ПОСТРОЕНИЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОТСТУПЛЕНИЙ ЖЕЛЕЗНОДОРОЖНОГО ПУТИ НА ОСНОВЕ СТАТИСТИЧЕСКОГО АНАЛИЗА БОЛЬШИХ ДАННЫХ

Енин А.В., Дубицкий И.С., Владова А. Ю.

Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная д.65
avladova@ipu.ru

Аннотация: Железнодорожный путь - это комплекс инженерных сооружений и обустройств, расположенных в полосе отвода, образующих дорогу с направляющей рельсовой колеёй. В процессе эксплуатации на железнодорожном пути возникают различные виды неисправностей. В результате проведенных исследований получен перечень отступлений для построения прогнозирующей модели, оценены зависимости между отступлениями по корреляции. Сделано предположение о возможности создания прогнозирующей модели.

Ключевые слова: отступления, железная дорога, предобработка данных, корреляция, граф зависимостей, модель прогнозирования.

Введение

Техническое состояние железнодорожного полотна оценивают на основе данных, собранных подвижной лабораторией [1]. Лаборатория ежемесячно проезжает по всему пути и фиксирует отступления в автоматическом режиме. Под отступлениями понимаем изменение параметров верхнего строения пути [2] от незначительного (1 степень) до критического (4 степень). В настоящей работе используются четыре столбца из выгружаемых данных по отступлениям - дата, пикет, вид отступления и число зафиксированных отступлений (таблица 1).

Таблица 1. Образец обрабатываемых данных

Месяц	Пикет	Вид отступления	Число отступлений
2021-02-01	64.6	П	3
2019-03-01	111.3	У	2
2020-01-01	323.2	Пр.Л	4
2021-02-01	609.7	П	1
2020-12-01	296.2	У	2

Отступления подразделены на 12 видов. В Таблице 1 присутствуют «Пр.П» - Просадка правая; «У» - Плавное отклонение по уровню от нулевой линии; «П» – Перекос. Для построения модели прогнозирования необходимо проанализировать связанность числа отступлений с числом в прошлом. Кроме того необходимо оценить связи отступлений каждого вида. Таким образом, цель работы сформулирована как построение модели прогнозирования отступлений в будущем на основе данных о текущих и прошлых отступлениях.

Прежде чем построить прогнозную модель необходимо оценить имеющиеся данные для ее построения. Если зависимости будут слабы, придется отказываться от текущей методики и искать другую.

1 Метод исследования

Построение прогнозной модели проведено в несколько этапов в соответствии с Рисунком 1.

На этапе отбора высокочастотных видов отступлений выполнен расчет общего числа отступлений для каждого вида. Отобраны отступления с общим числом зарегистрированных отступлений более 5000. Далее данные по каждому виду отступлений нормированы по следующей формуле:

$$< \text{Новое значение} >=< \text{Текущее значение} > \cdot \frac{<\text{Среднее значение в } 2018-2019 \text{ гг}>}{<\text{Среднее значение в текущем году}>} \quad (1)$$

Исследование связанности видов отступлений внутри одного месяца и одного пикета проведено с помощью корреляционного анализа между различными видами отступлений. Исследование зависимостей отступлений от отступлений в предыдущем месяце в рамках одного пикета проведено также при помощи корреляционного анализа. Анализ проводился между всеми возможными парами отступлений.

Рис. 1. Метод проведения исследований

Зависимости рассматривались на основе коэффициента линейной корреляции Пирсона для числа отступлений и их логарифмов. Как будет показано ниже, частота фиксации отступлений от вида к виду меняется на порядки. Редкие виды отступлений были отброшены. Дискретность времени выбрана равной одному месяцу. Выбор обусловлен периодичностью работы подвижной лаборатории и работы ремонтных бригад. Дискретность дистанции выбрана из сложившейся на железной дороге практики измерять все пикетами. Поскольку использован коэффициент корреляции Пирсона для исследования связности числовых рядов, первая предлагаемая модель прогнозирования линейна. Она строит прогноз для текущего и соседнего пикетов на основе данных прошлых месяцев.

2 Результаты

2.1 Отбор высокочастотных видов отступлений

Для 18 видов отступлений построена гистограмма (рис. 2), по которой из всего разнообразия были выбраны наиболее часто встречающиеся семь видов: «Уш» - Уширение; «Суж» - Сужение; «У» - Плавное отклонение по уровню от нулевой линии; «Пр.л» - Просадка левая; «Пр.п» - Просадка правая; «П» - Перекос; «Р» - Отступление в плане по рихтовочной рельсовой нити.

