

Прогнозирование объемов спроса на грузоперевозки для стационарных временных рядов*

А. Д. Воронов¹, А. Н. Громов², А. С. Инякин², А. А. Замковой³
voronov.ad@phystech.edu; gromlogin@gmail.com; inyakin@forecsys.ru;
Zamkovoy.al.an@gmail.com

¹Московский физико-технический институт, 141701, Московская область, г. Долгопрудный, Институтский пер., 9; ²Федеральный исследовательский центр Информатика и Управление Российской Академии Наук, 119333, Москва, Вавилова, д.44, кор.2; ³Акционерное общество Институт Экономики и Развития Транспорта, 105066, Москва, Новорязанская, д.24

Исследуются свойства прогностических моделей объемов спроса на грузовые железнодорожные перевозки с целью структуризации процессов в области управления и планирования грузовых железнодорожных перевозок. В работе предлагаются четыре модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающая специфику измеряемых данных, бизнес-процессов и нормативов индустриального партнера. При построении моделей используется многофакторный статистический анализ и прогнозирование взаимозависимых временных рядов. Анализируются свойства построенных моделей. Строятся прогнозы в разрезах день, неделя, месяц для станций и регионов. Предложенные прогностические модели сравниваются по критериям средней абсолютной и средней процентной ошибки.

Ключевые слова: *временные ряды, прогнозирование, суперпозиция, прогностические модели, ассимметричное распределение, регрессионные остатки*

DOI: 10.21469/22233792.4.1.05

1 Введение

Грузовые перевозки обеспечивают свыше 80% общей выручки железнодорожного транспорта [8]. В связи с этим, планирование грузовых перевозок имеет большое практическое значение для производственно-хозяйственного планирования и управления в данной отрасли. В рамках реформы железнодорожного транспорта [12] была проведена отмена предварительных заявок грузоотправителей, грузоотправителям была предоставлена возможность выбора поставщиков и видов транспорта. Эти изменения привели к необходимости прогнозирования спроса на перевозки грузов при планировании перевозок, то есть замене оперативного планирования перевозок их прогнозированием для определения реальных потребностей грузоотправителей в перевозке грузов [8].

В рамках исследования рассматриваются методы краткосрочного прогнозирования временных рядов с целью создания модели модели прогнозирования объемов спроса на грузовые железнодорожные перевозки. Разделение прогнозов на краткосрочные, среднесрочные и долгосрочные зависит от особенностей анализируемого ряда и целей прогноза [6]. Под краткосрочным прогнозированием подразумевается выполнение прогноза на период времени от суток до года. Краткосрочное прогнозирование связано в основном с оперативным и текущим планированием производства [7], поэтому при прогнозе учитываются в первую очередь микроэкономические показатели, такие как цены на перевозимую продукцию. В железнодорожной отрасли выделяют долгосрочные прогнозы (на 5–10 и

*Работа выполнена при финансовой поддержке РФФИ (проекты 17-20-01212, 17-20-01184).

более лет), среднесрочные (на 3–5 лет), текущие (на 1 год) и оперативные (квартальные и месячные) [8]. В зависимости от выбранного типа прогноза меняется номенклатура планируемых грузов и степень детализации планов. Рассматриваемые при долгосрочном и среднесрочном прогнозировании данные содержат информацию по ограниченной групповой номенклатуре грузов. Текущие планы основываются на годовых прогнозах перевозок. Такие прогнозы более детальны и предусматривают разработку плана по основным массовым грузам. Наиболее подробными и точными являются оперативными планы перевозок, в рамках которых выполняется прогнозирование объемов перевозок на квартал и на месяц.

Долгосрочные и среднесрочные прогнозы применяются для стратегического планирования, поэтому в них особое внимание уделяется макроэкономическому анализу товарного и транспортного рынков. Макроэкономические параметры используются, например, в [9–11] при создании транспортно-экономических моделей с целью выполнения прогнозов глубиной более года. Например, авторы [9] используют поиск конкурентно-транспортного равновесия для создания новых моделей транспортного планирования, включающих модели роста транспортной инфраструктуры городов. В [10] предложена модификация модели конкурентного равновесия для анализа проблем формирования тарифной и инвестиционной политики управления железнодорожными грузоперевозками. В [11] для анализа спроса на ГЖДП с учетом тарифов и инвестиций в развитие инфраструктуры предложена транспортная модель, учитывающая функции спроса и предложения перевозимых товаров. Модель разработана для данных, агрегированных по годам, и предназначена для прогнозирования с глубиной не менее года. При стратегическом планировании широко используются системы экспертных оценок, основанные на применении различных аналитических матриц для исследования альтернатив возможного стратегического развития. Важным инструментом стратегического планирования являются транспортно-экономические балансы, обеспечивающих сбалансированность объемов произведенной продукции с размерами их в рассматриваемых территориальных единицах.

Задача прогнозирования спроса на грузовые перевозки поставлена для оперативного планирования перевозок. Рассмотрена задача прогнозирования нестационарных временных рядов в случае несимметричных функций потерь, учитывающих экспертные оценки потерь при недорознозе и перепрогнозе. Один из широко используемых методов прогнозирования нестационарных временных рядов, авторегрессионное интегрированное скользящее среднее ARIMA [13], позволяет прогнозировать временные ряды с трендом, а также при модификации и ряды с сезонной компонентой. Настройка параметров этого алгоритма осуществляется путем минимизации квадратичной функции потерь для обеспечения несмещенности прогнозов, а также гомоскедастичности, нормальности (с нулевым матожиданием) и некоррелированности регрессионных остатков. Свойства прогнозов временных рядов при использовании несимметричных функций потерь были исследованы в работе [14], авторы которой отмечают смещенность прогнозов при несимметричных потерях и делают вывод о необходимости разработки специальных методов прогнозирования временных рядов в условии несимметричности функции потерь. В работах [14, 15] сделан вывод о том, что модель ARIMA не подходит для решения задачи прогнозирования в случае несимметричной функции потерь.

В работах [16, 17] были предложены модификации модели, позволяющие учесть несимметричность функции потерь при настройке параметров алгоритма. Однако обе предложенные модификации сложны в реализации, не позволяют использовать пакеты для прогнозирования временных рядов, в которых есть стандартные реализации, и требуют для каждой функции потерь создания и обучения индивидуальной модели, что неприемлемо

в промышленных задачах. Еще одним методом, предложенным для работы с несимметричными функциями потерь, является квантильная регрессия [18]. Она позволяет находить оптимальный смещенный прогноз для несимметричных функций потерь кусочно-линейного вида, но не дает возможности работать с функциями потерь других видов, а также применима только для стационарных временных рядов.

В работе предложен двухэтапный алгоритм прогнозирования, на первом этапе которого отслеживаются свойства временного ряда, обуславливающие его нестационарность, такие как тренд и сезонность. На втором этапе вычисляется поправка, обеспечивающая оптимальность прогноза в случае несимметричной функции потерь. На втором этапе алгоритма в качестве временного ряда выступают регрессионные остатки, однако их плотность распределения не известна. В качестве оценки плотности используется гистограмма значений регрессионных остатков, как предложено в [21]. В алгоритме используется ряд упрощений задачи минимизации свертки функции потерь с оценкой плотности распределения регрессионных остатков, которые приводят к задаче приближенного нахождения минимума путем перебора конечного числа значений, из которых выбирается то, которое обеспечивает наименьшее значение свертки. Предлагаемый алгоритм строит прогнозы с минимальным математическим ожиданием потерь при использовании несимметричных функций потерь для различных временных рядов, в том числе имеющих тренд и сезонную компоненту, то есть не являющихся стационарными. При этом не накладываются ограничения на класс функций потерь, которые можно использовать в задаче прогнозирования: функции потерь могут быть несимметричными либо симметричными, отличными от квадратичной или модуля.

В качестве базового алгоритма для сравнения полученных прогнозов в [1–3] используется модель авторегрессионного скользящего среднего ARMA. Для построения прогнозов данных с выраженной периодичностью в [4] используется модель sARIMA. В [5] для решения подобных задач применяются нейронные сети.

Прогнозы строятся для номенклатуры, учитывающей 41 наименование грузов. Основным источником данных являются учетные системы промышленного партнера. Данные содержат информацию об отправлениях грузов: дату погрузки, станцию отправления, станцию назначения, количество вагонов, которые прошли по маршруту от станции отправления до станции назначения, код груза, род вагонов, суммарный вес груза в тоннах и признак маршрутной отправки.

2 Программа вычислительных экспериментов

Цель вычислительного эксперимента – сравнение результатов прогнозирования ОСГП, получаемых с использованием макета модуля прогнозирования ОСГП и аналогичных прогнозов, получаемых по модели ARMA, со значениями контрольной выборки данных об объемах спроса на ГП. Прогнозирование выполняется с использованием моделей Ex VAR, Regularized ARMA, Mean by 5 days, Median by 70 days [19, 20]. Сокращение Ex VAR используется в качестве обозначения двухуровневого алгоритма прогнозирования ARIMA+hist [19, 20]. На первом этапе этого ПНИ [19] был предложен двухэтапный алгоритм прогнозирования ARIMA+hist, на первом шаге которого отслеживаются свойства временного ряда, обуславливающие его нестационарность, такие как тренд и сезонность. На втором шаге вычисляется поправка, обеспечивающая оптимальность прогноза в случае несимметричной функции потерь. На этом, втором, шаге в качестве временного ряда выступают регрессионные остатки, однако их плотность распределения не известна. В качестве оценки плотности используется гистограмма значений регрессионных остатков,

как предложено в [21]. В алгоритме hist используется ряд упрощений задачи минимизации свертки функции потерь с оценкой плотности распределения регрессионных остатков, которые приводят к задаче приближенного нахождения минимума путем перебора конечного количества значений, из которых выбирается то, которое обеспечивает 13 наименьшее значение свертки. Предложенный алгоритм ARIMA+hist (далее в этом отчете обозначается как Ex VAR) строит прогнозы с минимальным математическим ожиданием потерь при использовании несимметричных функций потерь для различных временных рядов, в том числе имеющих тренд и сезонную компоненту, то есть не являющихся стационарными. При этом не накладываются ограничения на класс функций потерь, которые можно использовать в задаче прогнозирования: функции потерь могут быть несимметричными либо симметричными, отличными от квадратичной или модуля.

