Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТКИ 174 ГРУППЫ

«Построение композиций прогностических моделей и оценка качества прогноза временных рядов»

Выполнила:
студентка 4 курса 174 группы
Газизуллина Римма Камилевна

Научный руководитель:
д.ф-м.н.
Стрижов Вадим Викторович

Москва, 2015

Содержание

Аннотация

Работа посвящена исследованию алгоритма непараметрического прогнозирования объемов железнодорожных грузоперевозок. Решается задача прогнозирования количества вагонов с различными грузами, следующих по различным маршрутам. Задана топология железнодорожной сети — для всех возможных пар железнодорожных районов дана информация о всех блоках вагонов, совершивших переезд из одного района в другой, включая количество вагонов в блоке, вид груза и дату прохождения маршрута. Для построения прогноза используется алгоритм, основанный на свертке эмпирической плотности распределения значений временного ряда с функцией потерь. Ранее прогноз выполнялся для каждого железнодорожного узла в отдельности. Предлагается повысить качество прогноза за счет учета топологии, то есть прогнозирования по парам районов вместо прогнозирования отправления всех вагонов с данного узла. Алгоритм проиллюстрирован посуточными данными за полтора года о перевозках 38 типов грузов.

Ключевые слова: прогнозирование, непараметрический метод, загруженность железнодорожного узла, функция потерь, эмпирическое распределение, свертка.

1 Введение

Повышение эффективности транспортировки грузов по железнодорожным путям требует решения задачи прогнозирования объемов и времени перевозок грузов с ветки на ветку и загруженности железнодорожных узлов на основании данных о времени прибытия и отправления вагонов. Решение этой задачи необходимо для повышения качества управления в сложных инфокоммуникативных средах [1]. Управленческие решения принимаются на основании работы систем сбора статистической информации высокой точности и доступности [2, 3]. Данная работа является расширением и уточнением метода решения задачи прогнозирования, изложенном в статье [4]. Используемый в ней алгоритм основан на алгоритме квантильной регрессии [5, 6], модифицированным сверткой гистограммы с функцией потерь, что позволяет учитывать стоимость ошибок в реальной задаче.

Для получения прогноза по временному ряду предлагается построить гистограмму. Прогнозируемое значение (последующий элемент временного ряда) ищется как значение, соответствующее оптимальному значению свертки гистограммы и функции потерь, которая задана исходя из экспертных предположений об экономических потерях при недостаточном или избыточном наличии составов на железнодорожном узле. В работе [4] показано, что данное решение находит более точные прогнозы, в сравнении с алгоритмом ARMA [7, 8], также используемым для решения подобных задач. Кроме того данное решение позволяет в некоторых случаях решить задачу, когда алгоритм ARMA не может работать.

Согласно работе [9], в настоящее время насчитывается свыше 100 классов моделей прогнозирования, в связи с чем возникает задача выбора моделей, которые давали бы адекватные прогнозы для изучаемых процессов или систем. Одним из основных методов для решения подобных задач является метод непараметрической регрессии и модификации, такие как ядерное сглаживание, сглаживание сплайнами, авторегрессия, скользящее среднее и другие, описаные в [10, 11, 12, 13]. Они заключаются в комбинации взвешенных значений элементов временного ряда для получения прогноза. Также для решения подобных задач применяют нейронные сети [14, 15].

Данные, описывающие объемы грузоперевозок, имеют следующий формат. Для каждой станции для каждого дня запись содержит количество вагонов с грузами

различного типа, с указанием ветки прибытия и отправления. Для построения прогнозов используются временные ряды двух типов:

1) количество отправленных с ветки вагонов с грузом конкретного типа,

2) количество отправленных на заданную ветку вагонов с грузом конкретного типа.

Основной идеей и новизной работы является использование для прогноза временных рядов грузоперевозок по парам веток в отличие от ранее предложенного метода, использовавшего грузоперевозки по одному узлу. Это позволяет точнее отслеживать особенности грузоперевозок поездов между различными парами веток и, как следствие, точнее предсказывать будущие перевозки. В данной работе предполагается, что благодаря рассмотрению более детализированных данных по перевозкам, удастся выявить тренды и периодичность. С другой стороны, решение этой задачи может сделать весь прогноз в целом более точным.

2 Постановка задачи гистограммного прогнозирования

Задан набор временных рядов, в котором временной ряд $\mathbf{x} = \{x_i\}_{i=1}^{T}$, и горизонт отсрочки прогноза h (число отсчетов от конца временного ряда до точки прогноза, включительно). Требуется спрогнозировать следующую точку x_{T+1} временного ряда $\mathbf{x}$ так, чтобы выполнялось условие оптимальности функции потерь (2) и свертки гистограммы (1), построенной по значениям временного ряда.

