УДК 004.9
DOI: 10.26102/2310-6018/2026.62.11.002

Моделирование процесса кластеризации адресов в блокчейне на основе байесовской оптимизации весов эвристик

С.А. Мишин[1]✉, Г.И. Зверев[2]

[1]Центральный филиал Российского государственного университета правосудия имени В.М. Лебедева, Воронеж, Российская Федерация
[2]Воронежский институт Министерства внутренних дел Российской Федерации, Воронеж, Российская Федерация

Резюме. В статье рассматривается задача кластеризации адресов в блокчейне – ключевой метод криптовалютной форензики, используемый при расследовании преступлений с цифровыми активами. Предложен подход к комбинированию трех эвристик кластеризации: множественного входа (Multi-Input Heuristic, MIH), адреса сдачи (Change Address Heuristic, CAH) и временных паттернов (Temporal Pattern Heuristic, TPH). Для комбинирования используется взвешенное голосование, а для подбора весовых коэффициентов и порога объединения применяется байесовская оптимизация с гауссовским процессом в качестве суррогатной модели и функцией сбора ожидаемого улучшения (Expected Improvement). Формализована математическая модель кластеризации с использованием AT-графов, A-графов и E-графов, описаны этапы перехода от графа транзакций к графу сущностей. Проведены три эксперимента на реальных данных из датасета maltemoeser. В первом эксперименте (500 транзакций, 1562 адреса) комбинирование эвристик позволило достичь значения F_1-меры (гармоническое среднее между точностью и полнотой) 0,9284, что соответствует уровню наилучшей отдельной эвристики (CAH: 0,9284). Во втором эксперименте (775 транзакций, 2850 адресов) получен $F_1 = 0{,}8251$, что превышает показатели отдельных эвристик (MIH: 0,8151, CAH: 0,8184, TPH: 0,0838). В третьем эксперименте (990 транзакций, 3445 адресов) получен $F_1 = 0{,}8146$, что также превышает показатели отдельных эвристик. Результаты экспериментов свидетельствуют в пользу наличия синергетического эффекта предлагаемого подхода. Комбинирование в ряде случаев позволяет частично компенсировать недостатки отдельных эвристик и достичь более высокого качества кластеризации. Изменение структуры данных при увеличении объема выборки приводит к перераспределению весов эвристик, что отражает адаптивность предлагаемого метода. Обоснована практическая применимость метода в реальных расследованиях: эталонная разметка требуется только на этапе калибровки, после чего веса фиксируются для применения на новых данных. Проведен сравнительный анализ с альтернативными подходами машинного обучения.

Ключевые слова: кластеризация адресов, блокчейн-аналитика, байесовская оптимизация, эвристика множественного входа, гауссовский процесс, криптовалютная форензика, теория графов, цифровые активы, синергетический эффект эвристик.

Для цитирования: Мишин С.А., Зверев Г.И. Моделирование процесса кластеризации адресов в блокчейне на основе байесовской оптимизации весов эвристик. *Моделирование, оптимизация и информационные технологии.* 2026;14(11). URL: https://moitvivt.ru/ru/journal/article?id=2587 DOI: 10.26102/2310-6018/2026.62.11.002

Modeling the process of address clustering in a blockchain based on Bayesian optimization of heuristic weights

S.A. Mishin[1]✉, G.I. Zverev[2]

[1]Central branch of the Russian State University of Justice named after V.M. Lebedev, Voronezh, the Russian Federation

© Мишин С.А., Зверев Г.И., 2026

[2]*Voronezh Institute of the Ministry of Internal Affairs of the Russian Federation, Voronezh, the Russian Federation*

Abstract. This article examines the problem of blockchain address clustering, a key method in cryptocurrency forensics used in investigating digital asset crimes. An approach combining three clustering heuristics is proposed: Multi-Input Heuristic (MIH), Change Address Heuristic (CAH), and Temporal Pattern Heuristic (TPH). Weighted voting is used for the combination, and Bayesian optimization with a Gaussian process as a surrogate model and an Expected Improvement function is applied to select the weights and the merging threshold. A mathematical clustering model is formalized using AT-graphs, A-graphs, and E-graphs, and the stages of transition from a transaction graph to an entity graph are described. Three experiments are conducted on real data from the maltemoeser dataset. In the first experiment (500 transactions, 1562 addresses), combining the heuristics allowed us to achieve an F_1-measure (harmonic mean between precision and recall) of 0.9284, which corresponds to the level of the best individual heuristic (CAH: 0.9284). In the second experiment (775 transactions, 2850 addresses), $F_1 = 0.8251$ was obtained, which exceeds the performance of individual heuristics (MIH: 0.8151, CAH: 0.8184, TPH: 0.0838). In the third experiment (990 transactions, 3445 addresses), $F_1 = 0.8146$ was obtained, which also exceeds the performance of individual heuristics. The experimental results support the presence of a synergistic effect of the proposed approach. In some cases, combining the heuristics makes it possible to partially compensate for the shortcomings of individual heuristics and achieve higher clustering quality. Changes in the data structure with an increase in the sample size lead to a redistribution of the heuristic weights, which reflects the adaptability of the proposed method. The method's practical applicability to real-world investigations is substantiated: reference labeling is required only during the calibration stage, after which the weights are fixed for application to new data. A comparative analysis with alternative machine learning approaches is conducted.

Keywords: address clustering, blockchain analytics, Bayesian optimization, multi-input heuristic, Gaussian process, cryptocurrency forensics, graph theory, digital assets, synergistic effect of heuristics.

For citation: Mishin S.A., Zverev G.I. Modeling the process of address clustering in a blockchain based on Bayesian optimization of heuristic weights. *Modeling, Optimization and Information Technology.* 2026;14(11). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2587 DOI: 10.26102/2310-6018/2026.62.11.002

Введение

Стремительное развитие криптовалют и децентрализованных финансовых систем поставило перед правоприменительными и правоохранительными органами принципиально новые вызовы. Псевдонимность блокчейн-транзакций, возможность создания неограниченного количества адресов одним пользователем и появление сервисов повышения анонимности создают существенные препятствия для идентификации лиц, совершающих преступления в сфере оборота цифровых активов [1].