2.1 Модель ARMA для стационарного временного ряда

Стационарный временной ряд со средним значением μ описывается моделью ARMA(p, q), если выполняется

$$\nabla^d x_t = \alpha + \varphi_1 \nabla^d x_{t-1} + \dots + \varphi_p \nabla^d x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (1)$$

где $\alpha = \mu(1 - \varphi_1 - \dots - \varphi_p)$, $\varphi_1, \dots, \varphi_p, \theta_1, \theta_q$ — константы, ε_t — гауссов шум с нулевым средним и постоянной дисперсией.

2.2 Задача прогнозирования стационарных временных рядов с известной плотностью распределения

Нахождение прогноза \hat{x} следующего значения x_{T+1} временного ряда $x = \{x_1, \dots, x_T\}$, $x_t \in \mathbb{R}_+$, минимизирующего математическое ожидание $El(c, x_{T+1})$ потерь:

$$\hat{x} = \arg \min_{c \in \mathbb{R}_+} El(c, x_{T+1}) \quad (2)$$

где $l(c, x_{T+1})$ — заданная функция потерь. При разработке базовой версии прогностического алгоритма были сделаны следующие предположения. Временной ряд является стационарным, то есть все его значения x_1, \dots, x_t являются реализациями стационарного случайного процесса. Для любых $t, \tau \in [1, \dots, T]$ выполняется

$$P(x_t < \xi) = P(x_\tau < \xi), \quad \xi \in \mathbb{R}_+ \quad (3)$$

и прогнозируемое значение x_{T+1} временного ряда генерируется из того же распределения, что и все наблюдаемые значения x_τ . Случай нестационарных временных рядов рассмотрен в разделе 3. Для x_t — непрерывной случайной величины с плотностью распределения $p(u)$,

$$\lim_{dt \rightarrow 0} P(u < x_t \leq u + du) = p(u), \quad (4)$$

функция плотности распределения $p(u)$ имеет конечное число точек разрыва:

$$\tilde{u}_{\min} = u_0, \dots, \tilde{u}_C = u_{\max}, \quad C \in \mathbb{N}_0 : p(u) \in C(\tilde{u}_{i-1}, \tilde{u}_i), i = 1, \dots, C. \quad (5)$$

Основанием для этого предположения служит тот факт, что функции плотности распределения большинства непрерывных случайных величин, рассматриваемых при решении задач, связанных с прогнозированием временных рядов, носят непрерывный характер. При

известной плотности распределения $p(u)$, на основании которой генерируются прогнозируемое значения временного ряда x_{T+1} , математическое ожидание потерь выражается аналитически в виде

$$\text{El}(c, x_T + 1) = \int_{u_{\min}}^{u_{\max}} l(c, u)p(u)du. \quad (6)$$

Интеграл (6) существует в силу предположения (5). Задача прогнозирования формулируется так:

$$\hat{x} = \arg \min_{c \in \mathbb{R}_+} \int_{u_{\min}}^{u_{\max}} l(c, u)p(u)du = \arg \min_{c \in \mathbb{R}_+} L(c), \quad (7)$$

где $L(c) = \text{El}(c, u)$. В ряде случаев прогноз можно найти аналитически. В частности, получены значения оптимального прогноза (10) и (11) для квадратичной и абсолютной функций $l(\hat{x}, x)$, используемых для оценки потерь при прогнозировании. Для этих функций производная ожидаемых потерь $L(c)$ имеет простой вид, анализируя который можно получить выражение для оптимального прогноза. Для корректности операции дифференцирования интеграла (7) по параметру требуется выполнение следующих условий [22]: 1) непрерывность подынтегральной функции $l(c, u)p(u)$, 2) непрерывность ее частной производной $p(u)\frac{\partial l(c, u)}{\partial c}$. По предположению (7), функция $p(u)$ имеет конечное число $C + 1$ точек разрыва, следовательно интеграл (8) может быть представлен в виде суммы C интегралов

$$L(c) = \sum_{i=1}^C \int_{\tilde{u}_{i-1}}^{\tilde{u}_i} l(c, u)p(u)du, \quad (8)$$

для каждого из которых условия 1) и 2) выполняются. Ввиду конечности суммы (8), дифференцирование функции $L(c)$ корректно. Для квадратичной функции потерь $l(\hat{x}, x_{T+1}) = (\hat{x} - x_{T+1})^2$ производная ожидаемых потерь имеет вид

$$\frac{dL}{dc} = \int_{u_{\min}}^{u_{\max}} 2(c - u)p(u)du = 2c \int_{u_{\min}}^{u_{\max}} p(u)du - 2 \int_{u_{\min}}^{u_{\max}} up(u)du = 2c - 2Ex. \quad (9)$$

Отсюда следует, что оптимальный прогноз, минимизирующий L , равен математическому ожиданию x_t

$$\frac{dL}{dc} = 2c - 2Ex \rightarrow \hat{x} = Ex. \quad (10)$$

Для абсолютной функции потерь $l(\tilde{x}, x_{T+1}) = |\tilde{x} - x_{T+1}|$ производная функции L имеет вид $\frac{dL}{dc} = - \int_{u_{\min}}^c p(u)du + \int_c^{u_{\max}} p(u)du$, то есть, оптимальный прогноз совпадает с медианой распределения $p(u)$:

$$\tilde{x} = \text{med } p(u). \quad (11)$$

При прогнозировании объемов спроса на грузовые железнодорожные перевозки истинное распределение $p(u)$ неизвестно. На основе результатов первого этапа исследований было принято решение использовать непараметрические методы прогнозирования. В частности, оценка плотности $p(u)$ распределения x_t в базовой версии прогностического алгоритма проводилась с помощью гистограмм. Было также рассмотрено обобщение предложенного алгоритма с использованием ядерных оценок плотности.

2.3 Алгоритм прогнозирования hist

Алгоритм прогнозирования hist оценивает распределение значений временного ряда x , находит функцию $p(u)$, и ищет приближенное решение задачи минимизации (7) перебором. Вход алгоритма: стационарный временной ряд $x = \{x_1, \dots, x_T\}$ и функция потерь

$l(\hat{x}, x_{T+1})$ Выход: прогноз \hat{x} , минимизирующий математическое ожидание потерь. Порядок вычислений: 1) задание количества n столбцов гистограммы, 2) вычисление ширины столбцов гистограммы

$$b = \frac{\max(x) - \min(x)}{n} \quad (12)$$

и координат концов отрезков постоянства u_0, u_1, \dots, u_n для функции $\hat{p}(u)$, 3) построение гистограммы; нахождение функции $\hat{p}(u)$; нормирование гистограммы; вычисление значений функции на отрезках постоянства y_1, \dots, y_n , 4) вычисление значений свертки

$$\sum_{i=1}^n h_i l(c, \frac{u_i + u_{i-1}}{2}) \quad (13)$$

для всех $c \in \{\frac{u_1+u_0}{2}, \dots, \frac{u_n+u_{n-1}}{2}\}$; выбор c^* , при котором достигается минимальное значение свертки; вычисление соответствующего прогнозируемого значения \hat{x} :

$$\hat{x} = c^* \in \{\frac{u_1 + u_0}{2}, \dots, \frac{u_n + u_{n-1}}{2}\}. \quad (14)$$

Функция потерь l в каждом конкретном случае выбирается с учетом особенностей прикладной задачи и стоимости ошибки прогноза в ту или иную сторону. Функцию потерь задавают эксперты. При заданных выборке данных и функции потерь результат прогнозирования зависит только от числа столбцов гистограммы n . При малых n оценка плотности распределения $\hat{p}(u)$ получается огрубленной, при больших n – более детальной, однако при увеличении n возрастает вероятность переобучения прогностической модели.

