АЛГОРИТМ ИМИТАЦИИ ДИНАМИЧЕСКИХ ХАРАКТЕРИСТИК ТРАФИКА ВЕБ-СЕРВИСА

Горбачёв А. А.[1], Лысенко Д. Э.[2]

DOI: 10.21681/2311-3456-2024-4-104-115

Цель исследования: *разработка алгоритма, основанного на классе ARIMA (autoregressive integrated moving average) – интегрированной модели авторегрессии – скользящего среднего, а также оценка алгоритма для решения задачи имитации сетевого трафика веб-сервиса, позволяющего с одной стороны обеспечить заданный уровень степени сходства динамических свойств реальных узлов вычислительных сетей с ложными, а с другой стороны – приемлемый уровень вычислительной сложности алгоритма.*

Используемые методы: *критерий Акаике, метод максимального правдоподобия, расширенный тест Дики – Фуллера, тест Филипса – Перрона, градиентный спуск, тест Дарбина – Уотсона.*

Результат исследования: *разработан алгоритм, который позволяет синтезировать временной ряд моментов генерации ложного веб-трафика, имеющий относительно низкую ошибку аппроксимации динамических характеристик реального сетевого трафика в условиях приемлемой вычислительной сложности процесса структурно-параметрической идентификации модели и расчета временного ряда для имитации веб-трафика.*

Научная новизна: *заключается в применении интегрированной модели авторегрессии – скользящего среднего с учетом ее адаптивной структурной идентификации по критерию Акаике, параметрической идентификации методом максимального правдоподобия, гиперпараметрической оптимизации длины обучающей выборки и длительности структурно-параметрической идентификации модели для моделирования временного ряда задержек между пакетами ложного трафика веб-сервиса информационных систем.*

Ключевые слова: *временной ряд, моделирование, маскирование, веб-сервис, ложные сетевые информационные объекты, имитация трафика.*

ALGORITHM FOR SIMULATING DYNAMIC TRAFFIC CHARACTERISTICS WEB SERVICE

Gorbachev A. A.[3], Lysenko D. E.[4]

The purpose of the study: *The aim of the work is to develop a model and algorithm based on the class of the integrated autoregression model – the moving average (hereinafter referred to as the ARIMA model), as well as to assess their quality to solve the problem of generating false dynamic properties of real nodes of computer networks when generating false network traffic of a web service, allowing on the one hand to provide a given level of similarity dynamic properties of real nodes of computer networks with false ones, and on the other hand, an acceptable level of computational complexity of the mathematical apparatus.*

The methods used are: *Akaike criterion, maximum likelihood method, extended Dickey – Fuller test, Philips – Perron, gradient descent, Darbin – Watson test, Harke – Bera, Cochrane criterion, direct and iterative time series generation method.*

Result: *the presented model makes it possible to synthesize a time series of moments of generating false web traffic, which has a relatively low error in approximating the dynamic characteristics of real network traffic in conditions of acceptable computational complexity of the process of structural parametric identification of the model and calculation of a time series for generating false web traffic. The presented algorithm makes it possible to increase the effectiveness of protecting computer network nodes by reducing the ability of an attacker to uncover the fact of generating false network traffic in terms of its dynamic characteristics.*

1 Горбачёв Александр Александрович, кандидат технических наук, Краснодарское высшее военное орденов Жукова и Октябрьской Революции Краснознаменное училище имени генерала армии С. М. Штеменко, г. Краснодар, Россия. E-mail: infosec23.00@mail.ru

2 Лысенко Дмитрий Эдуардович, Краснодарское высшее военное орденов Жукова и Октябрьской Революции Краснознаменное училище имени генерала армии С. М. Штеменко, г. Краснодар, Россия. E-mail: dmitrii.Lysencko@yandex.ru

3 Alexander A. Gorbachev, Ph.D., Krasnodar Higher Military Orders of Zhukov and the October Revolution Red Banner School named after General of the Army S. M. Shtemenko, Krasnodar, Russia. E-mail: infosec23.00@mail.ru

4 Dmitry E. Lysenko, Krasnodar Higher Military Orders of Zhukov and the October Revolution Red Banner School named after General of the Army S. M. Shtemenko, Krasnodar, Russia. E-mail: Dmitrii.Lysencko@yandex.ru

Scientific novelty: *it consists in the application of an integrated autoregression model – a moving average, taking into account its adaptive structural identification according to the Akaike criterion, parametric identification by the maximum likelihood method, hyperparametric optimization of the length of the training sample and the duration of the structural-parametric identification of the model to simulate a time series of delays between packets of false traffic of a web service of military information systems.*

Keywords: *time series, modeling, masking, web service, false network information objects, traffic simulation.*

Введение

В связи с ростом объемов обработки данных и предоставления информационных сервисов через Интернет, становится критически важным обеспечение их информационной безопасности. Атаки на веб-приложения – один из наиболее популярных методов кибератак.

По данным исследования центра реагирования на инциденты информационной безопасности промышленных инфраструктур «Лаборатории Касперского» (Kaspersky ICS CERT), 17% от общего числа атак пришлось на эксплуатацию уязвимостей и недостатков защиты веб-приложений[5].

Злоумышленники могут использовать скомпрометированные сайты в различных целях: для распространения вредоносного программного обеспечения, кражи конфиденциальных данных, несанкционированного внедрения информации, для мошенничества или проникновения во внутреннюю инфраструктуру организации [1].

Наряду с наиболее распространенными мерами защиты, включающими в себя методы и средства предотвращения вторжений [2], обнаружения и реагирования на инциденты[6], средства криптографической защиты[7], резервного копирования и восстановления[8], целесообразными для применения в общей системе защиты информационных систем являются методы и средства маскирования информационных направлений (трафика) [3, 4]. Маскирование трафика потенциально позволяет обеспечить выполнение требований к мерам защиты информационных систем, включающих: скрытие архитектуры и конфигурации систем [5, 6]; создание фиктивных систем или компонентов для обнаружения и анализа действий атакующих[9]; имитацию или сокрытие настоящих информационных технологий и структурных особенностей системы [7, 8].

Маскирование трафика осуществляется посредством имитации реального трафика между узлами вычислительной сети. В настоящее время разработан ряд научно-технических предложений по имитации реального трафика в локальных вычислительных сетях[10]. Однако, задача по обеспечению заданного уровня степени сходства динамических свойств реальных узлов вычислительных сетей с ложными при имитации сетевого трафика веб-сервиса остается актуальной.

