ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ "ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "ИНФОРМАТИКА И УПРАВЛЕНИЕ" РОССИЙСКОЙ АКАДЕМИИ НАУК " (ФИЦ ИУ РАН)

УДК 65.012.226

ВГК ОКП 506190

№ госрегистрации

Инв. №

УТВЕРЖДАЮ
Директор ФИЦ ИУ РАН

_____ Соколов И. А.

____._____.2015 г.

ОТЧЕТ

О ПРИКЛАДНЫХ НАУЧНЫХ ИССЛЕДОВАНИЯХ

Исследование и разработка математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования объемов спроса на грузовые железнодорожные перевозки)

по теме:

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МАТЕМАТИЧЕСКОЙ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОБЪЕМОВ СПРОСА НА ГРУЗОВЫЕ ЖЕЛЕЗНОДОРОЖНЫЕ ПЕРЕВОЗКИ

(промежуточный)

Этап второй

Соглашение о предоставлении субсидии от 19 июня 2014 г. № 14.604.21.0041

ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса на 2014-2020 годы»

Приоритетное направление «Транспортные и космические системы»

Руководитель работ _____ К.В. Рудаков

Москва 2015

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель работ, член-корреспондент РАН	_____ (подпись, дата)	К.В. Рудаков (все разделы – рук.)
Исполнители темы:		
Ведущий научный сотрудник, доктор физ.-мат. наук	_____ (подпись, дата)	Стрижов В.В. (все разделы)
Студент	_____ (подпись, дата)	Мотренко А. П. (все разделы)
Аспирант	_____ (подпись, дата)	Кузнецов М.П. (все разделы)
Инженер-исследователь	_____ (подпись, дата)	Каширин Д.О. (разделы: 4)
Нормоконтролёр ведущ.научн.сотр., д.т.н., проф.	_____ (подпись, дата)	А.И. Эрлих

РЕФЕРАТ

Отчёт 110 с., 44 рис., 10 табл., 67 источников.

АСИММЕТРИЧНАЯ ФУНКЦИЯ ПОТЕРЬ, ВРЕМЕННОЙ РЯД, ГИСТОГРАММНОЕ ПРОГНОЗИРОВАНИЕ, ГРУЗОВЫЕ ЖЕЛЕЗНОДОРОЖНЫЕ ПЕРЕВОЗКИ, НЕПАРАМЕТРИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ, НЕСТАЦИОНАРНЫЙ ВРЕМЕННОЙ РЯД, ПРОГНОЗИРОВАНИЕ, РЖД, СМЕСИ ГИСТОГРАММ, ЭКЗОГЕННЫЙ ВРЕМЕННОЙ РЯД.

Объект прикладных научных исследований – объемы спроса на грузовые железнодорожные перевозки, оценка влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки и структура процессов в области управления и планирования грузовых железнодорожных перевозок.

Цель ПНИ: Разработка математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера – ОАО «РЖД».

Методология проведения ПНИ – многофакторный статистический анализ и прогнозирование взаимозависимых временных рядов.

Основные результаты, полученные на втором этапе ПНИ:

- Разработана математическая модель прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающая влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера.
- Разработан макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки и выполнена серия вычислительных экспериментов по прогнозированию объёмов спроса на грузовые железнодорожные перевозки на модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов.
- Разработан генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов.
- Разработаны программа и методика тестирования генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенные факторы.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

Временной ряд	Последовательность значений наблюдаемого процесса, измеренных через равные промежутки времени
Гистограмма	Приближение плотности распределения вероятности случайной величины, построенное по выборке данных
Железнодорожный узел	Пункт пересечения или примыкания не менее трех железнодорожных линий, ряд связанных соединительными ходами станций, работающих по единой технологии
Непараметрическое прогнозирование	Способ вычисления прогноза, при котором модель не описывается конечным набором параметров
Обучающая выборка	Часть выборки данных, используемая при построении модели для оценки параметров
Прогнозирование	Расчет будущих значений наблюдаемого процесса на основе математической модели
Пропускная способность сети	Предельное количество единиц транспорта, проходящих через сеть в единицу времени
Регрессионная модель	Функция от набора независимых переменных, принадлежащая некоторому параметрическому семейству. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные
Ретроспективный прогноз	Вид прогнозирования, используемый для оценки качества метода прогнозирования. Прогноз выполняется на участке ряда, значения которого известны. Качество прогноза оценивается сравнением спрогнозированных значений, с известными значениями
Стационарность	Под стационарным временным рядом понимается ряд, математическое ожидание и дисперсия которого постоянны во времени
Функция ошибки	Функция, задающая ошибку прогнозирования на основе сравнения прогноза и истинного значения спрогнозированной величины
Экзогенный фактор	Фактор, изменение которого происходит вне моделируемой системы. Временной ряд, описывающий поведение экзогенного фактора, не является производным от временных рядов, описывающих поведение эндогенных факторов