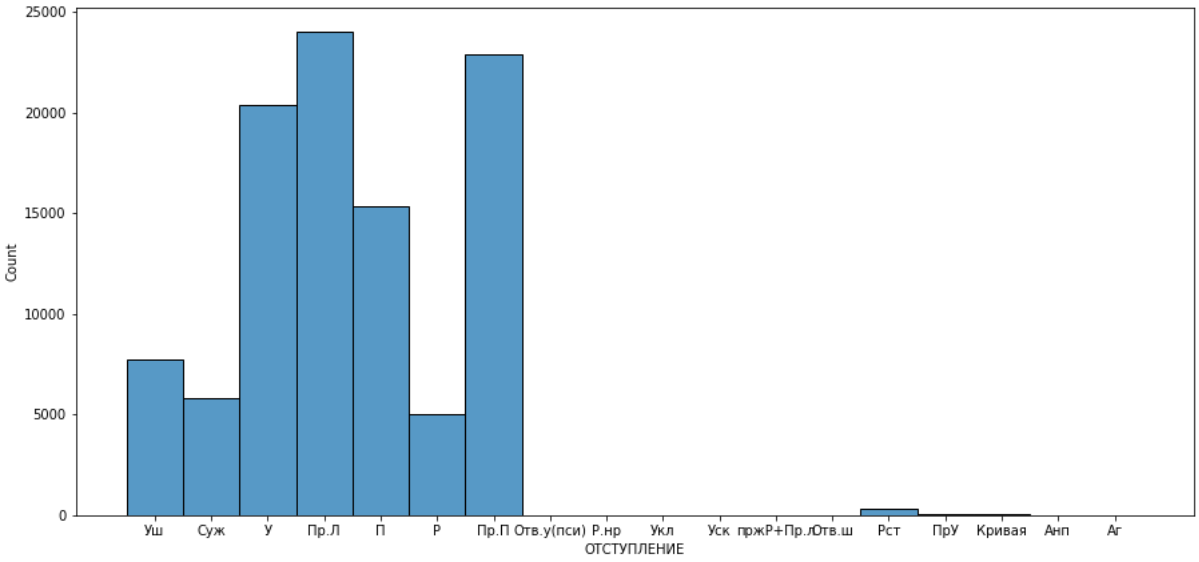


Рис. 2. Высокочастотные виды отступлений

2.2 Нормировка данных

При первичном визуальном анализе данных (Рисунок 3) замечено возрастание числа регистрируемых отклонений год к году в 2-4 раза. Сделано предположение, что это обусловлено переходом на новую инструкцию расчета степеней отступлений различных видов [3].

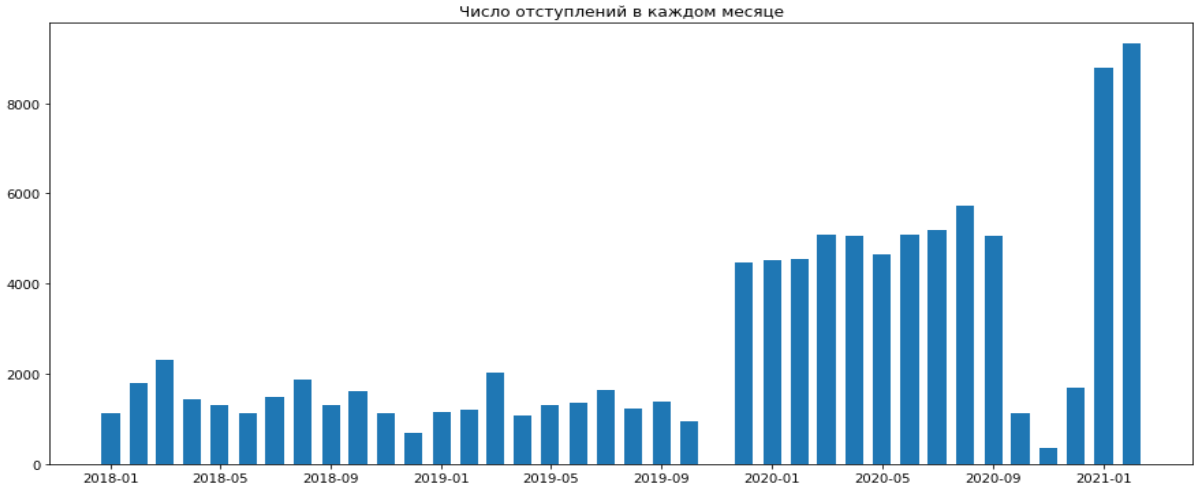


Рис. 3. Число отступлений в каждом месяце

Для устранения отличия ежемесячное число отступлений поделили на среднегодовое в текущем году и умножили на среднегодовое число отступлений в 2018-2019 годах. Такие данные уже можно подавать на модель, поскольку они распределены более равномерно (рисунок 4).