2.4 Непараметрическая оценка плотности распределения с помощью гистограммы при прогнозировании объемов спроса на грузовые железнодорожные перевозки

В этом пункте рассмотрена задача оценки плотности $p(u)$ распределения прогнозируемой величины x_t и вычисления свертки, аппроксимирующей математическое ожидание экспертно заданной функции потерь $l(c, u)$. Гистограмма как оценка плотности распределения является разрывной кусочно-постоянной функцией $\hat{p}(u)$ с n интервалами постоянства

$$(u_i, u_{i+1}) : u_0 = u_{\min} \leq u_1 \leq \dots \leq u_{n-1} \leq u_n = u_{\max} : \hat{p}(u) = h_i, u \in (u_{i-1}, u_i), i = 1, \dots, n \quad (15)$$

одинаковой длины

$$b = u_1 - u_0 = \dots = u_m - u_{n-1} = \frac{u_{\max} - u_{\min}}{2}. \quad (16)$$

Интеграл от функции $\hat{p}(u)$ равен единице:

$$\int_{u_{\min}}^{u_{\max}} \hat{p}(u) du = \sum_{i=1}^n h_i (u_i - u_{i-1}) = b \sum_{i=1}^n h_i = 1. \quad (17)$$

При использовании оценки плотности $\hat{p}(u)$ математическое ожидание потерь $El(c, u)$ приближается величиной $L_{\text{hist}}(c)$:

$$L \approx L_{\text{hist}} = \int_{u_{\min}}^{u_{\max}} l(c, u) \hat{p}(u) du = y_1 \int_{u_0}^{u_1} l(c, u) du + \dots + y_n \int_{u_{n-1}}^{u_n} l(c, u) du = \sum_{i=1}^n h_i \int_{u_{j-1}}^{u_i} l(c, u) du. \quad (18)$$

Точность этой оценки монотонно растет с увеличением количества отрезков постоянства функции $\hat{p}(u)$, то есть с увеличением числа столбцов гистограммы. С учетом данного приближения задача прогнозирования принимает вид:

$$\hat{x} = \arg \min_c L_{\text{hist}}(c) = \arg \min_c \sum_{i=1}^n h_i \int_{u_{j-1}}^{u_j} l(c, u) du. \quad (19)$$

Так как интегрирование функции потерь $l(\hat{x}, x_{T+1})$ является трудоемкой операцией, интеграл в постановке задачи прогнозирования (19) оценивается по формуле

$$\int_{u_{j-1}}^{u_i} l(c, u) du \approx l\left(c, \frac{u_i + u_{i-1}}{2}(u_i - u_{i-1})\right) = bl\left(c, \frac{u_i + u_{i-1}}{2}\right). \quad (20)$$

Точность этой оценки повышается с уменьшением длины отрезков постоянства функции $\hat{p}(u)$, достигаемой за счет увеличения их количества. Поскольку величина b является постоянной, то в задаче поиска минимума она опущена, и задача поиска оптимального прогноза принимает вид:

$$\hat{x} = \arg \min_c L_{\text{conv}}(c) = \arg \min_c \sum_{i=1}^n h_i l\left(c, \frac{u_i + u_{i-1}}{2}\right). \quad (21)$$

Приближенное решение выбирается из конечного множества точек $c \in \left\{\frac{u_1+u_0}{2}, \dots, \frac{u_n+u_{n-1}}{2}\right\}$, которое является набором середин отрезков постоянства функции $\hat{p}(u)$. В качестве прогноза выбирается та точка, которая дает минимальное значение свертки в (21). Поскольку в большинстве случаев количество отрезков постоянства функции $\hat{p}(u)$ (количество столбцов в гистограмме) невелико, искомый прогноз можно найти простым перебором.

При тестировании предложенного алгоритма оптимальные согласно (21) прогнозы сравнивались с прогнозами, оптимальными согласно (8), для квадратичной и абсолютной функций потерь, так как для этих функций существует аналитическое решение задачи (9). Для квадратичной функции потерь переход от вычисления интеграла к вычислению значения функции потерь в середине отрезка постоянства функции $\hat{p}(u)$ не влияет на полученный прогноз и оптимальными согласно (9) и (19), (21) прогнозами являются выборочное среднее

$$\hat{x} = \frac{1}{T} \sum_{t=1}^T x_t \quad \hat{x} = b \sum_{i=1}^n h_i \frac{u_i + u_{i-1}}{2}. \quad (22)$$

Для абсолютной функции потерь оптимальным согласно (19) прогноз задается соотношением

$$y_k(2c - (u_k + u_{k-1})) + b\left(\sum_{i:u_i < c} h_i - \sum_{i:u_{i-1} > c} h_i\right) = 0.$$

Оптимальный по (21) прогноз является оценкой медианы распределения, вычисленной по оценке распределения $\hat{p}(u)$.

2.5 Зависимость прогноза от ширины окна h

Согласно проведенным исследованиям, значения ширины окна h разбиваются на интервалы, в рамках которых точность прогнозов, полученных алгоритмом hist на выборках разной длины, значимо не меняется. При увеличении длины выборки T расширяется интервал значений ширины окна h , при которых точность прогнозов алгоритма hist_k, обеспечиваемой выборочными статистиками. Положение диапазона ширины окна h , в котором

качество прогнозов, даваемых алгоритмом hist_k , совпадает с качеством, которое обеспечивается выборочным средним при квадратичной функции потерь и выборочной медианой при абсолютной функции потерь, смещается вправо с ростом дисперсии распределения данных: с ростом дисперсии увеличивается оптимальная ширина окна.

На рис. 1,2 представлены графики отклонения прогнозов от оптимального для квадратичной и абсолютной функций потерь от ширины окна h при различной длине выборки $T = 1000$ оси абсцисс в логарифмической шкале отложена ширина окна h , по оси ординат – отклонение прогноза от математического ожидания. Синим цветом обозначен график для прогнозов алгоритма hist_k , красным – для оптимального согласно (19) прогноза.

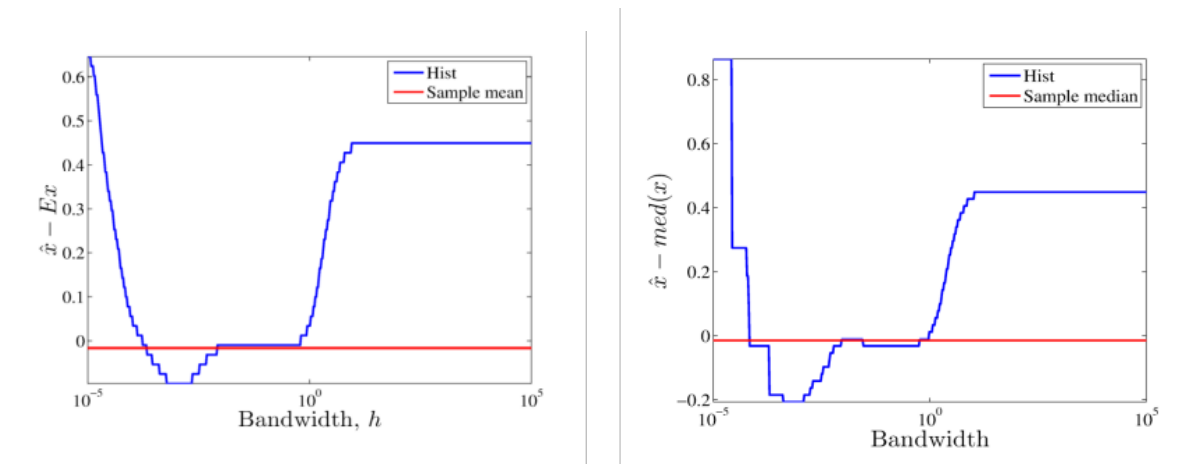


Рис. 1 Зависимость качества прогнозирования от ширины окна при длине выборки $T = 1000$ и количестве точек свертки $n = 300$ для квадратичной (слева) и абсолютной (справа) функции потерь.

3 Разработка и обоснование методов прогнозирования нестационарных временных рядов

В этом пункте описан двухэтапный алгоритм ARIMA + hist прогнозирования (9) временных рядов при несимметричной функции потерь, который позволяет строить прогнозы для нестационарных рядов, а также позволяет использовать функции потерь любого вида. Алгоритм основан на идее работы [23], в которой для построения прогноза используется авторегрессионная модель с минимизацией квадратичной функции потерь для получения несмещенного прогноза, и анализ регрессионных остатков для оценки оптимального смещения прогноза. При разработке алгоритма использован результат из работы [24] о зависимости смещения прогноза исключительно от функции потерь и дисперсии временного ряда.

Алгоритм ARIMA + hist, вычисляющий прогноз \hat{x} для временного ряда $x = \{x_1, \dots, x_T\}$ и функции потерь $l(\hat{x}, x_{T+1})$, состоит из следующей последовательности шагов: 1) подбор подходящей для временного ряда модели ARIMA по методологии Бокса-Дженкинса [13], 2) вычисление прогноза нестационарной компоненты \hat{x}^{ns} на основании выбранной модели ARIMA, 3) вычисление регрессионных остатков $r = \{r_1, \dots, r_T\}$ для выбранной модели ARIMA, 4) задание количества столбцов в гистограмме для алгоритма hist, вычисление прогноза стационарной компоненты \hat{x}^s с помощью алгоритма hist, 5) вычисление прогноза $\hat{x} = \hat{x}^{ns} + \hat{x}^s$ суммированием прогнозов, полученных алгоритмами ARIMA и hist.

В предложенном алгоритме прогноз \hat{x} нестационарного временного ряда складывался из двух частей: прогноза нестационарной компоненты \hat{x}^{ns} и прогноза стационарной компоненты \hat{x}^s : $\hat{x} = \hat{x}^{ns} + \hat{x}^s$.