Эмпирическое распределение	Кусочно-постоянная аппроксимация функции распределения случайной величины, построенная по выборке данных
Эндогенный фактор	Фактор, изменение которого происходит внутри моделируемой системы. Здесь это – объёмные показатели групп перевозимых грузов
ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ	
$A \otimes B$	Декартово произведение множеств A и B
ACF_τ	Автокорреляционная функция с лагом τ
b	Ширина интервалов гистограммы
d	Количество единичный корней (порядок дифференцирования модели ARIMA)
$\mathbf{E}X$	Математическое ожидание случайной величины X
h	Ширина окна, используемая при ядерных оценках плотности
$h_1, \dots, h_n$	Высота столбцов гистограммы
$\mathrm{KL}(\hat{p}\|\|p)$	Расстояние Кульбака-Лейблера между распределениями $\hat{p}$ и p
$K(r)$	Ядерная функция
L	Оператор дифференцирования временного ряда
$\mathbb{N}$	Множество натуральных чисел
n	Количество столбцов гистограммы
p	Порядок лагирования авторегрессионной модели AR
$PACF_\tau$	Частичная автокорреляционная функция с лагом τ
$\hat{p}(u)$	Кусочно-постоянная функция, задающая гистограмму
$\hat{p}_{\mathbf{x},\mathbf{Y}}(u, v_1, \dots, v_N)$	Многомерная гистограмма
$\hat{p}_{\mathbf{x}\|\mathbf{Y}}(u, v_1, \dots, v_N)$	Срез многомерной гистограммы

$\hat{p}_{\mathbf{x}\|\mathbf{y}^j}(u,v)$	Гистограмма ряда **x**, условная по *j*-му экзогенному ряду $\mathbf{y}^j$
q	Порядок модели скользящего среднего MA
$\mathbb{R}_+$	Множество неотрицательных действительных чисел
$r_t = x_t - \hat{x}_t^{ns}$	Регрессионные остатки
$T_{\min}$	Минимальная длина предыстории, необходимая для применения алгоритма прогнозирования
$u_0 = u_{min}, \dots, u_i, \dots, u_n = u_{max}$	Концы интервалов гистограммы
$\mathbf{w} = [w_0, \dots, w_N]^T$	Вектор весов компонент смеси гистограмм, задающих априорное распределение компонент смеси
w_{jt}, $j \in [1, \dots, n]$, $t \in [1, \dots, T]$	Апостериорное распределение компонент смеси гистограмм
$\mathbf{x} = \{x_1, \dots, x_T\}$	Временной ряд объёмов спроса на грузовые железнодорожные перевозки
$\hat{x}$	Прогноз для следующего значения x_{T+1} временного ряда $\mathbf{x} = \{x_1, \dots, x_T\}$
$\hat{x}^{ns}$	Прогноз нестационарной компоненты временного ряда,
$\hat{x}^{s}$	Прогноз стационарной компоненты временного ряда
$\tilde{\mathbf{x}}$	Синтетический временной ряд, сгенерированный из распределения, задаваемого смесью гистограмм
$\mathbf{y}^j$	Экзогенный Временной ряд *j*-го экзогенного фактора
$\Phi_P(\mathrm{L}^s), \Theta_Q(\mathrm{L}^s)$	Операторы лагирования временного ряда в модели SARIMA, связанные с наличием сезонной компоненты
∇^d	Оператор лагирования временного ряда в модели ARIMA
Abs	Функция ошибки прогнозирования. При ретроспективном прогнозе вычисляется как абсолютное (по модулю) отклонение прогноза от реальных значений прогнозируемого временного ряда

ARMA	AutoRegressive Moving Average, модель авторегрессии-скользящего среднего
ARMAX	AutoRegressive Moving Average model including eXogenous covariates, модель авторегрессии-скользящего среднего, учитывающая экзогенные факторы
ARIMA	AutoRegressive Integrated Moving Average, интегрированная модель ARMA

SARIMA	Seasonal AutoRegressive Integrated Moving Average, модель авторегресии-скользящего среднего с учетом сезонной составляющей
hist	Разработанный в ВЦ РАН метод гистограммного прогнозирования (обеспечивает оптимальность свертки построенной гистограммы и функции потерь).
MAPE	Mean Average Percentage Error, функция ошибки прогнозирования. При ретроспективном прогнозе вычисляется как среднее относительное отклонение прогноза от реальных значений прогнозируемого временного ряда
MSE	Mean Squared Error, функция ошибки прогнозирования. При ретроспективном прогнозе вычисляется как среднее квадратичное отклонение прогноза от реальных значений прогнозируемого временного ряда
SSE	Sum of Squared Error, функция ошибки прогнозирования. При ретроспективном прогнозе вычисляется как сумма квадратичных отклонений прогноза от реальных значений прогнозируемого временного ряда
ГЖДП	Грузовые железнодорожные перевозки
РЖД	ОАО «Российские железные дороги»

ВВЕДЕНИЕ

Современное состояние научно-технической проблемы. В рамках исследования были рассмотрены существующие и созданы новые методы краткосрочного прогнозирования временных рядов с целью создания модели модели прогнозирования объемов спроса на грузовые железнодорожные перевозки. Разделение прогнозов на краткосрочные, среднесрочные и долгосрочные в различных иточниках носит условный характер и зависит от смыла отсчета времени (день, неделя, год) особенностей анализируемого ряда и целей прогноза [1]. В этом ПНИ под краткосрочным прогнозированием подразумевается выполнение прогноза на период времени от суток до года. Краткосрочное прогнозирование связано в основном с оперативным и текущим планированием производства [2], 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 более лет), среднесрочные (на 3–5 лет), текущие (на 1 год) и оперативные (квартальные и месячные) [3]. 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 [4–6] при создании транспортно-экономических моделей с целью выполнения прогнозов глубиной более года. Например, авторы работы [4] используют поиск конкурентно-транспортного равновесия для создания новых моделей транспортного планирования, включающих модели роста транспортной инфраструктуры городов. В работе [5] предложена модификация модели конкурентного равновесия для анализа проблем формирования тарифной и инвестиционной поли-

тики управления железнодорожными грузоперевозками. В работе [6] 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 продукии с размерами их в рассматриваемых территориальных единицах.

