Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным*

Вальков А. С.[1]***, Кожанов Е. М.***[2]***, Медведникова М. М.***[3]***, Хусаинов Ф. И.***[4]
valkov@forecsys.ru, vinger4@gmail.com, medvmasha@rambler.ru, f-husainov@yandex.ru
1 — Вычислительный центр РАН, 2 — Московский государственный технический университет имени Н. Э. Баумана, 3 — Московский физико-технический институт, 4 — Российская открытая академия транспорта Московского государственного университета путей сообщения (МИИТ)

Предложен алгоритм непараметрического прогнозирования загруженности железнодорожных узлов РЖД по историческим данным. Алгоритм основан на свертке эмпирической плотности распределения значений временного ряда с функцией потерь. В работе исследуются свойства авторегрессионной прогностической модели. Алгоритм проиллюстрирован данными загруженности железнодорожных узлов Омской области за 2007 и 2008 годы.

Ключевые слова: *временные ряды, прогнозирование, загрузка железнодорожного узла, непараметрический метод, эмпирическое распределение.*

Valkov A. S.[1]***, Kozhanov E. M.***[2]***, Medvednikova M. M.***[3]***, Husainov F. I.***[4]
1 — Computing Center of the Russian Academy of Sciences; 2 — Bauman Moscow State Technical University; 3 — Moscow Institute of Physics and Technology; 4 — Moscow State University of Railway Engineering

The authors propose a method of non-parametric forecasting of railroad stations occupancy according to historical data. The algorithm is based on convolution of empirical density of distribution of time series values and loss function. The features of autoregressive prognostic model are investigated. The algorithm is illustrated by railroad stations occupancy data in Omsk region in 2007 and 2008.

Keywords: *time series, forecasting, railroad station occupancy, non-parametric method, empirical distribution.*

Введение

Прогнозирование потребностей в вагонах у заказчиков РЖД в узлах погрузки/разгрузки с учетом временных интервалов доставки, а также использование загруженности железнодорожных узлов является проблемой, которую необходимо решить для повышения эффективности транспортировки грузов. Данная работа посвящена решению задачи прогнозирования загруженности железнодорожных узлов. Прогноз выполняется на основании исторических знаний о приходящих на станцию и уходящих со станции вагонах. При этом в качестве единицы учета рассматривается блок вагонов, неделимый на всем протяжении маршрута.

Используется формат данных, образец которых представлен в табл. 1. Каждая строка (запись в базе данных) содержит информацию о дате погрузки, станции отправления, станции назначения, количестве вагонов, которые прошли по маршруту от станции

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-13154.

отправления до станции назначения, коде груза, роде вагонов, суммарном весе груза и признаке маршрутной отправки.

Согласно используемым данным железнодорожный узел рассматривается без детализации по путям и по очередности отправления блоков вагонов. Для прогноза не используются внешние данные.

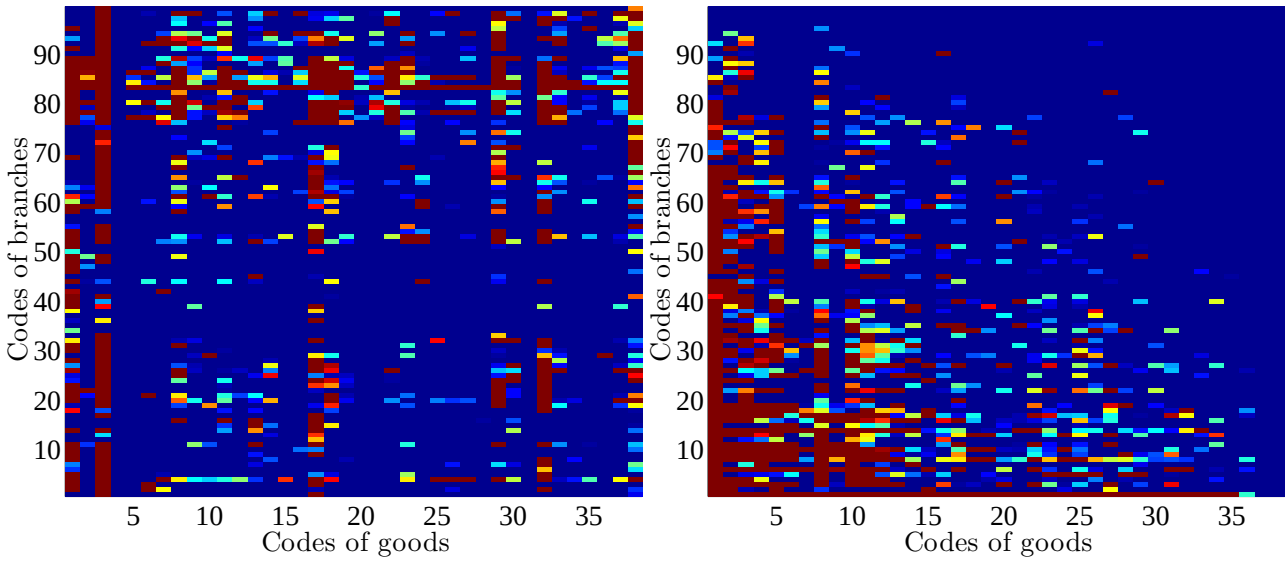
Для решения задачи требуется сформировать прогноз отправления/погрузки грузов в заданном периоде:

1) на месяц посуточно;
2) на месяц подекадно;
3) на квартал помесячно;
4) на год помесячно;
5) на год поквартально;
6) на период больше года;

и прогноз отправления/погрузки грузов с разложением:

1) по группам грузов;
2) по родам подвижного состава;
3) по комбинированному разложению, учитывающему перечисленные варианты.

Для получения адекватного прогноза должна быть решена задача определения рационального уровня детализации прогноза (по станции или по группе станций). Существенные внешние ограничения возникают в связи с тем, что в одном составе перемещаются вагоны, принадлежащие различным собственникам, и с тем, что возникают запреты на движение товарных поездов из-за необходимости обеспечения возможности высокоскоростного передвижения.



(a) Без сортировки (b) С сортировкой

Рис. 1: Суммарная матрица перевозок

Используемый в данной работе алгоритм основан на алгоритме квантильной регрессии [10, 11], модифицированным сверткой гистограммы с функцией потерь. Главное преиму-

щество такого подхода заключается в возможности учета стоимости ошибки прогнозирования в прикладной задаче.

Основные методы непараметрической регрессии, такие как ядерное сглаживание, сглаживание сплайнами, авторегрессия, скользящее среднее и др., описаны в [1, 2, 3, 4]. Они заключаются в присвоении имеющимся значениям временного ряда некоторых весов и комбинации взвешенных значений для получения прогноза. Также для решения подобных задач применяют нейронные сети [8, 9].

Для построения прогностической модели предлогается использовать непараметрические методы прогнозирования. В частости, предполагая временной ряд локально-стационарным (выполнено условие Дики-Фуллера [4]), предлагается построить гистограмму распределения его значений и вычислить свертку гистограммы с экспертно заданной функцией потерь для каждого возможного прогнозируемого значения. Оптимальным прогнозом является то значение центра сегмента гистограммы, которое доставляет минимальное значение свертки. Также проверяется применимость на практике данной модели к прогнозированию нестационарных временных рядов.

В качестве базового алгоритма для сравнения полученных прогнозов используется модель авторегрессионного скользящего среднего ARMA [5, 6, 7].

Алгоритм непараметрического прогнозирования временного ряда

Задан 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 временного ряда $\mathbf{x}$ так, чтобы выполнялось условие оптимальности свертки гистограммы, построенной по значениям временного ряда, и функции потерь. Предполагается, что ряд стационарен, иными словами, распределение прогнозируемого значения в точке равно распределению точек заданного временного ряда.