Таким образом, прежде чем минимизировать ожидаемые потери в задаче (6), все нестационарные особенности оценивались и исключались из временного ряда. Для этого вычислялся прогноз нестационарной компоненты \hat{x}^{ns} . Алгоритм прогнозирования нестационарной компоненты временного ряда должен быть таким, чтобы регрессионные остатки при прогнозе доступной для обучения истории $\{x_1, \dots, x_T\}: r_t = x_t - \hat{x}_t^{ns}, t = 1, \dots, T$ были стационарным временным рядом, значения которого сгенерированы из одного распределения с плотностью $p(u)$. После получения прогноза нестационарной компоненты временного ряда \hat{x}^{ns} прогноз стационарной компоненты \hat{x}^s был получен при помощи оценки плотности распределения $p(u)$ регрессионных остатков $r = \{r_1, \dots, r_T\}$ и решения для этой плотности задачи минимизации ожидаемых потерь (6). Стационарность остатков обеспечивалась выбором подходящей модели ARIMA для прогнозирования нестационарной компоненты.

4 Выбор модели ARIMA и оценка ее параметров

Временной ряд описывается моделью ARIMA(p, d, q), если ряд его разностей $\nabla^d x_t = (1-L)^d x_t$, где L – оператор дифференцирования временного ряда, описывается моделью ARMA(p, q). При необходимости модель ARIMA можно расширить мультипликативной сезонностью. Временной ряд описывается моделью SARIMA, если

$$\Phi_P(L^s)\varphi(L)\nabla_s^D \nabla^d x_t = \alpha + \Theta_Q(L^s)\theta(L)\varepsilon_t, \quad \Phi_P(L^s) = 1 - \Phi_1 L^s - \Phi_2 L^{2s} - \dots - \Phi_P L^{Ps}, \quad (23)$$

$$\Theta_Q L^s = 1 + \Theta_1 L^s + \Theta_2 L^{2s} + \dots + \Theta_Q L^{Qs}. \quad (24)$$

Для оценки параметров модели ARIMA была использована методология Бокса-Дженкинса анализа временных рядов, согласно которой порядок дифференцирования временного ряда d выбирается так, чтобы ряд разностей был стационарным. Параметры p и q выбираются при помощи анализа автокорреляционной функции:

$$\text{ACF}_\tau = \frac{\sum_{i=1}^{T-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sum_{i=1}^T (x_i - \bar{x})^2} \quad (25)$$

(с лагом автокорреляции τ) и частичной автокорреляционной функции

$$\text{PACF}_\tau = \begin{cases} \text{corr}(x_{t+1}, x_t), & \tau = 1; \\ \text{corr}(x_{t+\tau} - x_{t+\tau}^{\tau-1}, x_t - x_t^{\tau-1}), & \tau \geq 2, \end{cases} \quad (26)$$

$$x_t^{\tau-1} = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \dots + \beta_{\tau-1} x_{t+\tau-1},$$

$$x_{t+\tau}^{\tau-1} = \beta_1 x_{t+\tau-1} + \beta_2 x_{t+\tau-2} + \dots + \beta_{\tau-1} x_{t+1},$$

где $\beta_1, \dots, \beta_{\tau-1}$ – коэффициенты линейной регрессии. Выбор параметров p и q осуществлялся из следующих соображений: в модели ARIMA($p, d, 0$) автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а частичная автокорреляционная функция значимо отличается от нуля при лагах, не больших p ; в модели ARIMA($0, d, q$) частичная автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а автокорреляционная функция значимо отличается от нуля при лагах, не больших q . Выбор параметров для модели SARIMA в сезонной компоненте также осуществлялся с помощью анализа автокорреляционной и частичной автокорреляционной функций.

При наличии сезонной компоненты у временного ряда на графиках этих функций наблюдались характерные максимумы в лагах, соответствующих периоду сезонной компоненты. После выбора всех параметров модели, необходимые коэффициенты настраивались путем минимизации квадратичной функции потерь. После обучения модели проводился анализ остатков. Регрессионные остатки проверялись на несмещенность, стационарность, некоррелированность, нормальность, гомоскедастичность (стационарность дисперсии).

5 Анализ качества прогнозов ARIMA + hist

Экспериментальные исследования предложенного алгоритма проводились для трех различных функций потерь: квадратичной

$$l(\hat{x}, x_{T+1}) = (\hat{x} - x_{T+1})^2, \quad (27)$$

абсолютной

$$l(\hat{x}, x_{T+1}) = |x_{T+1} - \hat{x}|, \quad (28)$$

асимметричной функции потерь

$$l(\hat{x}, x_{T+1}) = \begin{cases} 0,5|x_{T+1} - \hat{x}|, & x_{T+1} \leq \hat{x}; \\ 2|x_{T+1} - \hat{x}|, & x_{T+1} > \hat{x}. \end{cases} \quad (29)$$

Все три функции выпуклые, достигают минимума при совпадении прогноза и действительного значения временного ряда. Первые две функции (27) и (28) симметричные, последняя (29) – несимметричная кусочно-линейная функция. Ниже описано применение предложенного алгоритма к прогнозированию временного ряда цен на сахар. Для прогнозирования нестационарной части была выбрана модель ARIMA(1, 0, 0), остатки которой затем были спрогнозированы с помощью алгоритма hist.

Для симметричных функций потерь (27), (28) предельное значение для прогнозов близко к нулю, что означает, что для симметричных функций потерь алгоритм hist не дает существенной поправки к прогнозу нестационарного компонента, полученному с помощью модели ARIMA. В то же время для несимметричной функции потерь (29) предельное значение прогнозов существенно больше нуля. Это значит, что суммарный прогноз значительно превышает прогноз нестационарного компонента, поскольку рассматриваемая функция потерь штрафует недопрогноз гораздо сильнее, чем перепрогноз. При сравнении качества прогнозов модели ARIMA и композиции ARIMA + hist 20% последних точек каждого временного ряда использовались как контрольные. Для каждой контрольной точки по доступной истории временного ряда (все точки от первой до предшествующей рассматриваемой контрольной) оптимизировались параметры выбранной для временного ряда модели ARIMA, и вычислялся ряд регрессионных остатков. По ряду регрессионных остатков оптимизировались параметры модели hist с заданной функцией потерь и заданным количеством столбцов в гистограмме. Прогноз для контрольной точки складывался из прогноза ARIMA и hist. Эксперимент проводился для функций потерь (27)–(29) и вариантов алгоритма hist с 20, 50, 300, 500 столбцами в гистограмме. Средние потери для каждой функции потерь приведены для всех вариантов алгоритма в Табл. 1.

Как видно из Табл 1, при использовании асимметричной функции потерь двухэтапный алгоритм прогнозирования ARIMA + hist позволяет получать среднюю ошибку прогноза существенно ниже, чем прогнозирование с помощью модели ARIMA. При этом для симметричных функций потерь использование двухэтапного алгоритма прогнозирования не приводит к значительным изменениям по сравнению с прогнозом модели ARIMA.

6 Методика оценки результатов вычислительных экспериментов

Оценка результатов вычислительных экспериментов производится следующим образом.

1. Вычислить значения агрегированных ошибок MAE и MAPE для каждого из алгоритмов с детализацией по дням, неделям, месяцам, по станциям и регионам по формулам Табл. 2, где прогнозируемые и исходные значения временных рядов агрегированы как $x(t_W, i, k) = \sum_{t \rightarrow t_W} x(t, i, k)$ – значение временного ряда объемов грузоперевозок, агрегированное по неделям для пары станций,
 $\hat{x}(t_W, i, k) = \sum_{t \rightarrow t_W} \hat{x}(t, i, k)$ – значение прогноза временного ряда объемов грузоперевозок, агрегированное по неделям для пары станций,
 $x(t, I, k) = \sum_{i \in I} x(t, i, k)$ – значение временного ряда объемов грузоперевозок, агрегированное для пары регионов,
 $\hat{x}(t, I, k) = \sum_{i \in I} \hat{x}(t, i, k)$ – значение прогноза временного ряда объемов грузоперевозок, агрегированное для пары регионов.
2. Вычислить усредненные значения агрегированных ошибок MAE и MAPE по парам станций и регионов для всех типов грузов следующим образом:

$$MAE_{ST/REG}^t = \frac{1}{|K|} \sum_{k \in K} MAE_{ST/REG}^t(k), \quad MAPE_{ST/REG}^t = \frac{1}{|K|} \sum_{k \in K} MAPE_{ST/REG}^t(k). \quad (30)$$

Таблица 1 Функция потерь: средние потери прогнозирования для временного ряда Sugar price

Алгоритм	Квадратичная	Абсолютная	Асимметричная
ARIMA	0,127	0,265	0,340
ARIMA+hist(20)	0,128	0,267	0,260
ARIMA+hist(50)	0,127	0,266	0,267
ARIMA+hist(300)	0,127	0,265	0,266
ARIMA+hist(500)	0,127	0,265	0,266

7 Анализ результатов и обсуждение

Значения ошибок прогнозирования разработанного алгоритма Ex VAR (с учетом фактора, оказывающего наибольшее влияние на прогнозируемый временной ряд перевозок) статистически не превосходят значения ошибок прогнозирования алгоритма ARMA. В Табл. 5 приводится оценка ошибок прогнозов, полученных с помощью модуля прогнозирования и по модели ARMA, относительно контрольной выборки данных об объемах спроса на ГП.