Ключевым инструментом преодоления этих препятствий является кластеризация адресов – процесс группировки множества адресов, которые с высокой вероятностью контролируются одним субъектом (владельцем). Как отмечается в научной литературе, кластеризация адресов является краеугольным камнем блокчейн-аналитики и использует эвристики для разбиения множества адресов на кластеры, предположительно принадлежащие одному участнику [2, 3].

В настоящее время в криптовалютной форензике наиболее широко применяются две основные эвристики: эвристика множественного входа (Multi-Input Heuristic, MIH), которая предполагает, что все адреса, одновременно выступающие входами одной транзакции, принадлежат одному владельцу, и эвристика одноразового адреса сдачи (Change Address Heuristic, CAH), основанная на том, что в модели UTXO (Unspent Transaction Output) излишек средств при транзакции возвращается отправителю на

новый адрес [3, 4]. Однако, как показано в работе [2], использование отдельных эвристик имеет ограничения, и комбинирование нескольких эвристик позволяет повысить полноту и точность кластеризации.

Цель настоящего исследования – разработка и обоснование метода комбинирования эвристик кластеризации адресов с оптимизацией их весовых коэффициентов на основе байесовской оптимизации, позволяющего адаптивно настраивать вклад каждой эвристики в зависимости от структуры данных.

Задачи исследования:

1) формализовать математическую модель кластеризации адресов с использованием теории графов;

2) разработать метод комбинирования эвристик с оптимизацией весов на основе байесовской оптимизации;

3) провести экспериментальную валидацию метода на реальных данных;

4) обосновать практическую применимость метода в расследованиях преступлений с цифровыми активами.

Материалы и методы

Модели учета средств в блокчейн-индустрии. В блокчейн-индустрии используются две основные модели учета средств:

– модель UTXO применяется в биткоине и многих других криптовалютах. В этой модели средства не хранятся на счетах в виде баланса, а существуют как множество «монет» (UTXO), каждая из которых представляет собой определенную сумму, закрепленную за конкретным адресом. Пользователь может потратить UTXO только целиком, создавая новую транзакцию, в которой указываются входы (ссылки на ранее полученные UTXO) и выходы (новые UTXO для получателей). При превышении суммы входов над суммой, которую пользователь хочет отправить, разница возвращается отправителю в виде сдачи (change) на новый адрес;

– аккаунт-ориентированная модель используется в Ethereum и совместимых сетях. В этой модели каждый адрес имеет баланс, который изменяется при совершении транзакций, аналогично банковским счетам. Различие моделей учета существенно влияет на методы кластеризации: в UTXO-блокчейнах ключевой является эвристика множественного входа, тогда как в аккаунт-ориентированных системах большее значение приобретает анализ паттернов взаимодействия адресов.

Математическая модель кластеризации адресов.

Определение 1 (AT-граф). AT-графом (Address-Transaction graph) называется двудольный ориентированный граф $G_{AT} = (V, R)$, где $V = A \cup T$ – множество вершин, состоящее из адресов $A = \{a_1, a_2, \dots, a_n\}$ и транзакций $T = \{t_1, t_2, \dots, t_m\}$; $R \subseteq (A \times T) \cup (T \times A)$ – множество дуг, где дуга (a_i, t_j) означает, что адрес a_i является входом транзакции t_j, а дуга (t_j, a_k) – что адрес a_k является выходом транзакции t_j.

Определение 2 (A-граф). A-графом называется ориентированный граф $G_A = (A, R_A)$, где A – множество адресов, $R_A \subseteq A \times A$ – множество дуг между адресами, определяемое как $R_A = \{(a_i, a_j) \in A \times A \mid \exists t \in T: (a_i, t) \in R \text{ и } (t, a_j) \in R\}$, т. е. дуга (a_i, a_j) существует, если существует транзакция $t \in T$ такая, что a_i является входом *t*, а a_j – выходом *t*.

Определение 3 (кластер адресов). Кластером $C_q \subseteq A$ $(q = 1, \dots, k)$ называется максимальное по включению множество адресов, которые, согласно применяемым эвристикам, с высокой вероятностью контролируются одним владельцем. Множество всех кластеров образует разбиение $C = \{C_1, C_2, \dots, C_k\}$ множества адресов *A*.

Определение 4 (E-граф). Результатом кластеризации является переход от AT-графа к агрегированному E-графу (Entity Graph) $G_E = (\mathcal{E}, R_\mathcal{E}, \rho)$, где множество сущностей $\mathcal{E} = \{e_1, e_2, \dots, e_k\}$ находится во взаимно-однозначном соответствии с множеством кластеров $C = \{C_1, C_2, \dots, C_k\}$ (каждый элемент e_q интерпретируется как реальный владелец, контролирующий соответствующий кластер адресов C_q); $R_\mathcal{E} \subseteq \mathcal{E} \times \mathcal{E}$ – множество дуг между сущностями (дуга $(e_x, e_y) \in R_\mathcal{E}$ существует, если между адресами, входящими в кластеры e_x и e_y, были транзакции); $\rho: R_\mathcal{E} \to \mathbb{R}$ – весовая функция дуг, отражающая агрегированные параметры транзакций (суммарный объем переведенных средств, количество транзакций и др.).

Эвристики кластеризации адресов. Эвристика множественного входа (MIH). Если два или более адреса выступают в качестве входов одной транзакции, то все эти адреса с высокой вероятностью контролируются одним владельцем (пользователем). *Обоснование*: для создания транзакции с несколькими входами необходимо располагать частными ключами, соответствующими каждому из входных адресов. Эмпирические исследования показывают, что одна только эвристика множественного входа позволяет идентифицировать более 69 % адресов в биткоин-блокчейне [4].

Эвристика одноразового адреса сдачи (CAH). В модели UTXO каждая транзакция может иметь множество входов и выходов. Если сумма входов превышает сумму, которую пользователь намерен отправить получателю, избыток возвращается отправителю в виде отдельного выходного UTXO, направляемого на адрес сдачи. Критерии идентификации адреса сдачи включают: время создания адреса (новый адрес, не имеющий истории), правило возврата к отправителю (выходной адрес принадлежит тому же кластеру, что и входные) [4, 5].