При использовании обманных систем (*deception systems*, ложных сетевых информационных объектов) для маскирования трафика, системы могут стать менее уязвимыми к атакам, направленным на их дестабилизацию или выведение из строя. Это особенно важно для критических инфраструктур, таких как финансовые учреждения, здравоохранение и государственные службы, где последствия кибератак могут иметь значительные масштабы.

Имитация трафика веб-сервиса ложными сетевыми информационными объектами потенциально позволяет снизить эффективность сетевой разведки по отношению к узлам вычислительной сети.

Основная идея состоит в создании ложного трафика, который затрудняет выделение и анализ важных данных за счет:

5 Второе полугодие 2023 года — краткий обзор основных инцидентов промышленной кибербезопасности // Официальный информационный ресурс АО «Лаборатория Касперского» [Электронный ресурс]. 2023. – URL: https://ics-cert.kaspersky.ru (дата обращения 17.04.2024).

6 Канев А. Н. Мониторинг событий и обнаружение инцидентов информационной безопасности с использованием SIEM-систем // Международный студенческий научный вестник. 2015. №. 3-1. С. 122–123.

7 Авдошин С. М., Савельева А. А. Криптографические методы защиты информационных систем // Известия АИН им. А. М. Прохорова. Бизнес-информатика. 2006. Т. 17. №. 2. С. 91–99.

8 Даниленко А. Ю. Защита данных в сложных информационных системах // Труды Института системного анализа Российской академии наук. 2007. Т. 29. С. 49–58.

9 Иванов И. И., Максимов Р. В. Спецификация функциональной модели для расширения пространства демаскирующих признаков в виртуальных частных сетях // Инновационная деятельность в Вооруженных Силах Российской Федерации: Труды всеармейской научно-практической конференции, Санкт-Петербург. ВАС, 2017. С. 138–147.

10 Татарникова Т. М. Статистические методы исследования сетевого трафика // Информационно-управляющие системы. – 2018. – №. 5 (96). – С. 35–43.

- имитации активности пользователей и системных процессов путём имитации трафика, не несущего реальной информационной нагрузки, что создаёт дополнительный объём данных, среди которого злоумышленнику сложнее выделить значимую информацию;
- изменения параметров сетевого трафика, включая интервалы времени между пакетами, размеры пакетов и порядок их следования, для создания случайного характера трафика;
- применения алгоритмов машинного обучения для адаптации ложного трафика под текущие условия сети и активность пользователя, обеспечивая высокий уровень реалистичности и эффективности маскирования.

В данной статье разработан алгоритм, который может значительно повысить результативность защиты информационных систем от внутренних и внешних угроз. Сформулирована и решена задача структурно-параметрической идентификации интегрированной модели авторегрессии – скользящего среднего (*ARIMA*) для синтеза временного ряда моментов имитации сетевого трафика веб-сервиса от узлов вычислительной сети.

Идентификация сетевого трафика является одной из задач в области безопасности, защиты и управления трафиком сетей передачи данных. Решение данной задачи осуществляется с использованием методов классификации и моделирования сетевого трафика.

В основе ряда моделей трафика лежат стационарные случайные процессы, с помощью которых воспроизводятся характеристики трафика (количество пакетов, полученных или отправленных в течение определенного промежутка времени; интервалы между пакетами и т.д.). Основными из них являются:

- модели на основе законов распределения[11];
- модели на основе теории фракталов [9];
- регрессионные и авторегрессионные модели[12];
- модели экспоненциального сглаживания[13];
- модели, основанные на алгоритмах машинного обучения[14];
- модели на базе цепей Маркова [10];
- классификационные модели и др. [11, 12].

Наиболее популярными и широко используемыми являются классы авторегрессионных и нейросетевых моделей[15]. В рамках рассматриваемой задачи имитации трафика, с целью выделения достаточного ресурса для хранения трафика в сетевом оборудовании в краткосрочном промежутке времени, авторегрессионная модель имеет ряд преимуществ перед моделями глубокого обучения. В условиях малого количества исходных данных, высокой изменчивости статистических и динамических свойств временных рядов, отсутствия значительных временных и вычислительных ресурсов на обучение алгоритмов глубокого обучения, данная модель способна генерировать с приемлемой ошибкой аппроксимации временной ряд задержек между пакетами сетевого трафика на относительно коротких промежутках времени. В данной статье будут рассматриваться модели на основе стохастических временных рядов.

Модель *ARIMA* широко используется для анализа и генерирования значений временных рядов благодаря возможности моделировать различные типы данных, включая нестационарные временные ряды. Для построения модели *ARIMA* достаточно использовать информацию, содержащуюся в самих анализируемых данных временного ряда. Если ряд после взятия d последовательных разностей сводится к стационарному, то для генерирования новых значений временного ряда можно применить комбинированную модель авторегрессии и скользящего среднего, обозначаемую как ARIMA (p, d, q). Сокращение I в данной аббревиатуре означает «интегрированный» [13, 14].

Модель ARIMA (p, d, q) *– (модель Бокса-Дженкинса или модель интегрированной авторегрессии – скользящего среднего)* позволяет работать с зависимостями, имеющими тренд (1), параметры модели приведены в таблице 1.

$$\Delta^d Y_t = c + \sum_{i=1}^{p} \alpha_i \Delta^d Y_{t-i} + \sum_{j=1}^{q} \beta_j \Delta^d \varepsilon_{t-j} + \varepsilon_t, \quad (1)$$

11 Тырсин А. Н. Метод подбора наилучшего закона распределения непрерывной случайной величины на основе обратного отображения //Вестник Южно-Уральского государственного университета. Серия: Математика. Механика. Физика. – 2017. – Т. 9. – №. 1. – С. 31–38.

12 Селиверстова А. В. Сравнительный анализ моделей и методов прогнозирования // Современные научные исследования и инновации. 2016. № 11(67). С. 241–248.

13 Калекар П. С. Прогнозирование временных рядов с использованием экспоненциального сглаживания Холта-Уинтерса // Школа информационных технологий имени Канвала Рекхи. 2004. №.13. С. 1–13.

14 Гладышев А. И., Жуков А. О. Достоинства и недостатки имитационного моделирования с использованием нейронных сетей // Вестник Российского нового университета. 2013. №. 4. С. 53.

15 Тихонов Э. Е. Методы прогнозирования в условиях рынка: учебное пособие. Невинномысск, 2006. – 221 с.