Вычисленные значения оценки точности результатов прогнозирования разработанного алгоритма Ex VAR относительно аналогичных результатов алгоритма ARMA для контрольной выборки данных представлен в Табл. 2, 4 и на Рис. 5--10. для пар станций и регионов с детализацией по дням, неделям и месяцам. Относительная точность алгоритмов определена в соответствии с моделью:

$$\delta_{ST}^t = \left(1 - \frac{MAE_{ST}(ExVAR)}{MAE_{ST}(ARMA)} \right), \quad \delta_{REG}^t = \left(1 - \frac{MAE_{REG}(ExVAR)}{MAE_{REG}(ARMA)} \right), \quad t = \{D, M, W\} \quad (31)$$

Таблица 2 Формулы для вычисления агрегированных ошибок MAE и MAPE с детализацией по дням, неделям и месяцам

По парам станций	По парам регионов
$mae_{ST}(t_D, k) = mean_i(x(t_D, i, k) - \hat{x}(t_D, i, k))$	$mae_{REG}(t_D, k) = mean_i(x(t_D, I, k) - \hat{x}(t_D, I, k))$
$mape_{ST}(t_D, k) = \frac{mae(t_D, k)}{\max(1, mean_i(x(t_D, i, k)))}$	$mape_{REG}(t_D, k) = \frac{mae(t_D, k)}{\max(1, mean_i(x(t_D, I, k)))}$
$mae_{ST}(t_W, k) = mean_i(x(t_W, i, k) - \hat{x}(t_W, i, k))$	$mae_{REG}(t_W, k) = mean_i(x(t_W, I, k) - \hat{x}(t_W, I, k))$
$mape_{ST}(t_W, k) = \frac{mae(t_W, k)}{\max(1, mean_i(x(t_W, i, k)))}$	$mape_{REG}(t_W, k) = \frac{mae(t_W, k)}{\max(1, mean_i(x(t_W, I, k)))}$
$mae_{ST}(t_M, k) = mean_i(x(t_M, i, k) - \hat{x}(t_M, i, k))$	$mae_{REG}(t_M, k) = mean_i(x(t_M, I, k) - \hat{x}(t_M, I, k))$
$mape_{ST}(t_M, k) = \frac{mae(t_M, k)}{\max(1, mean_i(x(t_M, i, k)))}$	$mape_{REG}(t_M, k) = \frac{mae(t_M, k)}{\max(1, mean_i(x(t_M, I, k)))}$
$MAE_{ST}^D(k) = mean_{t_D}(mae_{ST}(t_D, k))$	$MAE_{REG}^D(k) = mean_{t_D}(mae_{REG}(t_D, k))$
$MAPE_{ST}^D(k) = mean_{t_D}(mape_{ST}(t_D, k))$	$MAPE_{REG}^D(k) = mean_{t_D}(mape_{REG}(t_D, k))$
$MAE_{ST}^W(k) = mean_{t_W}(mae_{ST}(t_W, k))$	$MAE_{REG}^W(k) = mean_{t_W}(mae_{REG}(t_W, k))$
$MAPE_{ST}^W(k) = mean_{t_W}(mape_{ST}(t_W, k))$	$MAPE_{REG}^W(k) = mean_{t_W}(mape_{REG}(t_W, k))$
$MAE_{ST}^M(k) = mean_{t_M}(mae_{ST}(t_M, k))$	$MAE_{REG}^M(k) = mean_{t_M}(mae_{REG}(t_M, k))$
$MAPE_{ST}^M(k) = mean_{t_M}(mape_{ST}(t_M, k))$	$MAPE_{REG}^M(k) = mean_{t_M}(mape_{REG}(t_M, k))$

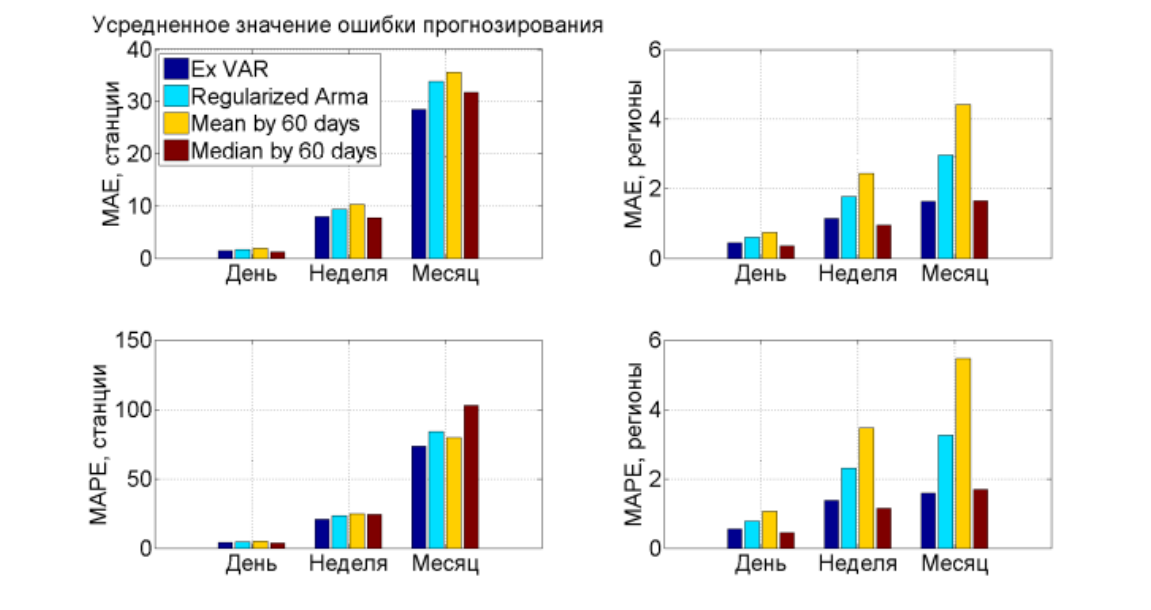


Рис. 2 Значения ошибок прогнозирования MAE и MAPE

Ранжированные значения точности разработанного алгоритма Ex VAR относительно ARMA для пар станций и регионов с детализацией по дням, неделям, месяцам приведены в Табл. 2, 4 соответственно. Положительным значениям точности соответствуют меньшие значения ошибки прогнозирования алгоритма Ex VAR по сравнению с алгоритмом ARMA.

Результаты для типов грузов «33 Сахарная свекла и семена» и «4 Торф» исключены из представленных в виду их низкой репрезентативности, связанной с наличием значительного количество нулевых значений объемов перевозок данного типа груза.

Из результатов, представленных в Табл. 5 следует, что точность алгоритма Ex VAR превышает точность алгоритма ARMA с детализацией по дням:

- для пар станций – для 30 типов грузов из 36 (83%);
- для пар регионов – для 30 типов грузов из 36 (83%).

Таблица 3 Значения ошибок MAE и MAPE, усредненные ошибки прогнозирования для всех типов груза

MAE, Станции	Д	Н	М	MAPE, Станции	Д	Н	М
Ex VAR	1,49	7,96	28,46	Ex VAR	3,96	20,90	73,67
Regularized ARMA	1,70	9,40	33,83	Regularized ARMA	4,39	23,59	83,97
Mean by 60 days	1,93	10,27	35,65	Mean by 60 days	4,89	24,64	79,42
Median by 60 days	1,28	7,70	31,77	Median by 60 days	3,95	24,21	102,76
MAE, Регионы	Д	Н	М	MAPE, Регионы	Д	Н	М
Ex VAR	0,45	1,14	1,64	Ex VAR	0,55	1,37	1,60
Regularized ARMA	0,60	1,77	2,97	Regularized ARMA	0,79	2,30	3,25
Mean by 60 days	0,74	2,43	4,42	Mean by 60 days	1,06	3,47	5,47
Median by 60 days	0,36	0,95	1,65	Median by 60 days	0,44	1,14	1,69

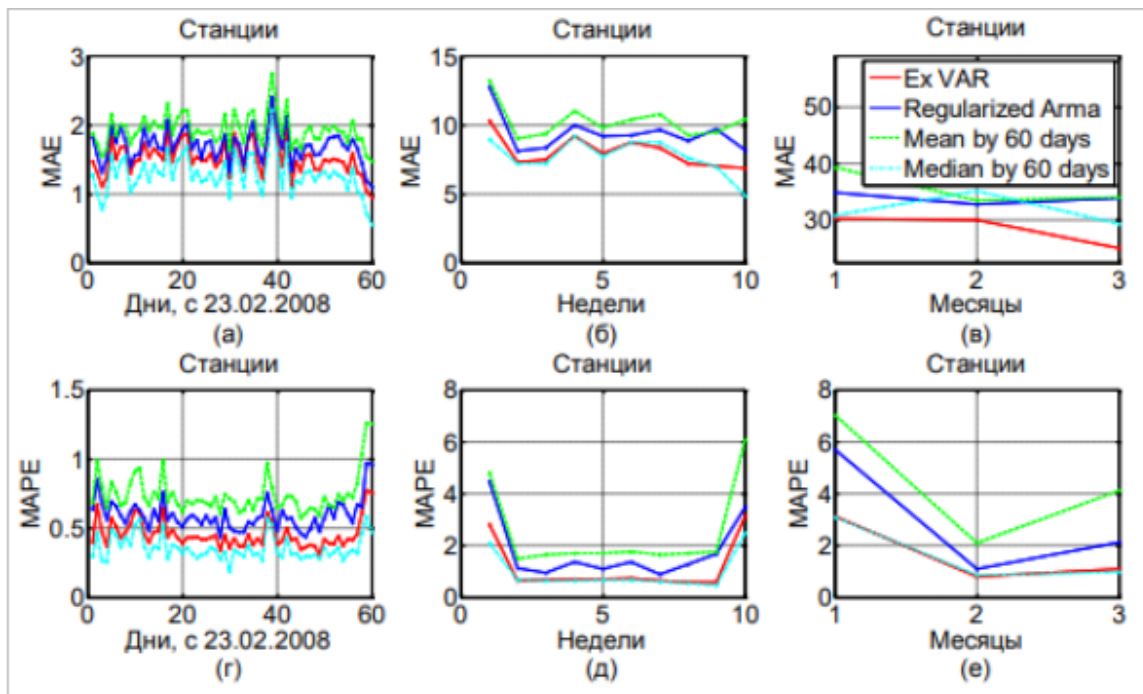


Рис. 3 Графики значений ошибок прогнозирования MAE и MAPE в ретроспективном разрезе, усредненные по всем типам грузов для пар станций

Из результатов, представленных в Табл.5 следует, что точность алгоритма Ex VAR превышает точность алгоритма ARMA с детализацией по неделям:

- для пар станций – для 32 типов грузов из 36 (89%);
- для пар регионов – для 29 типов грузов из 36 (81%).