Для построения прогностической модели используются элементы квантильной регрессии. По временному ряду $\mathbf{x}$ построим гистограмму $\mathcal{H}$ — набор пар

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

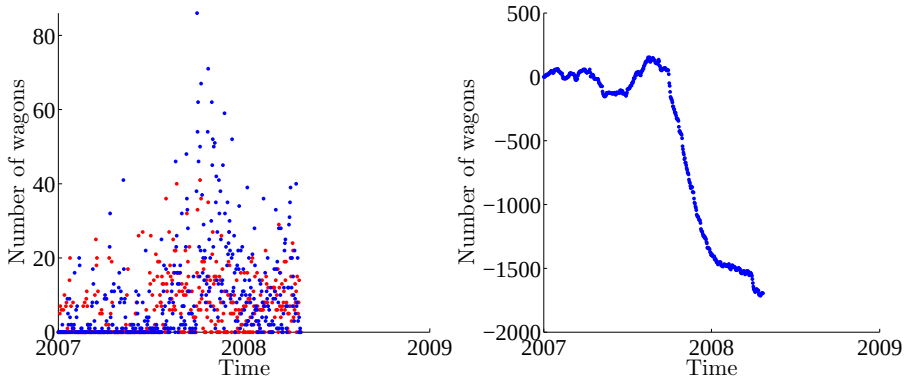
где K — число интервалов $[y_k^{\min}, y_k^{\max}]$ со средним значением y_k, на которые разбита ось значений ряда, g_k — высота столбца гистограммы на интервале y_k, которая равна взвешенной сумме количества точек ряда, попавших в этот интервал. В алгоритме квантильной регрессии [10, 11] для прогноза используется значение y_k, соответствующее самому высокому столбцу (моде) гистограммы. В данной работе предлагается модификация алгоритма квантильной регрессии.

Введем функцию потерь $L(\hat{y}, y)$ — штраф за несоответствие прогнозируемого значения $\hat{y}$ историческому значению y. Далее будем использовать одну из трех функций потерь:

1) $L(z, x) = (z - x)^2$;
2) $L(z, x) = |z - x|$;
3) $L(z, x) = \begin{cases} 0, & \text{если } |z - x| < a; \\ |z - x| - a, & \text{если } |z - a| \geqslant a, \text{ где } a > 0 \text{ — экспертно заданный параметр.} \end{cases}$

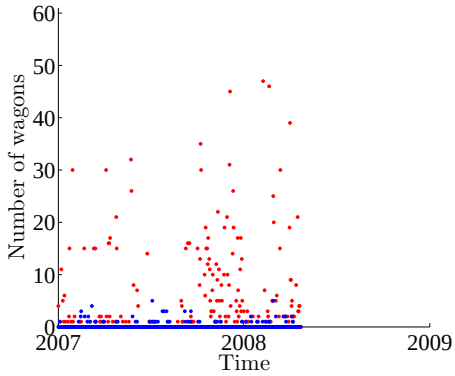
Построение гистограммы. Для каждой точки i временного ряда $\mathbf{x}$ определим ее вес как произведение $w_i = w_i^F w_i^H$. Сомножитель w_i^F задает показательную весовую функцию

$$w_i^F = v^{\frac{-i+T+h}{F}} \in (0, 1], \tag{2}$$

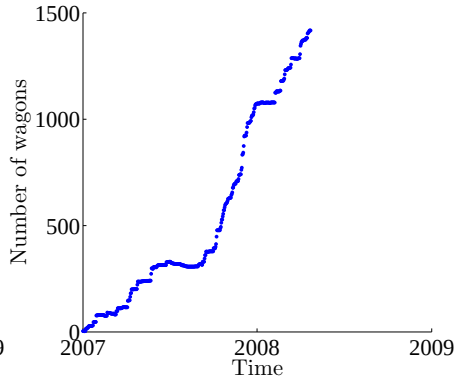


(a) Ветка 0

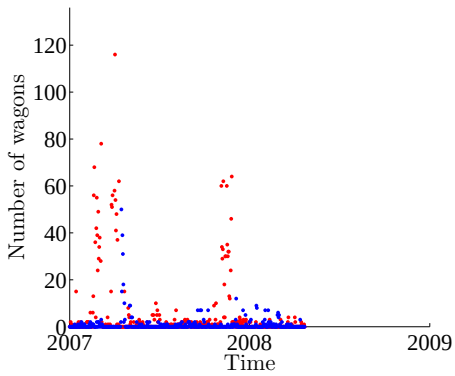
(b) Ветка 0



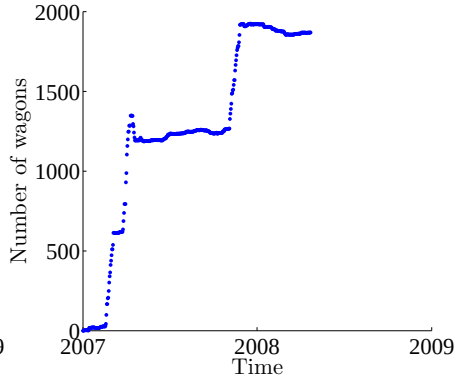
(c) Ветка 6



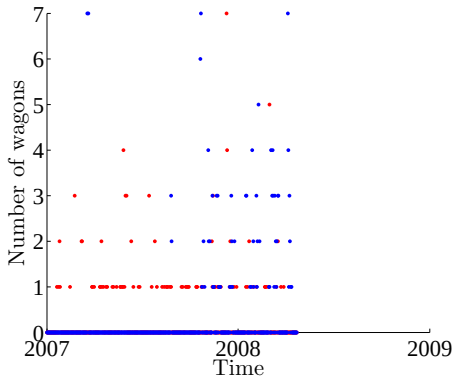
(d) Ветка 6



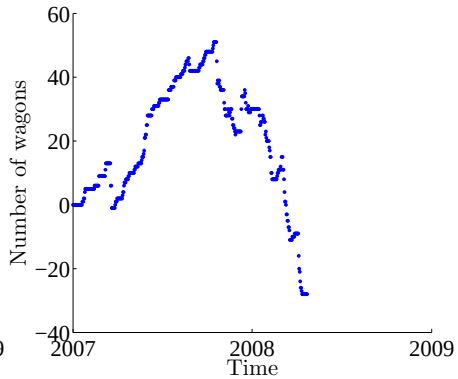
(e) Ветка 24



(f) Ветка 24



(g) Ветка 75



(h) Ветка 75

Рис. 2: Прибывшие (красный) и отправленные (синий) вагоны и число вагонов на ветке

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

$$w_i^H = \begin{cases} K(i_H, PH), & \text{если } H > 0; \\ 1, & \text{если } H = 0, \end{cases} \tag{3}$$

где индекс H вычисляется в результате решения оптимизационной задачи

$$i_H = \min_{n=0,\ldots,\mathrm{floor}\left(\frac{T+h}{P}\right)} |T + H - nP - i|. \tag{4}$$

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 гистограммы $\mathcal{H}$ (1).

Настраиваемые параметры: $v \in [0, 1]$ в выражении (2) — параметр показательного взвешивания точек ряда, параметр «забывания»; $H \in [0, 0.5]$ в выражениях (3) и (4) — параметр ядра весовой функции для годовой сезонности, половина ширины «шапки» годовой сезонности.

Ненастраиваемые параметры: P в выражениях (3) и (4) — длина годового сезонного периода (обычно $P = 365$); $w^{\min}$ в выражении (5) — минимальный допустимый вес; F в выражении (2) — 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 [12]) $K = \lceil 3\sqrt[3]{n}\,\rceil$, если $K < 5$, то $K = 5$, если $K > 100$, то $K = 100$;
3) границы $y_1 = \min\limits_{i:w_i > w^{\min}}(x_i)$, $y_k = \max\limits_{i:w_i > w^{\min}}(x_i)$;
4) столбцы выбираются равной ширины.