Таблица 1.

Параметры модели ARIMA

Параметры	Описание
Y_t	уровень временного ряда в момент времени t (зависимая переменная)
Y_{t-i}	уровни временного ряда в моменты времен, соответственно (независимые переменные)
α_i	оцениваемые коэффициенты авторегрессии
ε_t	случайное возмущение, описывающее влияние переменных, не учтенных в модели
ε_{t-j}	значения остатков j временных периодов назад (независимые переменные)
β_j	оцениваемые коэффициенты скользящего среднего
d	порядок модели *ARIMA*, характеризующий степень интегрирования
Δ^d	оператор взятия конечной разности порядка d
p	параметр обозначает количество лагов (задержек) временного ряда, используемых в качестве предикторов
q	параметр указывает на количество лагов ошибок сгенерированных новых значений, используемых в модели

В общем виде модель (2) представляет собой отображение входных характеристик в выходные. Входные характеристики представляют собой множество управляемых факторов-аргументов A (3-4) и множество неуправляемых параметров S (5). Выходные характеристики модели Z (6) представляют собой временной ряд пауз между пакетами генерируемого ложного сетевого трафика веб-сервиса, то есть модель позволяет реализовать имитацию трафика в смысле идентичности динамических характеристик реального и ложного трафика без имитации содержимого.

$$F : \{A,S\} \longrightarrow Z, \quad (2)$$

В качестве неуправляемых и управляемых факторов выступают:

$$A = \{p,q,\Theta,l\}, \quad (3)$$

$$\Theta = \{c,\alpha_i,\beta_j,\varepsilon_\tau\},\ i \in [1,...,p],\ j \in [1,...,q],\ d \in [0,1], \quad (4)$$

$$S = \{\tau_1,...,\tau_K\}, \quad (5)$$

$$Z = \{F(\tau_1,p,q,\Theta,n),...,F(\tau_K,p,q,\Theta,n)\}. \quad (6)$$

где: Θ – параметры модельного оператора, l – длина аппроксимируемого ряда, S – множество неуправляемых параметров (входной имитируемый временной ряд пауз между Опакетами реального сетевого трафика веб-сервиса), Z – выходные характеристики модели, представляющие собой временной ряд пауз между пакетами имитируемого (ложного сетевого трафика).

Область допустимого множества факторов модели формализована выражением 7:

$$Q = \begin{cases} \tau \in R;\ n \in [1,...,10^5]; \\ p \in [0,...,5];\ q \in [0,...,5];\ d \in [0,1]; \\ c \in R;\ \varepsilon_t \in [0,+\infty]; \\ \alpha_i \in [-1,1];\ \beta_j \in [-1,1]. \end{cases} \quad (7)$$

Модель применяется для генерации новых значений временного ряда Z.

Таблица 2.

Основные исходные данные

Параметры	Описание
l	длина считываемого временного ряда, задержек между пакетами сетевого трафика вычислительной сети, необходимых для идентификации математической модели *ARIMA* или устанавливается временной интервал, в течение которого будут собираться данные.
K^*_{st}	критическое значение критерия стационарности временного ряда для уровня значимости с целью определения стационарности временного ряда и структурной идентификации математической модели *ARIMA*.

Описание алгоритма имитации динамических характеристик трафика веб-сервиса

Основной задачей алгоритма является поиск оптимальных параметров интегрированной модели авторегрессии – скользящего среднего для генерации новых значений пауз между пакетами ложного трафика, основанного на динамических характеристиках реального трафика информационных систем, тем самым снижая эффективность сетевой разведки злоумышленников и последующей реализации компьютерных атак в вычислительной сети.

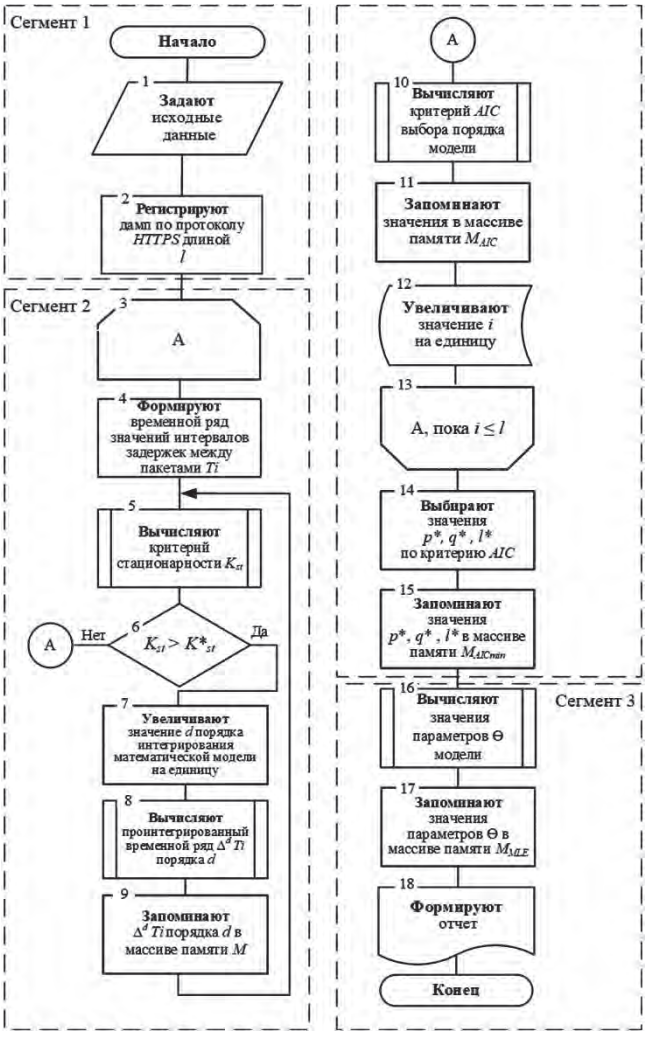

Рис. 1. Алгоритм структурной и параметрической идентификации модели

Создание модифицированных потоков данных обеспечивает среду, в которой атакующие сталкиваются с препятствиями при попытке идентифицировать реальные (отличить от ложных) информационные ресурсы, в связи с чем снижается качество анализа сетевого трафика злоумышленниками.

В таком случае особенно важно не только разработать адекватный алгоритм генерации, но и учесть природу трафика, протокол передачи, архитектуру сети, степень загруженности и характер нагрузки. В зависимости от перечисленных выше факторов динамические свойства трафика будут различаться.

