

# АГРЕГИРОВАНИЕ ДЕФЕКТОВ ЖЕЛЕЗНОДОРОЖНОГО ПУТИ С УЧЕТОМ ИХ ФИЗИЧЕСКОЙ ПРИРОДЫ

Владова А.Ю.

Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия  
avladova@ipu.ru

*Аннотация.* Данные диагностики железнодорожного пути преобразованы в пространственно-временные ряды изменения параметров дефектов. Установлены редкие и часто встречающиеся типы, наличие сезонности и несбалансированности данных. Предложенный алгоритм балансировки учитывает физическую природу дефектов при объединении в агрегаты, что повышает точность прогнозирования размеров сравнительно редких типов дефектов.

*Ключевые слова:* оптимизация, агрегирование, минимизация отклонений.

## Введение

Современные системы мониторинга железнодорожной инфраструктуры генерируют огромные массивы данных, содержащих информацию о геометрических параметрах пути. Однако существующие методы анализа не в полной мере учитывают пространственно-временную природу этих данных [1–3]. В работе представлен подход к обработке диагностической информации, позволяющий прогнозировать развитие дефектов.

Собираемые эксплуатирующей организацией данные содержат технические и эксплуатационные характеристики железнодорожного пути, которые могут существенно влиять на анализ и прогноз размеров дефектов [4]. К техническим характеристикам относятся ПЧ (Путевая часть) – идентификатор участка инфраструктуры, ПС (Путевая структура) – тип конструкции пути, СТРЕЛКА, ОБК, МОСТ – флаги особых элементов пути, БАЛЛ – оценка опасности (чем выше, тем опаснее дефект). К эксплуатационным характеристикам относятся СК\_УСТ\_ПАСС/ГРУЗ – установленные скорости для пассажирских/грузовых поездов, КОЛИЧЕСТВО – частота появления дефекта на участке, УСЛРАСЧЕТА – условия расчёта (климатические/нагрузочные). К инфраструктурным характеристикам относятся КР, ДЗ, Т+, З, ИС – коды состояний пути, PR\_PREDUPR – флаги предупредительных мероприятий. На рис. 1 представлены дефекты типа Р.нр со степенью опасности равной 4 (наиболее опасные), обнаруженные на путевой части № 4.

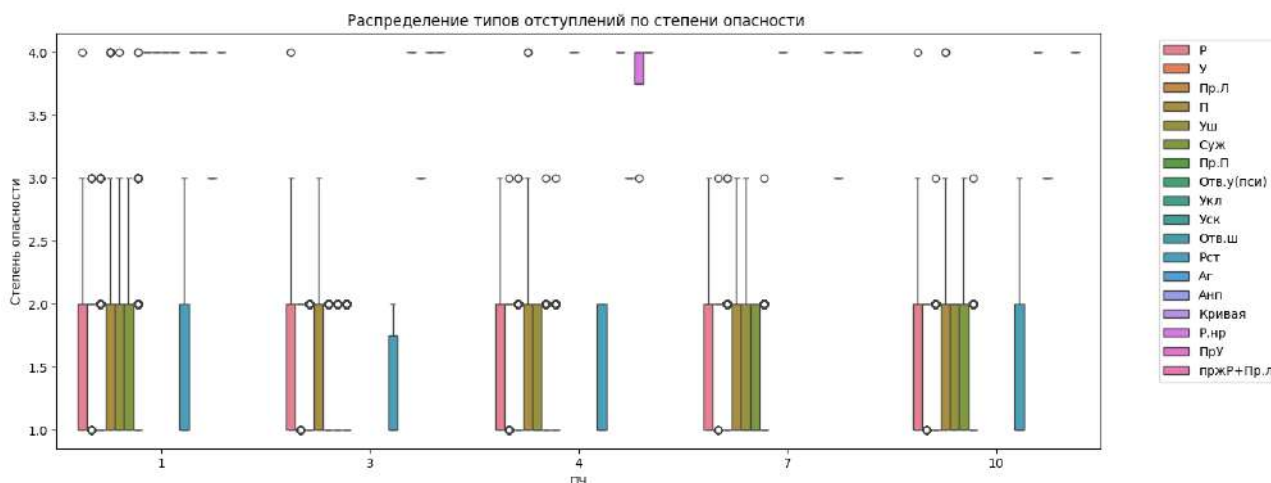


Рис. 1. Распределение типов дефектов по степени опасности и путевым участкам

Рисунок 2 демонстрирует, что на мостах развиваются в основном перекосы, просадки, рихтовки и уклонения. Известно о повышенных динамических нагрузках в зонах въезда на мост по сравнению с открытыми участками пути.