Из результатов, представленных в Табл. 5 следует, что точность алгоритма Ex VAR превышает точность алгоритма ARMA с детализацией по месяцам:

- для пар станций – для 29 типов грузов из 36 (81%);
- для пар регионов – для 26 типов грузов из 36 (72%).

Из сравнения результатов прогнозирования с контрольной выборкой данных и аналогичными результатами по модели ARMA следует, что алгоритм Ex VAR целесообразно использовать для прогнозирования объемов перевозок для пар станций и регионов с детализацией по дням.

Таблица 4 Сводная таблица ошибок прогнозирования MAE и MAPE для каждого из алгоритмов с детализацией по дням, неделям, месяцам, агрегированных по парам станций и регионов

1 Каменный уголь								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	5,25	32,20	106,32	Ex VAR	1,14	1,62	1,41	
Regularized ARMA	5,80	34,13	108,40	Regularized ARMA	1,29	1,80	1,47	
Mean by 60 days	6,79	39,64	125,66	Mean by 60 days	1,54	1,99	1,74	
Median by 60 days	4,97	31,69	110,22	Median by 60 days	1,07	1,56	1,42	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	16,96	102,76	331,05	Ex VAR	1,31	1,52	1,21	
Regularized ARMA	18,40	104,90	316,60	Regularized ARMA	1,49	1,67	1,20	
Mean by 60 days	20,12	107,99	321,61	Mean by 60 days	1,71	1,73	1,33	
Median by 60 days	16,92	108,50	365,73	Median by 60 days	1,27	1,54	1,31	
2 Кокс								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	8,73	46,04	129,11	Ex VAR	2,35	9,20	0,74	
Regularized ARMA	8,15	43,66	121,48	Regularized ARMA	1,58	5,18	0,63	
Mean by 60 days	9,35	49,04	162,48	Mean by 60 days	2,28	8,32	1,02	
Median by 60 days	8,46	44,20	147,67	Median by 60 days	2,02	7,36	0,89	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	11,83	60,20	160,53	Ex VAR	2,82	12,22	0,64	
Regularized ARMA	11,09	58,82	161,09	Regularized ARMA	1,85	6,75	0,61	
Mean by 60 days	12,59	64,91	210,76	Mean by 60 days	2,71	11,03	0,94	
Median by 60 days	11,53	59,00	181,98	Median by 60 days	2,42	9,78	0,76	
3 Нефть								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	14,80	70,02	229,75	Ex VAR	1,22	1,64	1,40	
Regularized ARMA	14,49	71,19	243,73	Regularized ARMA	1,17	1,59	1,45	
Mean by 60 days	16,35	77,01	248,14	Mean by 60 days	1,35	1,82	1,53	
Median by 60 days	12,11	71,99	296,33	Median by 60 days	0,95	1,43	1,69	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	47,82	235,42	717,26	Ex VAR	0,85	1,26	0,91	
Regularized ARMA	50,31	256,51	864,15	Regularized ARMA	0,87	1,23	1,01	
Mean by 60 days	51,47	247,24	693,99	Mean by 60 days	0,92	1,39	0,89	
Median by 60 days	51,96	320,62	1367,81	Median by 60 days	0,86	1,32	1,66	
4 Торф								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	0,00	0,00	0,00	Ex VAR	0,00	0,00	0,00	
Regularized ARMA	0,02	0,13	0,56	Regularized ARMA	0,02	0,13	0,56	
Mean by 60 days	0,58	4,08	18,08	Mean by 60 days	0,58	4,08	18,08	
Median by 60 days	0,00	0,00	0,00	Median by 60 days	0,00	0,00	0,00	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	0,00	0,00	0,00	Ex VAR	0,00	0,00	0,00	
Regularized ARMA	0,02	0,17	0,74	Regularized ARMA	0,02	0,17	0,74	
Mean by 60 days	0,78	5,44	24,11	Mean by 60 days	0,78	5,44	24,11	
Median by 60 days	0,00	0,00	0,00	Median by 60 days	0,00	0,00	0,00	
6 Флюсы								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	0,58	3,48	14,12	Ex VAR	0,08	0,20	0,80	
Regularized ARMA	1,83	11,05	42,07	Regularized ARMA	1,41	8,58	21,31	
Mean by 60 days	1,24	7,76	21,14	Mean by 60 days	0,73	4,19	8,00	
Median by 60 days	0,58	3,48	14,12	Median by 60 days	0,08	0,20	0,80	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	0,92	5,52	22,35	Ex VAR	0,08	0,20	0,80	
Regularized ARMA	2,89	17,49	66,62	Regularized ARMA	2,18	13,45	33,00	
Mean by 60 days	1,94	11,89	25,08	Mean by 60 days	1,10	6,45	11,90	
Median by 60 days	0,92	5,52	22,35	Median by 60 days	0,08	0,20	0,80	
7 Руда железная								
MAE - Станции	Д	Н	М	MAPE - Станции	Д	Н	М	
Ex VAR	4,49	26,40	90,55	Ex VAR	1,67	1,91	2,09	
Regularized ARMA	7,48	48,50	185,08	Regularized ARMA	3,19	3,36	3,89	
Mean by 60 days	4,49	26,76	90,20	Mean by 60 days	1,67	1,95	2,19	
Median by 60 days	3,23	25,54	106,15	Median by 60 days	0,90	1,73	2,28	
MAE - Регионы	Д	Н	М	MAPE - Регионы	Д	Н	М	
Ex VAR	7,44	42,44	148,59	Ex VAR	1,84	1,68	1,84	
Regularized ARMA	12,97	80,10	300,63	Regularized ARMA	3,78	3,01	3,35	
Mean by 60 days	7,50	41,07	138,02	Mean by 60 days	1,87	1,62	1,80	
Median by 60 days	5,75	45,50	189,14	Median by 60 days	0,92	1,73	2,28	

Таблица 5 Сводная таблица значений оценки точности алгоритма Ex VAR относительно ARMA (в %) для пар станций и регионов с детализацией по дням, неделям, месяцам

Тип груза	Оценка точности для пар станций и регионов					
	Станции			Регионы		
	Д	Н	М	Д	Н	М
1 Каменный уголь	9,44	5,66	1,92	7,83	2,04	-4,57
2 Кокс	-7,04	-5,44	-6,28	-6,63	-2,35	0,35
3 Нефть	-2,12	1,65	5,74	4,94	8,22	17,00
4 Торф	100,0	100,0	100,0	100,0	100,0	100,0
6 Флосы	68,23	68,46	66,44	68,23	68,45	66,44
7 Руда железная	39,93	45,56	51,07	42,63	47,02	50,57
8 Руда цветная	15,26	14,37	6,26	13,72	10,50	-5,96
9 Черные металлы	27,14	31,09	32,58	24,29	27,48	15,03
10 Машины и оборудование	37,33	34,74	26,31	37,16	33,48	19,82
11 Металлические конструкции	28,59	41,01	31,54	26,27	33,16	12,91
12 Метизы	1,81	1,26	5,14	4,25	-0,09	3,57
13 Лом черных металлов	15,00	20,20	16,94	-3,84	-3,19	2,23
14 Сельскохозяйственные машины	2,02	8,22	6,25	1,45	7,68	7,44
15 Автомобили	81,41	83,98	69,59	81,39	83,89	69,10
16 Цветные металлы	18,08	21,94	21,51	18,05	21,83	21,31
17 Химические и минеральные удобрения	29,73	17,26	-7,98	24,13	-5,00	-40,45
18 Химикаты и сода	-10,91	-18,77	-35,81	-10,00	-18,44	-36,24
19 Строительные грузы	3,18	0,94	-9,86	2,19	-8,84	-19,00
20 Промышленное сырье	29,48	19,00	14,11	29,47	18,94	14,00
21 Шлаки гранулированные	20,44	35,28	21,60	20,44	35,28	21,60
22 Огнеупоры	20,99	29,32	21,92	20,95	29,12	21,36
23 Цемент	24,02	21,43	17,77	20,90	12,84	-8,37
24 Лесные грузы	39,58	42,05	36,95	38,82	40,16	35,47
25 Сахар	-2,16	2,77	-1,90	-2,05	2,58	-1,37
26 Мясо и масло животное	15,20	19,85	28,79	14,50	16,34	24,00
27 Рыба	52,56	57,46	63,18	51,64	52,36	58,46
28 Картофель, овощи и фрукты	36,78	29,16	-8,57	35,71	26,07	-34,47
29 Соль поваренная	32,70	34,29	21,87	32,70	34,29	21,87
30 Остальные продовольственные товары	-8,54	-3,62	6,36	-3,77	2,44	8,95
31 Промышленные товары народного потребления	-8,61	-18,46	-47,11	-6,09	-6,16	-34,58
34 Зерно	6,70	6,72	10,68	4,11	5,23	4,16
35 Продукты перемола	11,25	13,19	14,50	11,42	12,15	10,79
36 Комбикорма	30,25	22,25	25,46	30,23	22,16	25,24
38 Жмыхи	86,71	87,65	78,95	86,69	87,57	78,11
39 Бумага	10,44	11,44	6,00	10,42	11,27	5,66
42 Грузы в контейнерах	43,64	38,27	43,21	43,56	37,99	42,29
43 Остальные и сборные грузы	8,58	9,99	10,74	7,25	2,60	-1,74