Эвристика временных паттернов (TPH) основана на анализе временных интервалов между транзакциями. В отличие от эвристики множественного входа (*MIH*) и эвристики адреса сдачи (*CAH*), которые опираются на структурные признаки транзакции (состав входов и выходов), *TPH* использует метаданные о времени подтверждения транзакций в блоках блокчейна. Логика эвристики основана на предположении, что адреса, принадлежащие одному владельцу, могут демонстрировать устойчивые временные паттерны взаимодействия. Регулярные переводы между одними и теми же адресами в пределах короткого промежутка времени могут свидетельствовать о принадлежности одному лицу (владельцу) [4]. Данное предположение основано на поведенческих особенностях пользователей криптовалютных кошельков. Операции консолидации UTXO, подготовки к обмену, вывода средств на биржу, регулярные выплаты, как правило, выполняются в рамках ограниченных временных интервалов и могут повторяться с определенной периодичностью. Формально, для пары адресов (a_i, a_j) фиксируется множество временных меток $\mathcal{T}_{ij} = \{t_1, t_2, \dots, t_z\}$, соответствующих транзакциям, в которых участвовали оба адреса. Индикаторная функция принимает значение 1, если $\mathbb{I}_{TPH}(a_i, a_j) = 1 \Leftrightarrow |\mathcal{T}_{ij}| \geq \tau \; и \; \exists t_p, t_{p+1} \in \mathcal{T}_{ij}: |t_{p+1} - t_p| < \Delta T$, где τ – минимальное количество взаимодействий (в работе $\tau = 2$), $\Delta\mathbb{T}$ – максимальный временной интервал между последовательными взаимодействиями (в работе $\Delta\mathbb{T} = 7$ суток). Порог $\tau = 2$ выбран как минимально достаточный для исключения случайных единичных совпадений. Два адреса, единожды взаимодействовавшие в одной транзакции, могут оказаться несвязанными. Однако повторное взаимодействие в течение ограниченного временного окна существенно повышает вероятность принадлежности одному владельцу. Окно $\Delta\mathbb{T} = 7$ суток отражает типичный период планирования операций при перемещении средств между собственными адресами. Эмпирически установлено, что операции консолидации UTXO, подготовки к обмену и вывода на

биржу, как правило, выполняются в течение нескольких часов или дней, но редко превышают недельный интервал.

Для примера, рассмотрим *алгоритм кластеризации на основе эвристики множественного входа*.

Вход: множество транзакций $T = \{t_1, t_2, \dots, t_m\}$, каждая из которых содержит в том числе множество входных адресов $inputs(t) \subseteq A$. Выходные адреса в этом алгоритме не участвуют так как эвристика множественного входа основана на совместном использовании адресов во входах транзакции.

Выход: разбиение множества адресов *A* на кластеры $C = \{C_1, C_2, \dots, C_k\}$.

Кластеризация основана на принципе транзитивности: если адрес a_1 объединен с a_2, а a_2 объединен с a_3, то все три адреса принадлежат одному кластеру. Формально, пусть $\sim$ – отношение эквивалентности на множестве адресов *A*, где $a_i \sim a_j$ означает, что эвристика связывает a_i и a_j. Тогда кластер C_q $(q = 1, \dots, k)$ определяется как класс эквивалентности: $C_q = \{a' \in A: a' \sim a\}$, где $a \in C_q$.

Для эффективной реализации кластеризации используется структура данных *Union-Find* (Disjoint Set Union, DSU), поддерживающая разбиение множества элементов на непересекающиеся подмножества и выполняющая две основные операции:

1) *Find(x)* – определить, к какому множеству (кластеру) принадлежит элемент. Обычно возвращает идентификатор (корень) представителя множества.

2) *Union(x, y)* – объединить множества, содержащие элементы *x* и *y*, в одно.

Шаг 1. Инициализация. Для каждого адреса $a \in A$ создать отдельный кластер $C_a = \{a\}$, где *a* – адрес, а не индекс кластера. Инициализировать структуру *Union-Find*.

Шаг 2. Объединение по эвристике множественного входа. Для каждой транзакции $t \in T$:

$$\text{Если } |inputs(t)| \geq 2, \text{то } \forall a_i, a_j \in inputs(t): Union(a_i, a_j). \quad (1)$$

Шаг 3. Транзитивное замыкание. Повторять *шаг 2* до тех пор, пока не будет выполнено ни одного нового объединения.

Шаг 4. Формирование результата. Для каждого корневого элемента в структуре *Union-Find* сформировать кластер как множество всех адресов, имеющих данный корень.

После завершения работы алгоритма полученные кластеры нумеруются как $C_1, C_2, \dots, C_k$, где *k* – итоговое число кластеров.

Предлагаемый метод: комбинирование эвристик с оптимизацией весов. В отличие от использования отдельных эвристик, предлагаемый метод основан на комбинировании нескольких эвристик с оптимизацией их весовых коэффициентов. Каждая эвристика представляет собой один «голос», но голоса имеют разный вес ω_h, отражающий их надежность.

Формальная модель комбинирования. Общая оценка связи $S(a_i, a_j)$ для пары адресов вычисляется как взвешенная сумма показаний всех эвристик:

$$S(a_i, a_j) = \sum_{h \in \mathcal{H}} \omega_h \cdot \mathbb{I}[\, a_i \sim_h a_j], \quad (2)$$

где $\mathcal{H} = \{h_1, h_2, \dots, h_l\}$ – множество используемых эвристик; *l* – количество эвристик; $\omega_h \in [0, 1]$ – весовой коэффициент, отражающий значимость эвристики *h;* $\mathbb{I}[a_i \sim_h a_j]$ – индикаторная функция, возвращающая 1, если эвристика *h* связывает адреса a_i и a_j, и 0 в противном случае.

Значение индикаторной функции определяется внутренней логикой конкретной эвристики. Например, для эвристики множественного входа $\mathbb{I}_{MIH}(a_i, a_j) = 1$, если существует хотя бы одна транзакция, где a_i и a_j одновременно являются входами. Для

эвристики адреса сдачи $\mathbb{I}_{CAH}(a_i, a_j) = 1$, если в транзакции, где a_i является входом, среди выходов есть адрес a_j, идентифицированный как адрес сдачи.

Правило принятия решения. Адреса a_i и a_j объединяются в один кластер, если $S(a_i, a_j) \geq \theta$, где $\theta \in [0, \sum_{h\in\mathcal{H}} \omega_h]$ – пороговое значение.

Алгоритм подбора весовых коэффициентов на основе байесовской оптимизации.