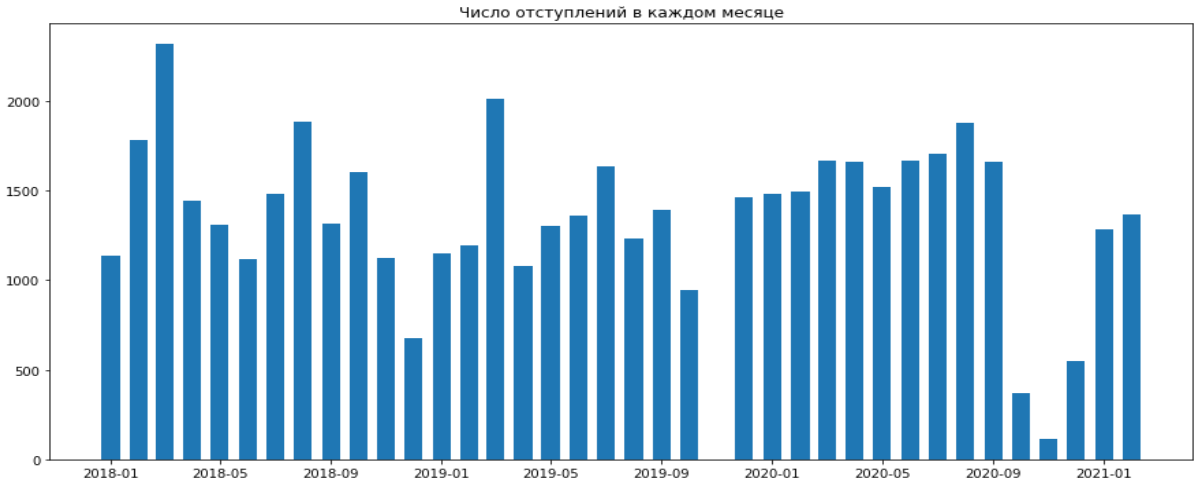


Рис. 4. Число отступлений в каждом месяце после общей нормировки

Однако при анализе числа среднемесячных отступлений в разрезе видов обнаружены отличия в средних на тех же временных участках для отдельных видов отступлений. Применение нормировки отдельно по каждому виду исправило ситуацию. Гомоскедастичность данных в разных годах можно наблюдать на диаграмме размаха для двух отступлений (см. Рисунок 5).