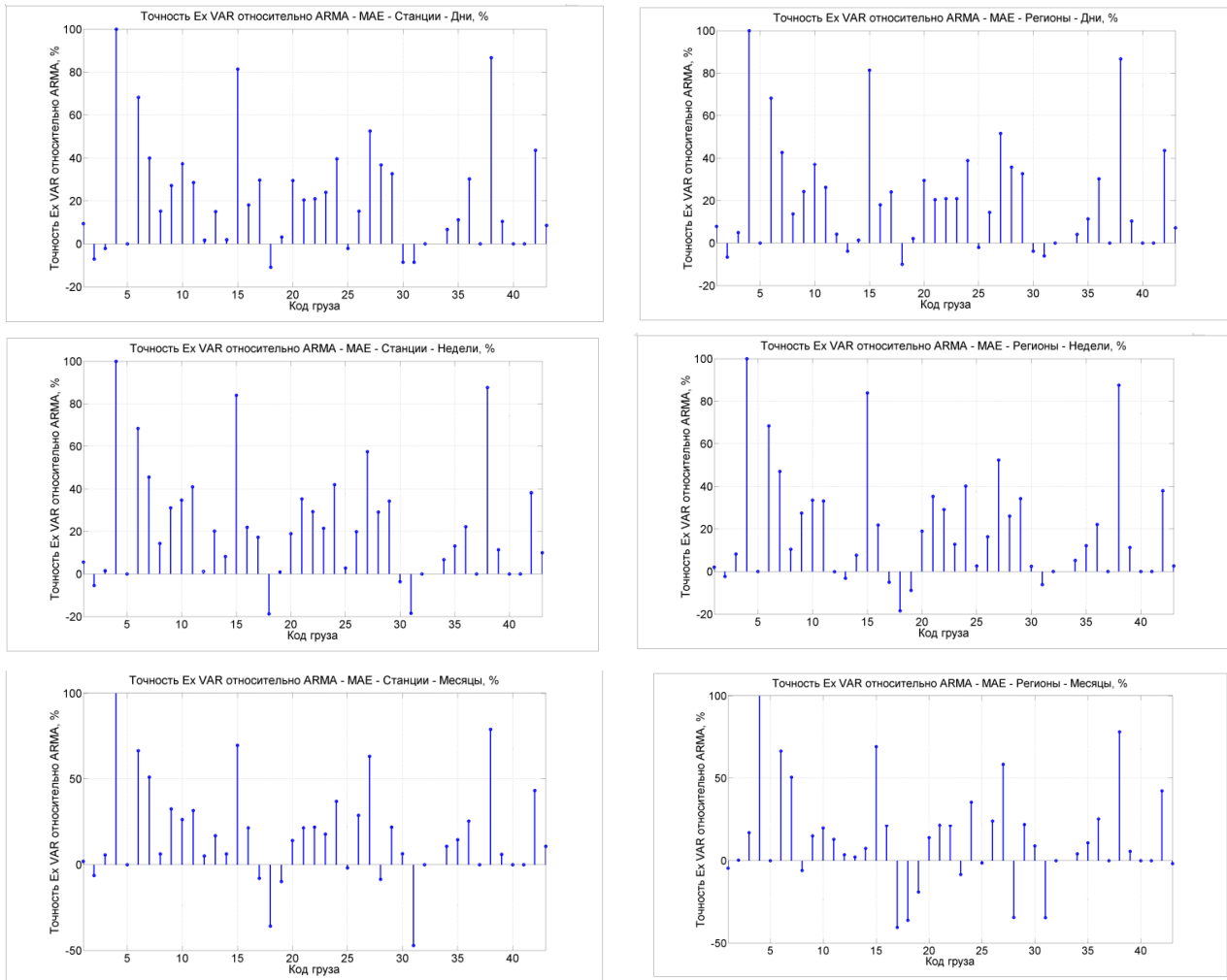


Рис. 4 Оценка точности Ex VAR относительно ARMA для пар станций и регионов с детализацией по дням, неделям, месяцам, %

Положительный результат при сравнении ошибок прогнозирования для алгоритма Ex VAR достигнут для 83% и 83% исходных данных для пар станций и регионов с детализацией по дням соответственно. Положительный результат при сравнении ошибок прогнозирования для алгоритма Ex VAR достигнут для 89% и 81% исходных данных для пар станций и регионов с детализацией по неделям соответственно. Положительный результат при сравнении ошибок прогнозирования для алгоритма Ex VAR достигнут для 81% и 71% исходных данных для пар станций и регионов с детализацией по месяцам.

8 Заключение

Результаты сравнения полученных значений ошибок соответствуют ожиданиям экспертов. Значения ошибок прогнозирования и оценки точности алгоритмов Ex VAR и ARMA зависят от качества исходных данных, но в целом, по результатам вычислительного эксперимента можно утверждать о статистическом превосходстве точности разработанного алгоритма Ex VAR с учетом экзогенных факторов по сравнению с ARMA без учета экзогенных факторов.

Литература

- [1] Попов П. А. Стохастическое моделирование движения поездов в задачах интервального регулирования // Наука и транспорт, 2013, Т. 6, № 2. С. 8–19.
- [2] Кокурин И. М., Катцын Д. В., Малыгин И. Г. Организация экспорта каменного угля на основе концентрации логистической цепи поставок // Наука и транспорт, 2013, Т. 6, № 2. С. 15–17.
- [3] Иванченко В. Н. Современные информационные технологии управления сложными процессами расформирования–формирования поездов // Наука и транспорт, 2013, Т. 6, № 2. С. 64–69.
- [4] Tong M., Xue H. Highway Traffic Volume Forecasting Based on Seasonal ARIMA Model // J. of Highway and Transp. Res. and Develop., 2008. 3(2):109–112.
- [5] Duddu V. R., Pulugurtha S. S. Principle of Demographic Gravitation to Estimate Annual Average Daily Traffic: Comparison of Statistical and Neural Network Models // J. of Transportation Engineering, 2013. 139(6):585–595.
- [6] Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. — М.: ЮНИТИ-ДАНА, 2001.
- [7] Албегов М. М., Бурса Б. И., Истомина Р. П., Медведев В. Г и др. Краткосрочное прогнозирование регионального развития в условиях неполной информации. — М.: Едиториал УРСС, 2001.
- [8] Терёшина Н. П., Галабурда В. Г., Токарев В. А. и др. Экономика железнодорожного транспорта: Учебник для вузов ж.-д. транспорта. — М.: УМЦ ЖДТ, 2008.
- [9] Гасников А. В. Заметка об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях // Математическое моделирование, arXiv:1410.3123, 2015.
- [10] Ващенко М. П., Гасников А. В., Молчанов Е. Г. и др. Вычислимые модели и численные методы для анализа тарифной политики железнодорожных грузоперевозок // Сообщения по прикладной математике, С. 1–51. ВЦ РАН Москва, 2014.
- [11] Шананин А. А. Вычислимая модель железнодорожных грузоперевозок с учетом коммуникационных ограничений // Тезисы докладов 10-й Международной конференции Интеллектуализация обработки информации. Греция, о. Крит, 2014, С. 192.
- [12] Постановление Правительства РФ от 18.05.2001 № 384 “О программе структурной реформы на железнодорожном транспорте”.
- [13] Box G. E. P., Jenkins G. M., Reinsel G. C. Time Series Analysis: Forecasting and Control — Englewood Cliffs, 3rd edition, 1994.
- [14] Patton A. J. and Timmermann A. Properties of optimal forecasts under asymmetric loss and nonlinearity // Journal of Econometrics, 2007. 140(2), pp. 884–918.
- [15] Berk R. Asymmetric loss functions for forecasting in criminal justice settings // Journal of Quantitative Criminology, 2011. 27(1), pp. 107–123.
- [16] Cipra T. Asymmetric recursive methods for time series // Applications of Mathematics, 1994. 39(3), pp. 203–214.
- [17] Koenker R. and Zhijie Xiao Quantile autoregression // Journal of the American Statistical Association, 2006. 101(475), pp. 980–990.
- [18] Koenker R. Quantile regression. — Cambridge university press, 2005.
- [19] Рудаков К. В., Стрижов В. В. и др. Обоснование и выбор направления исследований. Исследование экзогенных факторов и их влияния на объемы спроса на грузовые железнодорожные перевозки // Отчет о ПНИ «Исследование и разработка математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования объемов

- спроса на грузовые железнодорожные перевозки)» Этап первый. — М.: ВЦ РАН, 2014, 124 с.
- [20] Рудаков К. В., Стрижов В. В. и др. Исследование и разработка математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки // Отчет о ПНИ «Исследование и разработка математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования объемов спроса на грузовые железнодорожные перевозки)» Этап второй. — М.: ФИЦИУ РАН, 2014, 112 с.
- [21] Biau G., Bleakley K., Györfi L., Ottucsák G. Nonparametric sequential prediction of time series // Journal of Nonparametric Statistics, 2010. 22(3), pp. 297–317.
- [22] Кудрявцев Л. Д. Курс математического анализа. В 3 томах. Т. 2. Ряды. Дифференциальное и интегральное исчисления функций многих переменных. — М.: Дрофа, 2004
- [23] Granger C. W. J. Prediction with a generalized cost of error function // Operational Research Society, 1969. 20(2), pp. 199–207.
- [24] Christoffersen P. F., Diebold F. X. Optimal prediction under asymmetric loss // Econometric theory, 1997. 13(06), pp. 808–817.