Актуальность и новизна темы. Грузовые перевозки обеспечивают свыше 80% общей выручки железнодорожного транспорта [3]. 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 [7] 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 [3].

Задача прогнозирования спроса на грузовые перевозки была поставлена для оперативного планирования перевозок. Рассмотрена задача прогнозирования нестационарных временных рядов в случае несимметричных функций потерь, учитывающих экспертные оценки потерь при недорогнозе и перепрогнозе. Один из широко используемых методов прогнозирования нестационарных временных рядов, авторегрессионное интегрированное скользящее среднее ARIMA [8], позволяет с хорошим качеством прогнозировать временные ряды с трендом, а также при небольшой 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 [9], 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 [9, 10] сделан вывод о том, что модель ARIMA не подходит для решения задачи прогнозирования в случае несимметричной функции потерь.

В работах [11, 12] были предложены модификации модели ARIMA, 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 ARIMA, 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 [13]. 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 этом ПНИ предложен двухэтапный алгоритм прогнозирования ARIMA + hist, на первом этапе которого отслеживаются свойства временного ряда, обусловливающие его нестационарность, такие как тренд и сезонность. На втором этапе вычисляется поправка, обеспечивающая оптимальность прогноза в случае несимметричной функции потерь. На втором этапе алгоритма ARIMA + hist 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 [14]. 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 наименьшее значение свертки. Предлагаемый алгоритм ARIMA + hist 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.

Исходные данные и научно-технические заделы. Прогнозы разрабатываются для номенклатуры грузов, учитывающей до 41 наименования грузов. Основным источников данных являются учетные системы индустриального партнера. Данные содержат информацию об отправках 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.

При разработке непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки был использован имеющийся у авторского коллектива научный задел. В частности, в основу модели лег алгоритм hist непараметрического прогнозирования загруженности железнодорожных узлов РЖД, основанный на свертке эмпирической плотности распределения значений временного ряда с функцией потерь, разработанный ранее специалистами ВЦ РАН, МФТИ и Российской открытой академии транспорта. Алгоритм hist является обобщением алгоритма квантильной регрессии [13] и находит приближенное решение задачи минимизации математического ожидания потерь.

При создании моделей, методов и алгоритмов прогнозирования объёмов спроса на ГЖДП в этом ПНИ будут учитывается как предыстория самих грузоперевозок в РЖД, так и предыстория экзогенных факторов, характер влияния которых на объемы спроса на ГЖДП были исследованы в рамках первого этапа этого ПНИ (раздел 3.1, Проведение экспертного анализа значимости и характера влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки). При разработке алгоритма выполнения алгебраических операций над гистограммми для учета экзогенных временных рядов в модели hist были использованы результаты

экспертного анализа значимости и характера влияния экзогенных факторов на объемы спроса на ГЖДП.

Основанием для проведения ПНИ в рамках мероприятия 1.2 приоритетного направления «Транспортные и космические системы» федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2017 годы», является Соглашение о предоставлении субсидии от 19 июня 2014 г. № 14.604.21.0041.

Сведения о планируемом научно-техническом уровне разработок. В рамках этого ПНИ в интересах Индустриального партнёра – РЖД – будут разработаны новые модели, методы и алгоритмы прогнозирования объёмов спроса на ГЖДП, нацеленные на повышения точности прогнозирования с учетом специфичных для РЖД условий выполнения железнодорожных грузоперевозок.

Разработка указанных моделей, методов и алгоритмов носит инновационный характер. Ожидаемые результаты ни в чём не уступают уровню современных зарубежных и отечественных исследований в этой области.

Сведения о выполненных патентных исследованиях и выводы из них. Объектом выполненных патентных исследований являлись, методы и системы прогнозирования объемов спроса на грузовые железнодорожные перевозки.

Среди выявленных в результате информационно-патентного поиска охранных документов нет патентов и заявок на изобретения, которые могут препятствовать применению результатов выполняемого ПНИ в Российской Федерации, а также препятствовать получению охранных документов в других странах посредством подачи PCT-заявок на изобретения и/или полезные модели.

Связь с другими научно-исследовательскими работами и разработками. Выполняемые в рамках этого проекта исследования и разработки связаны с пятью проектами Российского фонда фундаментальных исследований: «Разработка макета системы прогноза грузоперевозок на основе интеграции опыта специалистов ВЦ РАН и ПГК» 11-07-13154-офи-м-2011-РЖД, «Методы анализа взаимного влияния пассажирского и грузового трафиков РЖД» 13-07-13139-офи-м-РЖД, «Развитие теории индуктивного порождения и выбора моделей» 10-07-00422-а, «Высокоэффективные технологии имитационного моделирования взаимодействия подвижного состава и инфраструктуры железнодорожного транспорта» 12-07-13135-офи-м-РЖД,

«Методы анализа и прогнозирования нестационарных временных рядов в задачах мониторинга технического состояния подвижного состава» 12-07-13141-офи-м-РЖД.