Гистограмма $\mathcal{H}$ — набор пар

$$\mathcal{H} = \{(y_k, g_k)\}_{k=1}^{K}, \tag{1}$$

где K — число интервалов $[y_k^{\min}, y_k^{\max}]$ со средним значением y_k, на которые разбита ось значений ряда x, g_k — высота столбца гистограммы на интервале $\bar{y}_k$, равная взвешенной сумме количества точек ряда, попавших в этот интервал.

Предполагается, что ряд $\mathbf{x} = \{x_i\}_{i=1}^{T}$ стационарен, то есть совместное распределение вероятностей T наблюдений $x_1, x_2, x_3, \ldots, x_T$ совпадает с распределением $x_{1+\tau}, x_{2+\tau}, x_{3+\tau}..., x_{T+\tau}$ при любых T и τ.

Введем функцию потерь $L(\hat{y}, y)$ — штраф за несоответствие прогнозируемого значения $\hat{y}$ историческому значению y. Например:

1) $L(z, x) = (z - x)^2$;

2) $L(z, x) = |z - x|$;

3) $L(z, x) = \begin{cases} 0, & \text{если } |z - x| < a; \\ |z - x| - a, & \text{если } |z - a| \geq a, \text{ где } a > 0 \text{ — экспертно заданный параметр.} \end{cases}$

В данной работе используется функция

$$L(z, x) = |z - x|. \tag{2}$$

Задача нахождения прогнозируемого значения $\hat{y}$ временного ряда с использованием функции L имеет вид:

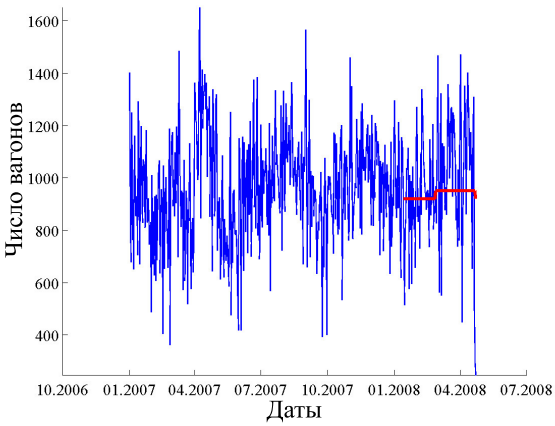
$$\hat{y} = \arg\min_{z \in \mathbb{N}} \sum_{k=1}^{K} L(z, y_k), \tag{3}$$

где связанная переменная z принадлежит множетсву натуральных чисел, так как требуется спрогнозировать число вагонов.

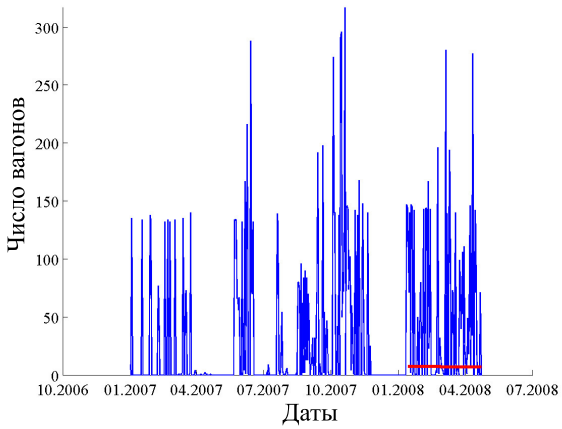
Пусть мы построили прогноз значений ряда $\{\hat{y_1}, \hat{y_2}..\hat{y_k}\}$ в некоторых точках, при этом реальные значения ряда x в этих же точках равны $\{y_1, y_2..y_k\}$. Эти значения не использовались при построении прогноза и нужны для получения ретроспективной оценки качества прогностической модели. Вычислим среднюю ошибку ретроспективного прогноза

$$\text{MeanError} = \frac{1}{k}\sum_{i=1}^{k} L(\hat{y_i}, y_i) \to \min_{\hat{y_1},\ldots,\hat{y_k}}. \tag{4}$$

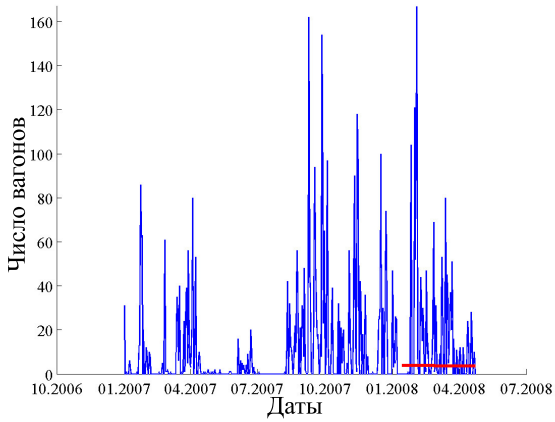
Чем меньше средняя ошибка, тем точнее прогноз. Таким образом, требуется решить задачу минимизации средней ошибки прогнозирования, то есть найти значения $\{\hat{y_1}, \hat{y_2}..\hat{y_k}\}$ такие, что значение MeanError минимально.