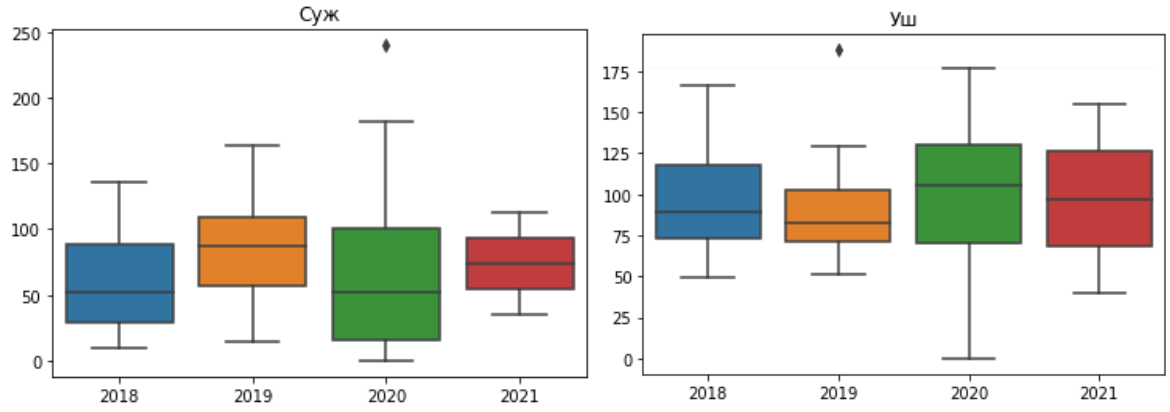
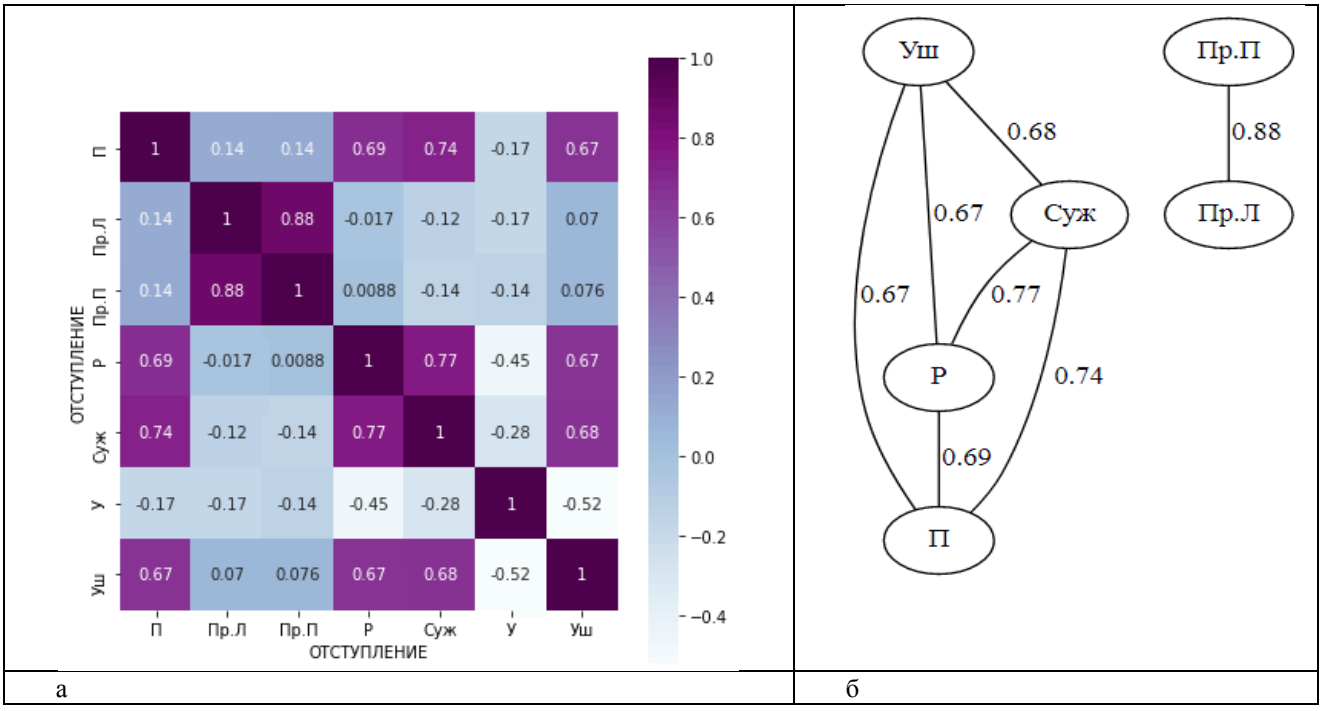


Рис. 5. Диаграмме размаха для отступлений «Суж» и «Уш»

2.3 Исследование связности видов отступлений внутри одного месяца и одного пикета

Связность отступлений анализируем по значениям коэффициента линейной корреляции Пирсона. Перенеся наиболее сильные корреляции на граф [4] можно увидеть два полносвязных графа из четырех и двух видов отступлений (Рисунок 6).

