ПРИМЕНЕНИЕ ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ ДЛЯ ИДЕНТИФИКАЦИИ ТЕХНИЧЕСКОГО СОСТОЯНИЯ ВЕРХНЕГО СТРОЕНИЯ ПУТИ

Владова А.Ю.

Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная д.65,
Финансовый университете при Правительстве РФ
Россия, г. Москва, пр. Ленинградский, 49
avladova@ipu.ru, avladova@fa.ru

Аннотация: Перегрузка автодорог, сокращение выбросов углекислого газа, возросшие скорости подвигают промышленные предприятия к активному использованию железнодорожного транспорта. Эффективность работы железных дорог повышают обновлением инфраструктуры и повышением качества технического обслуживания. Это достигается внедрением современных технологий сбора, хранения и обработки данных.

Ключевые слова: железная дорога, data science, кластеризация, статистика.

Введение

Железнодорожное хозяйство является крупномасштабной иерархической системой и традиционно выступает объектом прикладных исследований и разработок [1]. Проблема идентификации технического состояния этой системы с ростом наработки приобретает большое значение [2]. В [3] предложен математический аппарат, позволяющий прогнозировать развитие отступлений в вертикальной плоскости в зависимости от прошедшего тоннажа, конструкции и основания пути. В настоящее время мониторинг состояния верхнего строения пути (рельсы, шпалы, путевые скрепления, стрелочные переводы, земляное полотно, балласт и др.) осуществляется автономными диагностическими комплексами. Они представляют собой рельсовые вагоны, одновременно осуществляющие дефектоскопию рельсового полотна, видеоконтроль пути, путеизмерение, профилометрию, георадарный контроль балластной призмы и ряд других измерений различной физической природы [4]. наиболее эффективно осуществляет «быстрый» комплексный контроль железнодорожной инфраструктуры на перегонах большой протяжённости и плотным график движения поездов с малыми интервалами движения поездов. Собранные данные с привязкой к координатам пути передают операторам стационарной части, расшифровывают, формируют диагностическую карту участка рельсового пути и заносят в единую систему управления. В [5] представлены результаты прогнозирования ресурса рельсов по контактно-усталостным дефектам, основанные на вероятностном подходе и возможности «переноса» результатов наблюдений за выходом рельсов по дефектам на опытном участке на другой участок пути. Основными факторами влияния в исследовании названы свойства материалов, их микроструктура и остаточные напряжения при накоплении контактно-усталостных повреждений в поверхностных слоях материала рельса и развитии трещин.

Один из подходов к проблеме формирования дефектов предлагает статья [6], посвященная оценке воздействия деформированных колес на железнодорожный путь путем полноволнового численного моделирования на суперкомпьютерных системах. Для специалистов в сфере машинного обучения статья интересна тем, что перечисляет целый ряд факторов, влияющих на зависимую переменную: степень поврежденности колесной пары, локация повреждения, погодные условия, тип используемых при строительстве железнодорожных путей материалов, геолокационные условия, скорость движения подвижного состава, наличие исходных дефектов в рельсе, методика укладки шпал, тип стыка между рельсами, влияние пустот под шпалами, влияние акустических неоднородностей вблизи железнодорожных путей. Современные технологии анализа данных позволяют интегрировать математические и статистические методы с программированием и способами хранения данных. В [7]применительно к Московской железной дороге описана система предварительного автоматического ранжирования инцидентов, основанная на модели машинного обучения «ансамбль решающих деревьев». Она оценивает вероятность предотказного состояния по имеющимся низкоуровневым признакам и позволяет и адаптивно добавлять признаки к решающему правилу. К признакам отнесены логическая занятость, пониженное напряжение на лучевом реле, пониженное сопротивление изоляции кабеля, невозможность перевода стрелки, увеличенное время перевода, повышенное напряжение на путевом реле, выключение электропитания, понижение сопротивления изоляции источника питания, перегорание лампы запрещающего огня, уменьшенное время перекрытии, потери контроля стрелки, критическое понижение сопротивления изоляции, перегорание

красной лампы светофора, отсутствие основного или резервного питания, потеря контроля занятой стрелки, пониженное напряжение, неисправность контроллера или светофора и др. В [8] решалась задача классификации инцидентов на предотказы и ложные срабатывания по известным на момент классификации параметрам: места, времени, типа события, показания датчиков, а также контекста происхождения: погода, прошлые инциденты. Авторы статьи [9] предлагают проводить классификацию режима работы устройств железнодорожной автоматики и телемеханики (в частности, электрической рельсовой цепи, работающей в трех режимах) на основе логистической регрессии или метода опорных векторов с гауссовым ядром. При этом вектор признаков формируют по параметрам электрических сигналов на входе и выходе устройства.