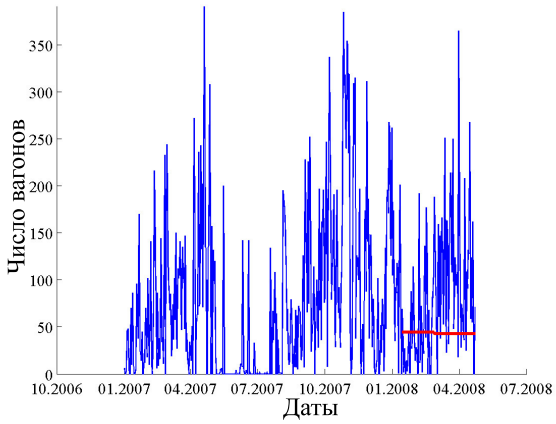
(a) На все ветки



(b) На ветку 96



(c) На ветку 97



(d) На ветку 98

Рис. 1: Графики временных рядов отправленного количества вагонов с нефтью с 83 ветки с прогнозом выполненным для промежутка в 16 месяцев.

3 Алгоритм прогнозирования

Первым шагом работы алгоритма является разбиение исходных данных для получения временных рядов. Для этого для каждого дня τ для всех возможных комбинаций типа груза G, веток отправления и прибытия L и A считается количество вагонов с грузом G, отправленных в день τ с ветки L на ветку A. При фиксированных значениях G, L, A получаем числовую зависимость количества вагонов от дня τ, то есть временной ряд x. Если при фиксированных значениях G, L просуммировать временные ряды для разных веток прибытия A, то получим временной ряд, соответствующий суммарному количеству отправленных грузов определенного типа с ветки L.

Вторым шагом работы является применение базового алгоритма к построенным на первом этапе временным рядам для построения прогнозов. Базовый алгоритм основан на алгоритме непараметрического прогнозированния временного ряда, описанного в [4]. Для построения гистограммы (1) для каждой точки i временного ряда $\mathbf{x}$ определим ее вес как произведение $w_i = w_i^F w_i^H$. Сомножитель w_i^F задает показательную весовую функцию

$$w_i^F = v^{\frac{-i+T+h}{F}} \in (0, 1], \tag{5}$$

убывающую к началу временного ряда и равную 1 в точке прогноза. Сомножитель w_i^H задан как

$$w_i^H = \begin{cases} K(i_H, PH), & \text{если } H > 0; \\ 1, & \text{если } H = 0, \end{cases} \tag{6}$$

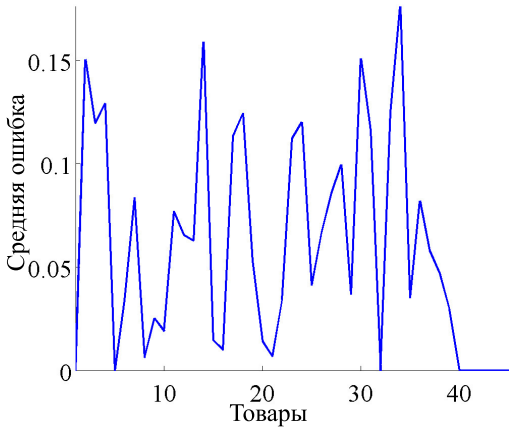
где индекс H вычисляется в результате решения оптимизационной задачи

$$i_H = \min_{n=0,\ldots,\mathrm{floor}\left(\frac{T+h}{P}\right)} |T + H - nP - i|. \tag{7}$$

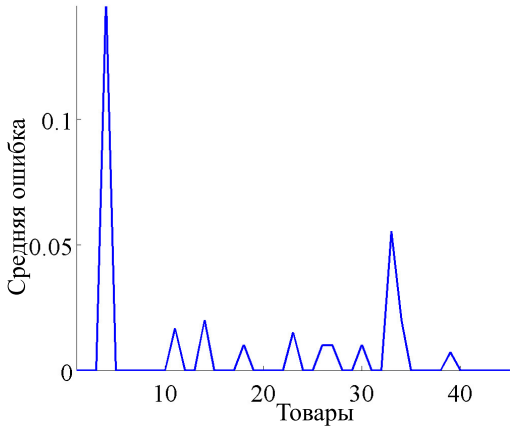
Эта формула задает вес i-той точки, соответствующий годовой сезонности. Ядро задается выражением:

$$K(x, z) = \begin{cases} \left(1 - \left(\frac{x}{z}\right)^2\right)^2, & \text{если } |x| < z; \\ 0, & \text{иначе.} \end{cases}$$

