирует поведение процессора (рис. 1).

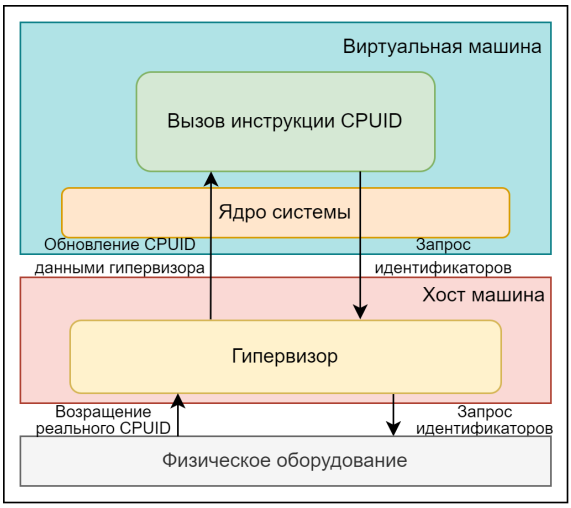


Рис. 1. Обработка инструкции CPUID в виртуализированной среде

Вторым идентификатором, связанным с выделением процессорных ресурсов, является замер времени работы RDTSC – функции подсчета тактов процессора. Вызывая эту функцию во время многочисленного подсчета хэш-функции или любой другой работы, можно заметить расхождение во времени: иногда функция срабатывает дольше, или наоборот – выполняется неестественно много процессорных команд. Подсчитав это время, можно найти большие «прыжки» во времени, что нетипично для физического оборудования.

Для процессорных методов реализован прототип низкоуровневых измерений на языке C++ со средствами сбора и принятия решения. Реализация включает сбор 31-го бита ECX инструкции CPUID и серийные замеры RDTSC при выполнении контролируемой нагрузки – циклический расчет хэш-функции с подсчётом аномальных скачков времени. Тестирование проводилось на гипервизорах KVM, VMware и VirtualBox. Признаки по отдельности оказались ненадёжными: CPUID скрывается KVM-гипервизором, два других гипервизора выводят ин-

формацию о виртуализации. RDTSC надежен при фоновой нагрузке, но иногда даёт ложные аномалии из-за системных прерываний.

2. Детектирование по метаданным DMI

Для определения работы в виртуальной машине на основе BIOS/DMI информации можно воспользоваться командами, извлекающими системные метаданные. С помощью утилиты dmidecode в Linux или SystemInfo на Windows считываются поля System Vendor, Product Name, BIOS Version, сохранённые в SMBIOS [5] — области, формируемой прошивкой. В физических машинах эти поля содержат сведения о производителе аппаратного обеспечения, в то время как в виртуальных средах они заполняются идентификаторами гипервизора. В полях могут встречаться строки, связанные с использованием виртуализации и созданные гипервизором. Помимо этого в SMBIOS присутствует специальный флаг – SMBIOS VM-bit, который прямо указывает на виртуализацию и может быть использован как однозначный признак работы внутри ВМ. Описанные признаки метаданных DMI представлены в табл. 1.

Таблица 1

Признаки виртуальной среды по данным DMI

Поле	**Физическая машина**	**Виртуальная машина**
sys_vendor	Dell Inc.	VMware, Inc.
product_name	Latitude 7280	VMware Virtual Platform
bios_version	A05	Phoenix 1.20 VIRTUALBOX 1.2
SMBIOS VM-bit	0	1
ACPI-таблицы	SSDT, DSDT	VMxxx, VBOX, QEMU

Для извлечения метаданных используется прямое чтение `/sys/class/dmi/id/` и разбор вывода `dmidecode`. Проверяются поля `sys_vendor,product_name,bios_version,` наличие специфичных ACPI-таблиц и SMBIOS VM-bit. Тесты проводились в аналогичной конфигурации на KVM, VMware и VirtualBox. Для каждой конфигурации выполнялась программа на C++, которая функцией `popen()` вызывала системный `dmidecode` и построчно анализировала консольный вывод; затем программа через `read_file()` читала содержимое директории `/sys/class/dmi/id/`. В стандартных настройках ВМ данный метод всегда срабатывает успешно — все тестовые машины имеют явные DMI-метки. Однако главный недостаток в том, что, зная расположение

и формат этих файлов, возможна целенаправленная подмена, поэтому метод следует сочетать с другими проверками.

3. Системные артефакты в виртуальных средах

Для детектирования можно использовать следы, оставленные виртуализацией в аппаратных характеристиках операционной системы. Гипервизоры оставляют свои идентификаторы в названии оборудования, системных файлах и регистрах. В гостевой машине автоматически назначаются сетевым адаптерам MAC-адреса с определенными начальными октетами, характерными для каждого гипервизора. В виртуальных машинах VMware это MAC-адреса с префиксами 00:50:56 или 00:0C:29, для Hyper V типичным будет 00:15:5D.

Кроме MAC-адресов можно ориентироваться на малый размер жесткого диска, нетипичный для физической машины. Ещё одним аппаратным следом гипервизора являются датчики температуры: виртуальные машины к ним не имеют доступа, и датчики эмулируются или отсутствуют.

Гостевые машины оставляют свои метки в служебных файлах: на Linux это наличие директорий `/sys/hypervisor/*`, файлов `/run/.containerenv` или `/.dockerenv`. А вот в Windows проверяются ключи реестра `HKLM\HARDWARE\DESCRIPTION\System\SystemBiosVersion`, разделы `HKLM\SOFTWARE\Oracle\VirtualBox Guest Additions` и `HKLM\SOFTWARE\VMware, Inc.\VMware Tools`, а также службы в `HKLM\SYSTEM\CurrentControlSet\Services`. Наличие служебных файлов и ключей реестров, описанных выше, указывает на работу внутри виртуальной машины.

Реализованная программа на C++ запускалась в KVM, VMware и VirtualBox и собирала признаки сетевых адаптеров, диска и датчиков. Для MAC-адресов читала файл `/sys/class/net/*/address`; размер диска определялся через `statvfs()`, а значения температур — чтением `/sys/class/thermal/`. На Windows аналогично использовались функции `RegOpenKeyEx()`, `RegQueryValueEx()` для проверки ключей реестра и служебных записей. Полученные программой значения показали, что префиксы MAC и отсутствие температурных датчиков являются надежными индикаторами в стандартной конфигурации ВМ.