Постановка задачи. Пусть имеется множество эвристик $\mathcal{H} = \{h_1, h_2, \dots, h_l\}$, каждая из которых определяет бинарное отношение на множестве адресов *A*. Требуется найти вектор весов $w = \{\omega_1, \omega_2, \dots, \omega_l\} \in [0,1]^l$ и пороговое значение $\theta \in [0, \sum_{h\in\mathcal{H}} \omega_h]$, максимизирующие качество кластеризации на валидационном множестве.

Метрики качества. Для оценки качества кластеризации используются три метрики: точность (*Precision*), полнота (*Recall*) и *F1-мера*. В качестве целевой функции *F1(w, θ)*, подлежащей максимизации в процессе байесовской оптимизации, используется *F1-мера*, поскольку она является сбалансированным показателем, учитывающим как точность, так и полноту кластеризации. *F1-мера* вычисляется как гармоническое среднее между точностью и полнотой:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad (3)$$

где TP – количество пар адресов, которые правильно объединены в один кластер; FP – количество пар адресов, которые неправильно объединены в один кластер; FN – количество пар адресов, которые не были объединены, но должны были быть объединены.

Шаг 1. Инициализация весов. Начальные веса для каждой эвристики вычисляются на основе ее *F1-меры* на валидационном множестве:

$$\omega_h^{(0)} = 2 \cdot \frac{Precision_h \cdot Recall_h}{Precision_h + Recall_h} = F_1(h), \quad (4)$$

где $Precision_h$ и $Recall_h$ – метрики качества для случая, когда кластеризация выполняется только на основе эвристики h.

Шаг 2. Байесовская оптимизация. Для поиска оптимальных значений w^* и θ^* используется байесовская оптимизация, состоящая из двух основных компонентов [6, 7]:

1) априорная модель (суррогатная функция) – гауссовский процесс (Gaussian Process, GP), который аппроксимирует неизвестную целевую функцию *F1(w, θ);*

2) функция сбора – определяет, в какой точке пространства параметров следует вычислить целевую функцию на следующей итерации. В качестве функции сбора используется ожидаемое улучшение (Expected Improvement, EI).

Гауссовский процесс (5) представляет собой распределение над функциями, полностью задаваемое средней функцией $\mu(x)$ и ковариационной функцией $\kappa(x, x')$, в качестве которой в работе используется ядро Матерн с параметром гладкости $\nu = 5/2$. Значение $\nu = 5/2$ обеспечивает дважды дифференцируемую аппроксимацию целевой функции *F1(w, θ)*, что соответствует предположению о ее гладкости.

$$f(x) \sim \mathcal{GP}\big(\mu(x), \kappa(x, x')\big), \quad (5)$$

где $x = (\omega_1, \omega_2, \omega_3, \theta)$ – вектор оптимизируемых параметров, а x' – другая точка в пространстве оптимизируемых параметров.

В многомерном случае ($d = 4$ – три веса эвристик и порог) с учетом $\nu = 5/2$ ковариационная функция имеет вид:

$$\kappa(x, x') = \sigma^2 \cdot \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \cdot \exp\left(-\frac{\sqrt{5}r}{\ell}\right), \quad (6)$$

где $r = \sqrt{\sum_{j=1}^{4} \frac{(x_j - x'_j)^2}{\ell_j^2}}$; σ^2 – амплитуда сигнала (*F_1-меры*); ℓ_j – масштаб длины для *j*-го измерения, определяющий скорость убывания корреляции с расстоянием *r* по данному параметру, *r* – евклидово расстояние между точками *x* и x'в пространстве параметров.

Гиперпараметры гауссовского процесса ($\mathcal{GP}$), включая масштабы длины ℓ_j для каждого из четырех оптимизируемых параметров, определяются автоматически в процессе максимизации маргинального правдоподобия. Это обеспечивает адаптивность модели к структуре данных и позволяет $\mathcal{GP}$ автоматически выявлять относительную значимость каждого параметра. Выбор гауссовского процесса в качестве суррогатной функции обоснован следующими факторами:

1) естественная количественная оценка неопределенности. Гауссовский процесс предоставляет не только точечную оценку значения функции $\mu(x)$, но и меру уверенности в этой оценке $\sigma^2(x)$;

2) непараметрический характер. Гауссовский процесс не требует априорного задания параметрической формы функции (например, полинома определенной степени). Вместо этого он задает распределение на пространстве всех функций, согласованных с наблюдаемыми данными. Это особенно важно в задачах оптимизации «черного ящика», где аналитический вид целевой функции неизвестен;

3) непрерывность пространства поиска. Пространство параметров (веса ω_1, ω_2, $\omega_3 \in [0, 1]$ и порог $\theta \in [0, 3]$) является непрерывным. Гауссовские процессы по своей природе предназначены для работы с непрерывными входными пространствами и предполагают гладкость целевой функции, что соответствует характеру *F_1-меры* от весов эвристик;

4) малая размерность пространства поиска. Размерность пространства параметров равна 4, что значительно ниже порога (d > 20), при котором гауссовские процессы начинают терять эффективность.

Функция ожидаемого улучшения (EI). Для гауссовского процесса функция *EI* имеет аналитическое выражение:

$$EI(x) = \begin{cases} (\mu(x) - F_{max})\Phi(Z) + \sigma(x)\phi(Z), \sigma(x) > 0 \\ 0, \qquad \sigma(x) = 0 \end{cases}, \tag{7}$$

где $\mu(x)$ – прогностическое среднее гауссовского процесса в точке *x*; $\sigma(x)$ – прогностическое стандартное отклонение; F_{max} – текущее наилучшее значение целевой функции; $Z = \frac{\mu(x) - F_{max}}{\sigma(x)}$; $\Phi(Z)$ – функция стандартного нормального распределения; $\phi(Z)$ – функция плотности стандартного нормального распределения.

Шаг 3. Кросс-валидация. Для оценки обобщающей способности полученных параметров используется *k*-кратная кросс-валидация. Верифицированное множество данных разделяется на *k* непересекающихся подмножеств (фолдов). Для каждого фолда выполняется *шаг 2* на обучающей выборке, и вычисляется *F_1-мера* на валидационной выборке. Итоговые веса w^* и порог θ^* определяются как средние значения по всем фолдам.