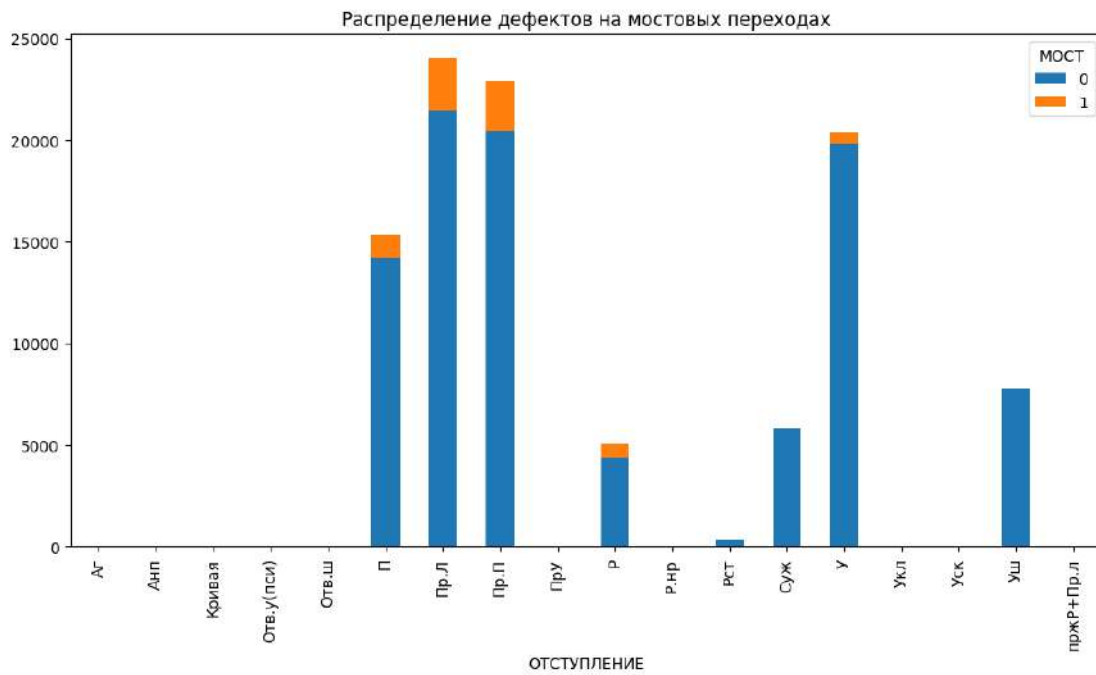


Рис. 2. Распределение дефектов на линейной части и мостах (П, Пр, Р, У)

В следующем разделе проведен анализ параметров дефектов по отклонениям от нормативных значений, типам и временным периодам, с использованием иерархических диаграмм.

### 1. Статистический анализ параметров дефектов

Для анализа дефектов пути по временным и геометрическим параметрам данные по амплитудам и длинам дефектов разделены на интервалы, характеризующие нормальные, малые, средние и большие отклонения от нормативных значений (рис. 3).