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

$$g_k = \sum_{i=1}^{T} w_i [x_i \in y_k][w_i > w^{\min}], \tag{5}$$

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

Алгоритм непараметрического прогнозирования. Полученная гистограмма $\mathcal{H}$ используется для построения прогноза. 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:

$$\hat{y} = \arg\min_{z \in \{y_1, \ldots, y_K\}} \sum_{k=1}^{K} g_k L(z, y_k). \tag{6}$$

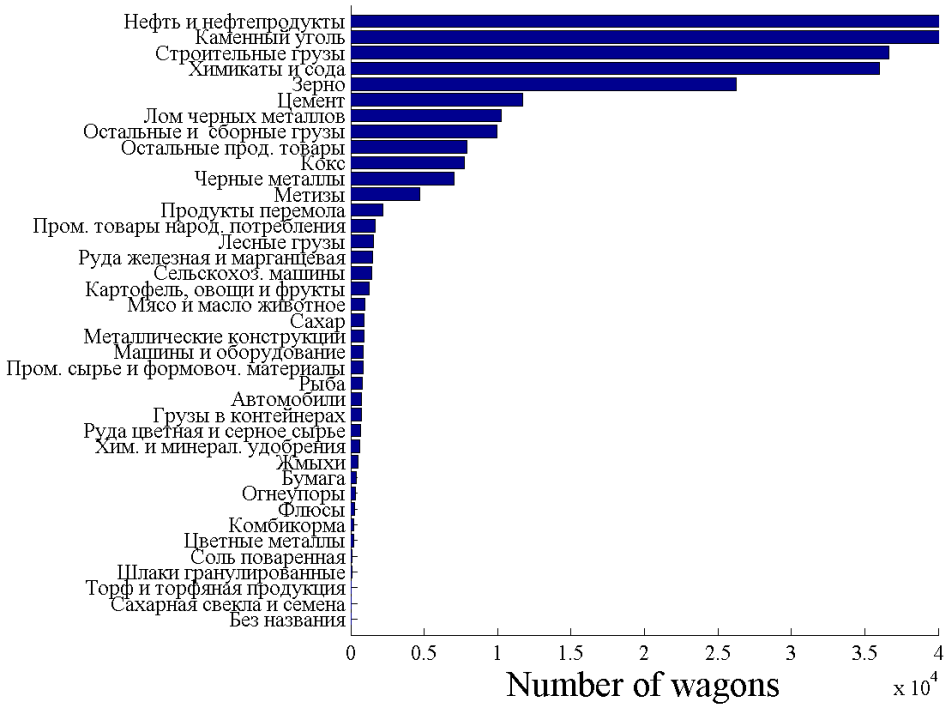


Рис. 3: Вагоны с разными типами грузов

Тест Дики-Фуллера на стационарность временного ряда является одним из тестов на единичные корни. Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд $I(0)$:

$$\Delta x_i = x_i - x_{i-1} \sim I(0).$$

При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1):

$$x_i = ax_{i-1} + \varepsilon_i,$$

где ε — ошибка. Если $a = 1$, то процесс имеет единичный корень и ряд $\mathbf{x}$ не стационарен. Если $|a| < 1$, то ряд стационарный. Приведенное авторегрессионное уравнение можно переписать в виде

$$\Delta x_i = bx_{i-1} + \varepsilon_i,$$

где $b = a - 1$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы $b = 0$ против альтернативы $b < 0$. Статистика теста (DF-статистика) — t-статистика для проверки значимости коэффициентов линейной регрессии $y = bx$:

$$t = \frac{\hat{b} - b}{S_b},$$

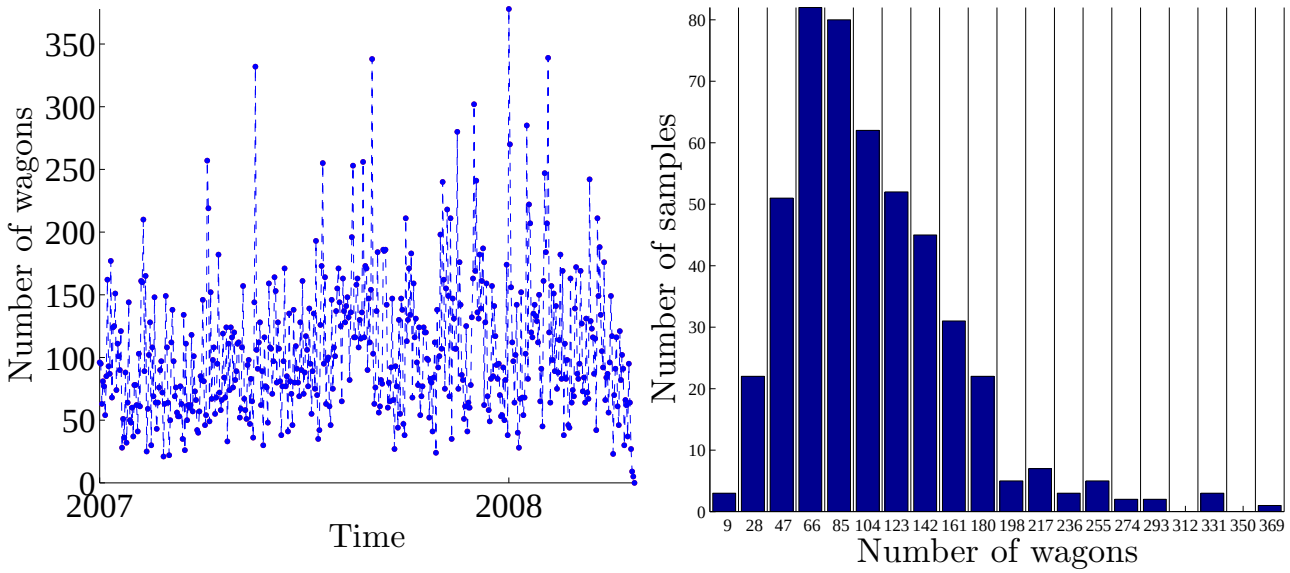
где $\hat{b}$ — оценка коэффициента регресии по выборке и

$$S_b = \frac{S}{S_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{b}x_i)^2;$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Распределение DF-статистики выражается через винеровский процесс и называется распределением Дики-Фуллера [4].

Считаем выполнение этого теста необходимым для применения прогностической модели.



(a) Временной ряд (b) Гистограмма

Рис. 4: Прибытие вагонов с нефтью и нефтепродуктами

Базовый алгоритм прогнозирования

В качестве базового алгоритма для проверки полученных результатов используется авторегрессионное скользящее среднее (ARMA) [5, 6, 7]. Пусть, как и ранее, задан временной ряд $\mathbf{x} = \{x_i\}_{i=1}^{T}$. Авторегрессионная модель (AR):

$$x_i = c + \sum_{\tau=1}^{p} \varphi_\tau x_{i-\tau} + \varepsilon_i,$$

где $\varphi_1, \ldots, \varphi_p$ — параметры модели, c — константа, ε_i — шум. Модель скользящего среднего (MA):

$$x_i = \mu + \varepsilon_i + \sum_{\tau=1}^{q} \theta_\tau \varepsilon_{i-\tau},$$

где $\theta_1, \ldots, \theta_q$ — параматры модели, μ — математическое ожидание x_i, $\varepsilon_i, \varepsilon_{i-1}, \ldots$ — шумы. Модель ARMA:

$$x_i = c + \varepsilon_i + \sum_{\tau=1}^{p} \varphi_\tau x_{i-\tau} + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j}.$$

Шумы ε_i обычно принимают независимыми одинакого нормально распределенными случайными величинами с нулевым математическим ожиданием: $\varepsilon_t \in \mathcal{N}(0, \sigma^2)$.