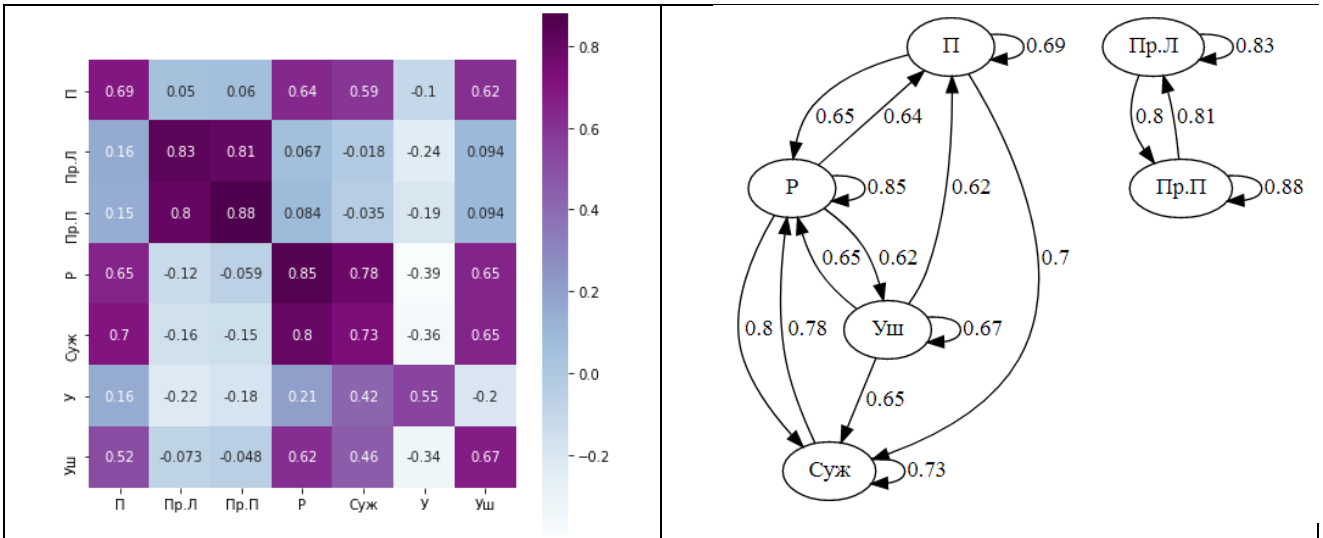


а б

Рис. 6. Связность отступлений внутри одного месяца и пикета: а) матрица корреляций; б) графы с коэффициентами корреляций

2.4 Исследование зависимостей отступлений от отступлений в предыдущем месяце в рамках одного пикета

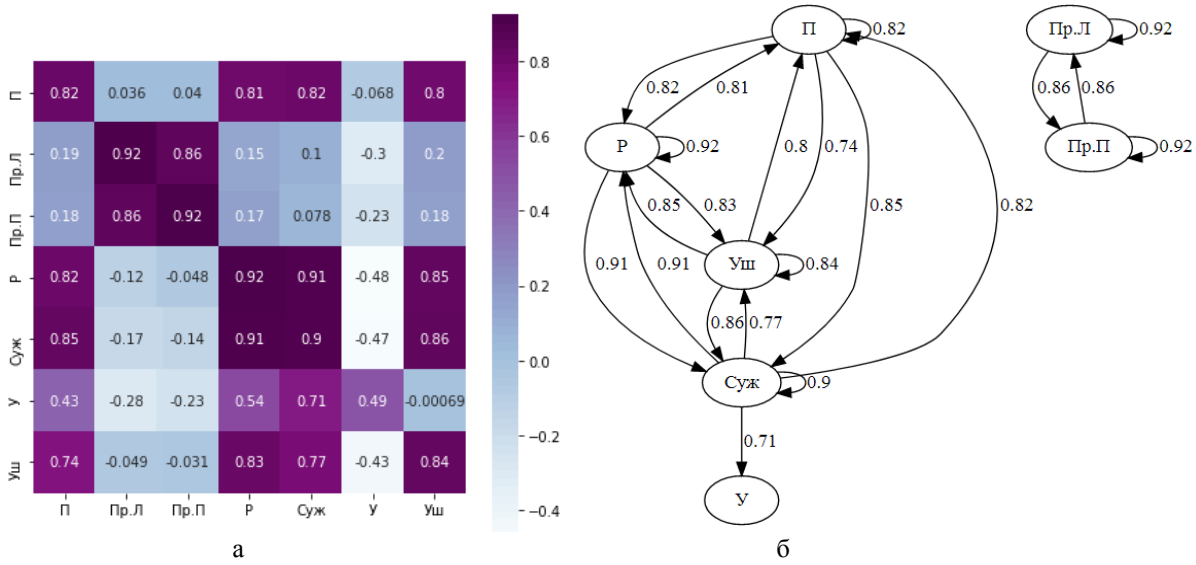
Обнаружены значимые зависимости между отступлениями из разных месяцев (текущий и предыдущий). На матрице корреляций можно видеть их значения (Рисунок 7а). Два ориентированных графа отражают переходы одних отступлений в другие в рамках одного пикета и отличаются от предыдущей пары наличием петель.



а б

Рис. 7. Связность отступлений внутри одного пикета и двух месяцев: а) матрица корреляций; б) графы с коэффициентами корреляций

Для повышения степени связности проведено логарифмирование данных. Результаты можно видеть на следующем графе зависимостей (рисунок 8).