В работе процедура кросс-валидации была реализована следующим образом:

– множество кластеров *ground truth* случайным образом разбивалось на 5 непересекающихся фолдов примерно равного размера;

– для каждого фолда q формировалась обучающая выборка (все пары адресов, принадлежащих кластерам вне фолда q), валидационная выборка (все пары адресов, принадлежащих кластерам фолда q), а также для обеспечения независимости начальных точек оптимизации между фолдами использовалось свое значение random_state;

– на обучающей выборке запускалась байесовская оптимизация (40 вызовов целевой функции, 15 начальных точек), в результате чего определялись веса ω_1, ω_2, ω_3 и порог θ;

– полученные параметры применялись к полному множеству адресов, и вычислялась *F_1-мера* на парах адресов, входящих в валидационный фолд; дополнительно вычислялась *F_1-мера* на обучающих фолдах, что позволило оценить разницу между качеством на обучающей и валидационной выборках (признак переобучения);

– по результатам 5 фолдов вычислялись среднее значение F_1, стандартное отклонение и 95 % доверительный интервал с использованием распределения Стьюдента.

Помимо кросс-валидации, для количественной оценки статистической неопределенности точечной оценки F_1 применялся метод бутстрэпа (bootstrap). Из множества адресов, полученных в ходе эксперимента, формировалось 1000 ресэмплированных выборок того же размера с возвращением. Для каждой выборки вычислялась *F_1-мера* при фиксированных оптимальных весах и пороге. По полученному эмпирическому распределению F_1 вычислялись среднее значение, стандартное отклонение и 95 % доверительный интервал. Бутстрэп-интервал, характеризуя устойчивость оценки F_1 по отношению к случайной вариации состава адресов, позволяет определить в каких пределах точечная оценка F_1 может изменяться при другой реализации выборки того же размера из той же генеральной совокупности.

Для оценки устойчивости процедуры байесовской оптимизации было выполнено 10 независимых запусков с различными значениями *random_state* (параметр, задающий начальное значение генератора псевдослучайных чисел, используемого алгоритмом байесовской оптимизации). В каждом запуске использовались одинаковые пространства поиска и гиперпараметры гауссовского процесса, но разные начальные точки, генерируемые методом латинского гиперкуба. По результатам 10 запусков вычислялись среднее значение, стандартное отклонение и 95 % доверительный интервал для каждого веса, порога и итоговой *F_1-меры*. Коэффициент вариации (*CV*) использовался как безразмерная мера устойчивости. Чем меньше значение коэффициента вариации, тем более стабильно метод находит решение. Значения $CV < 10$ % интерпретируются как высокая устойчивость, 10–25 % – как умеренная, $CV > 25$ % – как низкая.

Результаты

Для демонстрации адаптивности предлагаемого метода, его валидации на различных объемах данных, а также подтверждения синергетического эффекта проведены три эксперимента на реальных данных. Эксперименты проводились в среде Google Colab с использованием языка Python. В качестве эталонной разметки (*ground truth*) для оценки качества кластеризации использовался датасет *change-ground-truth.csv* (объем 2,2 ГБ), предоставленный *maltemoeser* и содержащий информацию о change-адресах в биткоин-блокчейне. Загрузка сырых (неразмеченных) данных о транзакциях, выполнялась через Blockstream API (*https://blockstream.info/api/tx/{hash}*). Для загрузки данных, построения эвристик, кластеризации и оптимизации использовались следующие библиотеки: *scikit-optimize* – реализация байесовской оптимизации (функция *gp_minimize*); *gdown* – загрузка *ground truth* датасета с Google Drive; *pandas*, *numpy* – обработка и анализ табличных данных; *requests* – выполнение HTTP-запросов к API биткоин-блокчейна (Blockstream); *tqdm* – отображение прогресса выполнения длительных операций; *matplotlib* – визуализация результатов (график эволюции F_{max}); *scikit-learn* (GaussianProcessRegressor) – реализация гауссовского процесса

(используется внутри scikit-optimize); *scipy.stats* – статистические вычисления для функции *Expected Improvement*. Код организован в виде последовательных блоков, реализующих загрузку данных, построение *ground truth* кластеров, вычисление эвристик, оценку отдельных эвристик, байесовскую оптимизацию и визуализацию.

Формирование экспериментальной выборки выполнялось в несколько этапов. На первом этапе из полного датасета *change-ground-truth.csv* были прочитаны первые 140000 записей. На втором этапе из 140000 записей была сформирована случайная выборка объемом 35000 транзакций с фиксированным значением random_state=42 для обеспечения воспроизводимости. Из этой выборки были извлечены уникальные хэши транзакций с последующей сортировкой для обеспечения детерминированного порядка при загрузке. На третьем этапе для каждого хэша выполнялась загрузка «сырых» данных о транзакции через Blockstream API. Из 2500 запрошенных транзакций удалось успешно получить данные только для 500 (эксперимент 1) и 775 (эксперимент 2) транзакций. Низкий процент успешной загрузки (20 % и 31 % соответственно) обусловлен тем, что часть транзакций не индексируется публичными API, а также существующими лимитами на частоту запросов. В третьем эксперименте из 1500 запрошенных транзакций (уменьшение числа запрашиваемых транзакций обусловлено попыткой преодоления лимитов на частоту запросов) успешно загрузились 990 транзакций. Эксперимент 2 является расширением эксперимента 1 за счет дополнительно успешно загруженных транзакций из того же списка из 2500 хэшей. Эксперимент 3 проведен на усеченном списке из 1500 хэшей, что обеспечило более высокую долю успешных загрузок (66 %). На четвертом этапе формировались эталонные кластеры (ground truth). Для каждой успешно загруженной транзакции все входные адреса объединялись с адресом сдачи, указанным в датасете, образуя один кластер. В результате эксперимента 1 было получено 499 кластеров, содержащих 1562 адреса; в эксперименте 2 получено 775 кластеров, содержащих 2850 адресов, в эксперименте 3 получено 990 кластеров, содержащих 3445 адресов. На пятом этапе для всех пар адресов, участвовавших в загруженных транзакциях, вычислялись три эвристики: MIH, CAH и TPH. Общее количество пар, для которых сработала хотя бы одна эвристика, составило 2470 (эксперимент 1), 16770 (эксперимент 2) и 18264 (эксперимент 3).

В Таблице 1 представлено сравнение ключевых параметров и результатов трех проведенных экспериментов.