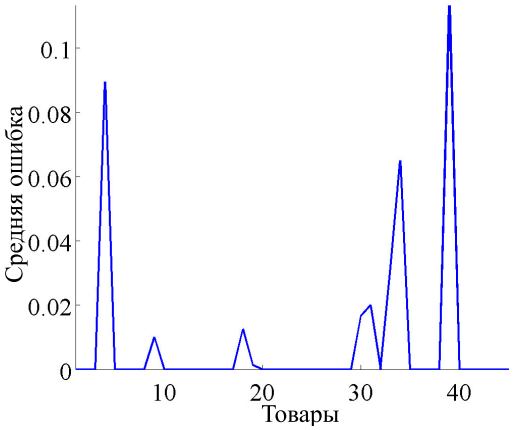
Взвешенные точки $x_i w_i$ используются для построения (1) гистограммы $\mathcal{H}$.



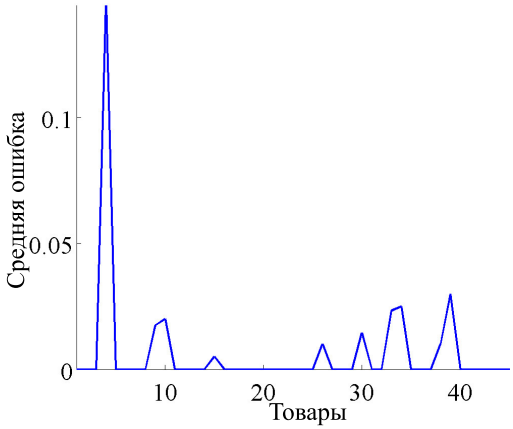
(a) Все ветки



(b) Ветка 96



(c) Ветка 97



(d) Ветка 98

Рис. 2: Средня ошибка прогноза для всех типов грузов.

Настраиваемые параметры: $v \in [0,1]$ в выражении (5) — параметр показательного взвешивания точек ряда, параметр «забывания»; $H \in [0, 0.5]$ в выражениях (6) и (7) — параметр ядра весовой функции для годовой сезонности, половина ширины «шапки» годовой сезонности.

Ненастраиваемые параметры: P в выражениях (6) и (7) — длина годового сезонного периода (обычно $P = 365$); $w^{\min}$ в выражении (8) — минимальный допустимый вес; F в выражении (5) — нормировочная константа «забывания». Предлагается выбрать F следующим образом $F = (T + H)\varepsilon \log_{10}(0.1), \quad \varepsilon = 10^{-3}$.

Выберем границы гистограммы, число столбцов и разбиение на столбцы следующим образом:

1) пусть n — число точек x_i, для которых $w_i > w^{\min}$;

2) выберем число столбцов (обоснование см. в [16]) $K = \lceil 3\sqrt[3]{n}\,\rceil$, если $K < 5$, то $K = 5$, если $K > 100$, то $K = 100$;

3) границы $y_1 = \min_{i:w_i>w^{\min}}(x_i)$, $y_k = \max_{i:w_i>w^{\min}}(x_i)$;

4) столбцы выбираются равной ширины.

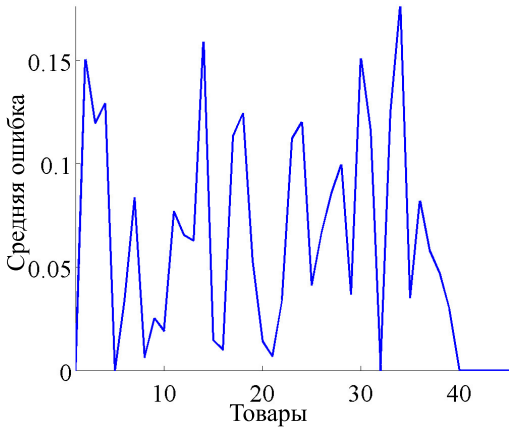
Для каждого $k = 1, \ldots, K$ высота столбца гистограммы g_k равна

$$g_k = \sum_{i=1}^{T} w_i [x_i \in y_k][w_i > w^{\min}], \tag{8}$$

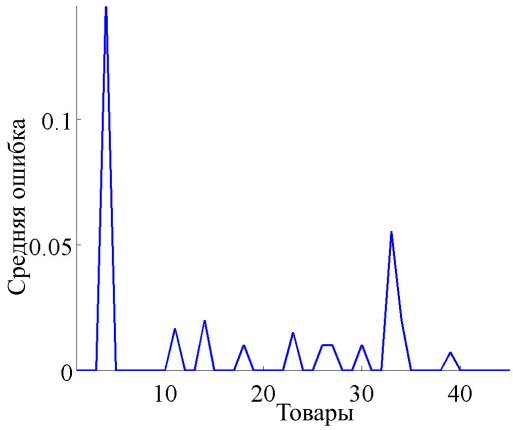
где выражение $[\cdot]$ равно 1, если в скобках стоит истинное логическое выражение, и 0 в противном случае.