Реализация предлагаемого алгоритма структурной и параметрической идентификации интегрированной модели авторегрессии – скользящего среднего для имитации трафика веб-сервиса поясняется блок-схемой последовательности действий, представленной на рис. 1 и включает следующие этапы:

1. *Задают исходные данные (блок 1), обозначение и описание которых приведены в таблице 2.*

Регистрируют дамп реального сетевого трафика по протоколу *HTTPS* (блок 2 на рис. 2).

Выполняют сбор сетевого трафика с помощью специализированного программного обеспечения и проводят фильтрацию данных, ограничиваясь пакетами, передаваемыми через порты, характерными для протокола *HTTPS* (*HyperText Transfer Protocol Secure*). Особое внимание уделяется пакетам с флагом *SYN*, которые инициируют *TCP*-соединения. К примеру, выборка из дампа сетевого трафика

вычислительной сети, снятого с сетевого устройства за 4 часа рабочего времени, содержит 7587 фактов регистрации поступления *TCP*-пакетов с флагом *SYN* на установление сетевого соединения по протоколу *HTTPS* (рис. 2). Для обеспечения точности и надежности результатов анализа была проведена предварительная обработка данных. В частности, была проведена очистка данных от поврежденных пакетов и пакетов, не относящихся к исследуемому потоку данных.

2. *Формируют временной ряд пауз между пакетами протокола HTTPS (рис. 3а) и проверяют*

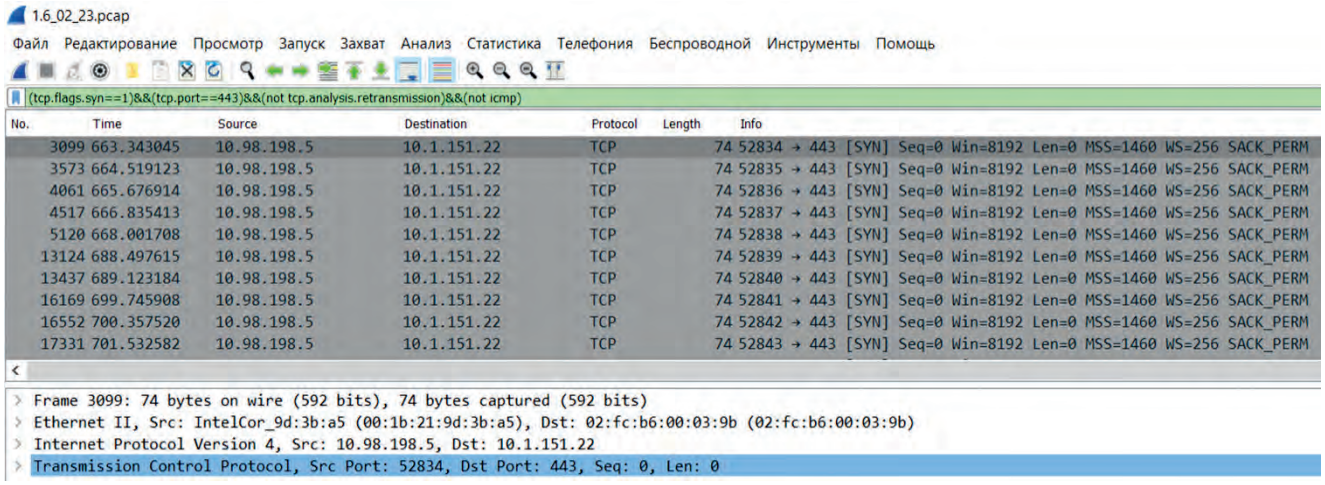

Рис. 2. Извлечение признаков из эмпирических данных о событиях, характеризующих временной ряд

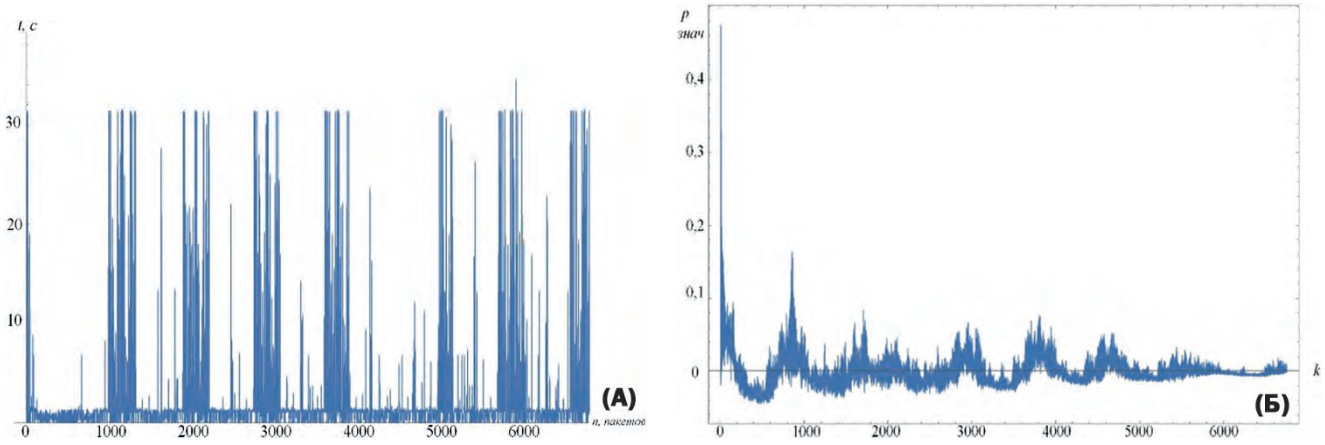

Рис.3. а) График временного ряда пауз между SYN–пакетами веб-трафика;
б) График автокорреляционной функции процесса регистрации TCP пакетов с флагом SYN на установление сетевого соединения

Таблица 3.

Результаты оценки стационарности исследуемого временного ряда с использованием тестов единичного корня

Наименования теста (критерия)	Значение статистики	p - value	Уровень значимости	Интерпретация результатов теста
Тест Дики – Фуллера	-87,840	1,6585×10-8	0,05	временной ряд стационарен
Расширенный Тест Дики – Фуллера	-6,892	6,0228×10-9	0,05	временной ряд стационарен
Тест Филипса – Перрона	-27,931	2,077×10-5	0,05	временной ряд стационарен

временной ряд на наличие явного тренда и гетероскедастичности (непостоянной дисперсии) (рис. 3б) для дальнейшего анализа и выбора модели (блок 4 на рис. 2).