Цели и задачи второго этапа, их место в выполнении проекта в целом. В соответствии с пп. 2.1.5–2.1.7 Технического задания на втором этапа выполняемого проекта достигнуты следующие цели:

- Разработана математическая модель прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающая влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера, в том числе:

а. разработана непараметрической модель прогнозирования объемов спроса на грузовые железнодорожные перевозки на железнодорожных узлах РЖД;

б. описаны свойства непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки ;

в. разработан и протестирован алгоритм построения гистограммы распределения значений объема спроса и вычисления свертки гистограммы с экспертно заданной функцией потерь для каждого возможного прогнозируемого значения временного ряда объемов спроса на грузовые железнодорожные перевозки;

г. разработан метод определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки;

д. разработан и протестирован алгоритм выполнения алгебраических операций с гистограммами распределения значений объемов спроса на грузовые железнодорожные перевозки;

е. разработан метод определения выполнения условия локальной стационарности временного ряда и реализован тест Дики-Фуллера.

- Разработан макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки и выполнена серия вычислительных экспериментов по прогнозированию объёмов спроса на грузовые железнодорожные перевозки на модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов.

- Разработан генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов и с использованием

макета модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки выполнена серия вычислительных экспериментов по прогнозированию объёмов спроса на грузовые железнодорожные перевозки для сравнения значений спрогнозированных объемов спроса на грузовые железнодорожные перевозки со значениями контрольной выборки данных об объемах спроса на грузовые железнодорожные перевозки и сравнения ошибки прогнозирования на основании предложенной модели с ошибкой прогнозирования модели ARMA на контрольной выборке данных.

В контексте выполняемого проекта в целом на втором этапе были поставлены следующие задачи:

- На основе проведенных исследований разработать математическую модель прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера.
- Провести тестирование разработанной модели, включающее тестирование всех алгоритмов, лежащих в ее основе, и описать свойства модели на основе серии вычислительных экспериментов прогнозированию объёмов спроса на грузовые железнодорожные перевозки.
- Разработать и протестировать генератор модельных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов.
- Разработать и протестировать макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки.

1 Исследования по разработке математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера

В этом разделе представлены результаты исследования по разработке математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, а также специфику бизнес-процессов и нормативов индустриального партнера, выполненного в соответствии с пп. 2.1.5 и 3.6 Технического задания.

В настоящем исследовании поставлена задача прогнозирования спроса на ГЖДП по историческим данным с временным интервалом от суток до года. Предложен алгоритм непараметрического прогнозирования, основанный на минимизации ожидаемых потерь, аппроксимируемых сверткой плотности распределения исторических значений временного ряда с функцией ошибки. Различные способы восстановления плотностей распределений описаны в [15]. В данном исследовании для этой цели использовался гистограммный подход. Впервые такой подход к аппроксимации распределений был предложен в [16]. Впоследствии эта тема развивалась в [17–21]. Были предложены, помимо одномерных гистограмм с равными отрезками разбиения, одномерные и многомерные гистограммы с одинаковой глубиной [17, 18], а также v-оптимальные гистограммы с минимальной дисперсией внутри каждого отрезка разбиения [20, 22]. В работах [22–26] освещается проблема размещения узлов гистограммы, работы [22–24] посвящены аппроксимации значений внутри отрезков разбиения. Также различные алгоритмы построения гистограмм можно найти в [23, 27–30]. В этом исследовании восстановление плотности распределения двумерной случайной величины проводилось путем построения двумерной гистограммы с интервалами одинаковой ширины с оптимизацией количества интервалов разбиения.

Для учета экзогенных временных рядов в данном ПНИ предложен метод уточнения гистограммы прогнозируемого временного ряда. Способ учета экзогенных факторов зависит от структуры модели. В наиболее простом случае, когда модель линейна, учет внешнего фактора заключается в аддитивном добавлении нес-

кольких значений (или их преобразований) экзогенного временного ряда в модель. Примером могут служить модель ARMA [31, 32], широко используемая при краткосрочном прогнозировании временных рядов [33], и ее экзогенная модификация ARMAX [34]. Модель ARMA содержит три аддитивных компоненты: авторегрессионную, скользящее среднее и ошибку. Модель ARMAX включает также комбинацию экзогенных временных рядов в качестве дополнительного аддитивного компонента.

Альтернативный способ повышения качества прогнозирования с учетом дополнительных временных рядов, предложен в работе [35]. Метод разработан для прогнозирования групп временных рядов. В частности, рассматривается задача прогнозирования спроса товары с учетом классификации товарных групп. При прогнозировании эндогенного временного ряда временные ряды спроса на товары из той же группы рассматриваются как экзогенные. В работе [36] продемонстрировано повышение качества учета сезонности при использовании информации о группе временных рядов, выделенных с помощью кластерного анализа [37]. Аналогично, если прогнозируемые временные ряды обладают иерархической структурой в качестве экзогенных могут быть использованы временные ряды с одного уровня иерархии. Под иерархией понимается наличие уравнения, задающего связь между рядами [38], в связи с чем ряды нельзя прогнозировать независимо. В работе [38] рассмотрен следующий пример иерархии: сумма прогнозов временных рядов должна совпадать с прогнозом их суммы. В работах [38, 39] предложены методы согласования независимых прогнозов с учетом иерархии, гарантированно неухудшающие качество прогноза.