Таким образом, анализ литературы показал, что сформированы подходы на основе статистических, вероятностных и обучающих методов по отдельным типам данных. Для решения проблемы идентификации технического состояния верхнего строения пути

1 Данные и методы

1.1 Данные

Исходными данными для идентификации технического состояния верхнего строения пути являются данные о рельсах, шпалах, промежуточных и стыковых скреплениях. Эти данные содержатся в перечнях отступлений и неисправностей пути, архивах капитальных ремонтов, рельсовых книгах, рельсо-шпало-балластных картах, активах земляного полотна, ведомостях дефектных рельсах и в других цифровых документах. Они могут быть обогащены информацией о состоянии элементов нижнего строения пути (насыпь, выемка, нулевое место, полунасыпь, полувыемка, полунасыпь-полувыемка, водоотводные сооружения).

Для манипулирования, визуализации и изучения данных использованы 14 библиотек Python с открытым исходным кодом. В частности, визуализация результатов анализа (корреляционные карты, графики, гистограммы) выполнена функциями библиотек Matplotlib, Pandas и Seaborn. Продвинутые математические и статистические операции на большом количестве данных осуществлены с помощью библиотек Numpy, Pandas и Stats-models. Набор библиотек для обучения моделей включает SciPy и Sklearn. Код создан в среде Google Colaboratory, позволяющей выполнять его прямо в браузере.

1.2 Методы

Подход, предлагаемый компаниями Mail.ru и Beeline по работе с большими данными, включает четыре этапа (рис. 1).

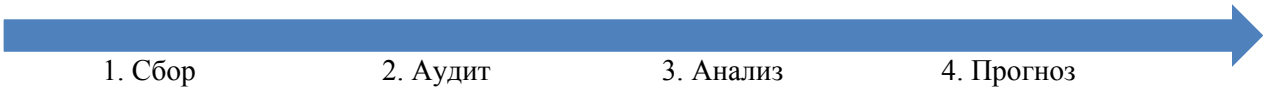


Рис. 1. Применение технологии анализа данных для идентификации технического состояния верхнего строения пути

В статье представлены результаты аудита паспортов рельсов, пропущенного тоннажа, износа.

2 Результаты

2.1 Аудит данных

Срок эксплуатации рельс определяется накатанностью в миллионах тонн брутто пропущенного по ним тоннажа. Рассматриваемые данные содержат информацию относительно самых распространенных двух типов рельс *Р65* и *Р50*, предназначенных для звеньевого и бесстыкового пути железных дорог широкой колеи, стрелочных переводов и приемо-отправочных станционных путей.

Распределение рельс по типу с учетом признака укладки («Н» - новый рельс, «П» - рельс переложен повторно) и назначения («ОН» - общего назначения, «ИК» - повышенной износостойкости и контактной выносливости) приведено на рисунке 2а. Наибольшее количество рельс для данного участка пути имеют длину 12,5 метра и произведены предприятием с литерой «К» (рисунок 2б).