Таблица 1 – Сравнение результатов экспериментов 1, 2 и 3
Table 1 – Comparison of the results of experiments 1, 2 and 3

Параметр	**Эксперимент 1 (500 транзакций)**	**Эксперимент 2 (775 транзакций)**	**Эксперимент 3 (990 транзакций)**
Количество адресов	1562	2850	3445
Количество кластеров (*ground truth)*	499	775	990
MIH F_1 (отдельно)	0,7164	0,8151	0,6036
CAH F_1 (отдельно)	0,9284	0,8184	0,6249
TPH F_1 (отдельно)	0,2269	0,0838	0,1000
ω_1 (MIH) ω_2 (CAH) ω_3 (TPH)	0,107 0,660 0,563	0,841 0,531 0,142	0,639 0,540 0,146
θ (порог)	0,572	0,673	0,339
F_1 *Precision* *Recall*	0,9284 0,9080 0,9497	0,8251 0,7346 0,9409	0,8146 0,7115 0,9527

Данные изменения отражают адаптацию метода к структуре данных. С ростом объема выборки MIH получает больше подтверждений и становится более надежной, тогда как CAH, напротив, дает больше ложных срабатываний. TPH оказывается наименее стабильной эвристикой. Увеличение порога свидетельствует о необходимости более строгой фильтрации связей при большем объеме данных. Во всех экспериментах комбинирование эвристик позволило достичь высоких значений F_1. В эксперименте 1 F_1 соответствует уровню наилучшей отдельной эвристики, а в экспериментах 2 и 3 превысило показатели отдельных эвристик, что может рассматриваться как подтверждение гипотезы о целесообразности объединения эвристик. На Рисунке 1 представлен график эволюции кумулятивного максимума *F_1-меры* (т. е. изменения F_{max} в процессе байесовской оптимизации) в эксперименте 3.

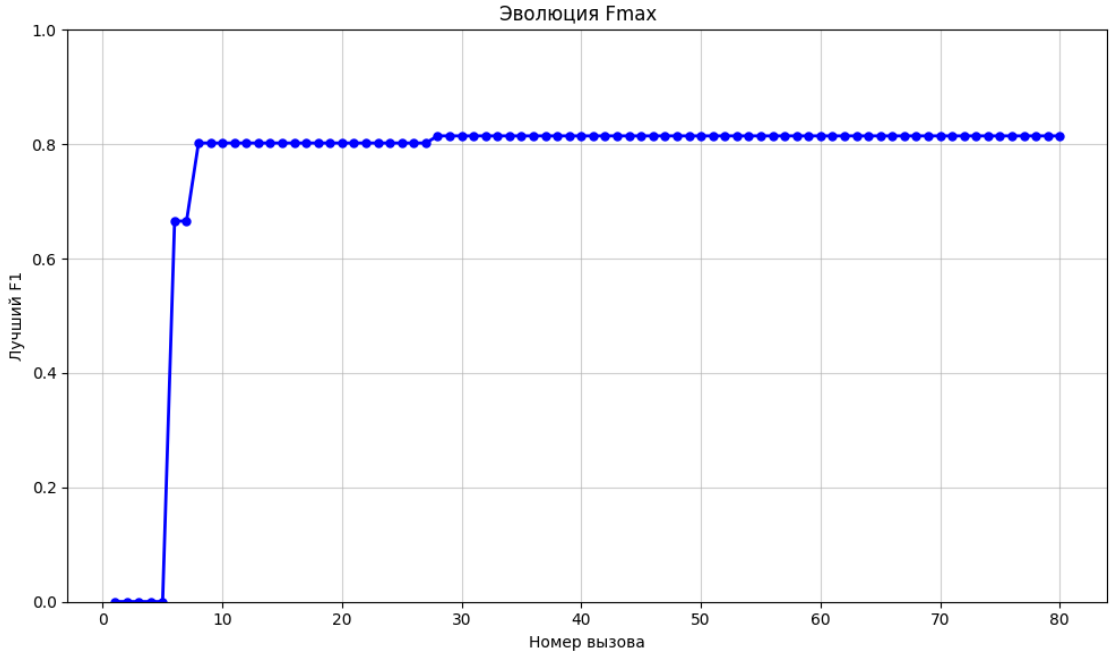


Рисунок 1 – Эволюция F_{max} в процессе байесовской оптимизации
Figure 1 – Evolution of F_{max} during Bayesian optimization

Точечная оценка *F_1-меры* в эксперименте 3 при оптимальных параметрах ($\omega_1 = 0{,}639$; $\omega_2 = 0{,}540$; $\omega_3 = 0{,}146$; $\theta = 0{,}339$) составила 0,8146. Для оценки статистической устойчивости этого результата были применены три независимых метода: бутстрэп (bootstrap), 5-кратная кросс-валидация на уровне кластеров и повторная оптимизация с различными начальными точками. Результаты представлены в Таблице 2.

Таблица 2 – Оценка статистической устойчивости *F_1-меры* (точечная оценка $F_1 = 0{,}8146$)
Table 2 – Assessment of statistical stability of *F_1-measure* (point estimate $F_1 = 0.8146$)

Метод оценки	**Среднее F_1**	**Стандартное отклонение**	**95 % доверительный интервал**	**Коэффициент вариации, %**
Бутстрэп (n = 1000)	0,8289	0,0180	[0,7928; 0,8647]	2,17
Кросс-валидация (k = 5)	0,9007	0,0704	[0,8133; 0,9881]	7,82
Повторная оптимизация (n = 10)	0,8028	0,0190	[0,7893; 0,8164]	2,36

Bootstrap-распределение *F1-меры* (эксперимент 3) показано на Рисунке 2.