Предложенная в данном исследовании модель имеет нелинейную структуру. Прогноз алгоритма hist основан на построении гистограммы эндогенного ряда, в связи с чем в данном ПНИ поставлена задача уточнения гистограммы эндогенного ряда с учетом экзогенных временных рядов. Для уточнения гистограммы используется взвешенная сумма гистограмм, условных по экзогенным временным рядам. Способ моделирования гистограммы с помощью смеси условных гистограмм основан на развитии подходов к моделированию распределения смесью компонент [40]. Смеси моделей используются для оценки распределений, не принадлежащих к какому-либо из основных параметрических семейств распределений. Плотность

распределения исследуемой величины при этом приближается взвешенной суммой плотностей, как правило, гауссовских: согласно [41], любую функцию плотности можно приблизить смесью гауссиан с произвольной точностью. Существуют различные модификации смесей моделей: позволяющие задавать матожидания каждого из компонентов [42], моделировать зависимость дисперсии компонента от ее истории [43], и моделировать веса компонентов с помощью марковских цепей [44]. Смеси гистограмм используются как более устойчивая альтернатива гауссовским смесям [40, 45] при распознавании объектов на изображениях и видео [46–48]. В перечисленных работах предполагается, что интервалы разбиения гистограмм совпадают. В работах [49, 50] вводятся арифметики над гистограммами, позволяющие, в частности, рассматривать сумму гистограмм с произвольным разбиением на интервалы. В работах [51–54] описаны методы выполнения алгебраических операций над гистограммами на основе численных методов нахождения сверток. Существует также и другой подход, основанный на восстановлении по гистограмме функции принадлежности нечеткого множества [55–57] и переходу к операциям над нечеткими множествами. Предложенный в данном исследовании метод уточнения гистограмм построен таким образом, что алгебраические операции применяется к гистограммам с одинаковым разбиением на интервалы.

1.1 Исследования по разработке непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки на железнодорожных узлах РЖД

В этом подразделе описаны результаты исследования по разработке непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки на железнодорожных узлах РЖД, выполненного в соответствии пп. 2.1.5.1, 3.6.1 Технического задания. Описанная модель основана на непараметрической модели hist [58].

1.1.1 Задача прогнозирования стационарных временных рядов с известной плотностью распределения

В этом пункте рассмотрена задача нахождения прогноза $\hat{x}$ следующего значения x_{T+1} временного ряда $\mathbf{x} = \{x_1, \dots, x_T\}$, $x_t \in \mathbb{R}_+$, минимизирующего математическое ожидание $\mathbf{E}\, l(c, x_{T+1})$ ожидаемых потерь:

$$\hat{x} = \arg\min_{c \in \mathbb{R}_+} \mathbf{E}\, l(c, x_{T+1}), \tag{1}$$

где $l(\hat{x}, x_{T+1})$ – заданная экспертами 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 случайного процесса, стационарного в узком смысле. Таким образом, для любых $t, \tau \in [1, \dots, T]$ выполненяется

$$\mathrm{P}(x_t < \xi) = \mathrm{P}(x_\tau < \xi) \quad \forall\, \xi \in \mathbb{R}_+. \tag{2}$$

и прогнозируемое значение x_{T+1} значение временного ряда генерируется из того же распределения, что и все наблюдаемые значения x_t. Случай нестационарных временных рядов рассмотрен в подразделе 1.6 этого отчета;

- x_t – непрерывная случайная величина с плотностью распределения $p(u)$:

$$\exists \lim_{dt \to 0} \mathrm{P}(u < x_t \le u + du) = p(u). \tag{3}$$

Основания для этоого предположения приведены в пункте 1.2.2;

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

$$\exists \tilde{u}_{\min} = u_0, \dots, \tilde{u}_C = u_{\max},\ C \in \mathbb{N}_0 \colon\ p(u) \in C(\tilde{u}_{i-1}, \tilde{u}_i), i = 1, \dots, C. \tag{4}$$

Это предположение введено для обеспечения корректности математических выкладок. Основанием для него 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 прогнозироуемое значения временного ряда x_{T+1}, математическое ожидание потерь выражается аналитически в виде

$$\mathbf{E}\, l(c, x_{T+1}) = \int_{u_{min}}^{u_{max}} l(c,u)\, p(u)\, du. \tag{5}$$

Интеграл (5) существует в силу предположения (4).

Задача прогнозирования формулируется так:

$$\hat{x} = \arg\min_{c \in \mathbb{R}_+} \int_{u_{min}}^{u_{max}} l(c,u)\, p(u)\, du = \arg\min_{c \in \mathbb{R}_+} L(c), \tag{6}$$

где $L(c) = \mathbf{E}\, l(c,u)$. В случае, когда функция потерь достаточна проста, прогноз можно найти аналитически. В частности, получены значения оптимального прогноза (см. (8) и (9)) для квадратичной и абсолютной функций $l(\hat{x}, x)$, часто используемых для оценки потерь при прогнозировании. Для этих функций можно производная ожидаемых потерь $L(c)$ имеет простой вид, анализируя который можно получить выражение для оптимального прогноза. Для корректности операции дифференцирования интеграла (5) по параметру требуется [59] выполнение следующих условий:

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 (4), функция $p(u)$ имеет конечное число $C+1$ точек разрыва, следовательно интеграл (5) может быть представлен в виде суммы C интегралов

$$L(c) = \sum_{i=1}^{C} \int_{\tilde{u}_{i-1}}^{\tilde{u}_i} l(c,u)\, p(u)\, du, \tag{7}$$