Проверяют временной ряд на стационарность (блок 5, 6 на рис. 2). Оценка стационарности временного ряда может быть проведена с использованием специфических параметрических тестов или статистических «тестов единичного корня» (*Unit root test*), которые позволяют оценить стационарность временного ряда (табл. 3).

В случае, если условие $K_{st} > K^*_{st}$ не выполняется, что соответствует сетевому трафику, обладающему свойством стационарности, вычисляют значения критерия Акаике (*AIC*) для выбора порядка математической модели для всех комбинаций параметров p, q порядка модели (блок 10 на рис. 2).

В случае если условие $K_{st} > K^*_{st}$ выполняется, что соответствует сетевому трафику, не обладающему свойством стационарности, увеличивают значение порядка интегрирования Δ^d математической модели на единицу. Затем вычисляют проинтегрированный временной ряд $\Delta^d T_i$ порядка Δ^d задержек между пакетами сетевого трафика длиной l. После чего запоминают проинтегрированный временной ряд порядка Δ^d в массиве памяти M. Затем вычисляют значение критерия стационарности K_{st} проинтегрированного временного ряда порядка Δ^d (блок 7–9 на рис.2).

Выбирается модель с параметрами порядка, соответствующими минимальному значению статистики[16]:

$$AIC(A,S) = 2k(A,S) - 2ln(\bar{L}(A,S)) \longrightarrow \min_{A,S \in Q} \qquad (8)$$

где k – количество оцененных параметров (включая p, q, константу c, если она включена, и дисперсию ошибок σ^2; $\bar{L}$ – максимальное значение логарифмической функции правдоподобия.

Данная процедура повторяется для каждой модели *ARMA*, и выбирается модель с наименьшим критерием *AIC*. Использование *AIC* в качестве инструмента для выбора модели *ARMA* позволяет автоматизировать процесс поиска оптимальных параметров, особенно при использовании программного обеспечения для статистического анализа, которое может быстро перебирать множество комбинаций параметров и автоматически вычислять *AIC* для каждой модели. Результаты запоминаются в массиве памяти M_{AIC} (блок 11 на рис. 2).

Итерационный процесс продолжается до тех пор, пока не будет найдена оптимальная модель по информационному критерию Акаике и оптимальная длина временного ряда, начиная с минимальной длины $l = 2$ и увеличиваясь с шагом $i = 1$ (блок 12–15 на рис. 2).

3. *Оценка параметров модели по методу максимального правдоподобия (блок 16 на рис.2).*

После структурной идентификации, проведенной на предыдущем шаге, осуществляется параметрическая идентификация модели.

В статистике применяются три основных метода оценивания:

1) метод наименьших квадратов;
2) метод моментов;
3) метод максимального правдоподобия [15].

Как правило, применение метода максимального правдоподобия для этих целей в моделях *ARIMA* (p, d, q) дает асимптотически несмещенную, состоятельную и эффективную оценку параметров. Его используют для любых моделей, задающих вид распределения наблюдаемых переменных. Два других метода можно использовать лишь тогда, когда распределение переменных можно представить в определенном виде. Если есть гипотеза о точном виде распределения, то всегда понятно, как получать оценки параметров, распределений параметров и различных статистик, как проверять гипотезы, хотя сами расчеты могут быть относительно трудоемкими. Если правильно выбрать параметризацию, то распределение оценок в малых выборках может быть близко к асимптотическому, если неправильно, то асимптотическое распределение будет неудовлетворительной аппроксимацией.

Оценка параметров модели представляет собой ключевой этап в анализе временного ряда, который требует точного и систематического подхода. На первом этапе строится логарифмическая функция правдоподобия $ln\ L$, которая представляет собой логарифм вероятности наблюдаемых данных

16 Носко В. П. Эконометрика. Элементарные методы и введение в регрессионный анализ временных рядов / В. П. Носко. – Москва: Фонд «Институт экономической политики им. Е. Т. Гайдара», 2004. – 501 с.

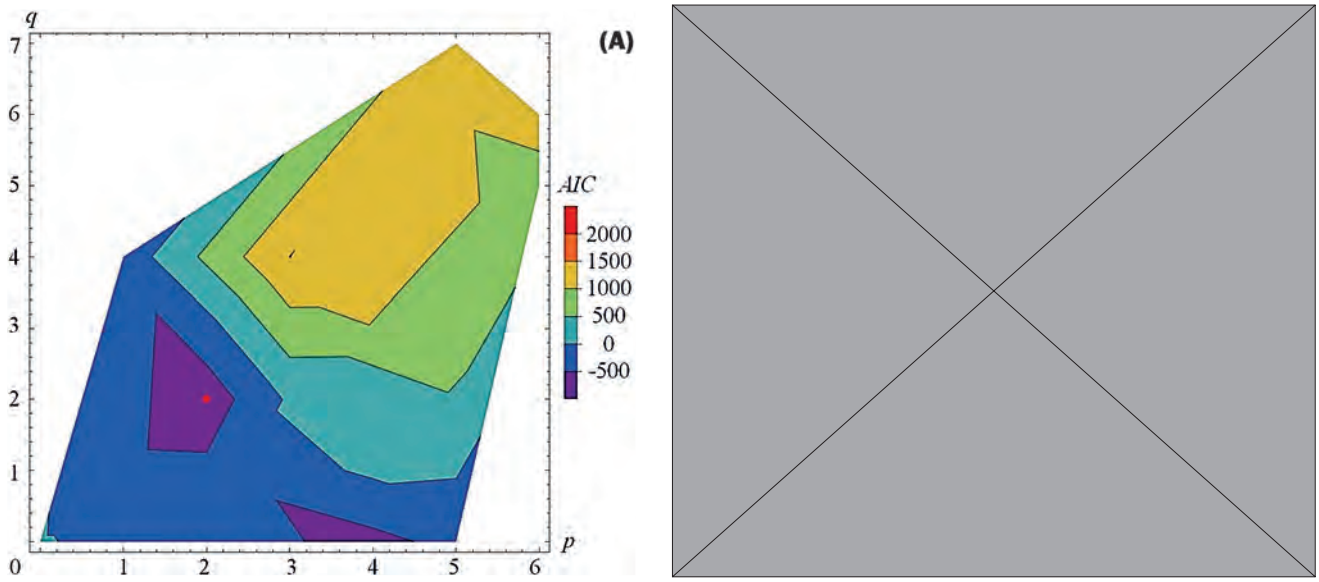

Рис. 4. Графическая интерпретация структурной идентификации модели:
структурная идентификация модели по критерию Акаике (а); оценка оптимальной длины обучающей выборки k (б)*

Таблица 4.