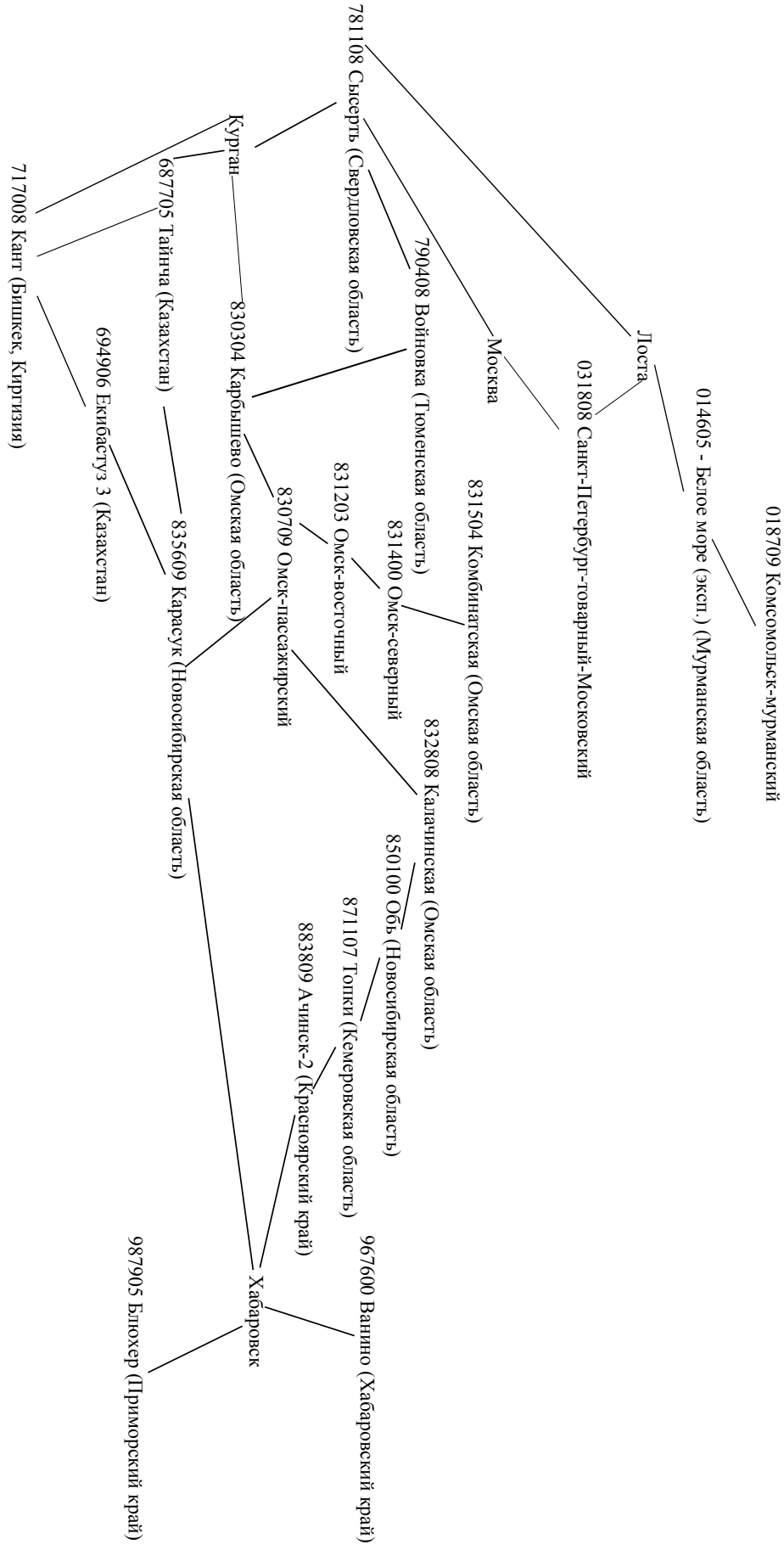


Рис. 5: Граф станций

Оптимизация параметров авторегрессионной модели описана в [7].

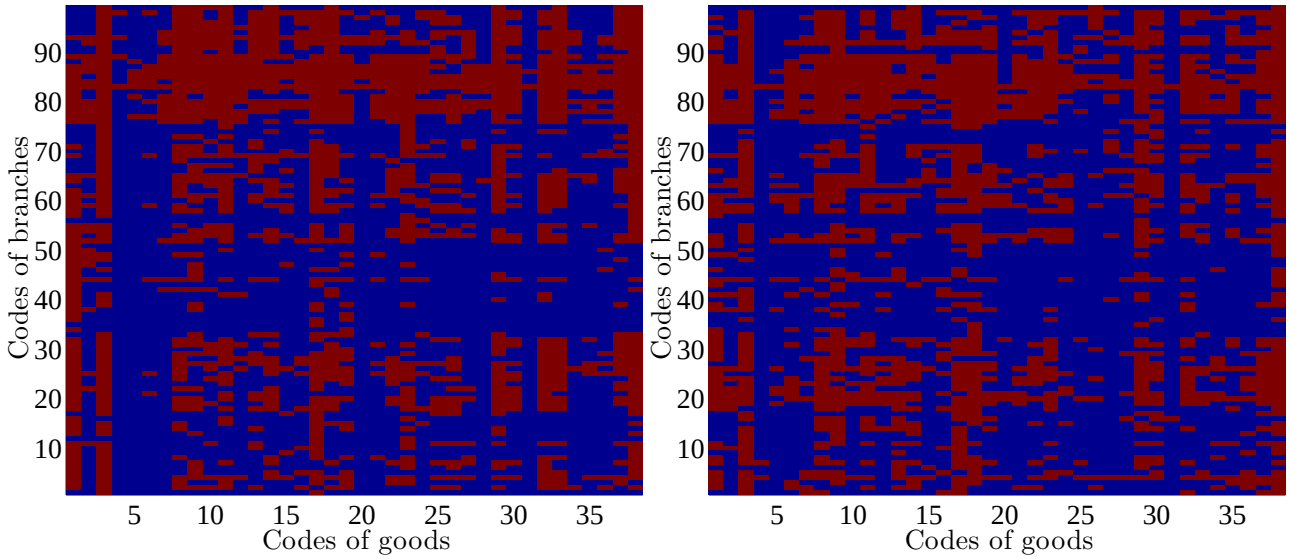
Ретроспективный прогноз и оценка качества прогностической модели

В ходе вычислительного эксперимента составлялся прогноз прибытия и отправления вагонов с различными типами грузов на день, неделю и месяц по всем железнодорожным веткам. В алгоритме непараметрического прогнозирования для построения гистограммы использовались H точек, предшествующие точке (интервалу) прогноза. Окно в H точек перемещалось по временному ряду с шагом в одну точку с построением прогноза для каждого шага.

Для оценки качества прогностической модели использовалась средняя доля ошибки:

$$MAPE = \frac{\sum_{i=1}^{h} \frac{|\hat{y}-y_i|}{y_i}}{h}, \tag{7}$$

где $\hat{y}$ — полученное при прогнозе значение, h — горизонт отсрочки прогноза. Затем ошибка прогнозирования усреднялась по всем интервалам. Полученные результаты сравнивались с результатами модели ARMA.