Прогнозируемое значение ряда x_{T+h} находится как значение $\hat{y} \in \{y_1, \ldots, y_K\}$, соответсвующее оптимальному значению свертки распределения $\{g_k\}_{k=1}^{K}$ и функции потерь L:

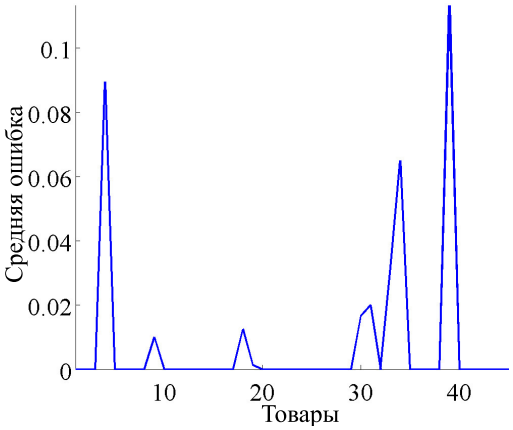
$$\hat{y} = \operatorname*{arg\,min}_{z \in \{y_1, \ldots, y_K\}} \sum_{k=1}^{K} g_k L(z, y_k). \tag{9}$$



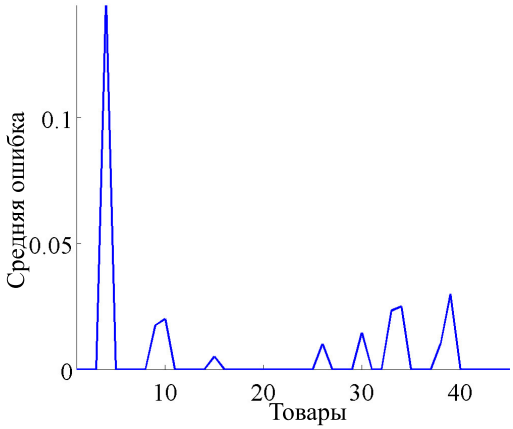
(a) Все ветки



(b) Ветка 96



(c) Ветка 97



(d) Ветка 98

Рис. 3: Средня ошибка прогноза для всех типов грузов.

4 Эксперимент

Цель эксперимента. Сделать прогноз для заданных комбинаций "пара веток-грузов", проверить, что разбиение данных по веткам повышает качество прогноза базового алгоритма.

В эксперименте использованы данные о посуточной загруженности железнодорожных узлов с 1 января 2007 года по 22 апреля 2008 года. В табл. 1 приведен пример записи.

Коды станций представляют собой шестизначные числа. Станции, в коде которых две первые цифры совпадают, входят в одну железнодорожную ветку. Станций отпраления 1566, станций назначения 1902, веток 99. Код груза — натуральное число от 1 до 43; также имеются перевозки, где код груза не указан. Род вагона — натуральное число, в имеющихся данных 75 различных типов вагонов.

Для экспериментов была выбраны данные об отправлении поездов с 83 ветки на все другие. Весь временной ряд разбивался так, что последние 100 дней образовывали контрольную выборку, а первые 300 – обучающую.

Было проведено два эксперимента, в которых обрабатывались одни и те же данные и выполнялись одни и те же алгоритмы, но в которых преследовались различные цели: в обоих исследовались данные о количестве грузов, отправленных с 83-й ветки на другие. Данные в обоих случаях были разбиты таким образом: временные ряды строились для количества вагонов с товарами, отправленных с 83-й ветки на другие конкретные ветки. Для них был запущен алгоритм и найдены значения прогноза и средней ошибки. В первом эксперименте исследовались данные об отправленных грузах на некоторые конкретные ветки для всех видов грузов (при этом графики отображающие некоторые зависимости представлены только для некоторых грузов), а также о суммарном количестве различных отправленных грузов на все ветки – задачей было визуализировать зависимость прогноза и ошибок для разных рядов. Второй эксперимент был проведен для всех видов грузов, отправленных на все ветки – целью было для каждого вида груза вычислить среднее и максимальное значения усредненной ошибки по всем веткам, и сравнить их со средней ошибкой для суммарного количества отправленных грузов для этого типа грузов.