а б

Рис. 8 Тепловая диаграмма и направленный граф корреляций после логарифмирования внутри одного пикета и двух месяцев

3 Анализ полученных результатов

Высокая связанность отступлений внутри одного месяца не позволит создать эффективную нейронную модель прогнозирования. Потребуется выделение главных компонент [5].

Первоначальная нормировка позволила получить гомоскедатичные ряды данных, которые можно подавать на модель.

В связи с наличием кластеров в отступлениях есть возможность задачу прогнозирования разбить на несколько отдельных.

Высокие коэффициенты корреляции позволяют надеяться на хорошую модель прогнозирования числа отступлений в будущем месяце.

Применение логарифмирования позволило выявить более сильные связи. Появилась зависимость для отступления вида “У” (плавное отклонение по уровню от нулевой линии), коэффициент корреляции Пирсона которой был недостаточно значим до логарифмирования.

Заключение

Поскольку число параметров не слишком велико (7 видов отступлений), для построения прогнозной модели можно использовать методы “Случайный лес” [6] и “Градиентный бустинг” [7]. Стандартные библиотечные реализации указанных методов позволят рассчитать важность параметров [8]. Отобранные параметры в дальнейшем уже можно будет использовать на более сложной нейронной сети.

В настоящей работе как параметр рассмотрено число регистрируемых отклонений. На практике более интересным для эксплуатирующей организации являются степени отклонений.

Не рассмотрены связи с отступлениями на соседних пикетах. Возможно это должно дать больше информации для будущей модели.

Рассмотрены зависимости в ближайшей окрестности текущего и предыдущего месяца и текущего пикета. Скорее всего должны существовать зависимости при углублении по времени (несколько предыдущих месяцев) и по дистанции (несколько соседних пикетов).

Далее предстоит проанализировать зависимости в широкой окрестности от текущего прогнозируемого месяца-пикета.

Ввиду разрастания количества входных параметров будет опробована сверточная нейронная сеть.

Литература

1. *Насибов Р.Э., Мехоношин С.А., Папировская Л.И.* ТЕХНОЛОГИЯ РАБОТЫ КОМПЬЮТЕРНОГО ВАГОНА-ЛАБОРАТОРИИ (КВЛ-П). В сборнике: МОЛОДЕЖНАЯ НАУКА В XXI ВЕКЕ: ТРАДИЦИИ, ИННОВАЦИИ, ВЕКТОРЫ РАЗВИТИЯ. материалы Международной научно-исследовательской конференции молодых ученых, аспирантов, студентов и старшеклассников. 2018. С. 184-185.
2. *Ворошилова А., Эльхутов С.Н.* НАЧАЛЬНЫЙ АНАЛИЗ ПАРАМЕТРОВ ВЕРХНЕГО СТРОЕНИЯ ЖЕЛЕЗНОДОРОЖНОГО ПУТИ. Сборник научных трудов Ангарского государственного технического университета. 2012. Т. 1. № 1. С. 031-034.
3. Инструкция по оценке состояния рельсовой колеи путеизмерительными средствами и мерам по обеспечению безопасности движения поездов, РЖД от 28.02.2020 года // *https://rzd-puteetz.ru/instruktsiya-po-otsenke-sostoyaniya-relsovoj-kolei-puteizmeritelnymi-sredstvami/*
4. *Виноградова М.Г., Федина Ю.А., Папулов Ю.Г.* ТЕОРИЯ ГРАФОВ В КОРРЕЛЯЦИЯХ "СТРУКТУРА – СВОЙСТВО". Журнал физической химии. 2016. Т. 90. № 2. С. 234-239.
5. *Салимгареева Д.А.* МЕТОД ГЛАВНЫХ КОМПОНЕНТ. NovaInfo.Ru. 2017. Т. 3. № 58. С. 5-9.
6. *Курейчик В.М., Картиев С.Б.* АЛГОРИТМ КЛАССИФИКАЦИИ, ОСНОВАННЫЙ НА ПРИНЦИПАХ СЛУЧАЙНОГО ЛЕСА, ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ. Программные продукты и системы. 2016. № 2. С. 11-15.
7. *Чорный Д.А.* СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ГРАДИЕНТНОГО БУСТИНГА ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ. В сборнике: Обработка информации и математическое моделирование. Материалы Российской научно-технической конференции. 2017. С. 208-215.
8. *Дьяконов А.Г.* Отбор (селекция) признаков. Цикл лекций «Прикладные задачи анализа данных». Москва, МГУ. 2020