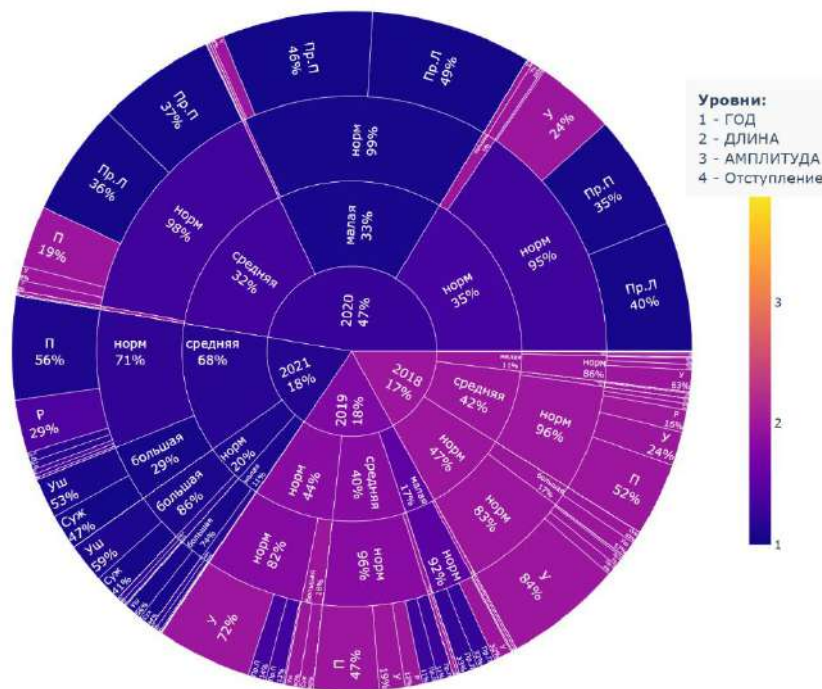


Рис. 3. Иерархическая круговая диаграмма распределения дефектов по годам, длинам, амплитудам, типам и степени опасности дефектов

На иерархической диаграмме [5] представлена структура распределения дефектов, где внутреннее кольцо соответствует годам наблюдений, демонстрируя разделение данных по временным периодам. Второе кольцо отражает длину дефектов, разделённую на четыре категории от "нормальной" до

"большой", с видимым преобладанием средних значений. Третье кольцо визуализирует амплитуду дефектов, где наиболее выражены секторы со нормальными значениями. Внешнее кольцо представляет типы дефектов, при этом дефекты, выделенные светлым оттенком, соответствуют второму уровню степени опасности (из четырех). Установлено, что дефекты третьей и четвертой степени опасности в 2021 году отсутствуют (рис. 4).

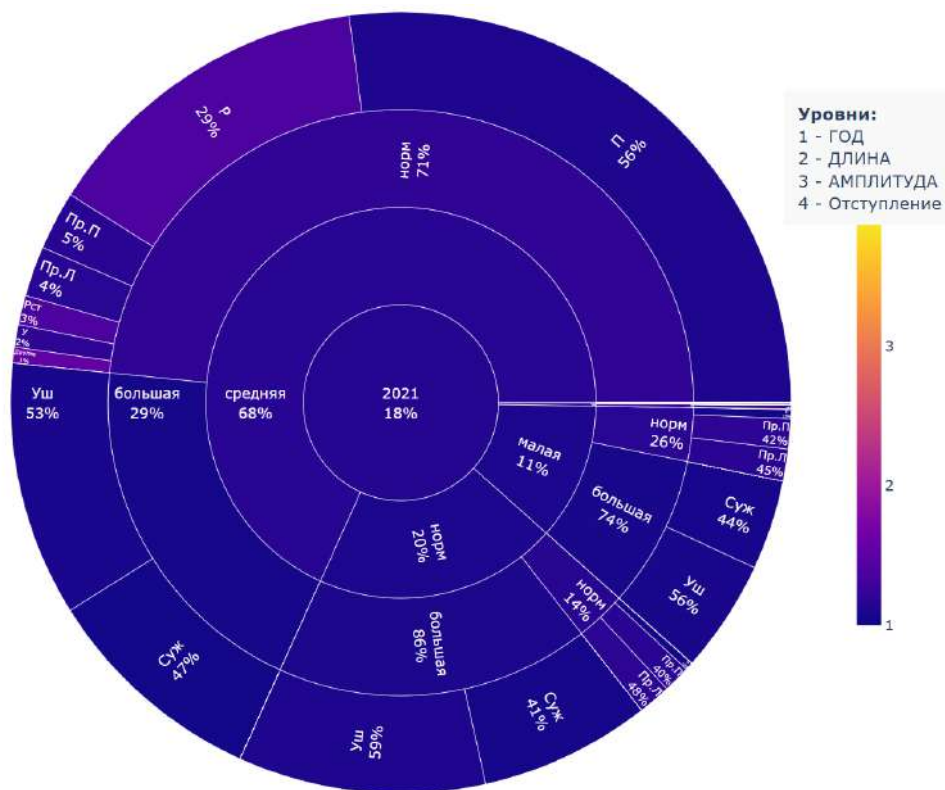


Рис. 4. Иерархическая круговая диаграмма распределения дефектов 2021 года по длинам, амплитудам, типам дефектов и степени опасности

В следующем разделе формализована задача балансировки данных путем объединения дефектов в агрегаты с учетом их физической природы.