Дата погрузки	Станция отправления	Станция назначения	Количество вагонов	Код груза	Род вагона	Суммарный вес груза	Признак маршрутной отправки
2007-01-01	020108	932902	1	1	216	56	9

Таблица 1: Вид записи базы данных железнодорожных перевозок

Тип груза	1	2	3	4	5	6	7	8
Минимальная ошибка	0,001	0,005	0,003	0,010	0,001	0,001	0,020	0,010
Средняя ошибка	0,003	0,010	0,010	0,040	0,002	0,009	0,050	0,013
Максимальная ошибка	0,010	0,150	0,130	0,190	0,006	0,040	0,080	0,030
Ошибка базового алгоритма	0,004	0,150	0,120	0,130	0,009	0,030	0,080	0,010

Таблица 2: Минимальная, средняя и максимальная ошибка по всем веткам в сравнении с ошибкой базового алгоритма.

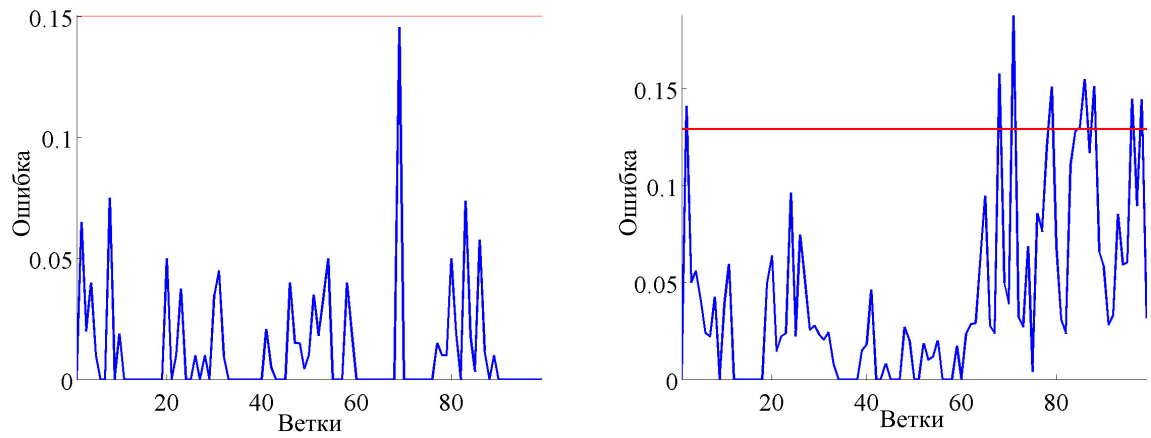
Перейдем к описанию полученных результатов. Начнем с первого эксперимента. На рис. 1 изображены графики временных рядов — с датой по оси X, и значением (количеством отправленных грузов) по оси Y. Графики приведены как для полного количества отправленных вагонов, так и для отправленных вагонов на некоторые конкретные ветки. Ветки выбирались таким образом, чтобы отразить различные варианты характеристик временных рядов.

Эксперименты также были проведены для всех грузов, отправленных с 83-й ветки. Для каждого вычисления (то есть каждой отдельной ветки прибытия груза, а также для суммарного количества отправленных грузов) была найдена средняя относительная ошибка (4). На рис. 3 представлены результаты этих вычислений — график 2(a) для полного количества отправленных грузов, и графики для трех различных веток. На них отображена зависимость средней относительной ошибки от вида груза.

Для сравнения результатов прогнозирования с базовым алгоритм (без разбивки по веткам) предлагается следующее. Для некоторых типов грузов были проделаны вычисления: найдена средняя ошибка при прогнозе количества вагонов с нефтью,

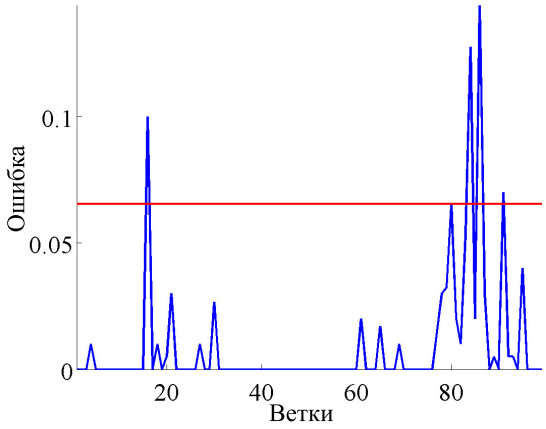
отправленных с 83-й ветки на все другие, а также средняя ошибка при прогнозе суммарного количества вагонов с нефтью, отправленных с 83-й ветки. Результаты этих вычислений представлены в виде графиков – зависимости ошибки от ветки. В качестве ориентира горизонтальной линией отображен уровень ошибки для суммарного количества грузов (красная линия). Результаты представлены на рис. 4.