для каждого из которых условия 1) и 2) выполняются. Ввиду конечности суммы (7), 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:

$$\frac{dL}{dc} = \int_{u_{min}}^{u_{max}} 2(c-u)p(u)du = 2c \int_{u_{min}}^{u_{max}} p(u)du - 2 \int_{u_{min}}^{u_{max}} u\,p(u)du = 2c - 2E\,x.$$

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

$$\frac{\mathrm{d}L}{\mathrm{d}c} = 2c - 2\mathbf{E}\,x \Rightarrow \hat{x} = \mathbf{E}\,x. \quad (8)$$

Для абсолютной функции потерь $l(\hat{x}, x_{T+1}) = |\hat{x} - x_{T+1}|$ производная функции L имеет вид

$$\frac{\mathrm{d}L}{\mathrm{d}c} = -\int_{u_{min}}^{c} p(u)du + \int_{c}^{u_{max}} p(u)du,$$

то есть, оптимальный прогноз совпадает с медианой распределения $p(u)$:

$$\hat{x} = \mathrm{med}\ p(u). \quad (9)$$

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.

1.1.2 Алгоритм прогнозирования hist

Алгоритм прогнозирования hist заключается в оценке распределения значений временного ряда x, то есть в нахождении функции $\hat{p}(u)$, и последующем поиске приближенного решения задачи минимизации (6) переборным алгоритмом.

Вход алгоритма: стационарный временной ряд $\mathbf{x} = \{x_1, \dots, x_T\}$ и функция потерь $l(\hat{x}, x_{T+1})$.

Выход: прогноз $\hat{x}$, минимизирующий математическое ожидание потерь.

Порядок вычислений:

Шаг 1: задание количества *n* столбцов гистограммы.

Шаг 2: вычисление ширины столбцов гистограммы $b = \frac{\max(\mathbf{x}) - \min(\mathbf{x})}{n}$ и координат концов отрезков постоянства $u_0, u_1, \dots, u_n$ для функции $\hat{p}(u)$.

Шаг 3: построение гистограммы; нахождение функции $\hat{p}(u)$; нормирование гистограммы; вычисление значений функции на отрезках постоянства $y_1, \dots, y_n$.

Шаг 4: вычисление значений свертки $\sum_{i=1}^{n} h_i l\left(c, \frac{u_i + u_{i-1}}{2}\right)$ 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 $\hat{x}$:

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

Функция потерь в каждом конкретном случае выбирается с учетом особенностей прикладной задачи и стоимости ошибки прогноза в ту или иную сторону. Функцию потерь могут задавать эксперты.

При заданных выборке данных и функции потерь результат прогнозирования зависит только от количества 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.

1.1.3 Модификация алгоритма с использованием ядерных оценок плотности

В этом пункте рассмотрен вариант прогностического алгоритма с оцениванием плотности распределения с помощью локальной непараметрической оценки Парзена-Розенблата [60] по окну ширины h

$$\hat{p}_h(u) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{u - x_i}{h}\right),$$

где ядро $K(r)$ – функция, удовлетворяющая следующим требованиям:

- четность;
- нормированность: $\int_{\mathbb{R}} K(r)\,dr = 1$;
- как правило, $K(r)$ – невозрастающая на положительной полуоси и неотрицательная функция.

В этом ПНИ рассмотрены следующие виды ядерных функций:

- гауссово ядро $K(r) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{r^2}{2}\right)$,
- треугольное ядро $K(r) = (1 - |r|)I_{|r|<1}$,
- прямоуольное ядро $K(r) = \frac{1}{2}I_{|r|<1}$,
- ядро Епанечникова $K(r) = \frac{3}{4}(1 - r^2)I_{|r|<1}$.

Ядерную оценку плотности $\hat{p}_h(u)$ можно вычислить в любой наперед заданной точке u, но аналитическая запись этой функции недоступна. Поэтому интеграл в задаче (6) берется численно. Для вычисления интеграла задаётся разбиение

$$u_0 = u_{min} < u_1 < \cdots < u_{n-1} < u_n = u_{max}$$

отрезка интегрирования $[u_{min}; u_{max}]$ на отрезки равной длины b. Математическое ожидание потерь $L(c) = \mathbf{E}l(c, u)$ приближается величиной $L_{kernel}(c)$

$$L(c) \approx L_{kernel}(c) = \int_{u_{min}}^{u_{max}} l(c,u)\, \hat{p}_h(u)\, du,$$

$$L_{kernel}(c) \approx \sum_{i=1}^{n} l\left(c, \frac{u_i + u_{i-1}}{2}\right) \hat{p}_h\left(\frac{u_i + u_{i-1}}{2}\right) (u_i - u_{i-1}) =$$

$$= b \sum_{i=1}^{n} l\left(c, \frac{u_i + u_{i-1}}{2}\right) \hat{p}_h\left(\frac{u_i + u_{i-1}}{2}\right).$$

Полученная сумма является дискретной сверткой функции потерь l с оценкой плотности распределения $\hat{p}_h$ по точкам $\frac{u_1+u_0}{2}, \frac{u_2+u_1}{2}, \dots, \frac{u_n+u_{n-1}}{2}$, которые являются серединами отрезков разбиения отрезка интегрирования.