(a) Прибытие (b) Отправление

Рис. 6: Тест Дики-Фуллера для временных рядов

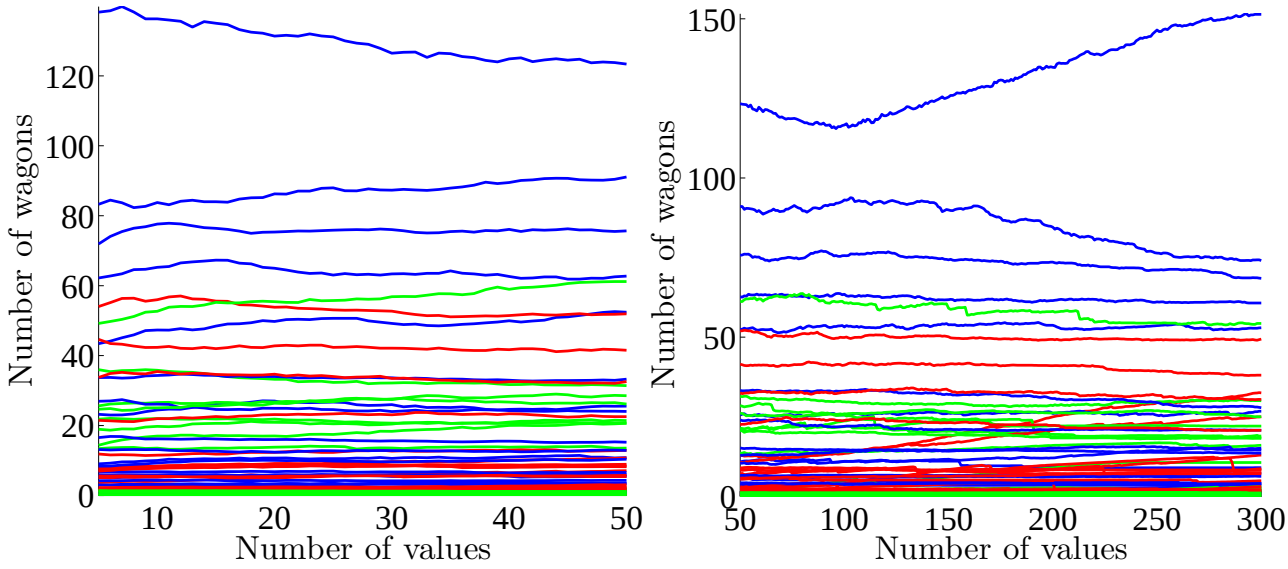
Вычислительный эксперимент

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

Таблица 1: Вид записи базы данных железнодорожных перевозок

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

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. Поскольку имеющихся данных недостаточно, для того чтобы проследить годовую периодичность временных рядов, то в ходе эксперимента наличие периодичности не учитывалось.



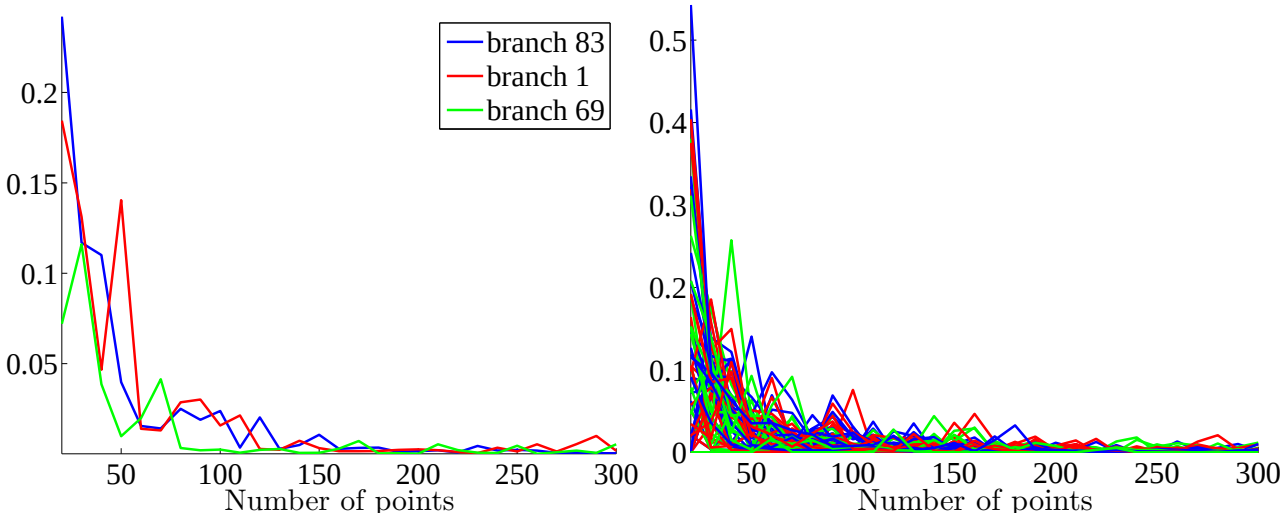
(a) Малое количество точек (b) Большое количество точек

Рис. 7: Ошибки прогнозирования

На рис. 1 изображена матрица перевозок. По горизонтали отложены коды грузов, по вертикали — номера железнодорожных веток. Цвет каждой ячейки соответствует числу вагонов с данным типом груза, прошедших за все время наблюдения через данную ветку. Большим значениям соответствуют ячейки красных оттенков, малым — синих оттенков. На рис. 1(a) показана матрица с неотсортированными столбцами и строками. Коды грузов и коды веток совпадают с отмеченными на осях числами. На рис. 1(b) строки и столбцы отсортированы по убыванию суммы их элементов (суммируются значения в столбцах, затем суммы сортируются по убыванию и переставляются столбцы; затем та же операция проводится со строками), и коды грузов и веток не совпадают с числами на осях.

Рис. 2 показывает посуточное перемещение всех типов вагонов по четырем веткам. По оси абсцисс отложены даты, по оси ординат — количество вагонов. На графиках в левом столбце красными точками отмечено количество прибывших на ветку вагонов в

течение суток, синими — число отправленных с ветки за тот же период. В правом столбце изображено количество оставшихся на ветке в течение суток вагонов в предположении, что изначально на ветке вагонов не было (этим объясняется возможность отрицательного числа вагонов). Графики показывают, что через разные ветки проходит различное число вагонов. Динамика числа вагонов на разных ветках также различна. Их число может возрастать, убывать или не иметь постоянного тренда. Причем резкие скачки на графике справа соответствуют пикам синего или красного цвета на графике слева, в зависимости от того, уменьшается или увеличивается количество вагонов.



(a) Ветки 1, 69, 83

(b) Все ветки

Рис. 8: Стабилизация гистограммы, прибытие вагонов с нефтью

На рис. 3 изображена гистограмма числа вагонов с разными типами грузов, прошедших через все станции полигона за все время наблюдения. По вертикали отмечены названия грузов, по горизонтали — количество вагонов. Самые большие столбцы, соответствующие перевозкам нефти и нефтепродуктов и каменного угля обрезаны на значении 40 000, чтобы более короткие столбцы были видны на диаграмме.

На рис. 4 изображен временной ряд и гистограмма прибытия на 83 ветку вагонов с нефтью и нефтепродуктами. По оси абсцисс гистограммы отложено число вагонов, по оси ординат — количество наблюдений, соответсвующее числу вагонов в интервале. По оси абсцисс временного ряда отложена дата, по оси ординат — число вагонов, пришедших на ветку за сутки.

На рис. 5 показана топология самых загруженных станций. Цифры обозначают коды станций, между соединенными станциями есть сообщение.

Поскольку стационарность временных рядов считается необходимой для использования рассматриваемой прогностической модели, для всех рядов был проведен тест на стационарность. На рис. 6 показаны результаты теста Дики-Фуллера для временных рядов для каждой ветки и каждого типа груза. Красным обозначены ряды, не прошедшие тест, синим — прошедшие. Использовалась реализация теста в среде MatLab. В ходе эксперимента предлагаемая в данной работе непараметрическая модель использовалась для прогноза всех рядов с целью проверки ее применимости на практике к нестационарным рядам.

Выбор параметров модели. Для построения прогностической модели выбиралась длина предыстории, используемая для построения гистограммы, и функция потерь для свертки.