Заключение

В работе представлены и реализованы методы детектирования исполнения программного кода в виртуальной среде на основе CPUID, RDTSC, анализа полей DMI и системных артефактов. Подготовлен прототип программы для обнаружения виртуализации. Проведён экс-

периментальный анализ и апробация программы на KVM, VMware и VirtualBox. По результатам запуска обнаружено, что не каждый из исследуемых методов надежен, и для успешного обнаружения виртуализации требуется совместное использование признаков. Текущая реализация позволяет блокировать запуск защищённого ПО, что уже может использоваться для защиты лицензии. Но такой подход сильно ограничивает пользователя, поэтому дальнейшие исследования этой темы будут направлены на поиск идентификаторов виртуальных машин, что позволит отличить каждую созданную машину от другой.

СПИСОК ЛИТЕРАТУРЫ

1. *Нестеров Д. А.* Исследование современных средств виртуализации // Сборник статей «Российская наука в современном мире» XXX Международная научно-практическая конференция, 2020. С. 125-128.
2. *Алексеев А. Л., Красноперова Е. А., Вахрушева Е. А.* Гипервизоры и виртуальные машины // Сборник статей «Информационные технологии в науке, промышленности и образовании» Всероссийская научно-техническая конференция. 2019 С. 121-128.
3. *Peter Ferrie* Attacks on Virtual Machine Emulators // Symantec Advanced Threat Research 2006.
4. *Вишневская Т. И., Макаренко О. К.* Метод маскировки виртуальной машины от её обнаружения вредоносным ПО // Образовательные ресурсы и технологии, 2020, №2.
5. *Rin N.* Virtual Machines Detection Enhanced: учебно-методическое пособие, 2013.
6. *Thomas Raffetseder, Christopher Kruegel, Engin Kirda* Detecting System Emulators // Information Security Conference (ISC), Lecture Notes in Computer Science, Springer, 2007.
7. *Liston T, Skoudis E* On the Cutting Edge: Thwarting Virtual Machine Detection // Intelguardians, SANS Internet Storm Center, 2006.

Чалых Григорий Валерьевич — студент Национального исследовательского Томского государственного университета. E-mail: *chalyhgrisha01@gmail.com*

Твардовский Александр Сергеевич — канд.ф.-м.н., доцент Национального исследовательского Томского государственного университета.

CHAINED PREDICTION OF UNIVERSITY COURSE SCORES USING MACHINE LEARNING

S. Erkinov[1], S. Bazarov[1]

[1] *Inha university in Tashkent, Tashkent, Uzbekistan*

This study investigates the feasibility of predicting the final score of a student in the third-year Probability and Statistics (PAS) course using scores and derived metrics from second-year mathematics courses: Engineering Mathematics (EM), Discrete Mathematics (DM) and Linear Algebra (LA), which are taught on the 2nd year. The prediction model consists of 2 models chained together in a linear workflow. One model predicts the all other components of score except a final exam, and second predicts the final exam score. Separation is done to improve the accuracy and counter the corruption of data due to difference in the exam format. The study demonstrates that reliable grade prediction is feasible even with a limited amount of data. **Keywords:** *grade prediction, educational data mining, machine learning, performance analysis*

Introduction

Early identification of students who may struggle academically allows universities to deploy timely interventions, allocate learning resources more effectively, and ultimately improve retention and graduation rates. Furthermore, the ability to predict future scores allows students to understand whether they fall behind much earlier and therefore take some actions. Most prior work on grade prediction assumes access to large, multi-modal datasets that include behavioural logs, (for instance the attendance statistics), demographic variables, or detailed assessment histories [1–3, 7]. Also the datasets are usually collected using online platforms, that allows to capture many details, like the average amount of time between posting an assignment and submitting solution to it, time spent on watching educational materials or how many clicks students make before answering. These features can give vital insights into students' behaviour. However, in practice, many institutions, with smaller enrolments or in developing countries do not possess technologies to gather such data, and usually have small number of students, which causes data scarcity. Therefore, as a proof of concept, this study reveals that grade prediction is possible to some extent even when have limited access to data. We decided to predict the PAS grades based on the DM, EM and LA grades for few reasons. Firstly, they are all taught by

the same professor. Secondly, as they are all maths related courses, there is a stronger correlation between them. We were given a table that contains all scores for midterm examination, final examination and the total score for each subject of 262 students. Initially, we started with naive approaches using simple regression algorithms, which yielded poor results. Then we moved on some advanced methods: we made a deep dive into the data, analysed correlations and engineered features. In the end, we came up with the approach discussed in this paper. First, the total score is split into 3 parts that contribute to it: midterm exam score (mid) – 30% of total score, final exam score (final) – 40% of total score and all other activities that contribute to total score (etc) – 30%. On top of that we did not use any data about the attendance of students, as it is not reliable. We identified high correlation between EM, DM, LA mid and etc and PAS mid and etc at around 0.7. However the EM, DM, LA final scores and PAS final had correlation only at around 0.3, which is explained by the fact that it was held in different conditions compared to other exams. Therefore we use a second model to predict accurate final scores. Overall, our approach lets predict the score with approximately 70% accuracy. In view of the data scarcity, absence of additional features, the obtained result can be regarded as significant.

1. Related Work

Most grade-prediction studies assume that assessments like exams and home assignments share a uniform, proctored delivery format and therefore exhibit stable cross-subject correlations. In addition, they assume access to reliable behavioural information like: attendance, weekly study time, number of past class failures and etc. On top of that, some researches had access to personal information of students: living conditions, relationship status, health and some subjective data – interest in subject, in higher education and so on [1–5, 7]. Under this premise, researchers have explored linear regression, tree-based ensembles, and kernel methods to forecast end-of-semester marks from continuous coursework signals or cumulative GPA snapshots. In contrast, comparatively few investigations use the smaller datasets, that contain only grades, for a few amount of subjects and for small number of students [6]. As gathering such information on constant basis, to analyse the performance of students each year is a burden for majority of educational institutions, there should be a model or an approach that would work with low amount of data. Also, none of the works we reviewed addressed the reliability gap that emerges when one high-stakes exam is conducted *online* (in different format compared to other exams)

while the remaining subjects retain traditional, in-person examination. And our research aims to cover exactly these difficulties in grade prediction.

2. Methodology

Feature Engineering.