Таким образом, исходная задача (6) с использованием ядерной оценки плотности сведится к задаче:

$$\hat{x} = \underset{c}{\operatorname{argmin}}\, L_{kernel}(c) \approx \underset{c}{arg\ min} \sum_{i=1}^{n} l\left(c, \frac{u_i+u_{i-1}}{2}\right) \hat{p}_h\left(\frac{u_i+u_{i-1}}{2}\right). \qquad (10)$$

Чем мельче разбиение $u_0 < u_1 < \cdots < u_n$ отрезка интегрирования, тем более точно [59] риманова сумма приближает значение интеграл. Аналогично гистограммному прогнозу, точность оптимального по (10) прогноза повышается с увеличением точек в свертке.

1.2 Исследования свойств непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки

В этом подразделе описаны результаты исследования свойств непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, выполненного в соответствии с пп. 2.1.5.2, 3.6.2 Технического задания.

1.2.1 Зависимость от параметров ядерной оценки плотности

Алгоритм прогнозирования $\mathtt{hist}_K$ с использованием ядерных оценок плотности распределения является обобщением исходного алгоритма с использованием гистограммы как оценки плотности распределения значений временного ряда. В

случае использования ядерных оценок плотности алгоритм имеет большее количество параметров – ядро, ширина окна и количество точек свертки.

Ядерные оценки плотности распределения чувствительны к ширине окна h. При больших h оценка плотности получается сглаженной, при малых h – более детальной, склонной к описанию свойств конкретной выборки. Так, в одном из предельных случаев $h \to \infty$, то есть h достаточно велико, чтобы в окно попадали с большим запасом все точки выборки. В этом случае оценка плотности $\hat{p}_h(u)$ получается максимально сглаженной, практически константной:

$$\hat{p}_h(u) \approx const(u),$$

и оптимальный прогноз не зависит от оценки плотности, а зависит только от размера шага разбиения отрезка интегрирования:

$$\hat{x} \approx \operatorname*{argmin}_c \sum_{i=1}^{n} l\left(c, \frac{u_i + u_{i-1}}{2}\right).$$

С другой стороны, при достаточно малых h, чтобы в окно попадало не более одной точки выборки, оценка плотности распределения в точке либо равна нулю, либо некоторой величине α

$$\hat{p}_h(u) = \begin{cases} \alpha, & \exists i: |u - x_i| \le h; \\ 0, & \text{иначе.} \end{cases}$$

При этом прогноз очень чувствителен к конкретной выборке. Такой алгоритм с высокой точностью прогнозирует обучающую выборку, но демонстрирует неадекватные результаты на контрольной выборке.

1.2.2 Удовлетворение прогноза физическим ограничениям

Физические ограничения $0 \le \hat{x} \le X_{\max}$, накладываемые на прогноз $\hat{x}$, связаны со спецификой прогнозируемых временных рядов. Прогноз должен быть неотрицателен и не должен принимать слишком больших значений, превышающих пропускную способность железнодорожной сети. Эти ограничения выполняются

автоматически, так значение $\hat{x}$ согласно (13) выбирается из фиксированного набора значений, каждое из которых не меньше минимального исторического значения ряда $\mathbf{x}$ и не превышает его максимального исторического значения.

Кроме того прогноз должен быть интерпретируем в рамках рассматриваемой предметной области. Постановка задачи (1) записана в предположении (3), что значения x_t прогнозируемого временного ряда $\mathbf{x}$ есть реализации непрерывного случайного процесса, принимающего неотрицательные действительные значения, $x_t \in \mathbb{R}_+$. В соответствии с этим предположением прогноз $\hat{x}$, полученный как решение задачи (10), также принимает значения из $\mathbb{R}_+$. Это допущение принято в связи нацеленностью на прогноз объема грузоперевозок, выраженного весом перевозимых грузов в тоннах. Предполагается, что такой прогноз наиболее информативен, так как прибыль индустриального партнера в большей мере определяется весом перевезенных грузов, чем количеством вагонов. Таким образом, получаемые рациональные значения прогнозов не противоречат физическим ограничениям. При необходимости спрогнозировать значение дискретного случайного процесса, такое как количество вагонов, проходящих между пунктами отправления и назначения, результат $\hat{x}$ решения задачи оптимизации (10) необходимо привести к соответствующему виду путем округления.

1.3 Разработка и тестирование алгоритма построения гистограммы распределения значений объема спроса и вычисления свертки гистограммы с экспертно заданной функцией потерь для каждого возможного прогнозируемого значения временного ряда объемов спроса на грузовые железнодорожные перевозки

В этом подразделе описаны результаты исследования и тестирования алгоритма построения гистограммы распределения значений объема спроса и вычисления свертки гистограммы с экспертно заданной функцией потерь, выполненных в соответствии с пп. 2.1.5.3, 3.6.3 Технического задания.

1.3.1 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}):

$$u_0 = u_{min} \le u_1 \le \cdots \le u_{n-1} \le u_n = u_{max}:$$

$$\hat{p}(u) = h_i, \quad u \in (u_{i-1}, u_i), \ i = 1, \dots, n$$

одинаковой длины b:

$$b = u_1 - u_0 = \cdots = u_n - u_{n-1} = \frac{u_{max} - u_{min}}{n}.$$

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

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

При использовании оценки плотности $\hat{p}(u)$ математическое ожидание потерь $\mathbf{E}l(c,u)$ приближается величиной $L_{hist}(c)$

$$L \approx L_{hist}(c) = \int_{u_{min}}^{u_{max}} l(c,u)\hat{p}(u)du = y_1 \int_{u_0}^{u_1} l(c,u)du + \cdots + y_n \int_{u_{n-1}}^{u_n} l(c,u)du = \\ = \sum_{i=1}^{n} h_i \int_{u_{i-1}}^{u_i} l(c,u)\, du. \tag{11}$$