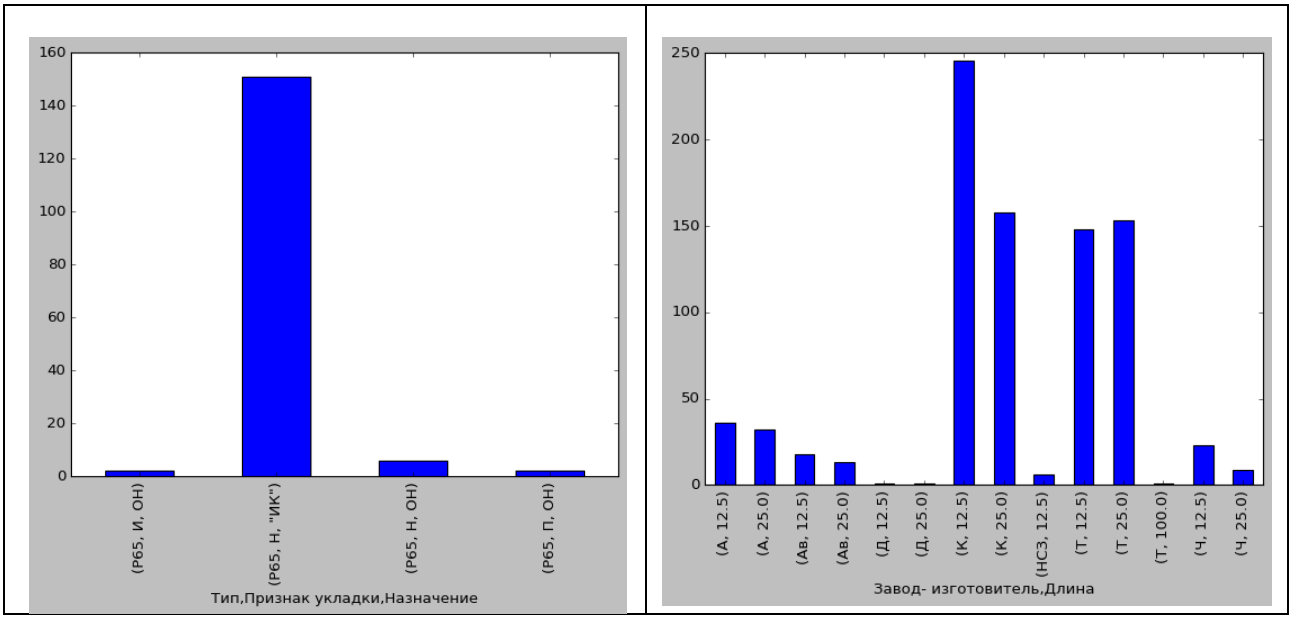


Рис. 2. Распределение рельс: а - по типу с учетом признака укладки и назначения; б – по длине и заводу-изготовителю

Анализ пропусков позволил выявить невозможность восстановления данных в поле волнообразный износ рельсов (рис. 3а). Этот износ влечет интенсивный шум, ухудшает плавность движения поездов и сокращает срок службы элементов верхнего строения пути и ходовой части подвижного состава. Поля *Назначение* с 161 ненулевым значением и *Длина* с 845 ненулевым значением удалось восстановить методом KNNImputer библиотеки sklearn за счет ограниченного набора значений этих полей. Недостающие значения каждого образца определены по среднему значению 3 ближайших соседей, найденных в обучающей выборке. Два образца близки, если их присутствующие признаки близки.