1) **Residual (*etc*) scores**: {prefix}_etc_score = {prefix}_score − ({prefix}_mid + {prefix}_final). Where prefix is one of EM,LA,DM or PAS.
2) **Scaled mid–final differences**: Exam columns were min–max scaled to $[0, 1]$; per-student changes $\Delta = \textit{final} - \textit{mid}$ were stored as {prefix}_diff.
3) **Percentile ranks**: Column-wise percentile ranks (%-score in cohort) give {prefix}_mid_pct_rank and {prefix}_final_pct_rank.
4) **Mathematics halo**: student_overall_avg_diff $= \frac{1}{3}$(EM_diff + +DM_diff + LA_diff) captures overall changes in each students scores from midterm to final examination.

Table 1

Mid-term Pearson correlations ($n = 262$).

	EM_mid	DM_mid	LA_mid	PAS_mid
EM_mid	1.000	0.673	0.762	0.674
DM_mid	0.673	1.000	0.721	0.706
LA_mid	0.762	0.721	1.000	0.705
PAS_mid	0.674	0.706	0.705	1.000

Table 2

Final-exam Pearson correlations.

	EM_final	DM_final	LA_final	PAS_final
EM_final	1.000	0.679	0.747	0.317
DM_final	0.679	1.000	0.716	0.330
LA_final	0.747	0.716	1.000	0.278
PAS_final	0.317	0.330	0.278	1.000

The Pearson correlations reveal a clear pattern across the courses. As shown in Table 1, all mid-term exam scores are strongly correlated ($r = 0.67$–0.76), indicating a shared quantitative aptitude and validating their use for early forecasting. In contrast, the final exam correlations in Table 2 show a significant divergence. While the EM, DM, and LA finals remain aligned, correlations with the online PAS_final drop sharply ($r < 0.34$), suggesting score inflation from the unproctored online format

Table 3

Correlation of residual (*etc*) components.

	EM_etc	DM_etc	LA_etc	PAS_etc
EM_etc	1.000	0.669	0.684	0.625
DM_etc	0.669	1.000	0.692	0.653
LA_etc	0.684	0.692	1.000	0.664
PAS_etc	0.625	0.653	0.664	1.000

rather than a shift in academic content. This anomaly is confirmed to be localized to the PAS final exam; after subtracting the exam components, the residual scores in Table 3 regain a uniform, moderate correlation ($r \approx 0.65$) across all subjects. Taken together, these views motivate our two-part strategy: (i) trusting the midterm and residual scores for capturing student knowledge, and (ii) correcting the anomalous `PAS_final` score before predictive modelling.

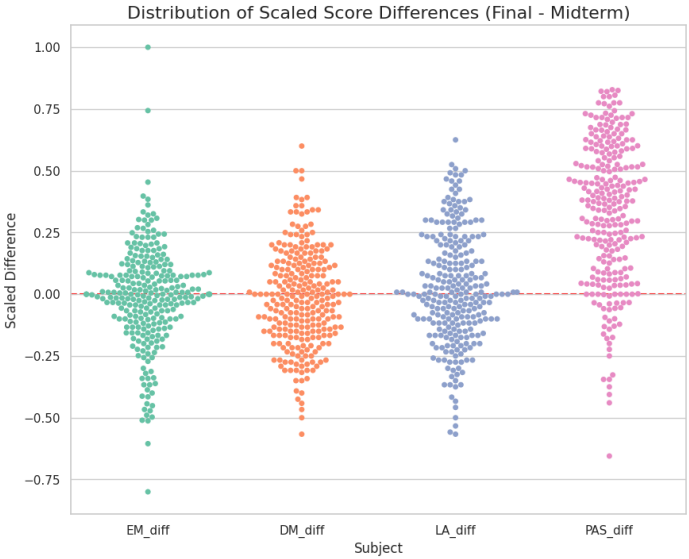


Figure 1. Dot/violin distributions of Δ for each subject.

What Fig. 1 shows. The violin kernel density for PAS is visibly fatter in the positive domain, while the other subjects display symmetric, slim profiles centred on zero. Individual dot scatter confirms more extreme pos-

itive Δ values in PAS, underscoring heterogeneity introduced by the online unmonitored test environment.

Statistical confirmation. Welch two-sample t-tests corroborate the visuals: `PAS_diff` significantly exceeds each of {EM, DM, LA}_diff ($p < 0.05$), and exceeds their pooled mean ($t = 11.2$, $p < 10^{-20}$). Hence the PAS final must be corrected before any predictive modelling.

Predictive Modelling. The modelling pipeline contains two stages, each solved by a best suiting ML model. The best model for each step was chosen by systematically evaluating the performance of different regression algorithms.

Stage A – Predicting `PAS_mid_etc`.

- **Target** `PAS_mid_etc`, i.e. the portion of the PAS course mark except for final component (as it is not reliable).
- **Predictors** {EM,DM,LA}_mid_etc + {EM,DM,LA}_mid_pct + {EM,DM,LA}_final_pct (9 features total).
- **Model** SVR [8] with following parameters: $C = 100.0, \varepsilon = 0.1, \gamma = 0.01$, kernel=rbf
- **Validation** Three-fold cross-validation. Mean $R^2 = 0.6633 \pm 0.02$.

Stage B – Predicting the offline-equivalent PAS final.

- **Target** $\widehat{\text{PAS_final_offline}}$ The true score that students would get if the final exam was held offline.
- **Predictors** Just two: {PAS_mid, `student_overall_avg_diff`}.
- **Model** Simple linear regression.
- **Validation** Three-fold cross-validation ($\overline{R^2} = 0.7276$)

Why these models With only 262 students and features deliberately engineered for linearity, other higher-capacity models (Random Forest [10], XGBoost [9], etc.) saw a strong undertraining. The two-stage approach solved this issue, and resolved problem of inconsitent finals.

Runtime & reproducibility. Both stages were complete end-to-end on the cloud environment provided by Google Colab. All pseudo-random generators were fixed at `RANDOM_STATE = 42`.

3. Experimental Results

Magnitude and Significance of Mid–Final Shifts. Figure 1 visualises $\Delta = \textit{final} - \textit{mid}$ for each course. Density confirms heavier skew for PAS compared with EM, DM and LA. Welch t-tests verify that `PAS_diff` differs significantly ($p < 0.05$) from every other subject, and from their pooled mean ($t = 11.2$, $p < 10^{-20}$). These findings justify treating the PAS final as an outlier when constructing the offline proxy target.

Predictive Performance.