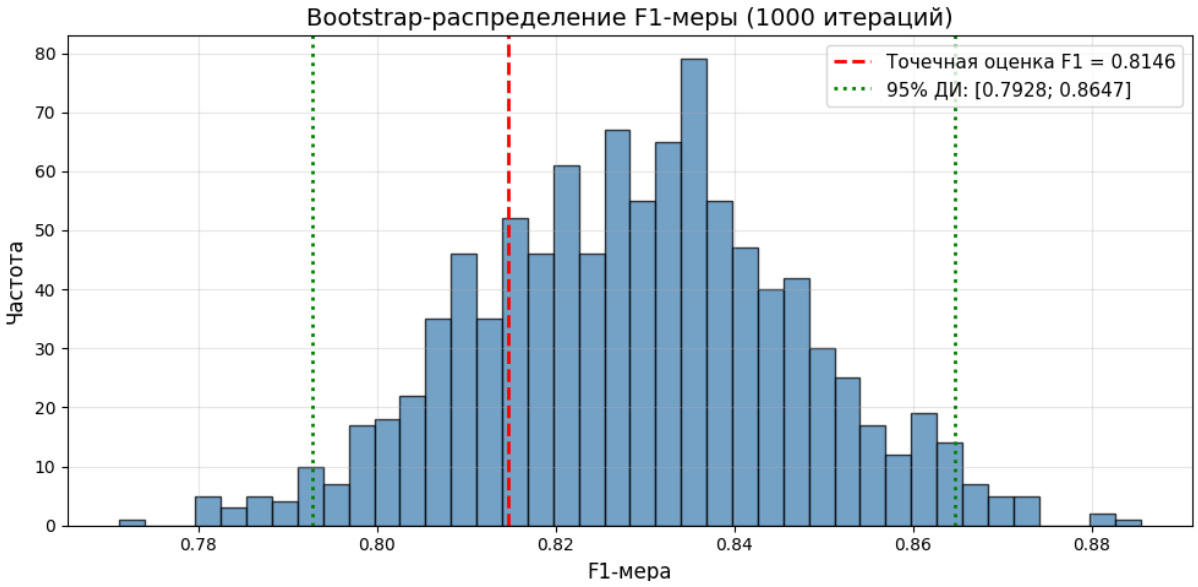


Рисунок 2 – Bootstrap-распределение *F1-меры*
Figure 2 – Bootstrap distribution of the *F1-measure*

Важно отметить, что полученные веса *не являются универсальными* и зависят от структуры конкретного набора данных. Это означает, что для практического применения метода необходима калибровка на репрезентативном эталонном наборе данных (*ground truth*), после чего полученные веса могут быть зафиксированы и использованы в реальных расследованиях без дополнительной настройки.

Обсуждение

Интерпретация результатов:

– Доминирование MIH и CAH. Обе эвристики показывают высокие значения F_1, что подтверждает их надежность как инструментов криптовалютной форензики. Оптимизация отдает предпочтение одной из них в зависимости от структуры данных (в экспериментах 1 и 3 – MIH, в эксперименте 2 – CAH). Доминирование MIH объясняется более строгим криптографическим обоснованием этой эвристики, так как совместное использование входов требует наличия нескольких приватных ключей. Доминирование CAH, напротив, наблюдается при наличии большого числа транзакций с одним входом и двумя выходами, где change-адрес идентифицируется с высокой точностью.

– Низкая эффективность TPH. Эвристика временных паттернов получила минимальный вес. Это указывает на то, что в наборе данных временные паттерны не являются информативными, что может быть связано с недостаточным количеством повторяющихся взаимодействий между адресами.

– Синергетический эффект. Полученные результаты свидетельствуют в пользу наличия синергетического эффекта от комбинирования эвристик. Однако следует отметить, что синергетический эффект не является универсальным и в значительной степени определяется структурой данных. Комбинирование эвристик повышает устойчивость результата к вариациям структуры данных, но не гарантирует монотонного улучшения F_1 относительно каждой отдельной эвристики. При наличии доминирующей эвристики (как CAH в эксперименте 1) комбинирование не дает существенного прироста; при сопоставимой информативности нескольких эвристик (как MIH и CAH в

экспериментах 2 и 3) комбинирование позволяет получить более сбалансированный результат по точности и полноте.

– высокий *Recall* (0,9527) при умеренной *Precision* (0,7115) указывает на то, что метод успешно обнаруживает большинство реальных связей, допуская некоторое количество ложных объединений.

Сравнение с альтернативными подходами. Классические алгоритмы кластеризации (k-means, DBSCAN) оперируют объектами в признаковом пространстве, тогда как адреса блокчейна образуют графовую структуру без естественного метрического пространства. В работе [8] показано, что distance-based clustering methods, включая k-means и DBSCAN, не дают удовлетворительных результатов на сетях биткоин-транзакций. Значения модулярности для DBSCAN оказываются на уровне случайного разбиения.

Графовые нейронные сети (GNN) являются перспективным направлением для классификации адресов, однако они решают задачу классификации (отнесение адресов к категориям), а не кластеризации (группировки по владельцам). Обучение GNN требует размеченного набора данных, что ограничивает их применение в условиях расследования, где эталонная разметка отсутствует [9, 10]. Тем не менее, в литературе отмечены гибридные подходы, где GNN используются для классификации change-адресов как промежуточного этапа кластеризации [10]. Предлагаемый метод, сохраняя преимущества эвристического подхода (интерпретируемость, отсутствие необходимости в размеченных данных для применения), может быть интегрирован с такими подходами. Выход GNN (вероятностная оценка change-адреса) может выступать в качестве дополнительной эвристики в комбинированной модели.

Наиболее близкий класс методов применяет эвристику *MIH* или *CAH* по отдельности, без комбинирования. Экспериментальное сравнение (Таблица 1) показывает, что предлагаемый метод не уступает лучшей отдельной эвристике в эксперименте 1 и превосходит каждую отдельную эвристику в экспериментах 2 и 3, тем самым подтверждает целесообразность комбинирования.

Применимость метода в реальных расследованиях. *Ground truth* (эталонная разметка) требуется только на этапе калибровки метода, а не в каждом конкретном расследовании. После калибровки веса фиксируются и используются для кластеризации новых, неразмеченных данных. Это делает метод практически применимым в условиях реальной правоприменительной деятельности. Вариативность весов при изменении объема данных не является недостатком, а отражает адаптивность метода. Метод может быть адаптирован для различных блокчейнов: для UTXO-модели (Bitcoin) – с использованием классических эвристик; для аккаунт-ориентированной модели (Ethereum) – с введением адаптированных эвристик.