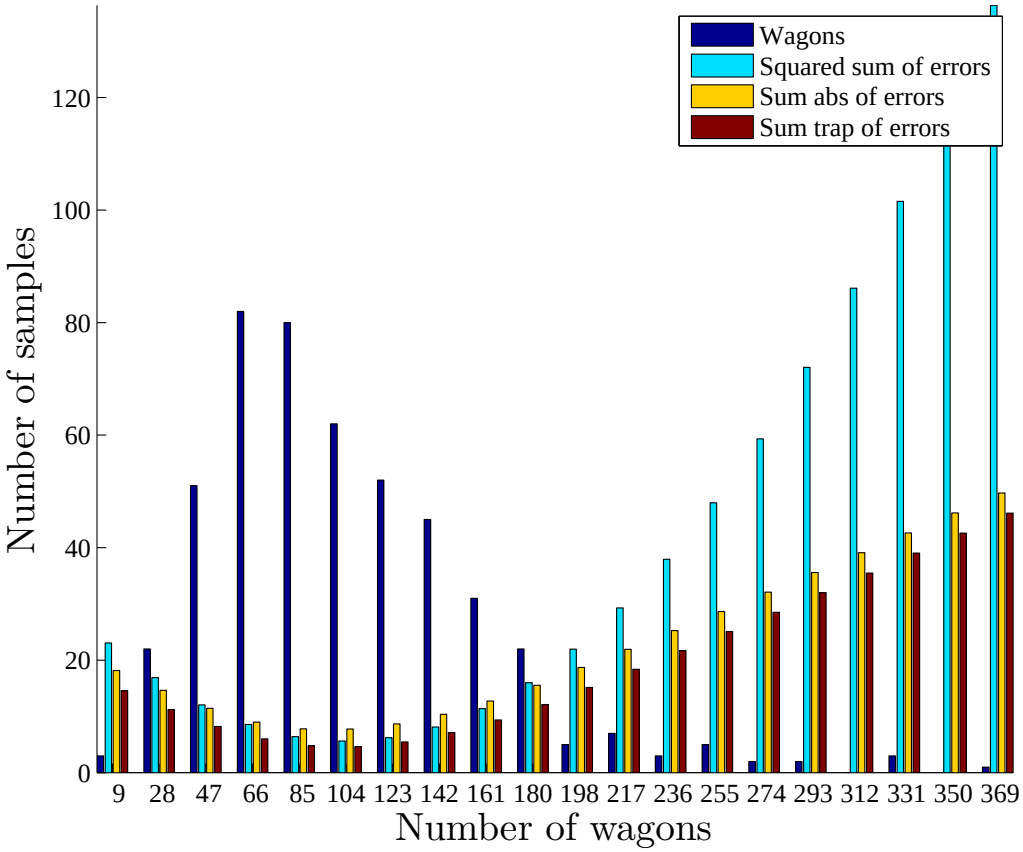


Рис. 9: Свертка гитограммы с различными функциями потерь

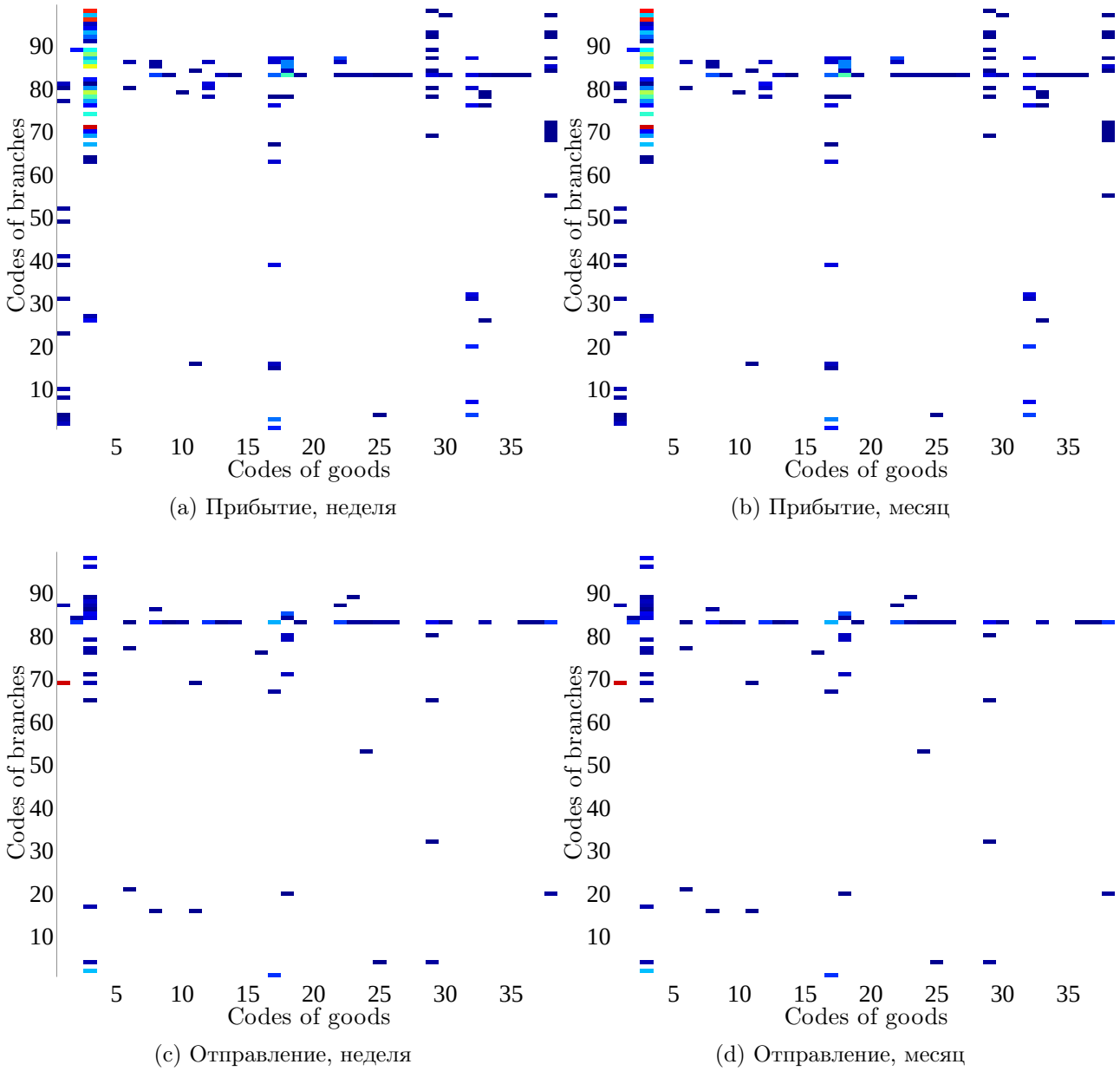
Для выбора длины предыстории были использованы два способа. В первую очередь исследовалась зависимость средней ошибки прогнозирования от длины предыстории. Результаты исследования представлены на рис. 7. По оси абсцисс графиков отложено количество точек, использованных для построения гистограммы, по оси ординат — средняя ошибка в количестве вагонов. Прогноз был сделан на один день для точек с номерами $301, \ldots, 478$. На графиках показано среднее по всем прогнозам значение модуля отклонения от реального значения ряда (в количестве вагонов) в зависимости от количества точек, использованных при построении гистограммы. Число столбцов гистограммы фиксируется равным оптимальному числу столбцов для гистограммы, построенной по всему временному ряду (см. описание алгоритма построения гистограммы). Использована квадратичная функция потерь. Эксперимент проводился для рядов, описывающих прибытие вагонов с нефтью и нефтепродуктами.

Из графиков следует, что средняя ошибка прогноза не падает с увеличением длины предыстории. Поэтому в качестве критерия для выбора этого параметра была использована стабилизация распределения, описываемого гистограммой. Для этого вычислялось расстояние Кульбака-Лейблера между парой распределений, построенных по наборам то-

чек, отличающихся на 10 точек:

$$dist(p_1, p_2) = \sum_{i=1}^{k} p_1(i) \ln \left(\frac{p_1(i)}{p_2(i)} \right),$$

где p_1, p_2 — плотности дискретных распределений, задаваемых гистограммами, i — принимаемые случайной величиной значения.