1) **Linear Regression dominates.** A mean $R^2 = 0.728$ (SD 0.017) outperforms every regularised, tree-based or kernel alternative, confirming that the engineered features linearly explain the adjusted PAS score.
2) **Non-linear models underperform.** SVR [8] (0.663), XGBoost [9] (0.640) and Random Forest [10] (0.631) do not capture additional variance, suggesting limited curvature in the feature–target relationship.
3) **Stability.** All standard deviations remain below 0.024, indicating consistent generalisation across folds.

Error Profile. Prediction residuals (not shown for brevity) were centred and approximately homoscedastic, with no visible pattern against fitted values—further evidence that a simple linear model suffices once the PAS anomaly is corrected.

Summary of Findings.

- Midterm results and scores for other activities (_etc) form a coherent block; the online PAS final disrupts this structure.
- Scaled mid–final differences confirm a statistically significant uplift exclusively in PAS.
- After adjusting the PAS target, a transparent Linear Regression reaches $\approx 73\%$ variance explained–adequate for early warning systems without sacrificing interpretability.

These results set the stage for the Discussion, where we interpret practical implications for instructors and outline limitations.

Conclusion

This work set out to demonstrate a possibility to create systems designed to predict scores in future courses based on previous scores, despite the lack of behavioural data and low enrolment. Crucially, this work highlights the importance of considering the mode of examination (e.g., online vs. in-person) in predictive modelling. Exploratory analysis revealed an anomalous inflation in the online conducted `PAS_final` exam, verified by t-tests on scaled mid–final differences.

Code & Data Availability. All scripts are available at the GitHub page: https://github.com/BSA44/grade_prediction_research. The authors express special gratitude to Dr. Anna Tomskova for providing us scores of students. The student grade data is private and cannot be shared due to privacy restrictions.

REFERENCES

1. *Al-Tameemi G., Xue J., Ali I. H., Ajit S.* A new hybrid machine learning approach for predicting student performance using multi-class educational datasets // Procedia Computer Science. 2024. Vol. 238. P. 888–895. Proc. 15th Int. Conf. on Ambient Systems, Networks and Technologies (ANT) / 7th Int. Conf. on Emerging Data and Industry 4.0 (EDI40), 23–25 April 2024, Hasselt University, Belgium. doi: 10.1016/j.procs.2024.06.108.
2. *Chen Y., Sun J., Wang J., Zhao L., Song X., Zhai L.* Machine learning-driven student performance prediction for enhancing tiered instruction // arXiv preprint. 2025. arXiv:2502.03143. `https://arxiv.org/abs/2502.03143`.
3. *Alharbi A. A., Allohibi J.* A new hybrid classification algorithm for predicting student performance // AIMS Mathematics. 2024. Vol. 9, № 7. P. 18308–18323. doi: 10.3934/math.2024893.
4. Student Grade Prediction [Data set] // Kaggle. 2025. Version 2025-07-07. `https://www.kaggle.com/datasets/dipam7/student-grade-prediction`.
5. *Ali I.* Machine learning models and feature relevance for student grade prediction // Journal of High School Science. 2024. Vol. 8, № 1. P. 180–197. doi: 10.64336/001c.115815.
6. *Hidalgo Suarez C. G., Llanos J., Bucheli Vol. A.* Predicting the final grade using a machine learning regression model: insights from fifty percent of total course grades in CS1 courses // PeerJ Computer Science. 2023. Vol. 9. Article e1689. doi: 10.7717/peerj-cs.1689.
7. *Sokkhey P., Okazaki T.* Hybrid machine learning algorithms for predicting academic performance // International Journal of Advanced Computer Science and Applications. 2020. Vol. 11, № 1. P. 32–40. doi: 10.14569/IJACSA.2020.0110104.
8. *Drucker H., Burges C. J. C., Kaufman L., Smola A., Vapnik V.* Support vector regression machines // Advances in Neural Information Processing Systems. 1997. Vol. 9. P. 155–161.
9. *Chen T., Guestrin C.* XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). 2016. P. 785–794. doi: 10.1145/2939672.2939785.
10. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45, № 1. P. 5–32. doi: 10.1023/A:1010933404324.

Erkinov Samobek — Student, School of Computer and Information Engineering. E-mail: *s.erkinov@student.inha.uz*

Bazarov Sarvar — Student, School of Computer and Information Engineering. E-mail: *s.bazarov@student.inha.uz*

ANALYSIS OF THE MPTCP PROTOCOL WITH ITS IMPLEMENTATION IN PNETLAB

A. Molchanov[1], E. Pakulova[1]

[1] *Southern Federal University, Taganrog, Russian Federation*

This paper presents an experimental analysis of the Multipath TCP (MPTCP) protocol and its implementation within the PNETLab network emulation environment. The study evaluates MPTCP performance under both homogeneous and heterogeneous network conditions, comparing it against traditional single-path TCP. Metrics such as throughput, goodput, round-trip time (RTT), jitter, retransmission rate, and out-of-order packet occurrence are investigated. Experimental results demonstrate that MPTCP can improve data transmission efficiency by leveraging multiple network paths, providing higher goodput and reduced retransmission rates, although with slightly increased latency and packet reordering. The findings highlight the protocol's potential benefits in optimizing bandwidth utilization in modern multi-interface networks.

Keywords: *Multipath TCP, MPTCP, TCP, PNETLab, Network Emulation, Throughput, Goodput, RTT, Jitter, Packet Reordering*

Introduction

Today, more and more end devices are equipped with multiple interfaces. Parallel multipath data transmission attracts the attention of telecommunication technology specialists. For these purposes, multipath data transfer protocols are used. It could be SCTP, MPRTP, MPTCP, MPQUIC etc.

Multipath TCP (MPTCP) is an extension of standard TCP that allows the simultaneous use of multiple network paths between two endpoints of a single connection. This could increase bandwidth, reliability, and resilience to infrastructure failures [1]. This protocol continues to be actively developed and is supported at the Linux kernel level (from version 5.6 and above), including major distributions such as Red Hat and Ubuntu. In addition, this protocol is integrated into Apple solutions (for Siri, Apple Maps, Wi-Fi Assist, and other services), enabling smooth transitions between Wi-Fi and cellular networks without disconnections. The protocol is also open to third-party developers of iOS apps.