Далее, рассмотрим результаты второго эксперимента. Для каждого вида груза вычислена среднее, минимальное (ненулевое) и максимальное значение усредненной ошибки по всем веткам прибытия, построены графики зависимости этих значений от вида груза (изображение 5 – красным цветом нарисована ошибка для суммарного количества отправленных грузов в зависимости от вида груза, зеленым, синим и черным – минимальная, усредненная и максимальная ошибка по всем веткам). Для некоторых грузов эти данные выписаны в таблице 2.

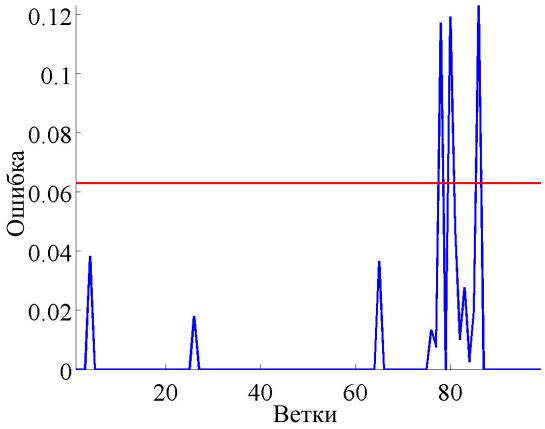


(a) Груз 2

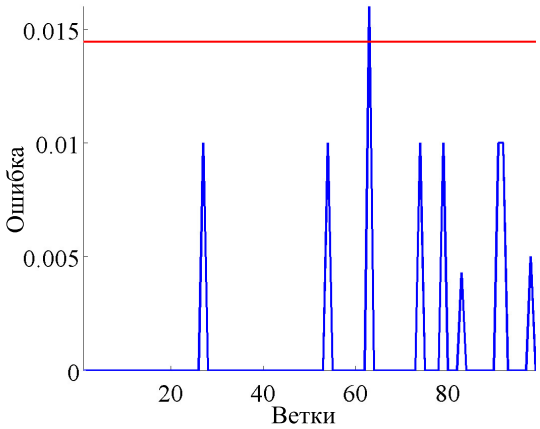
(b) Груз 4



(c) Груз 12



(d) Груз 13



(e) Груз 15

Рис. 4: Сравнение средней ошибки при прогнозе с узла на узел (синим) со средней ошибкой при прогнозе суммарного отправления вагонов с одного узла района на все узлы

5 Оценка результатов

На рис. 3 показано, что средняя относительная ошибка прогноза для разных веток почти всегда меньше, чем ошибка прогноза для суммарного количества вагонов. Это означает, что прогноз количества отправленных грузов на определенную ветку в среднем точнее, чем прогноз суммарного количества вагонов. Однако, есть и грузы, для которых средняя ошибка прогноза для некоторой конкретной ветки превосходит ошибку прогноза суммарного количества отправленных вагонов.

Это же подтверждают графики на рис. 4, которые дают более подробную информацию о средней ошибке прогнозирования для некоторых грузов. В целом, средняя относительная ошибка для веток меньше, чем уровень ошибки для суммарного количества. Тем не менее, есть и значения ошибок, большие уровня ошибки для суммарного количества. С другой стороны, видно, что веток, для которых средняя ошибка прогноза превосходит ошибку для прогноза суммарного количества грузов, немного (таких нет для груза 2, и 5-10 для остальных рассмотренных типов грузов).

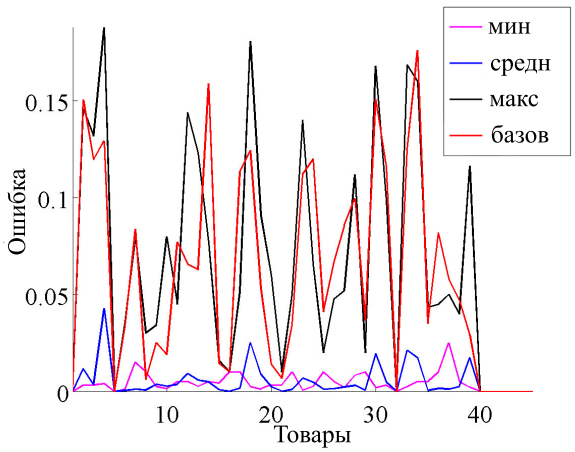


Рис. 5: Минимальная, средняя и максимальная ошибка по всем веткам в сравнении с ошибкой базового алгоритма.