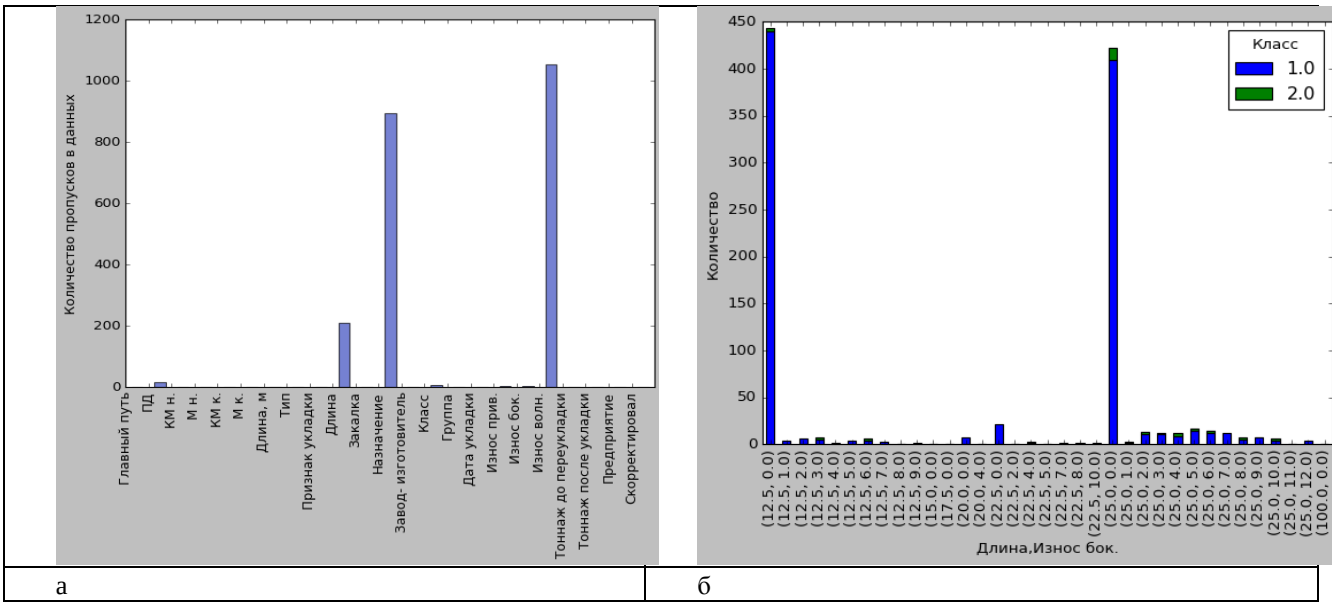


Рис. 3. Аудит данных: а – гистограмма отсутствующих значений; б – группировка по классам после восстановления значений

Статистический анализ после восстановления данных показал, что наибольшее количество исследуемых рельс имеет класс 1, тип Р65 с повышенной износостойкостью и контактной выносливостью.

2.2 Многофакторный анализ и кластеризация данных

Рисунок 4 показывает, что у I дороги преобладают небольшие уклоны на первых 5 километрах дистанции.

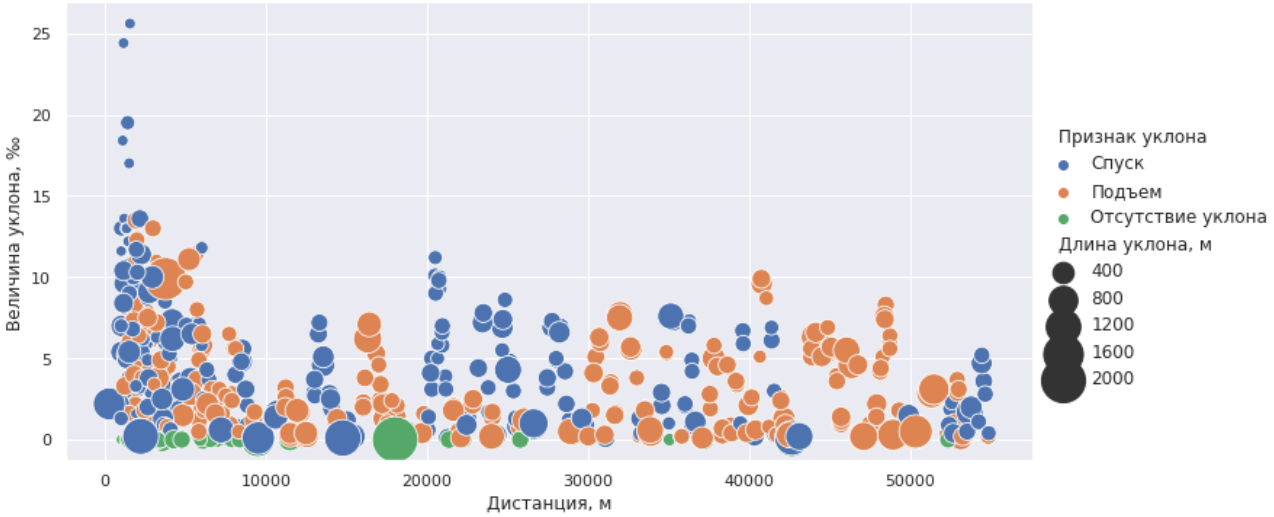
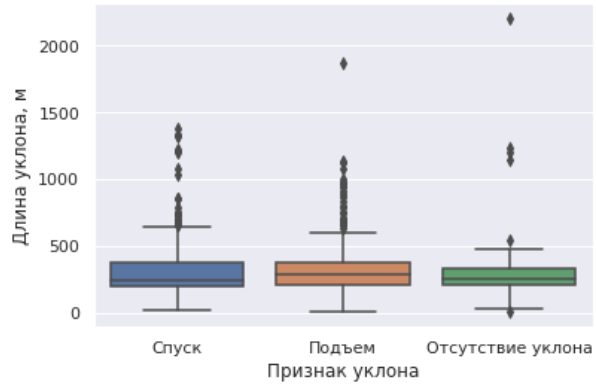
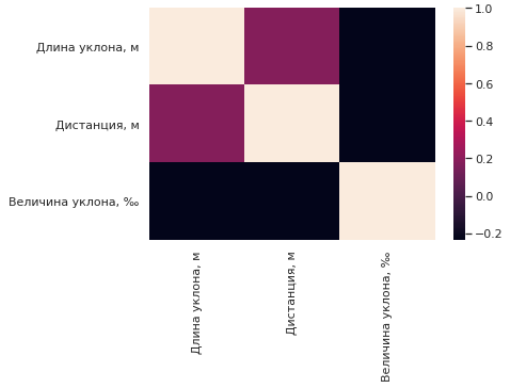