However, MPTCP is an end-to-end protocol that does consider the network state at lower levels. For example, different subflows of a single con-

nection can be distributed along the same path in the network. In such a situation, network resources are used inefficiently, and competition between subflows on the same path can lead to network congestion. In addition, due to different delays between subflows, data packets can arrive at the receiver in the wrong order. Packet reordering significantly reduces transmission performance. Another problem is that the number of MPTCP subflows is fixed and cannot be automatically and correctly adapted depends on service requirements and available network resources.

Simulation modeling is the most widely used method for analyzing the characteristics of data networks because it helps to overcome the limitations of analytical models. There are many tools available for network simulation, such as GPSS World, NS3, OPNET Modeler, Mininet, Pnetlab and others. The result of the modeling is a description of the processes of servicing a multidimensional input data stream, which allows determining Quality of Service (QoS) parameters and selecting the processing order. Pnetlab is a modern platform for network modeling and emulation that enables the creation of complex virtual network environments with support for many network devices and protocols [2].

The contribution of this paper is to analyze the operation of the Multipath TCP protocol implemented in Linux (Ubuntu) and to study the features of its implementation in PNETLab environment [3]. Moreover, our objective is to evaluate the efficiency of using MPTCP in various network conditions.

1. Research methodology

This study investigates the performance of the MPTCP protocol under diverse network conditions through a systematic experimental approach. The testing methodology employs traffic generation and transmission combined with comprehensive metric analysis using custom Python scripts. The experimental framework evaluates MPTCP behavior in two distinct network environments. First, controlled wired network conditions establish baseline performance metrics. Second, emulated heterogeneous network conditions replicate real-world scenarios with multiple connection paths. For comparative analysis, parallel measurements are conducted using both traditional TCP and MPTCP protocols under identical network configurations.

1.1. Performance Metrics

The chosen metrics are essential for evaluating MPTCP performance compared to TCP.

Throughput serves as a fundamental metric for assessing data transfer speed. Higher throughput values indicate more efficient protocol perfor-

mance, reflecting optimized data transmission logic. While MPTCP can potentially increase throughput by utilizing multiple paths, this requires more complex mechanisms compared to traditional TCP [4].

Goodput measures the actual useful data rate consumed by applications. MPTCP may improve this metric by reducing the impact of retransmissions, while TCP might suffer from the lack of alternative paths when packet loss or delays occur on the primary path [5].

Round-trip time (RTT) represents the time it takes for a packet to travel from the sender to the receiver and back. It represents delays associated with packet retransmission. This metric is particularly important for TCP connections, which are prone to packet retransmission. RTT reflects protocol reliability and efficiency in handling packet loss, directly impacting overall performance [5].

Jitter measures delay variability. Jitter helps assess a protocol's ability to handle inconsistent network conditions [4].

Retransmission Rate indicates the frequency packets need to be retransmitted. Lower retransmission rates typically suggest better protocol efficiency and reliability. This metric helps evaluate the stability of each protocol [6].

TCP Out-of-Order quantifies packets arriving at the receiver in a different sequence than transmitted. For order-sensitive applications, packet misordering can degrade performance. MPTCP may face greater challenges maintaining packet order due to its multipath nature. This metric helps assess the trade-offs between multipath utilization and packet ordering maintenance [6].

1.2. Experimental Testbed Configuration

All experiments were performed within the PNETLab environment. It is a network emulation platform that enables the creation of flexible virtual topologies.

In our testbed environment, we used two virtual machines (a server and a client) with the following characteristics:

- OS – Ubuntu 24.04
- Linux kernel image: 6.8.0-59-generic
- Hardware resources: 4 CPU, 4 GB RAM

The machines have two interfaces assigned to different subnets: eth1 and eth2. In our experiments we generated an 8.7 GiB test file for transmission from the client to the server. Next, we established MPTCP connection using the mptcpize run utility [7], which forces the use of the MPTCP protocol. To monitor the MPTCP connection we used the ip mptcp monitor command, launched before the start of the file transfer using netcat.

2. RESULTS OF EXPERIMENTS

2.1. Goodput

The utility netcat enable the measurement of the connection's goodput. This metric is calculated by dividing the transferred file size by the total transmission time.

All 10 experiments without any restriction on the channels showed a significant difference in file transfer time - MPTCP consistently outperformed TCP, completing the transfer 32 to 53 seconds faster in each case as shown in the diagram in the fig. 1.

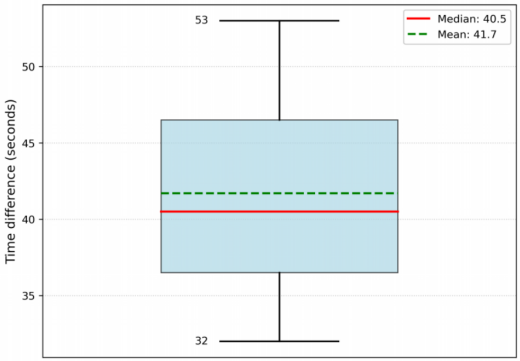


Figure 1. Test file transfer time difference between MPTCP and TCP

2.2. Round-Trip Time

RTT (Round-Trip Time) measurements were collected using the tshark utility [8] on the server side.This produced an output file containing RTT values, specifically the time differences between data transmission and receipt of ACK flags. For calculation average RTT and jitter we used the python scripts.

Table 1 illustrates the RTT values recorded during transmission of the test file over the primary subflow. The average RTT for eth1 stayed close to 2.5 ms. The mean jitter can be calculated by analyzing the differences between the RTT peaks using the script, resulting in an average jitter of 3.45 ms. Statistics for eth2 shows how this metric changes over time in the MPTCP subflow. The average observed RTT was 2.38 ms, with a mean jitter of 3.39 ms.

2.3. Additional Metrics

To gather information on retransmitted packets (Retransmission Rate) and packets arriving out of sequence (TCP Out-of-Order) we used the Wireshark filters tcp.analysis.retransmission and tcp.analysis.out_of_order respectively. Consequently, the following results were obtained during the experiment:

- Interface eth1: Retransmission Rate: 44 packets out of 41,698; TCP Out-of-Order: 7 packets out of 41,698;
- Interface eth2: Retransmission Rate: 55 packets out of 40,715; TCP Out-of-Order: 67 packets out of 40,715.

2.4. TCP metrics collection

During the experiments we also calculated the TCP metrics.The average RTT was 0.83 ms, with a mean jitter of 1.02 ms

Retransmission Rate = 1,000 packets out of 48,899.

TCP Out-of-Order = 0 packets out of 48,899.