Поступила в редакцию 28.05.2018

Forecasting amount of demand for cargo transportation for stationary time series*

*A. D. Voronov*¹, *A. N. Gromov*², *A. S. Inyakin*,² *A. A. Zamkovoy*³
voronov.ad@phystech.edu; gromlogin@gmail.com; inyakin@forecsys.ru;
Zamkovoy.al.an@gmail.com

¹Moscow institute of physics and technology, 141701 9 Institutskiy per., Dolgoprudny, Moscow Region, Russian Federation; ²Federal research center Informatics and Management of the Russian Academy of Science, 119333 Moscow Ulitsa Vavilova 44; ³Corporation Institute of Economics and Transport Development, 105066 Moscow Novoryazanskaya ulitsa 24

The properties of prognostic models of volumes of demand for freight rail transportation with the purpose of structuring processes in the field of management and planning of freight rail transportation are investigated. The paper proposes four models for forecasting the volumes of demand for freight rail transportation, taking into account the specificity of the measured data, business processes and standards of the industrial partner. When constructing models, multivariate statistical analysis and forecasting of interdependent time series are used. The properties of the constructed models are analyzed. Forecasts are made in the sections of day, week, month for stations and regions. The proposed prognostic models are compared by the criteria of the average absolute and average percentage error.

Keywords: *time series; forecasting; superposition; forecasting models; asymmetrical distribution; regression residues*

DOI: 10.21469/22233792.4.1.05

References

- [1] Popov P. 2013. Stokhasticheskoe modelirovanie dvizheniya poezdov v zadachakh interval'nogo regulirovaniya [Stochastic modeling of train traffic in interval control problems]. Nauka i transport [Science and transport] 6(2):8–19. (In Russian)

*The research was supported by the Russian Foundation for Basic Research (projects 17-20-01212, 17-20-01184).

- [2] Kokurin I., Katsyn D., Malygin I. 2013. Organizatsiya eksporta kamennogo uglya na osnove kontsentratsii logisticheskoy tsepi postavok [The organization of the export of coal on the basis of the concentration of the logistics supply chain]. *Nauka i transport*[Science and transport] 6(2):15–17. (In Russian)
- [3] Ivanchenko V. 2013. Sovremennye informatsionnye tekhnologii upravleniya slozhnymi protsessami rasformirovaniya–formirovaniya poezdov [Modern information technologies for managing complex processes of disbanding-forming trains]. *Nauka i transport*[Science and transport] 6(2):64–69. (In Russian)
- [4] Tong M., Xue H. Highway Traffic Volume Forecasting Based on Seasonal ARIMA Model // *J. of Highway and Transp. Res. and Develop.*, 2008. 3(2):109–112.
- [5] Duddu V. R., Pulugurtha S. S. Principle of Demographic Gravitation to Estimate Annual Average Daily Traffic: Comparison of Statistical and Neural Network Models // *J. of Transportation Engineering*, 2013. 139(6):585–595.
- [6] Aivazyan S., Mkhitarjan V. 2001. *Prikladnaya statistika i osnovy ekonometrii* [Applied statistics and principles of econometrics]. oscar: UNITI-DANA. (In Russian)
- [7] Albegov M., Bursa B., Istomina R., Medvedev V. 2001. *Kratkosrochnoe prognozirovanie regional'nogo razvitiya v usloviyakh nepolnoy informatsii* [Short-term forecasting of regional development in conditions of incomplete information]. Moscow: Editorial URSS. (In Russian)
- [8] Tereshina N., Galaburda V., Tokarev V. 2008. *Ekonomika zheleznodorozhnogo transporta: Uchebnik dlya vuzov zh.-d. transporta* [Economics of Railway Transport: A Textbook for Higher Educational Institutions of Railway Transport]. Moscow: UMCZDT. (In Russian)
- [9] Gasnikov A. 2015. *Zametka ob effektivnoy vychislivosti konkurentnykh ravnovesiy v transportno-ekonomicheskikh modelyakh* [A Note on the Effective Computability of Competitive Equilibria in Transport-Economic Models]. *Matematicheskoe modelirovanie*[Mathematical Modeling] arXiv:1410.3123 (In Russian)
- [10] Vaschenko M., Gasnikov A., Molchanov E. 2014. *Vychislimye modeli i chislennye metody dlya analiza tarifnoy politiki zheleznodorozhnykh gruzoperevozok* [Computable models and numerical methods for the analysis of the tariff policy of rail freight]. *Soobshcheniya po prikladnoy matematike*[Reports on applied mathematics]. Place of publishing: Moscow, CC RAS. 1–51 (In Russian)
- [11] Shanenin A. 2014. *Vychislimaya model' zheleznodorozhnykh gruzoperevozok s uchetom kommunikatsionnykh ogranicheniy* [Computable model of rail freight transport taking into account communication restrictions]. *Tezisy dokladov 10-y Mezhdunarodnoy konferentsii Intellektualizatsiya obrabotki informatsii*[Abstracts of the 10th International Conference Intellectualization of information processing]. 192–192. (In Russian)
- [12] *Postanovlenie Pravitel'stva RF ot 18.05.2001 384 "O programme strukturnoy reformy na zheleznodorozhnom transporte"*[Decree of the Government of the Russian Federation of May 18, 2001 384 "On the program of structural reform in railway transport"]. (In Russian)
- [13] Box G. E. P., Jenkins G. M., Reinsel G. C. 1994. *Time Series Analysis: Forecasting and Control* — Englewood Cliffs, 3rd edition.
- [14] Patton A. J. and Timmermann A. 2007. Properties of optimal forecasts under asymmetric loss and nonlinearity // *Journal of Econometrics*. 140(2):884–918.
- [15] Berk R. 2011. Asymmetric loss functions for forecasting in criminal justice settings // *Journal of Quantitative Criminology*. 27(1):107–123.
- [16] Cipra T. 1994. Asymmetric recursive methods for time series // *Applications of Mathematics*. 39(3):203–214.

- [17] Koenker R. and Zhijie Xiao 2006. Quantile autoregression // Journal of the American Statistical Association. 101(475):980–990.
- [18] Koenker R. 2005. Quantile regression. — Cambridge university press.
- [19] Rudakov K., Strizhov V. 2014. Obosnovanie i vybor napravleniya issledovaniy. Issledovanie ekzogenykh faktorov i ikh vliyaniya na ob”emy sprosna na gruzovye zheleznodorozhnye perezovzki [Rationale and choice of research direction. Investigation of exogenous factors and their impact on the volume of demand for freight rail transportation]. Otchet o PNI «Issledovanie i razrabotka matematicheskikh metodov i algoritmov dlya intellektual’noy sistemy analiza dannykh (podsistemy prognozirovaniya ob”emov sprosna na gruzovye zheleznodorozhnye perezovzki)» Etap pervyy.[Report on PNI ”Research and development of mathematical methods and algorithms for an intelligent data analysis system (subsystems for forecasting the volumes of demand for freight rail transportation)”Stage one.] Place of publication: Moscow, CC RAS, 124 p. (In Russian)
- [20] Rudakov K., Strizhov V. 2014. Issledovanie i razrabotka matematicheskoy modeli prognozirovaniya ob”emov sprosna na gruzovye zheleznodorozhnye perezovzki [Research and development of a mathematical model for forecasting the volumes of demand for freight rail transportation]. Otchet o PNI «Issledovanie i razrabotka matematicheskikh metodov i algoritmov dlya intellektual’noy sistemy analiza dannykh (podsistemy prognozirovaniya ob”emov sprosna na gruzovye zheleznodorozhnye perezovzki)» Etap vtoroy.[Report on PNI ”Research and development of mathematical methods and algorithms for an intelligent data analysis system (subsystems for forecasting the volumes of demand for freight rail transportation)”Stage two.] Place of publication: Moscow, FRC CSC RAS, 112 p. (In Russian)
- [21] Biau G., Bleakley K., Gy`rfi L., Ottucsák G. 2010. Nonparametric sequential prediction of time series // Journal of Nonparametric Statistics. 22(3):297–317.
- [22] Kudryavtsev L. 2004. Kurs matematicheskogo analiza. V tryokh tomakh. Tom 2: Ryady. Differentsial’noe i integral’noe ischislenie funktsiy mnogikh peremennykh[Course of mathematical analysis. In 3 volumes. Volume 2: Rows. Differential and integral calculus of functions of several variables]. oscow: Drofa. (In Russian)
- [23] Granger C. W. J. 1969. Prediction with a generalized cost of error function // Operational Research Society. 20(2):199–207.
- [24] Christoffersen P. F., Diebold F. X. 1997. Optimal prediction under asymmetric loss // Econometric theory. 13(06):808–817.

Received May 28, 2018