Рис. 4. Распределение уклонов по дистанции I дороги

Наиболее продолжительный участок без уклона начинается в районе 18 километра и составляет более двух километров. Этот участок не требует ограничений скоростного режима по этому параметру. С другой стороны, выявлен затяжной подъем длиной 2 км в районе 5 километров от начала дистанции с уклоном 10 промилле. Два наиболее протяженных спуска с небольшой величиной уклона приходятся на начало дистанции и район 16 км. Пять спусков с наибольшей величиной уклона находятся на начальных километрах дистанции. Оценить выбросы и медианы по признаку уклона можно при помощи графика ящик с усами (рис. 5а), а отсутствие связи между рассматриваемыми параметрами по корреляционной карте (рис. 5б).





А	б

Рис. 5. Статистическое исследование параметров уклонов I дороги: а – медианы и выбросы, б – корреляционная карта

Для этого же участка по другому источнику данных проведен анализ пропущенного тоннажа и допустимой скорости (рис. 6).

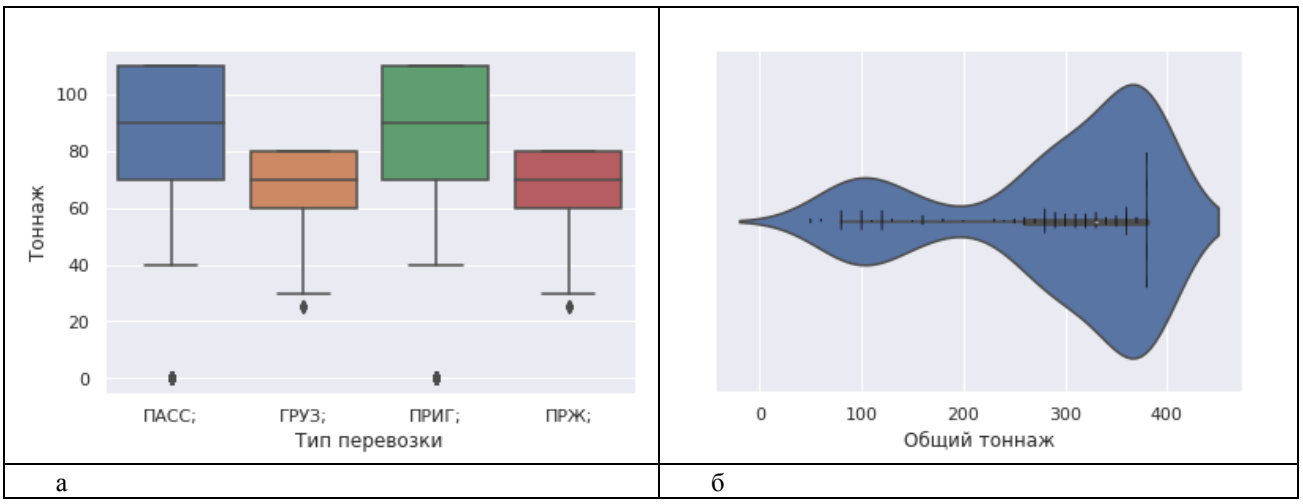


Рис. 6. Распределение тоннажа: а) по типам перевозок, б) по частоте

Гистограмма распределения тоннажа приведена на рис. 7а. Для изучения взаимосвязей между значениями переменных выполнена агломеративная кластеризация с помощью функции библиотеки seaborn. Она показывает каждый объект как одно скопление и в каждой итерации объединяет скопления до тех пор, пока не будет оставлено только одно. На рис. 7б приведен результат кластеризации тоннажа по участкам длины и дистанции. В качестве метода кластеризации использована одиночная связь, поскольку в данных имеем единичные выбросы и отсутствует шум. Метод относит объекты по кластерам, определяя минимум расстояния между ними:

$$d(u,v) = min(dist(u[i], v[j])), \qquad (1)$$

где $d(u,v)$- расстояние между объектами; u, v – массивы сравниваемых объектов.

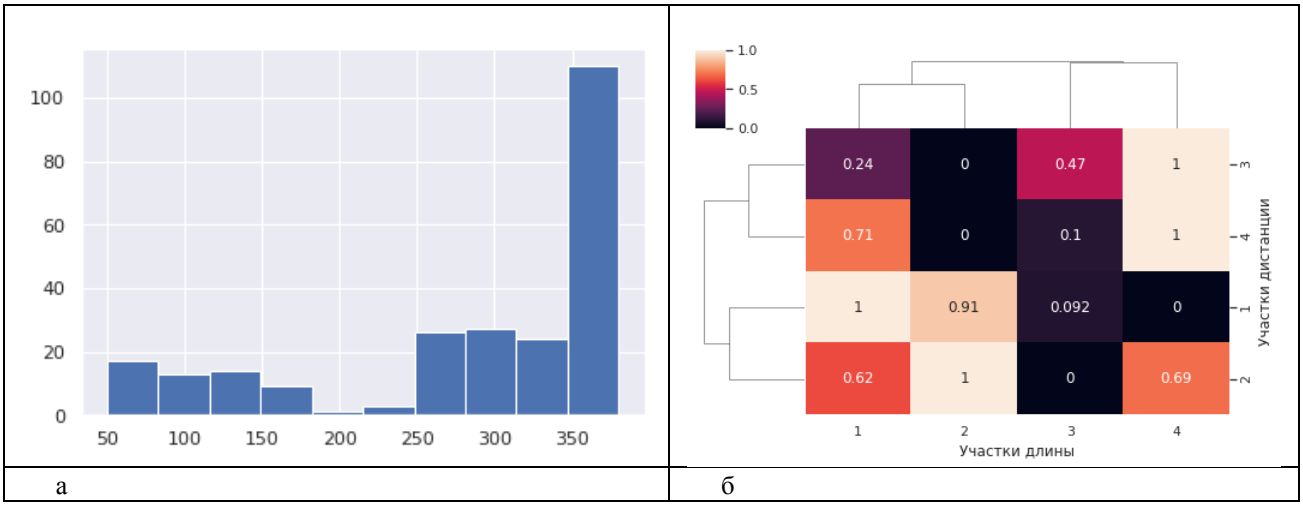


Рис. 6. Гистограмма распределения тоннажа и кластеризация

Для сравнения тоннажа по длинам участков и дистанции от нулевого километра, значения последних двух признаков разделены по квартилям и представлены как категориальные переменные *Участки длины* и *Участки дистанции*. Значения признака *Пропущенный тоннаж* нормированы с помощью метода *MinMaxScaler*. По рис. 6б определено, что наибольший тоннаж пропущен по третьему и четвертому участку дистанции (25,3 – 54,8) км с длиной (1,0 – 9,8) км, а также по сравнительно малым участкам длины (до 226 метров), находящимся на дистанции до 3,6 км от начала пути. То есть в начале и конце рассматриваемого участка верхнего строения пути. На рис. 7 приведены диаграммы размаха, полученные для признаков *Грузонапряженность* (параметр, характеризующий интенсивность использования железнодорожной сети, измеряемый количеством тонн, приходящихся на 1 км эксплуатационной длины линии) и *Осевой нагрузки* (количество тонн, приходящихся на ось вагона).