Table 1

Metrics of TCP and MPTCP

Metrics	**Protocols and interfaces**			
	MPTCP (eth1)	***MPTCP (eth2)***	***MPTCP (eth1+eth2)***	***TCP***
Throughput (Gbit/s)	3.8	3.3	7.1	7.8
Goodput (Mbit/s)	190.4	180.2	370.6	299.3
RTT (ms)	2.49	2.37	2.43	0.89
Jitter (ms)	3.41	3.29	3.35	1.14
Retransmission Rate (%)	0.06	0.04	0.1	0.3
TCP Out-of-Order (%)	0.2	0.1	0.3	0

2.5. COMPARATIVE ANALYSIS OF MPTCP AND TCP METRICS IN HOMOGENEOUS DATA TRANSMISSION

Based on a series of metric measurements, Table 1 shows the average results from 10 test iterations. MPTCP consistently demonstrated slightly

lower throughput and higher latency compared to TCP, though the difference in these metrics was marginal. While MPTCP experienced increased RTT, its connections demonstrated more stable packet retransmission counts, despite having slightly higher jitter compared to TCP. MPTCP also had a noticeable proportion of out-of-order packet deliveries, a factor that single-path TCP completely avoided. Nevertheless, MPTCP consistently achieved significantly greater useful throughput. The combined average RTT and jitter for MPTCP (using eth1 and eth2) were calculated as the arithmetic mean of both interfaces' values, since multipath transmission naturally favors the path with lower latency.

3. Investigation of Network Behavior through Emulation of Real Traffic

3.1. Limiting Bandwidth Using Token Bucket Filter (TBF)

When testing network applications or emulating real-world conditions at home, it is often useful to introduce controlled limitations on network bandwidth, delay, jitter, and packet loss. The Linux `tc` (traffic control) tool allows precise manipulation of these parameters on network interfaces.

To make the network conditions more realistic, we can add a NetEm qdisc as a child of the TBF. NetEm can emulate delay, jitter, and packet loss.

The experiments are conducted similarly to the previously described tests without channel restrictions. The test file is used to generate traffic, and the goodput is measured in the same way as before.

By applying the traffic shaping and network impairments described in the previous sections, we can observe how bandwidth limitations, delay, jitter, and packet loss affect the performance of the network. These experiments allow us to compare the impact of a restricted network environment on data transmission and application behavior. The experiment setup is as follows:

- **TCP traffic** flows exclusively through `eth1` (wired connection). The impact of the wired interface on goodput is measured.
- **MPTCP traffic** utilizes both `eth1` (wired) and `eth2` (Wi-Fi) simultaneously. This allows the multipath protocol to aggregate bandwidth and potentially improve overall goodput compared to TCP over a single wired link.

Airport Wired and Wi-Fi Network Scenario

In this experiment, we emulate a heterogeneous network environment similar to an airport scenario:

- **eth1 – Wired network**
  This interface represents a stable wired connection typically available in an airport.
- **eth2 – Wi-Fi network**
  This interface represents the wireless connection available in the same environment. In this scenario, the restriction parameters are as follows: rate limited to 30 Mbit/s, burst size set to 32 kbit, and latency configured to 100 ms.. As for `NetEm`: an average delay of 40 ms with a jitter of 10 ms and a packet loss rate of 0.5%.

This scenario demonstrates how MPTCP can leverage multiple interfaces to improve performance in a heterogeneous network environment, where one interface may be less reliable or slower. The results of the experiments (fig. 2) show that MPTCP provides better goodput, – faster data transmission even when it has additional slow channel. However, increasing delay

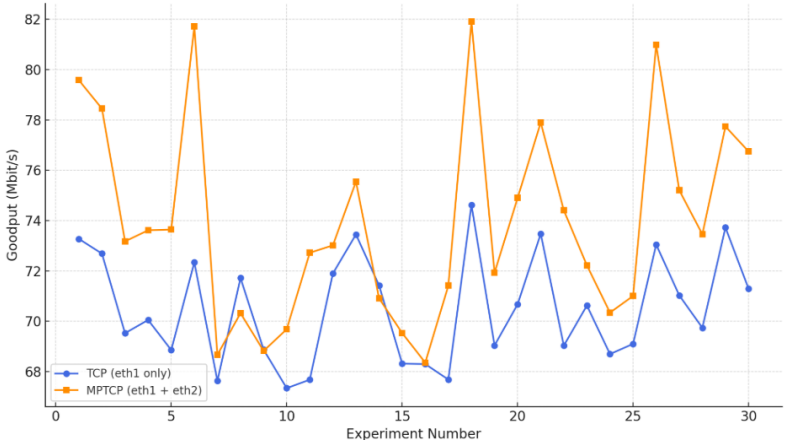


Figure 2. TCP vs MPTCP Goodput Across 30 Experiments (Airport Conditions)

on the secondary path causes the MPTCP goodput to decrease, resulting in overall performance degradation (fig. 3) .

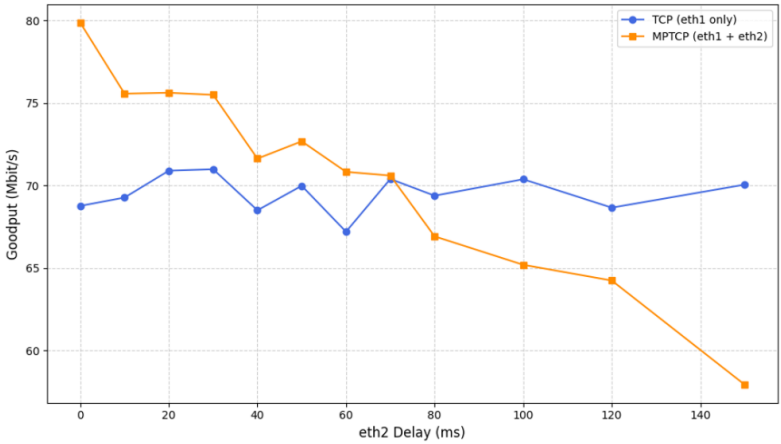


Figure 3. Impact of Wi-Fi (eth2) Delay on TCP vs MPTCP Goodput

4. Conclusion

The conducted study demonstrates that Multipath TCP (MPTCP) provides tangible advantages in network performance compared to traditional single-path TCP, particularly in scenarios where multiple network paths are available. Experimental results indicate that MPTCP consistently achieves higher goodput, enabling faster transmission of large files, although it exhibits slightly increased latency and jitter due to the need for packet synchronization and reordering across multiple subflows. Despite a higher proportion of out-of-order packets and modest resource overhead, MPTCP maintains lower retransmission rates, highlighting its reliability and efficiency in utilizing available network paths. The analysis within the PNETLab environment confirms that careful configuration of subflows and endpoint parameters allows effective exploitation of path diversity, resulting in improved overall data transfer performance. These findings suggest that MPTCP is a viable solution for enhancing bandwidth utilization offering significant benefits for modern applications.