Результаты значений параметров авторегрессии, скользящего среднего и дисперсии ошибок

Модель и ее порядок	**Длина обучающей выборки временного ряда k, знач.**	**Значение параметров**	**Значение *AIC***	**Время оценки параметров, сек**
ARMA (2,2)	1000	$c = 0{,}25986, \{\hat{a}_i = 0{,}17677; 0{,}47375\}, \{\hat{\beta}_j = -0{,}15073; -0{,}176258\}, \hat{\sigma}^2 = 0{,}446593$	–794	0,7757
ARMA (4,0)	800	$c = 0{,}34921, \{\hat{a}_i = 0{,}0174;0{,}2389;0{,}0808;0{,}1622\}, \hat{\sigma}^2 = 0{,}44329$	–638	0,0830
ARMA (2,1)	600	$c = 0{,}05116, \{\hat{a}_i = 0{,}95010; 0{,}00414\}, \{\hat{\beta}_j = -0{,}87796\}, \hat{\sigma}^2 = 1{,}2913$	–399	0,0720
ARMA (5,0)	1900	$c = 0{,}44921, \{\hat{a}_i = 0{,}0800; 0{,}1275; 0{,}1477; 0{,}1000; 0{,}054\}, \hat{\sigma}^2 = 0{,}8224$	–357	0,1724
ARMA (1,3)	1100	$c = 0{,}26701, \{\hat{a}_i = 0{,}76122\}, \{\hat{\beta}_j = -0{,}66714; 0{,}02674; 0{,}02489; 0{,}0531\}, \hat{\sigma}^2 = 1{,}31893$	–310	0,1190

Таблица 5.

Результаты работы алгоритма по поиску оптимальных значений

Размер обучающей выборки, шт.	1000
Модель	*ARMA* (2,2)
Значение AIC	–794
Параметры модели	$c = 0{,}25986, \{\hat{a}_i = 0{,}17677; 0{,}47375\}, \{\hat{\beta}_j = -0{,}15073; -0{,}176258\}, \hat{\sigma}^2 = 0{,}446593$
Время подбора параметров, с	0,0096975
Общее время работы алгоритма, с	67, 4401839

при заданных параметрах модели. Для модели *ARIMA* (p, d, q) эта функция может быть выражена как (9):

$$lnL\,(a_i,\beta_j,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t - \tilde{Y}_t)^2, \quad (9)$$

где Y_t – значение временного ряда в момент времени t; $\tilde{Y}_t$ – предсказанные значения модели; a_i – коэффициент *AR* части модели; β_j – коэффициент *MA* части модели; σ^2 – оценка дисперсии ошибок; n – общее количество наблюдений.

Для максимизации функции правдоподобия необходимо вычислить её частные производные по каждому из параметров модели: a_i (для $i = 1, \ldots, p$), β_j (для $j = 1,\ldots,q$) и σ^2. Это позволит определить направление наискорейшего роста функции правдоподобия.

Поскольку аналитическое решение задачи максимизации логарифмической функции правдоподобия часто недостижимо, применяются численные методы оптимизации, такие как градиентный спуск, метод Ньютона – Рафсона или алгоритмы квази-Ньютона. Эти методы итеративно корректируют оценки параметров, двигаясь в направлении градиента логарифмической функции правдоподобия.

Итерационный процесс продолжается до тех пор, пока не будет достигнут критерий сходимости (пока изменения в логарифмической функции правдоподобия или в значениях параметров не станут незначительными, в нашем случае это значение равно 10^{-4}). Это указывает на то, что были найдены параметры, максимизирующие функцию правдоподобия, и процесс оптимизации может быть остановлен.

После завершения процесса параметрической оптимизации, полученные значения параметров $\hat{a}_i$, $\hat{\beta}_j$, $\hat{\sigma}^2$ используются как оценки исходных параметров модели *ARIMA*. Эти оценки предоставляют информацию о взаимосвязях внутри временного ряда и могут быть использованы для дальнейшего анализа и генерации новых значений. В контексте модели *ARIMA* это может быть записано как (10):

$$(\hat{a}_i,\hat{\beta}_j,\hat{\sigma}^2) = argmax_{a_i,\beta_i,\sigma^2 \in Q}\, lnL(a_i,\beta_j,\sigma^2) \quad (10)$$

где: $\hat{a}_i$, $\hat{\beta}_j$, $\hat{\sigma}^2$ – оцененные значения параметров авторегрессии, скользящего среднего и дисперсии ошибок соответственно; $lnL(a_i,\beta_j,\sigma^2)$ – логарифмическая функция правдоподобия модели.

В процессе исследования была выполнена оценка и выбор оптимальных моделей, расчет их параметров и времени выполнения по информационному критерию Акаике. Изначально для каждого временного ряда, начиная с минимальной длины $l = 2$ и увеличиваясь с шагом $i = 1$, была подобрана модель (рис. 4).

После чего была сформирована таблица, содержащая данные о типе модели, параметрах авторегрессии и скользящего среднего, константе, дисперсии шума, порядках модели (p и q), времени выполнения расчетов и значения *AIC*. Для удобства анализа и дальнейшего обсуждения результаты были упорядочены по возрастанию значения критерия *AIC*, что позволило выделить модели с наилучшими значениями информационного критерия. Из полной таблицы были отобраны 5 наилучших моделей, которые демонстрируют наименьшее значение *AIC*, указывающее на оптимальное соотношение между качеством аппроксимации и сложностью модели. Результаты полученных значений представлены в таблице 4.

Формируют отчет (блок 18 на рис. 2). Формируется таблица (таблица 5), включающая результаты структурной и параметрической идентификации модели ARMA.

Вывод

В ходе исследования был разработан алгоритм структурной и параметрической идентификации интегрированной модели авторегрессии – скользящего среднего, который позволяет синтезировать временной ряд моментов имитации веб-трафика, имеющий низкую ошибку аппроксимации динамических характеристик реального сетевого трафика в условиях приемлемой вычислительной сложности процесса структурно-параметрической идентификации модели и расчета временного ряда для имитации веб-трафика. Подход базировался на итеративной процедуре структурной идентификации и оценки параметров моделей.