(a) Прибытие, неделя

(b) Прибытие, месяц

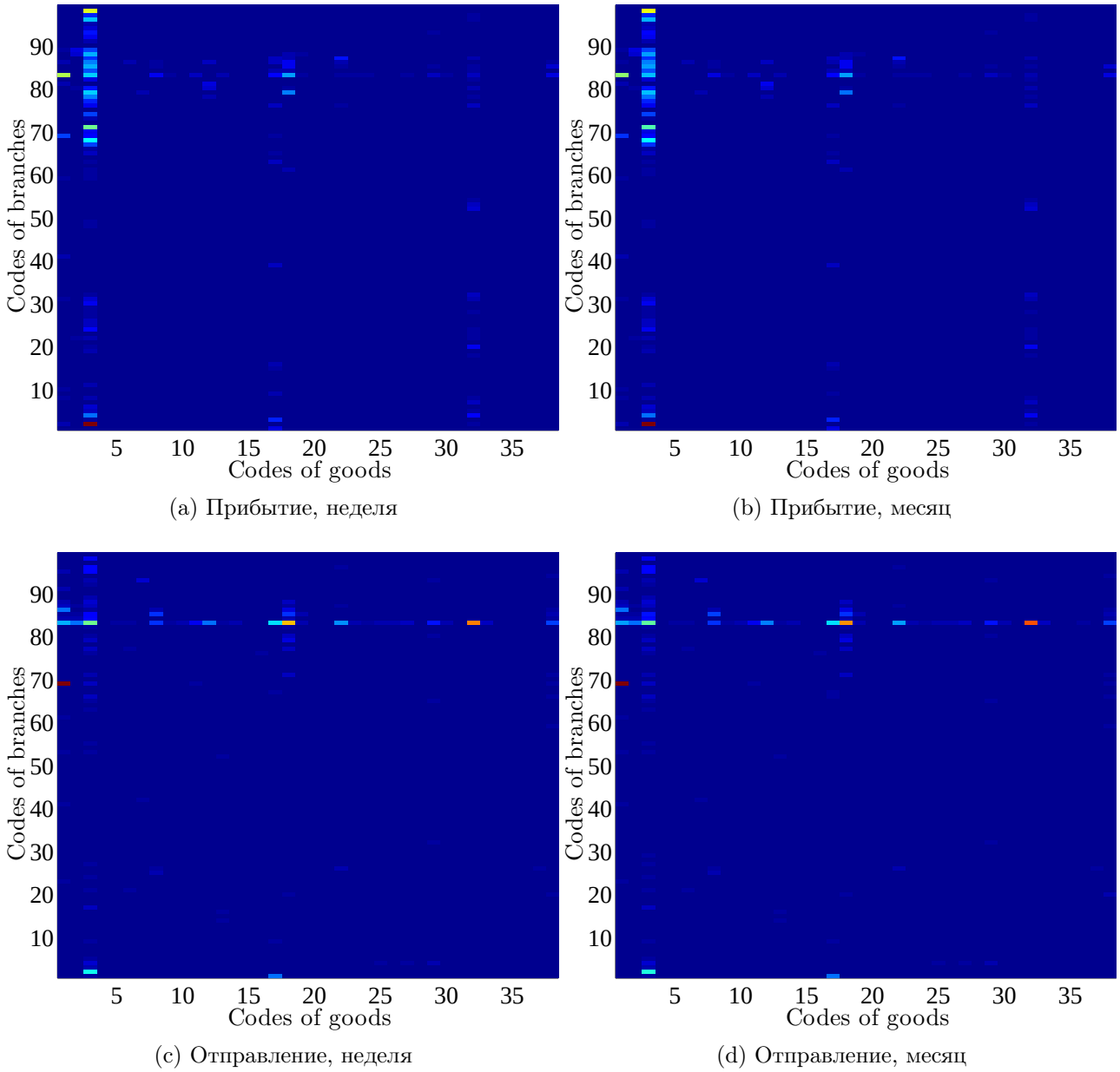
(c) Отправление, неделя

(d) Отправление, месяц

Рис. 10: Средняя ошибка при прогнозировании ARMA

На рис. 8 изображен график зависимости расстояния Кульбака-Лейблера между парой гистограмм от длины предыстории. По горизонтали отложено число точек, использованных при построении гистограммы, по вертикали — расстояние между двумя последовательно построенными гисторгаммами. Точки последовательно набираются, начиная с

конца временного ряда. Границы интервалов для гистограмм фиксируются по границам интервалов оптимальной гистограммы, построенной по всему временному ряду. Для дальнйших экспериментов была выбрана длина предыстории $H = 120$ точек, так как такого количества достаточно, как это следует из графиков, для получения расстояния между двумя последовательно построенными гистограммами не более 0.05.



(a) Прибытие, неделя

(b) Прибытие, месяц

(c) Отправление, неделя

(d) Отправление, месяц

Рис. 11: Средняя ошибка при непараметрическом прогнозировании с абсолютной функцией потерь

При построении прогнозов были использованы свертки с тремя функциями потерь:

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

3) $L(z,x) = \begin{cases} 0, & \text{если } |z-x| < a; \\ |z-x| - a, & \text{если } |z-a| \geqslant a, \text{ где } a = 19. \end{cases}$

Значение параметра a соответствует разности числа вагонов в соседних столбцах гистограммы.

На рис. 9 изображены свертки гистограммы с различными функциями потерь. По горизонтали отмечено количество вагонов, соответствующее среднему значению интервала гистограммы, по вертикали — количество значений ряда, попавших в интервал, и значения сверток гистограммы с функциями потерь. Столбцы, соответствующие значениям суммы квадратов, уменьшены в 10^4 раз, соответствующие сумме модулей и трапеции — в 10^2 раз.

Сравнение непараметрического алгоритма прогнозирования с алгоритмом ARMA. При выполнении вычислительного эксперимента были вычислены средние ошибки (7) прогнозирования непараметрического алгоритма для прогноза на день, неделю и месяц по каждой ветке и каждой категории груза. Результаты сравнивались со средней ошибкой, получаемых с помощью модели ARMA для рядов, в которых ненулевых значений не менее $\frac{1}{5}$ от числа всех значений ряда. При меньшем количестве ненулевых значений модель ARMA работает некорректно. Во избежании деления на ноль при вычислении средней ошибки ко всем элементам рядов было прибавлено число 100.

На рис. 10 показаны средние ошибки, даваемые при прогнозе алгоритмом ARMA для рядов, в которых не менее $\frac{1}{5}$ ненулевых значений. Ряды, для которых прогноз не строился, отмечены белым цветом. Значение ошибки показано цветом ячейки: чем больше значение, тем более красный оттенок. По горизонтали отложены коды грузов, по вертикали — коды веток.

На рис. 11 показаны средние ошибки, даваемые при прогнозе алгоритмом непараметрического прогнозирования с абсолютной функцией потерь. Значение ошибки показано цветом ячейки: чем больше значение, тем более красный оттенок. Оси соответствуют тем же величинам, что и на предыдущем рисунке.

Табл. 12 содержит средние ошибки прогнозирования в процентах для модели ARMA и непараметрической модели с тремя рассматриваемыми функциями потерь для рядов, описывающих прибытие вагонов с различными типами грузов на 83 ветку. Табл. 13 содержит аналогичную информацию для отправления вагонов с 83 ветки. Первый столбец обеих таблиц содержит коды грузов, второй — информацию о стационарности рядов. Если в ячейке стоит 1, то ряд нестационарный, если 0, то стационарный. Из анализа результатов, представленных в таблицах, можно сделать вывод, что применение непараметрической прогностической модели для прогнозирования нестационарных временных рядов возможно, но обеспечивает меньшую точность прогноза, чем при прогнозировании стационарных рядов.