REFERENCES

1. *Bikas S., Sayit M.* Providing QoS and Resiliency for Video Streaming over MPTCP //IEEE Access. 2025.
2. *Pakulova E., Ryndin A.* Simulation of priority multi-path data transmission in the AnyLogic environment. Proceedings 17th IEEE International Conference on Application of Information and Communication Technologies (AICT), 2023. P. 261–264.

3. Pnet.Lab is simple [Online]. Available: https://pnetlab.com/
4. Jinsung Lee, Jinsung Im, Youngbin Lee, David Joohyung. 2019. Modeling MPTCP Performance. IEEE Communications Letters. 10.1109/LCOMM.2019.2898664. [Online]. Available: `https://www.researchgate.net/publication/331027568_Modeling_MPTCP_Performance`
5. Yang Tao, Zhijun Yan. 2017. Goodput Improvement for MPTCP Based on Controlling the Difference of Delay [Online]. Available: `https://gvpress.com/journals/IJHIT/vol10_no4/1.pdf`
6. Atefeh Ahmadniai Khajekini, Hasan Amca, Ali Hakan Ulusoy, Enver Ever. (2024).
7. Matthieu B. Mptcpd 0.12-4 source package in Ubuntu [Online]. Available: https://code.launchpad.net/ubuntu/+source/mptcpd/0.12-4
8. Tshark - Dump and analyze network traffic [Online]. Available: https://www.wireshark.org/docs/man-pages/tshark.html

---

Molchanov Alexander — student of Southern Federal University. E-mail: *amolcha@sfedu.ru*

Pakulova Ekaterina — candidate of technical sciences, associate professor, Southern Federal University. E-mail: *epakulova@sfedu.ru*

СОДЕРЖАНИЕ

Прикладной вероятностный и статистический анализ 5
Войтишек А. В., Иванова Ю. А., Сейтмуратов У. П. Новые экономичные алгоритмы с уравниванием вероятностей для моделирования гауссовского распределения 6
Войтишек А. В., Сейтмуратов У. П., Шлымбетов Н. Х. Специальные показатели качества приближения при выборе ядерной функции в компьютерном функциональном алгоритме приближения вероятностной плотности по заданной выборке 12
Добронец Б. С., Кемельбаева А. Р. Регрессия в пространстве распределений, основанная на эрмитовых сглаживающих сплайнах 18
Ерилова А. Р., Семенова Д. В. Распределение суммы экспоненциальных величин со структурой зависимости, определяемой *FGM*-копулой 24
Кутоянц Ю. А. Об адаптивной Калмановской фильтрации систем с малым шумом 30
Мартынов Г. В. Критерии согласия для смеси распределений Вейбулла . 36
Мирза М. А. Анализ влияния начального шума на процесс генерации изображений с использованием диффузионных вероятностных моделей 42
Никифоров Н. И., Пергаменщиков С. М., Пчелинцев Е. А. Робастное эффективное непараметрическое оценивание в аддитивной регрессии 48
Рыбкина О. В., Карачанская Е. В. Стохастические модели защиты информационной системы с управлением 53
Соседова Н. И., Карачанская Е. В. Использование свойства фрактальности сетевого трафика для выявления аномалий 59
Степанов А. В. Об оценивании интервала охвата непрерывной случайной величины по выборочным данным 65
Чураков В. В., Фёдорова Е. А., Тюленина Я. А. Аппроксимация объема суточного потребления электроэнергии в жилых помещениях 71
Anzhin V. A method for experimental correlation distribution estimation for correlation-based digital image watermarks 78
Nurmukhamedova N. S., Berdimuradov M. B. Cox Model-Based Survival Analysis for Predicting Infrastructure Failures in Smart Cities under Right Censored Data Conditions 84

Dobronets B., Popova O. Fast numerical methods for stochastic modeling based on probabilistic extensions 90
Математическая теория телетрафика и теория массового обслуживания 97
Варакина А. А., Фёдорова Е. А. Численный анализ RQ-системы $M/M_k/1$ с деградацией и восстановлением прибора 98
Васильева Е. А., Танако И. Е., Станкевич Е. П. Стационарные характеристики системы массового обслуживания с групповым обслуживанием требований 104
Ву Х. З., Жарков М. Л., Казаков А. Л., Лемперт А. А. Применение управляемой сети массового обслуживания для моделирования перевалки грузов в порту 109
Даммер Д. Д., Гренц Е. А. Характеристики общей суммы образующих незатухающий поток взносов на накопительный счёт 115
Зорин А. В. Свойства вложенной цепи Маркова для процесса типа $M(t)/M(t)/1/\infty$ с периодическими интенсивностями 122
Измайлова Я. Е., Каныльбай З. Имитационное моделирование RQ-системы M|M|1 с искажением заявок и их восстановлением 128
Калинин Н. С., Пешкова И. В. Экстремальный индекс времени ожидания в системе обслуживания с суперпозицией входных потоков Парето 135
Костарева К. М., Измайлова Я. Е. Исследование RQ-системы с неисправными заявками и их восстановлением 140
Лапатин И. Л., Луценко Н. А. Алгоритм подбора параметров MAP-потока для моделирования on-off состояний многоядерных процессоров 146
Манаева В. Е., Маркова Е. В. Анализ возраста информации в системах с выраженной корреляцией во времени 152
Монгуш В. М., Королев Д. А.. Асимптотический анализ системы массового обслуживания M|M|∞ с MMPP-потоком отрицательных заявок 158
Назаренко Р. С., Тананко И. Е., Рогачко Е. С. Построение модели композиции веб-сервисов с ненадежными сервисами в виде сети массового обслуживания 165
Назаров А. А., Измайлова Я. Е. Асимптотически-диффузионный анализ RQ-системы с отрицательными заявками и восстановлением искаженных положительных заявок 171
Николаев Д. И., Гайдамака Е. В., Живцова А. А. Анализ производительности систем DECT-2020 на основе модели поллинга со шлюзовой дисциплиной 178