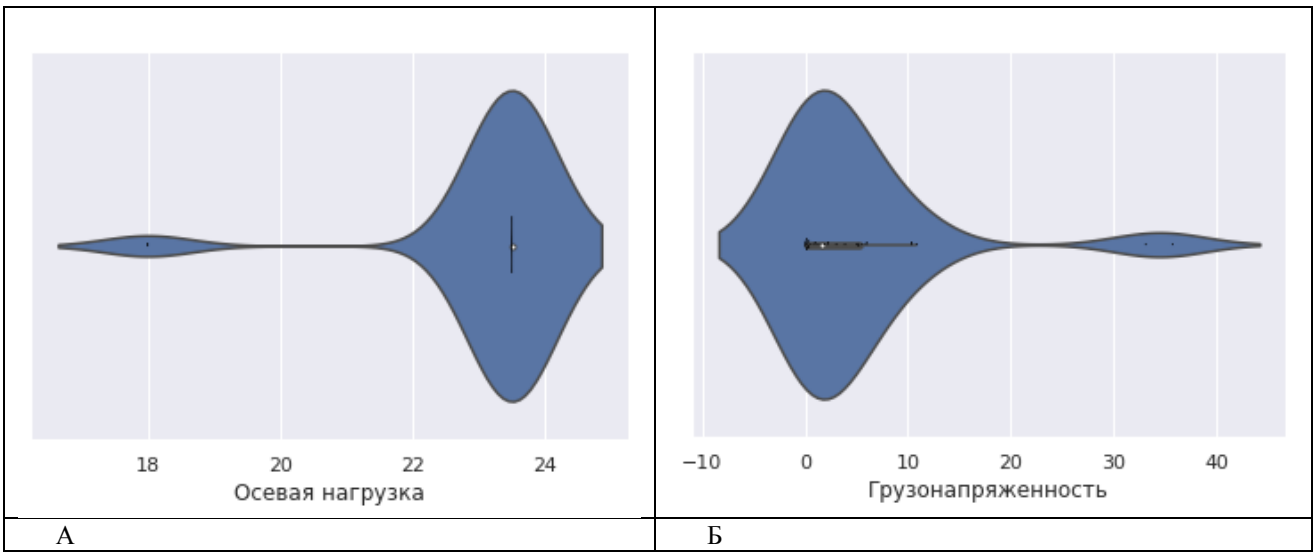


Рис. 7. Диаграммы размаха: а) осевой нагрузки, б) грузонапряженности

Центр масс осевой нагрузки приходится на 23,8 тонн на ось.

2.3 Нелинейная взаимосвязь

Для выявление нелинейной корреляции использован коэффициент корреляции Крамера [23]. Он базируется на статистике χ^2, и используется для порядковых и интервальных признаков:

$$\varphi_C = \sqrt{\frac{\chi^2}{N\,min(r-1,k-1)}}$$

где N- количество наблюдений; r (k) — это количество строк (столбцов) в contingency table.

На рисунке 8 приведена карта кластеров, полученных по значениям коэффициента корреляции Крамера.