Результаты второго эксперимента на рис. 5, табл. 2 также подтверждают тот факт, что средняя ошибка для прогноза количества грузов отправленных на конкретные ветки меньше, чем такая же ошибка для прогноза суммарного количества отправленных грузов. Однако средняя ошибка в некоторых случаях меньше минимальной — минимум брался по всем ненулевым значениям, иначе он был бы равен нулю для всех грузов, поскольку для каждого груза есть ветка, на которую такой

груз с 83 ветки не прибывает в течение рассматриваемого периода. Во-первых, рис. 5 наглядно демонстрирует, что усредненная по всем веткам подсчитанная ошибка существенно меньше, чем ошибка для суммарного количества грузов. Во-вторых, максимальная ошибка по веткам если и превосходит ошибку для суммы, то ненамного — более того, можно заметить что они зачастую совпадают или очень близки.

6 Заключение

Поставлен вычислительный эксперимент, который подтвердил предположение о том, что прогноз для отдельных веток точнее, чем прогноз суммарного количества отправленных вагонов. Приведенные результаты подтверждают выдвинутое предположение. Дальнейшие исследования данной задачи могут быть продолжены в следующих направлениях: во-первых, для прогноза количества отправленных грузов можно попытаться в рассматриваемых временных рядах выделять тренд и, если возможно, искать периодическую компоненту. Во-вторых, для построения прогноза предлагается применить различные варианты алгоритма. Например, использовать вариацию со скользящим средним.

Список литературы

[1] Зацаринный А.А., Шабанов А.П. Методический подход к управлению качеством информации в сложных инфокоммуникационных проектах // Системы и средства информатики, 2011. Т. 21, № 2. С. 3–20.

[2] Синицын И.Н., Шаламов А.С., Сергеев И.В., Синицын В.И., Корепанов Э.Р., Белоусов В.В., Агафонов Е.С., Шоргин В.С. Методы и средства анализа и моделирования стохастических систем интегрированной логистической поддержки // Системы и средства информатики, 2012. Т. 22, № 2. С. 3–28.

[3] Синицын И.Н., Корепанов Э.Р., Белоусов В.В., Шоргин В.С., Макаренкова И.В., Конашенкова Т.Д., Агафонов Е.С., Семендяев Н.Н. Развитие компьютерной поддержки статистических научных исследований систем высокой точности и доступности // Системы и средства информатики, 2011. Т. 21, №1. С. 3–33.

[4] Вальков А.С., Кожанов Е.М., Медведникова М.М., Хусаинов Ф.И. Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных, 2012. Т. 1, № 4. С. 448-465.

[5] Постникова Е. Квантильная регрессия // Новосибирск: НГУ, 2006.

[6] Koenker Jr., Bassett G. Regression Quantiles // Econometrica, 1978. Vol. 46(1) P. 33–50.

[7] Cortez P., Rotcha M., Neves J. Evolving time series forecasting ARMA models // Journal of Heuristics, 2004. Vol. 10(4) P. 419–429.

[8] Shumway R. H., Stoffer D. S. Time Series Analysis and Its Applications With R Examples // Springer, 2006.

[9] Тихонов Э.Е. Прогнозирование в условиях рынка. 2006.

[10] Джефри Р. Непараметрическая эконометрика: вводный курс // Машинное обучение и анализ данных, 2013. Т.1, № 5. С. 505–518.

[11] Хардле В. Прикладная непараметрическая регрессия // М: Мир, 1993.

[12] Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов // М: Финансы и статистика, 2003.

[13] Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз // М: Финансы и статистика, 2000.

[14] Gheyas I. A., Smith L. S. Neural Network Approach to Time Series Forecasting. Proceedings of the World Congress on Engineering, 2009. Vol. 2. P. 245–253.

[15] Thiesing F. M., Vornberger O. Sales Using Neural Networks // Lecture Notes in Computer Science, 1997. Vol. 1226. P. 321–328.

[16] Scott D. W. On optimal and data-based histograms // Biometrika, 1979. Vol. 66(3). P. 605—610.

[17] Мотренко А.П., Вальков А.С., Хусаинов Ф.И, Кожанов Е.М. Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла // Машинное обучение и анализ данных, 2013. Т. 1, № 5. С. 505–518.

[18] Кузнецов М.П., Мафусалов А.А., Животовский Н.К., Зайцев Е., Сунгуров Д.С. Сглаживающие алгоритмы прогнозирования // Машинное обучение и анализ данных, 2011. Т. 1, № 1. С. 104–112.

[19] Фирстенко А.Н., Кононенко Д.С., Кузнецов М.П., Морозов А.А., Сунгуров Д.С., Савинов Н.А., Корниенко А.И., Джамтырова Р.Б., Ивкин Н.П., Зайцев Е., Животовский Н.К., Кононенко Д.С., Быстрый Р.Б. Технологические карты разработки библиотеки алгоритмов прогноза временных рядов // Машинное обучение и анализ данных, 2011. Т.1, № 1. С. 113–121.