## 2. Математическая постановка задачи балансировки дефектов с учетом физической природы

В статьях [6–8] проанализирован имеющийся датасет с параметрами дефектов и установлено, что на значительном числе километров пути не зарегистрировано ни одного дефекта некоторых типов, и что существует значительная разница в долях дефектов разных типов. Проблема несбалансированности возникает, когда для некоторых типов дефектов  $c_k: P(c_k) \approx 0.5$  (часто встречающиеся дефекты), а для других типов  $c_m: P(c_m) \approx 0.05$  (редкие дефекты). Это приводит к смещению моделей машинного обучения в сторону частых классов [9]. Поэтому необходимо использовать методы работы с несбалансированными данными при обучении моделей [10], а также анализировать данные за длительный период времени, чтобы уловить редкие, но значимые закономерности.

Пусть имеется множество дефектов пути  $D = \{d_1, d_2, \dots, d_M\}$ , где дефект  $d_j$  задан местоположением, типом  $c \in C = \{\text{просадка, перекося, уширение и т.д.}\}$ , нормативной величиной амплитуды, временными метками и параметрами, изменяющимися во времени, такими как длина, амплитуда, степень опасности  $r \in R = \{1, 2, 3, 4\}$ . Исходное распределение типов несбалансировано, например, просадки составляют 50% данных, а рихтовочные дефекты – лишь 5%:

$$P(c_k) = \frac{1}{N} \sum_{j=1}^N I(d_j \text{ имеет тип } c_k), \quad (1)$$

где  $I(\cdot)$  – индикаторная функция (равна 1, если условие выполняется, иначе 0),  $N$  – общее число дефектов.

Необходимо сбалансировать агрегаты  $A = \{A_1, A_2, \dots, A_n\}$ , так чтобы для каждого агрегата  $A_m$  число дефектов было примерно одинаковым:

$$\sum_{c_k \in A_m} P(c_k) \approx \frac{1}{M}, \forall m \in \{1, 2, \dots, M\}, \quad (2)$$

где  $M$  – число агрегатов.

Для учета условия однородности типов дефектов внутри агрегата использовать метрику минимизации отклонения амплитуды дефектов от среднего в агрегате:

$$\min \sum_{m=1}^M \left( |A_m| - \frac{N}{M} \right)^2 \quad (3)$$

### 3. Анализ существующих алгоритмов оптимального разбиения

Задача оптимального разбиения множества элементов  $T$  на  $k$  агрегатов  $A = \{A_1, A_2, \dots, A_n\}$ , с минимизацией отклонения размеров параметров от среднего значения относится к классу NP-трудных задач [11], аналогичных задаче разбиения (Partition Problem) и задаче упаковки в контейнеры (Bin Packing Problem). Существующие подходы к её решению включают динамическое программирование (метод ветвей и границ) и эвристические методы (жадные алгоритмы), однако каждый из них имеет существенные ограничения. Жадные алгоритмы [12], такие как алгоритм "Первый подходящий по убыванию" (First-Fit Decreasing), обеспечивают приемлемое качество разбиения за полиномиальное время, но не гарантируют глобального оптимума, особенно при высокой вариативности размеров элементов. Динамическое программирование позволяет найти решение, но вычислительная сложность  $O(k \cdot n^2)$  делает его сложно реализуемым для больших  $n$  и  $k$ . Методы ветвей и границ [13] теоретически способны дать оптимальное разбиение, однако на практике их использование ограничено из-за экспоненциального роста числа вариантов при увеличении размерности задачи. Эвристики на основе кластеризации (например, модификации К-средних для одномерного случая) демонстрируют хорошую сходимость, но требуют тонкой настройки и могут застревать в локальных минимумах [14]. В следующем разделе приведен алгоритм и результаты его реализации на имеющихся данных.

### 4. Реализация

Предложенный алгоритм балансировки состоит из трех шагов, представленных на рисунке 5.