Структурная идентификация модели *ARIMA* осуществлялась посредством автоматического выбора порядков модели. Определены диапазоны значений для параметров авторегрессии (p) и скользящего среднего (q), которые для модели *ARIMA*(p, d, q)

находятся в следующих интервалах: p от 0 до 5, q от 0 до 3, порядок интегрирования d был определен в интервале от 0 до 2. Оценка моделей по информационному критерию Акаике выявила, что оптимальные значения *AIC* находятся в диапазоне от –794 до 2000 для различных комбинаций p и q.

Размеры обучающих выборок k были определены в интервале от 200 до 2800 значений, что демонстрирует гибкость подхода в адаптации к разнообразным объемам данных.

Числовые результаты для выбранных моделей *ARMA*(p, q) и *ARIMA*(p, d, q) отражают следующее: модель *ARMA*(2,2) с длиной обучающей выборки в 1000 значений и параметрами $c = 0{,}25986$, $\{\hat{a}_i = 0{,}17677;\ 0{,}47375\}$, $\{\hat{\beta}_j = -0{,}15073;\ -0{,}176258\}$, $\hat{\sigma}^2 = 0{,}446593$, показала *AIC* равный –794, что является оптимальным результатом в данном исследовании.

Время, затраченное на подбор параметров, составило 0,0096975 секунды, демонстрируя относительно высокую вычислительную эффективность разработанного алгоритма. Общее время работы алгоритма составило 67,4401839 секунды, подтверждая возможность его использования в условиях реального времени.

Сформированная на основе этого подхода процедура оценки моделей позволяет достичь оптимального баланса между точностью моделирования и вычислительной сложностью модели и алгоритма. Сложные и многоэтапные процедуры, такие как перебор параметров модели и оценка стационарности рядов, требуют значительных вычислительных ресурсов. Поэтому оптимизация алгоритмов подобных процессов играет ключевую роль в повышении эффективности общей системы анализа сетевого трафика.

Полученные результаты могут быть использованы для разработки алгоритма имитации сетевого трафика с целью введения в заблуждение злоумышленников и повышения защищенности информационных систем.

Литература

1. *Шерстобитов Р. С. Модель маскирования информационного обмена в сети передачи данных ведомственного назначения // Системы управления, связи и безопасности. 2024. № 1. С. 1–25. DOI: 10.24412/2410-9916-2024-1-001-025.*
2. *Фатеев А. Г. Применение средств защиты информации для реализации мер защиты, установленных специальными нормативными документами Федеральной службы по техническому и экспертному контролю // Инжиниринг и технологии. 2020. Т. 5, № 1. С. 24–29. DOI 10.21685/2587-7704-2020-5-1-6.*
3. *Москвин А. А., Максимов Р. В., Горбачев А. А. Модель, оптимизация и оценка эффективности применения многоадресных сетевых соединений в условиях сетевой разведки // Вопросы кибербезопасности. 2023. № 3(55). С. 13–22. DOI 10.21681/2311-3456-2023-3-13-22.*
4. *Горбачев А. А., Максимов Р. В. Проблема маскирования и применения технологий машинного обучения в киберпространстве // Вопросы кибербезопасности. 2023. № 5(57). С. 37–49. DOI 10.21681/2311-3456-2023-5-37-49.*
5. *Maximov R. V., Sokolovsky S. P., Telenga A. P. Methodology for substantiating the characteristics of false network traffic to simulate information systems // CEUR Workshop Proceedings : BIT 2021 – Selected Papers of 11th International Scientific and Technical Conference on Secure Information Technologies, Moscow. 2021. pp. 115–124.*
6. *Maximov R. V., Sokolovsky S. P., Telenga A. P. Honeypots network traffic parameters modelling // CEUR Workshop Proceedings: BIT 2021 – Selected Papers of 11th International Scientific and Technical Conference on Secure Information Technologies, Moscow. 2021. pp. 229–239.*
7. *Способ (варианты) защиты вычислительных сетей. Патент № 2307392 C1 Российская Федерация, МПК G06F 21/00, H04L 9/32. Выговский Л. С., Заргаров И. А., Кожевников Д. А., Максимов Р. В., Павловский А. В.,Стародубцев Ю. И., Худайназаров Ю. К., Юров И. А.; заявитель и патентообладатель Военная академия связи (RU). – № 2006114974/09 : заявл. 02.05.2006: опубл. 27.09.2007.*
8. *Способ контроля информационных потоков в цифровых сетях связи. Патент № 2267154 C1 Российская Федерация, МПК G06F 12/14, G06F 11/00. Андриенко А. А., Куликов О. Е., Костырев А. Л., Максимов Р. В., Павловский А. В., Лебедев А. Ю., Колбасова Г. С.; заявитель и патентообладатель Военная университет связи (RU). № 2004121529/09. заявл. 13.07.2004; опубл. 27.12.2005.*

9. Мельникова Ю. В., Лажаунинкас Ю. В. Компьютерное моделирование экономических процессов с применением методов фрактального анализа // Наука Красноярья. – 2022. – Т. 11. – №. 4. – С. 7–23.
10. Егоров И. К. Проверка прогнозирования посещения веб-страниц на основе цепи Маркова для моделирования профиля поведения пользователей / И. К. Егоров, В. Ю. Радыгин // Инновационные механизмы управления цифровой и региональной экономикой: Материалы V Международной студенческой научной конференции, Москва, 15-16 июня 2023 года. – Москва: Национальный исследовательский ядерный университет «МИФИ», 2023. – С. 429–437.
11. Микульский А. А. Обзор моделей прогнозирования // Dunărea–Nistru: Anuar. 2019. Т. 6. С. 284–304.
12. Хайндман Р. и Атанасопулос Дж. Прогнозирование: принципы и практика / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2023. – 458 с.: ил
13. Мирзакулова Ш. А. Исследование временного ряда на стационарность // Образовательная система: новации в сфере современного научного знания: сборник научных трудов. Казань: ООО «СитИвент», 2019. С. 318–333.
14. Фелькер М. Н., Чеснов В. В. Исследование влияния изменения параметров модели ARIMA на качество прогноза для коротких наборов данных // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. 2021. Т. 21. №. 3. С. 36–46.
15. Скоробогатых Е. Ю. К вопросу о методах нахождения оценок параметров регрессионных моделей / Е. Ю. Скоробогатых, С. Н. Мухина // Известия Балтийской государственной академии рыбопромыслового флота: психолого-педагогические науки. – 2023. – № 3(65). – С. 205-212. – DOI 10.46845/519.242071-5331-2023-3-65-205-212.