Заключение

Предложен алгоритм непараметрического прогнозирования загруженности железнодорожных узлов РЖД, основанный на свертке эмпирической плотности распределения значений временного ряда с функцией потерь. Проведен анализ структуры входных данных, на основе которого выбраны параметры модели. Проведен сравнительный анализ результатов предложенного алгоритма и алгоритма ARMA. Основным преимуществом непараметрического алгоритма по сравнению с алгоритмом ARMA является его применимость для прогнозирования стационарных временных рядов с большим количеством

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
0	0	NaN	0,05	0,05	0,05	NaN	0,05	0,05	0,05
1	1	NaN	34,63	33,23	33,46	NaN	34,05	33,02	33,28
2	0	NaN	0,33	0,29	0,29	NaN	0,34	0,30	0,30
3	1	NaN	20,88	19,87	20,02	NaN	20,75	20,11	19,89
4	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
6	0	NaN	0,52	0,38	0,38	NaN	0,53	0,38	0,38
7	0	NaN	0,58	0,38	0,38	NaN	0,58	0,38	0,38
8	0	NaN	0,63	0,42	0,42	NaN	0,64	0,41	0,41
9	1	12,92	8,27	5,88	12,76	15,43	8,94	6,04	13,50
10	0	2,04	1,23	1,01	6,62	2,26	1,26	1,02	7,07
11	0	NaN	0,40	0,26	0,26	NaN	0,40	0,26	0,26
12	0	NaN	3,25	3,51	6,88	NaN	3,26	3,57	7,04
13	0	NaN	0,87	0,79	0,79	NaN	0,88	0,79	0,79
14	0	4,33	2,43	2,31	2,64	4,80	2,42	2,34	2,43
15	0	1,10	0,69	0,49	0,49	1,22	0,71	0,51	0,51
16	0	NaN	0,20	0,15	0,15	NaN	0,20	0,15	0,15
17	0	NaN	0,54	0,41	0,41	NaN	0,52	0,39	0,39
18	0	15,00	8,33	7,47	10,19	16,22	8,45	7,56	10,23
19	1	30,10	19,45	16,89	17,77	34,86	20,53	17,77	18,71
20	0	2,12	1,42	1,17	7,01	2,37	1,43	1,17	7,44
21	0	NaN	0,19	0,12	0,12	NaN	0,20	0,12	0,12
22	0	NaN	0,44	0,33	0,33	NaN	0,45	0,34	0,34
23	0	2,85	1,66	1,33	1,33	2,87	1,67	1,33	1,30
24	0	2,24	1,22	1,11	1,21	2,39	1,21	1,11	1,21
25	0	2,33	1,35	1,18	1,26	2,60	1,36	1,18	1,28
26	0	1,61	0,97	0,65	0,65	1,79	0,98	0,66	0,66
27	0	1,25	0,78	0,53	0,53	1,39	0,79	0,55	0,55
28	0	2,74	1,56	1,09	1,09	2,83	1,56	1,08	1,08
29	0	NaN	0,32	0,23	0,23	NaN	0,33	0,23	0,23
30	0	6,66	3,90	3,57	9,54	7,48	3,99	3,63	9,77
31	0	2,92	1,56	1,51	1,99	3,19	1,55	1,52	2,04
33	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
34	0	6,38	3,60	3,01	3,92	6,96	3,61	2,90	4,00
35	0	1,85	0,94	0,89	0,87	1,95	0,92	0,88	0,86
36	0	0,99	0,67	0,47	0,47	1,04	0,67	0,47	0,47
38	0	1,77	1,07	0,94	0,95	1,83	1,04	0,93	0,93
39	0	1,25	0,71	0,55	0,55	1,34	0,72	0,56	0,56
42	0	NaN	0,55	0,51	0,51	NaN	0,52	0,50	0,50
43	0	NaN	5,63	5,16	10,09	NaN	5,93	5,30	10,45

Рис. 12: Средний процент ошибки при прогнозе прибытия

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
0	0	NaN	0,05	0,05	0,05	NaN	0,05	0,05	0,05
1	1	NaN	34,63	33,23	33,46	NaN	34,05	33,02	33,28
2	0	NaN	0,33	0,29	0,29	NaN	0,34	0,30	0,30
3	1	NaN	20,88	19,87	20,02	NaN	20,75	20,11	19,89
4	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
6	0	NaN	0,52	0,38	0,38	NaN	0,53	0,38	0,38
7	0	NaN	0,58	0,38	0,38	NaN	0,58	0,38	0,38
8	0	NaN	0,63	0,42	0,42	NaN	0,64	0,41	0,41
9	1	12,92	8,27	5,88	12,76	15,43	8,94	6,04	13,50
10	0	2,04	1,23	1,01	6,62	2,26	1,26	1,02	7,07
11	0	NaN	0,40	0,26	0,26	NaN	0,40	0,26	0,26
12	0	NaN	3,25	3,51	6,88	NaN	3,26	3,57	7,04
13	0	NaN	0,87	0,79	0,79	NaN	0,88	0,79	0,79
14	0	4,33	2,43	2,31	2,64	4,80	2,42	2,34	2,43
15	0	1,10	0,69	0,49	0,49	1,22	0,71	0,51	0,51
16	0	NaN	0,20	0,15	0,15	NaN	0,20	0,15	0,15
17	0	NaN	0,54	0,41	0,41	NaN	0,52	0,39	0,39
18	0	15,00	8,33	7,47	10,19	16,22	8,45	7,56	10,23
19	1	30,10	19,45	16,89	17,77	34,86	20,53	17,77	18,71
20	0	2,12	1,42	1,17	7,01	2,37	1,43	1,17	7,44
21	0	NaN	0,19	0,12	0,12	NaN	0,20	0,12	0,12
22	0	NaN	0,44	0,33	0,33	NaN	0,45	0,34	0,34
23	0	2,85	1,66	1,33	1,33	2,87	1,67	1,33	1,30
24	0	2,24	1,22	1,11	1,21	2,39	1,21	1,11	1,21
25	0	2,33	1,35	1,18	1,26	2,60	1,36	1,18	1,28
26	0	1,61	0,97	0,65	0,65	1,79	0,98	0,66	0,66
27	0	1,25	0,78	0,53	0,53	1,39	0,79	0,55	0,55
28	0	2,74	1,56	1,09	1,09	2,83	1,56	1,08	1,08
29	0	NaN	0,32	0,23	0,23	NaN	0,33	0,23	0,23
30	0	6,66	3,90	3,57	9,54	7,48	3,99	3,63	9,77
31	0	2,92	1,56	1,51	1,99	3,19	1,55	1,52	2,04
33	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
34	0	6,38	3,60	3,01	3,92	6,96	3,61	2,90	4,00
35	0	1,85	0,94	0,89	0,87	1,95	0,92	0,88	0,86
36	0	0,99	0,67	0,47	0,47	1,04	0,67	0,47	0,47
38	0	1,77	1,07	0,94	0,95	1,83	1,04	0,93	0,93
39	0	1,25	0,71	0,55	0,55	1,34	0,72	0,56	0,56
42	0	NaN	0,55	0,51	0,51	NaN	0,52	0,50	0,50
43	0	NaN	5,63	5,16	10,09	NaN	5,93	5,30	10,45

Рис. 13: Средний процент ошибки при прогнозе отправления

одинаковых значений, в том числе и нулевых. Проверена практическая применимость предложенного алгоритма для прогнозированияя нестационарных временных рядов.

Литература

[1] Хардле В. Прикладная непараметрическая регрессия. М: Мир. 1993.

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

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

[4] Магнус Я. Р., Катышев П. К., Персецкий А. А. Эконометрика: начальный курс. М: издательство «Дело». 2004.

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

[6] Nochai R., Nochai T. ARIMA model for forecasting oil palm price. Proceedings of the 2nd IMT-GT Regional conference of mathematics, statistics and applications. University Sains Malaisia, Penang. June 13–15, 2006.

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

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

[9] 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.

[10] Koenker Jr., Bassett G. Regression Quantiles. Econometrica, 1978. Vol. 46. №1. P. 33–50.

[11] Постникова Е. Квантильная регрессия. Новосибирск: НГУ. 2006.

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