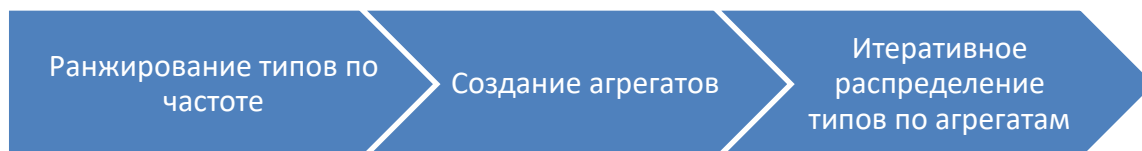


Рис. 5. Схема алгоритма балансировки

Алгоритм начинает балансировку с редких типов, добавляя их в агрегаты с наименьшим текущим размером, а часто встречающиеся типы организует в отдельные агрегаты. На этапе итеративного распределения если размер агрегата  $A_m$  превышает средний, часть его элементов перераспределяется в соседние агрегаты с учетом физической схожести. В результате работы алгоритма в агрегат "Профиль пути" вошли дефекты, характеризующие отклонения по вертикальной оси и связанные с деформациями продольного профиля пути, включая общие уровневые отклонения, просадки и ускоренный износ рельсов (У, ПрУ, Укл и Уск). Рихтовочные дефекты (Р, Р.нр, Рст) включены в эту категорию, так как неравномерная рихтовка часто приводит к волнообразным деформациям рельсового полотна с характерными перепадами высот. Комплексные случаи, сочетающие нескольких типов деформаций, ('пржР+Пр.л', 'Аг') объединяют как горизонтальные смещения, так и вертикальные просадки. Специфические дефекты вроде анизотропных проявлений ('Анп') включены в группу, поскольку их возникновение связано с усталостными процессами в рельсовой стали, которые проявляются через трещины и изломы. Агрегат "Геометрия колеи" сформирован из дефектов, влияющих на поперечные характеристики колеи – ее ширину, угловое положение рельсов и особенности в кривых участках (Уш, Суж, П, Отв.у(пси), Кривая), что критически важно для безопасности движения подвижного состава. Отдельные агрегаты для левых (Пр.Л) и правых (Пр.П) просадов позволили выделить наиболее часто встречающиеся дефекты.

В результате работы алгоритма сформировано четыре агрегата, статистика по которым представлена в таблице 1.

Таблица 1. Агрегатные статистики

Агрегат	Количество дефектов	Средняя амплитуда
Геометрия колеи	28843	718.826
Просадка левая	23982	9.012
Просадка правая	23009	13.218
Профиль пути	25779	9.088

Агрегат "Геометрия колеи" содержит более 50% перекосов и почти 47% сужений и уширений при отсутствии других типов дефектов. Агрегат "Просадка левая" полностью состоит из дефектов типа "Пр.Л". Агрегат "Просадка правая" на 99.6% состоит из дефектов "Пр.П" с включением типов "Кривая". Наиболее комплексным является агрегат "Профиль пути", где преобладают дефекты типа "У" (79%) и "ПрУ" (19.5%), с включениями рихтовочных дефектов. Таким образом, удалось добиться строгой специализации каждого агрегата по определенным типам деформаций, что подтверждает эффективность предложенного метода классификации (рис. 6).

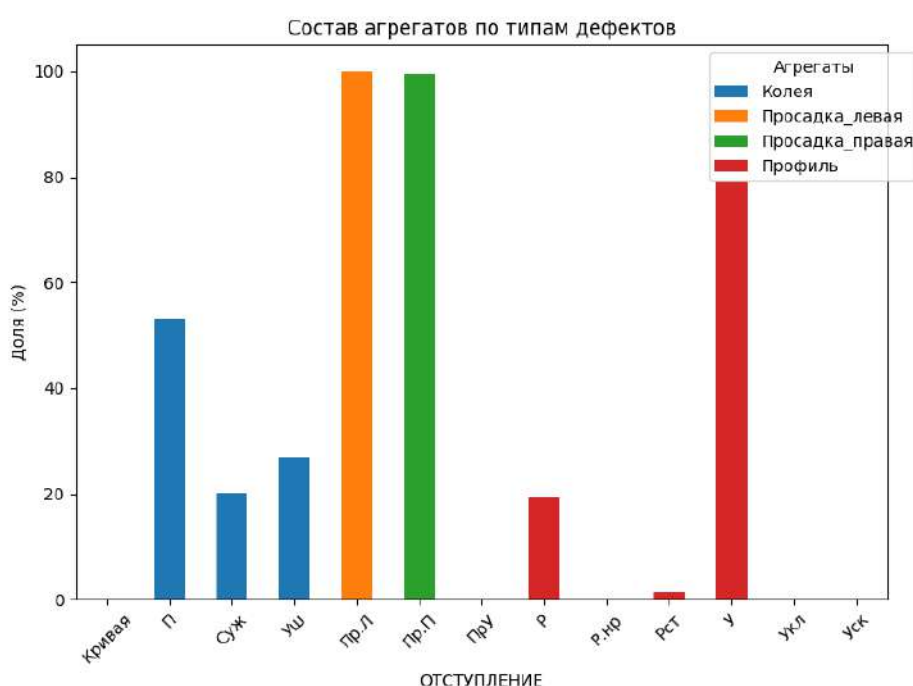


Рис. 6. Состав агрегатов по типам и количествам дефектов

## 5. Заключение

Задача балансировки формализована как оптимизационная проблема с ограничениями на физическую интерпретируемость агрегатов. Решение позволяет устранить перекоп в данных, повысить точность прогнозных моделей (например, избежать переобучения на частых классах), упростить анализ за счет сокращения числа категорий.