References

1. Sherstobitov R. S. Model maskirovaniya informatsionnogo obmena v seti peredachi dannikh vedomstvennogo naznacheniya [A model for organizing information exchange in a departmental data transmission network has been developed]. Management, communication and security systems. 2024. vol. 1. pp. 1–25. DOI: 10.24412/2410-9916-2024-1-001-025.
2. Fateev A. G. Primenenie sredstv zashchiti informatsii dlya realizatsii mer zashchiti, ustanovlennikh spetsialnimi normativnimi dokumentami Federalnoi sluzhbi po tekhnicheskomu i ekspertnomu kontrolyu [The use of information security tools for the implementation of protection measures established by special regulatory documents of the Federal Service for Technical and Expert Control]. // Inzhiniring i tekhnologii, 2020, vol. 1, pp. 24–29 (in Russia).
3. Moskvin A. A., Maksimov R. V., Gorbachev A. A. Model, optimizatsiya i otsenka effektivnosti primeneniya mnogoadresnikh setevikh soedinenii v usloviyakh setevoi razvedki [Model, optimization and evaluation of the effectiveness of multicast network connections in the context of network intelligence]. Cybersecurity issues. 2023. vol. 3(55). pp. 13-22. DOI 10.21681/2311-3456-2023-3-13-22 (in Russia).
4. Gorbachev A. A., Maksimov R. V. Problema maskirovaniya i primeneniya tekhnologii mashinnogo obucheniya v kiberprostranstve [The problem of masking and applying machine learning technologies in cyberspace]. Cybersecurity issues. 2023. vol. 5(57). pp. 37–49. DOI 10.21681/2311-3456-2023-5-37-49 (in Russia).
5. Maximov R. V., Sokolovsky S. P., Telenga A. P. Methodology for substantiating the characteristics of false network traffic to simulate information systems // CEUR Workshop Proceedings : BIT 2021 – Selected Papers of 11th International Scientific and Technical Conference on Secure Information Technologies, Moscow. 2021. pp. 115–124.
6. Maximov R. V., Sokolovsky S. P., Telenga A. P. Honeypots network traffic parameters modelling // CEUR Workshop Proceedings : BIT 2021 – Selected Papers of 11th International Scientific and Technical Conference on Secure Information Technologies, Moscow. 2021. pp. 229–239.
7. Sposob (varianti) zashchiti vichislitelnikh setei. Patent № 2307392 C1 Rossiiskaya Federatsiya, MPK G06F 21/00, H04L 9/32. Vigovskii L. S., Zargarov I. A., Kozhevnikov D. A., Maksimov R. V., Pavlovskii A. V.,Starodubtsev Yu. I., Khudainazarov Yu. K., Yurov I. A.; zayavitel i patentoobladatel Voennaya akademiya svyazi (RU). – № 2006114974/09 : zayavl. 02.05.2006: opubl. 27.09.2007 (in Russian).
8. Sposob kontrolya informatsionnikh potokov v tsifrovikh setyakh svyazi. Patent № 2267154 C1 Rossiiskaya Federatsiya, MPK G06F 12/14, G06F 11/00. Andrienko A. A., Kulikov O. E., Kostirev A. L., Maksimov R. V., Pavlovskii A. V., Lebedev A. Yu., Kolbasova G. S.; zayavitel i patentoobladatel Voennaya universitet svyazi (RU). № 2004121529/09. zayavl. 13.07.2004; opubl. 27.12.2005 (in Russian).
9. Melnikova Yu. V., Lazhauninkas Yu. V. Kompyuternoe modelirovanie ekonomicheskikh protsessov s primeneniem metodov fraktalnogo analiza [Computer modeling of economic processes using fractal analysis methods] // Science Krasnoyarsk. 2022. vol. 11. pp. 7–23.
10. Yegorov I. K. Proverka prognozirovaniya poseshcheniya veb-stranits na osnove tsepi Markova dlya modelirovaniya profilya povedeniya polzovatelei / I. K. Yegorov, V. Yu. Radigin // Innovatsionnie mekhanizmi upravleniya tsifrovoi i regionalnoi ekonomikoi : Materiali V Mezhdunarodnoi studencheskoi nauchnoi konferentsii, Moskva, 15–16 iyunya 2023 goda. – Moskva: Natsionalnii issledovatelskii yadernii universitet «MIFI», 2023. – S. 429–437. – EDN ATCIFV.
11. Mikulskii A. Obzor modelei prognozirovaniya [Overview of forecasting models]. Dunărea–Nistru: Anuar, 2019, vol. 6, pp. 284–304 (in Russia).

12. *Khaindman R. Dzh. i Atanasopulos Dzh. Prognozirovanie: printsipi i praktika. [Forecasting: principles and practice]: Melbourne, Australia. 2021 (in Russia).*

13. *Mirzakulova, Sh. A. Issledovanie vremennogo ryada na statsionarnost [Investigation of the time series for stationarity]. Obrazovatelnaya sistema: novatsii v sfere sovremennogo nauchnogo znaniya : sbornik nauchnikh trudov, Kazan, 2019, pp. 318–333 (in Russia).*

14. *Felker M. N., Chesnov V. V. Issledovanie vliyaniya izmeneniya parametrov modeli ARIMA na kachestvo prognoza dlya korotkikh naborov dannikh [Investigation of the effect of changing the parameters of the ARIMA model on the quality of the forecast for short data sets]. Vestnik Yuzhno-Uralskogo gosudarstvennogo universiteta. Seriya: Kompyuternie tekhnologii, upravlenie, radioelektronika, 2021, vol. 3, pp. 36–46 (in Russia).*

15. *Skorobogatikh Y. Y. K voprosu o metodakh nakhozhdeniya otsenok parametrov regressionnikh modelei [On the question of methods for finding estimates of the parameters of regression models] / Y. Y. Skorobogatikh, S. N. Mukhina // Izvestiya Baltiiskoi gosudarstvennoi akademii ribopromislovogo flota: psikhologo-pedagogicheskie nauki. – 2023. – vol. 3(65). pp. 205–212. – DOI 10.46845/519.242071-5331-2023-3-65-205-212.*