Пичугин С. Б. Модель системы связи со сканированием абонентской области обслуживающими приборами 184
Постнова О. С., Тананко И. Е., Рогачко Е. С. Оценка длительности сборки требований из фрагментов для сети массового обслуживания с делением и слиянием требований 190
Салимзянова Д. Д., Самойлов С. А., Лисовская Е. Ю. Оценка параметров рекуррентных потоков для анализа производительности многоядерных процессоров 196
Сергеева Н. В., Шабанова Д. Д., Тананко И. Е. Моделирование железнодорожных сортировочных систем сетью массового обслуживания 202
Старовойтова А. В., Моисеев А. Н., Назаров А. А. СМО с ненадежным прибором и неэкспоненциальным временем его восстановления 208
Тазаева А. А., Маркова Е. В. Схемы доступа к радиоресурсам беспроводной сети с двумя сетевыми слоями 214
Тюленина Я. А., Фёдорова Е. А. Среднее число заявок в СМО $M/M/\infty$ с переключением в нестационарном режиме 220
Хаустова Н. Г., Моисеева С. П. Характеристики показателя качества обслуживания систем с абсолютным и относительным приоритетами 226
Хомичков И. И. Система массового обслуживания с повторными обращениями и конечным бункером перед прибором 232
Цициашвили Г. Ш. Обобщение мультипликативной теоремы для сети с зависимыми от числа заявок переходными интенсивностями .. 238
Шкленник М. А., Лавриненко М. Д. Исследование гибридной СМО с мгновенной и отсроченной обратной связью 244
Balaji M. N., Dr. Jaya S Analysis of a Two-Server Queue with Main Server Interruptions, Vacation Policy and Backup Server Assistance 251
Binumon Joseph, K. P. Jose A k-out-of-n:G Retrial System Extending the Service to External Customers under Multiple Vacations and N-Policy 257
Bulinskaya E. Optimization of Multi-Dimensional Insurance Networks with Investment and Reinsurance 263
Chudinova A.A. Markov optimization strategy in distributed computer systems 269
Dinata Syalam Ali Wira Spatial Synchronous Markovian Arrival Process (Spatial SynchMAP) 278
Dudin A., Liu M. The multi-server retrial queueing system with diverse servers and batch Markov arrival process 286

Dudina O., Dudin S., Dudin A. Retrial queueing system with a finite number of sources and additional opportunistic flow of customers 292
Jaya S. Performance Analysis of a Fault Tolerence Queueing System with Interruption and Self Healing Mechanism in Random environment 298
Lapatin I., Nazarov A., Paul S., Lizyura O., Ivanova A. Modeling the CPU Performance with non-Poisson Workload 304
Melikov A., Jose K. P., Bobina J. M. A Production Inventory System with Deteriorating Items and MAP Arrivals: PH-Type Service, and Customer Retrials 310
Morozov E. V., Zhukova K. A. Busy period analysis of M/G/1-type retrial queue 316
Samaja S Naboothri, Jaya S. A Queuing Inventory Model with Batch Demands and Finite Buffer 320
Susmi Skaria, Salini S. Nair, Jose K. P. A MAP/PH/1 Production inventory system with service interruption and different rates of production 327
Tóth Á., Sztrik J. Simulation of a Retrial Queueing System with Two-Way Communication, an Unreliable Server and Impatient Customers Across Various Scenarios 335
Интеллектуальные системы, информационные технологии и программная инженерия 341
Блинова О. В., Таратухина Ю. В. Разработка интеллектуальных адаптивных программ в поликультурной информационно-образовательной среде 342
Голденок Е. Е., Лукьянова Н. А., Лихачев И. С. Эффективность адаптированных нейросетей в диагностике переломов 349
Ибрагимова Э. И. Вычислительная сложность алгоритмов решения задачи корреляционной кластеризации 355
Катаев С. Г., Катаева С. С., Миллер А. А., Ермолин Е. Ю. Обработка гамма-спектрометрических полевых измерений в геофизике с применением понятия интервалов стационарности 361
Корсун А. В., Сущенко С. П., Шкуркин А. С. Оптимизация быстродействия процессорного конвейера 367
Лазарев О. В., Змеев Д. О. Способы применения байесовской сети на основе стандарта OMG Essence для автоматического составления плана итерации 373
Лукьянова Н. А., Голденок Е. Е., Зарубин Е. Ю. Применение машинного обучения в анализе выживаемости при раке молочной железы 379

Маскалёв А. А., Баранова И. В. Исследование влияния параметров генеративно-состязательной нейросети на качество видео, создаваемого на основе набора видеозаписей 385
Пастушенко А. В., Семенова Д. В. Об алгоритмах генерации сетевого графика проектов для задачи RCPSP 392
Пчелинцев Е. А., Курдошев З. М. Метод выбора модели на основе нейронной сети со степенными параметрами для обработки сигналов 398
Солдатенко А. А. Разработка веб-сайта для поиска информации о научных изданиях 405
Хоанг В. Т. Сравнительный анализ методов цифрового предыскажения сигнала с использованием архитектуры косвенного обучения и различных адаптивных алгоритмов 410
Чалых Г. В., Твардовский А. С. Методы детектирования виртуальных машин на основе анализа CPUID, DMI и системных артефактов 417
Erkinov S., Bazarov S. Chained Prediction of University Course Scores Using Machine Learning 423
Molchanov A., Pakulova E. Analysis of the MPTCP protocol with its implementation in PNETlab 430

Научное издание

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2025)

МАТЕРИАЛЫ XXIV Международной конференции имени А. Ф. Терпугова 20–25 октября 2025 г.

Редактор *В. Г. Лихачева*
Корректор *О. А. Турчинович*
Компьютерная верстка *О. Д. Лизюра*
Дизайн обложки *Л. Д. Кривцовой*

Отпечатано на оборудовании
Издательства Томского государственного университета
634050, г. Томск, пр. Ленина, 36.
Тел. 8+(382-2)–52-98-49
Сайт: http://publish.tsu.ru
E-mail: rio.tsu@mail.ru

Подписано к печати 25.12.2025 г.
Формат $60 \times 84^1/_{16}$. Бумага для офисной техники. Гарнитура «Times».
Печ. л. 26. Усл. печ. л. 24.2. Тираж 500 экз. Заказ № 6591.

ISBN 978-5-908040-36-5