Баланс числа поагрегатных дефектов позволит построить эффективные прогностические модели, поскольку каждый агрегат содержит достаточное число примеров для обучения, что было бы невозможно при доминировании одного класса над другими. Ребалансировка устраняет проблему ложной точности, когда модель может демонстрировать высокую точность за счет правильного прогнозирования только частых классов [15, 16].

Численные примеры с фрагментами данных и кода доступны на GitHub автора: <https://github.com/avladova/Railway-track-deviations> или на сайте <http://vladova.ru/About>.

## Литература

1. Дубицкий И.С., Енин А.В., Владова А.Ю. Анализ динамики износа железнодорожных путей // Управление развитием крупномасштабных систем (MLSD'2021). – 2021. – С. 979–985.

2. *Енин А.В., Дубицкий И.С., Владова А.Ю.* Построение модели прогнозирования отступлений железнодорожного пути на основе статистического анализа больших данных // Управление развитием крупномасштабных систем (MLSD'2021): труды Четырнадцатой международной конференции. Москва: ИПУ РАН, 2021. – С. 993–998.
3. *Peinado Gonzalo A. и др.* Railway Track and Vehicle Onboard Monitoring: A Review // E3S Web of Conferences. EDP Sciences, 2023. – Т. 409. – С. 02014.
4. *Soleimanmeigouni I. и др.* Prediction of railway track geometry defects: a case study // Structure and Infrastructure Engineering. Taylor and Francis Ltd., 2020. – Т. 16, № 7. – С. 987–1001.
5. *Беринато С.* Сделай наглядно! Как визуализировать данные понятно и убедительно. Москва: Эксмо, 2021. – 264 с.
6. *Владова А.Ю.* Формирование пространства признаков и авторегрессионных моделей для прогноза отступлений железнодорожного полотна // Проблемы управления, 2023. – С. 54–64.
7. *Vladova A.Y.* Identification of the Railway Track Technical State // Proceedings of 2021 14th International Conference Management of Large-Scale System Development, MLSD 2021. Institute of Electrical and Electronics Engineers Inc., 2021.
8. *Владова А.Ю.* Определение ключевых признаков для регрессионного анализа отступлений железнодорожного полотна // XVI Всероссийская мультиконференция по проблемам управления (МКПУ-2023). Москва: Волгоградский государственный технический университет, 2023. – С. 260–264.
9. *García V. и др.* The class imbalance problem in pattern classification and learning, 2009.
10. *Владова А.Ю.* Регрессионные модели прогноза размеров отступлений железнодорожного полотна // Управление развитием крупномасштабных систем (MLSD'2021): труды Шестнадцатой международной конференции. Москва: ИПУ РАН, 2023. – С. 950–956.
11. *Стивенс Р.* Алгоритмы. Теория и практическое применение. Москва: Эксмо, 2021. – 544 с.
12. *Рафгарден Т.* Совершенный алгоритм. Жадные алгоритмы и динамическое программирование. Санкт-Петербург: Питер, 2022. – 256 с.
13. *Левитин А.В.* Алгоритмы: введение в разработку и анализ. Москва: Вильямс, 2006. – 576 с.
14. *Rosyid H., Mailok R., Lakulu M.M.* Optimizing K-Means Initial Number of Cluster Based Heuristic Approach: Literature Review Analysis Perspective // International Journal of Artificial Intelligence. Centre for Environment and Socio-Economic Research Publications, 2019. – Т. 6, № 2. – С. 120–124.
15. *Kim Y. geun, Kwon Y., Paik M.C.* Valid oversampling schemes to handle imbalance // Pattern Recognit Lett. North-Holland, 2019. – Т. 125. – С. 661–667.
16. *Khan S. и др.* Striking the Right Balance with Uncertainty // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2019. – Т. 2019-June. – С. 103–112.