Предлагаемый метод обладает рядом ограничений, которые необходимо учитывать при его применении в правоприменительной деятельности:

– зависимость от качества эталонной разметки. Веса эвристик, полученные в результате байесовской оптимизации, определяются структурой калибровочного набора данных. Если *ground truth* содержит систематические ошибки (например, ложные срабатывания MIH, использованные при его построении), то происходит смещение весов. Для минимизации риска рекомендуется использовать несколько независимых источников разметки и проводить перекрёстную валидацию;

– отсутствие интерпретируемости отдельных решений. Хотя метод в целом интерпретируем (веса эвристик имеют содержательный смысл), конкретное решение об объединении пары адресов является результатом порогового правила и не всегда может быть объяснено без обращения к исходным транзакциям. В судебной практике это

требует дополнительной верификации кластеров (например, через запросы к биржам или анализ открытых источников);

– уязвимость к CoinJoin-транзакциям и анонимизирующим сервисам. Проведенные эксперименты показали, что предложенный метод комбинирования эвристик демонстрирует устойчиво высокие результаты (F_1 = 0,81–0,83) на данных с низкой долей CoinJoin-транзакций и крупных консолидаций, однако его эффективность снижается при увеличении доли «шумных» транзакций. Данное ограничение не является специфическим для предложенного метода. Все эвристические подходы к кластеризации блокчейн-адресов в той или иной степени подвержены влиянию CoinJoin, консолидаций и пылевых атак. Практическое применение метода требует предварительной фильтрации аномальных транзакций, а также периодической перекалибровки весовых коэффициентов на репрезентативных выборках;

– требования к верификации результатов. Результаты кластеризации не могут использоваться как единственное доказательство в суде. Они должны сопровождаться документацией о происхождении данных и методике калибровки, экспертным заключением о допустимости примененных эвристик в конкретном деле.

Заключение

В статье предложен метод комбинирования эвристик кластеризации адресов в блокчейне с оптимизацией весовых коэффициентов на основе байесовской оптимизации. Проведена формализация математической модели кластеризации с использованием AT-графов, A-графов и E-графов. Разработан алгоритм подбора весов эвристик и порога объединения на основе байесовской оптимизации с гауссовским процессом и функцией сбора Expected Improvement. Экспериментальная валидация на реальных данных из датасета maltemoeser (990 транзакций, 3445 адресов) показала, что комбинирование эвристик позволяет достичь F_1 = 0,8146, что превышает показатели отдельных эвристик и демонстрирует синергетический эффект предлагаемого подхода.

Перспективные направления дальнейших исследований включают:

– интеграцию предложенного метода с графовыми нейронными сетями для гибридных подходов к кластеризации;

– разработку процедуры динамической адаптации весов в процессе работы;

– создание специализированных профилей калибровки для различных типов расследований;

– адаптацию метода для блокчейнов с аккаунт-ориентированной моделью.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Тисен О.Н. Механизм совершения преступлений с использованием криптовалют: по результатам анализа следственной и судебной практики. *Российский следователь*. 2024;(5):7–11. https://doi.org/10.18572/1812-3783-2024-5-7-11
Tisen O.N. The mechanism of committing crimes involving the use of cryptocurrencies: based on the results of an analysis of the investigative and judicial practice. *Russian Investigator*. 2024;(5):7–11. (In Russ.). https://doi.org/10.18572/1812-3783-2024-5-7-11
2. Harrigan M., Fretter Ch. The Unreasonable Effectiveness of Address Clustering. In: *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 18–21 July 2016, Toulouse, France*. IEEE; 2016. P. 368–373. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0071

3. Möser M., Narayanan A. Resurrecting Address Clustering in Bitcoin. In: *Financial Cryptography and Data Security: 26th International Conference (FC 2022), 02–06 May 2022, Grenada*. Cham: Springer; 2022. P. 386–403. https://doi.org/10.1007/978-3-031-18283-9_19
4. Caringella M., Violante F., De Lucci F., Galantucci S., Costantini M. BACH: A Tool for Analyzing Blockchain Transactions Using Address Clustering Heuristics. *Information*. 2024;15(10):589. https://doi.org/10.3390/info15100589
5. Liu F., Li Zh., Jia K., et al. Bitcoin Address Clustering Based on Change Address Improvement. *IEEE Transactions on Computational Social Systems.* 2024;11(6):8094–8105. https://doi.org/10.1109/TCSS.2023.3239031
6. Rasmussen C.E., Williams Ch.K.I. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press; 2005. 272 p. https://doi.org/10.7551/mitpress/3206.001.0001
7. Shahriari B., Swersky K., Wang Z., Adams R.P., de Freitas N. Taking the human out of the loop: a review of Bayesian optimization. *Proceedings of the IEEE*. 2016;104(1):148–175. https://doi.org/10.1109/JPROC.2015.2494218
8. Kamuhanda D., Cui M., Tessone C.J. Illegal Community Detection in Bitcoin Transaction Networks. *Entropy*. 2023;25(7):1069. https://doi.org/10.3390/e25071069
9. Huang Zh., Huang Y., Qian P., Chen J., He Q. Demystifying Bitcoin address behavior via graph neural networks. In: *2023 IEEE 39th International Conference on Data Engineering (ICDE), 03–07 April 2023, Anaheim, CA, USA*. IEEE; 2023. P. 1747–1760. https://doi.org/10.1109/ICDE55515.2023.00137
10. Schnoering H., Vazirgiannis M. Bitcoin research with a transaction graph dataset. *Scientific Data.* 2025;12(1):404. https://doi.org/10.1038/s41597-025-04684-8

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Мишин Сергей Александрович, кандидат технических наук, доцент, доцент кафедры правовой информатики, информационного права и естественнонаучных дисциплин, Центральный филиал Российского государственного университета правосудия имени В.М. Лебедева, Воронеж, Российская Федерация.
e-mail: samishin@bk.ru
ORCID: 0000-0002-8321-2759

Sergey A. Mishin, Candidate of Engineering Sciences, Docent, Associate Professor at the Department of Legal Informatics, Information Law and Natural Sciences, Central branch of the Russian State University of Justice named after V.M. Lebedev, Voronezh, the Russian Federation.

Зверев Георгий Игоревич, кандидат технических наук, заместитель начальника кафедры автоматизированных информационных систем органов внутренних дел, Воронежский институт Министерства внутренних дел Российской Федерации, Воронеж, Российская Федерация.
e-mail: georgiyzverev@gmail.com
ORCID: 0000-0001-7323-2425

Georgy I. Zverev, Candidate of Engineering Sciences, Deputy Head of the Department of Automated Information Systems of Internal Affairs Bodies, Voronezh Institute of the Ministry of Internal Affairs of the Russian Federation, Voronezh, the Russian Federation.

Статья поступила в редакцию 03.08.2026; одобрена после рецензирования 19.10.2026; принята к публикации 03.11.2026.

The article was submitted 03.08.2026; approved after reviewing 19.10.2026; accepted for publication 03.11.2026.