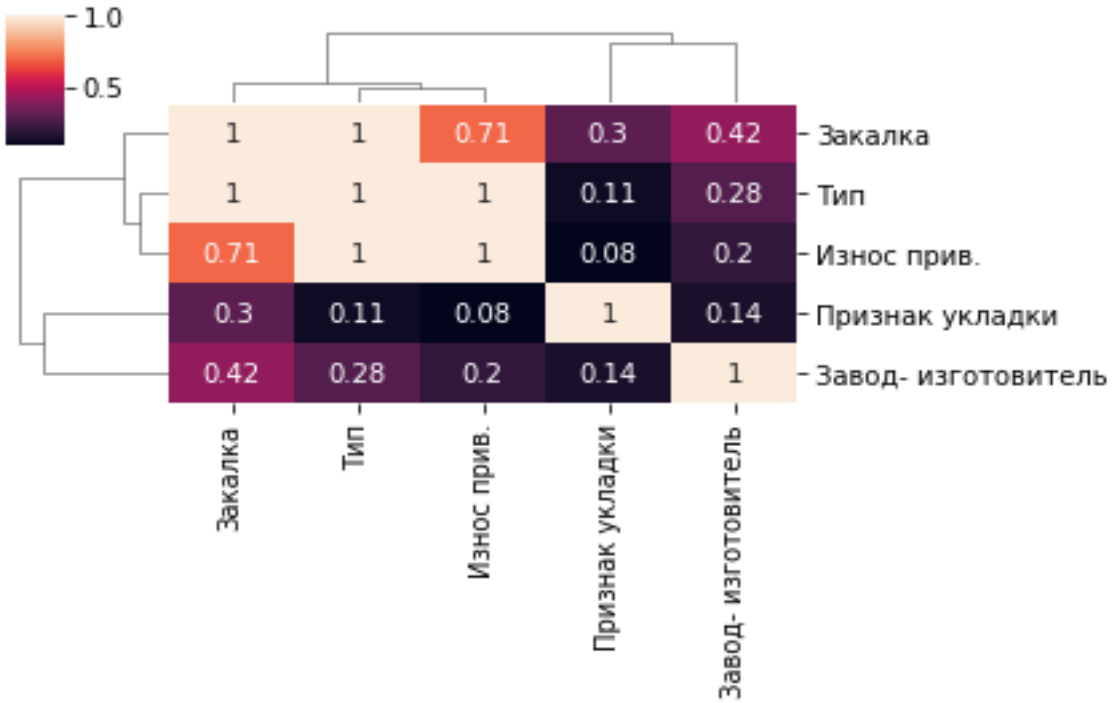


Рис. 8. Кластеризация категориальных признаков по коэффициенту корреляции Крамера

Из рисунка следует, что приведенный износ нелинейно зависит от типа закалки (Т1, ДТ370, Т2, ДТ350, Н, С) и типа рельс (Р65, Р50).

Заключение

Таким образом проведен аудит и восстановлены данные для нескольких цифровых полей, выполнено статистическое исследование ряда признаков, установлена нелинейная взаимосвязь между признаками.

Литература

1. *Цыганов В.В., Малыгин И.Г., Еналеев А.К., Савушкин С.А*. Большие транспортные системы: теория, методология, разработка и экспертиза. ИПТ РАН, СПб. (2016)
2. *Владова, А.Ю.* Построение информационной системы управления для оболочковых объектов. ИПК ГОУ ОГУ, Оренбург (2010)
3. *Соловьев, В.П., Анисин А. В., Надежин С. С., Певзнер В. О., Третьяков В. В., Третьяков И. В.* Моделирование процесса накопления остаточных деформаций пути с использованием суперЭВМ. In: Фундаментальные исследования для долгосрочного развития железнодорожного транспорта: сб. трудов членов и научных партнеров Объединенного ученого совета ОАО «РЖД». pp. 185–192. М.: Интекст (2013)
4. *Логинов А. Г., Никитина Т. С.* Многофункциональный автономный роботизированный комплекс диагностики и контроля верхнего строения пути и элементов железнодорожной инфраструктуры, https://findpatent.ru/patent/273/2733907.html, (2020)
5. *Горячева И. Г., Захаров С. М., Коган А. Я., Торская Е. В., Шур Е. А., Абдурашитов А. Ю., Борц А. И., Заграничек К. Л.* Комплексный подход к прогнозированию работоспособности и ресурса рельсов нового поколения. Бюллетень ОУС ОАО «РЖД». 5–6, 16–26 (2017)
6. *Фаворская А. В., Хохлов Н. И., Миряха В. А.* Разработка математических моделей, численных методов и расчетных программ для выявления дефектов элементов системы «колесо-рельс». Бюллетень Объединенного ученого совета ОАО РЖД. 1, 49–63 (2018)
7. *Бойко П.Ю., Быков Е.М., Соколов Е.И., Яроцкий Д.А.* Применение машинного обучения к ранжированию инцидентов на Московской железной дороге. Бюллетень Объединенного ученого совета ОАО РЖД. 6, 36–47 (2016)
8. *Яроцкий Д. А., Бойко П. Ю., Иванова Е. П.* Применение методов машинного обучения к задачам управления инфраструктурой ОАО «РЖД». Бюллетень Объединенного ученого совета ОАО РЖД. 6, 36–47 (2016)
9. *Prisukhina, I. V., Borisenko, D. V.* Machine state classification of electric track circuit by means of support vector machine. Omsk Scientific Bulletin. 126–130 (2018). https://doi.org/10.25206/1813-8225-2018-162-126-130