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 (10) принимает вид

$$\hat{x} = \underset{c}{\operatorname{argmin}}\, L_{hist}(c) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} h_i \int_{u_{i-1}}^{u_i} l(c,u)\, du. \qquad (12)$$

Так как интегрирование функции потерь $l(\hat{x}, x_{T+1})$ является трудоемкой операцией, интеграл в постановке задачи прогнозирования (12) оценивается по формуле

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

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. Таким образом, задача поиска оптимального прогноза принимает вид:

$$\hat{x} = \underset{c}{\operatorname{argmin}} L_{conv}(c) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} h_i l\left(c, \frac{u_i + u_{i-1}}{2}\right). \qquad (13)$$

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 (13). 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 (10) и (13) прогнозы сравнивались с прогнозами, оптимальными согласно (6) и (12), для квадратичной и абсолютной функций потерь, так как для этих функций существует аналитическое решение задачи (6). 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 (6) и (12), (13) прогнозами являются выборочное среднее

$$\hat{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

и

$$\hat{x} = b\sum_{i=1}^{n} h_i \frac{u_i + u_{i-1}}{2}.$$

Для абсолютной функции потерь оптимальным согласно (12) 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 (13) прогноз является оценкой медианы распределения, вычисленной по оценке распределения $\hat{p}(u)$.

1.3.2 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, практически 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 рисунках 1.1–1.6 представлены графики отклонения прогнозов от оптимального для квадратичной (рисунки 1.1–1.3) и абсолютной (рисунки 1.4–1.6) функций потерь от ширины окна h при различной длине выборки $T = 100,\ 1000,\ 10000$. По 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 (12) прогноза.

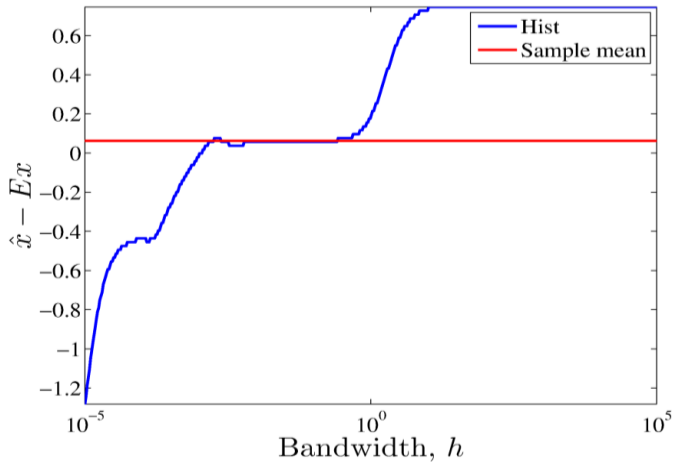

Рисунок 1.1 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 100$ и количестве точек свертки $n = 300$ для квадратичной функции потерь.

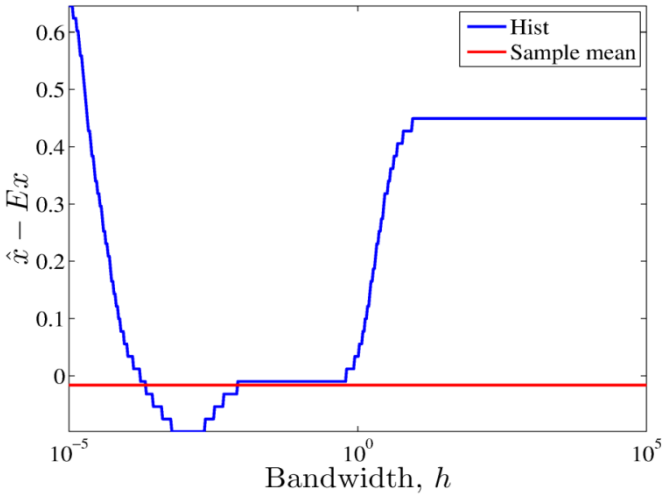

Рисунок 1.2 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 1000$ и количестве точек свертки $n = 300$ для квадратичной функции потерь.

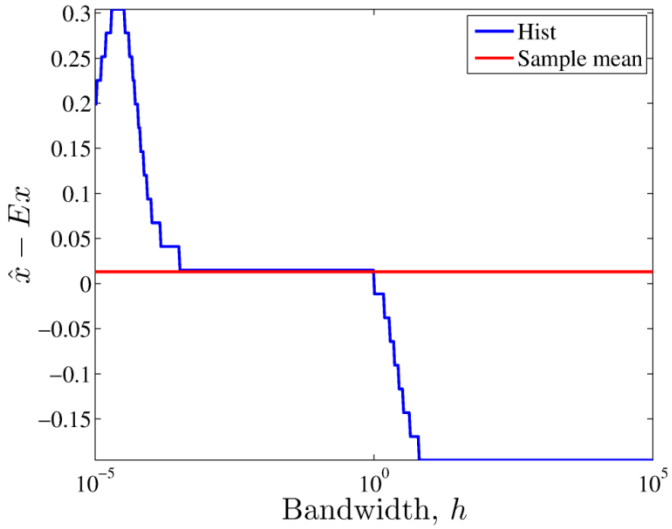

Рисунок 1.3 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 10000$ и количестве точек свертки $n = 300$ для квадратичной функции потерь.

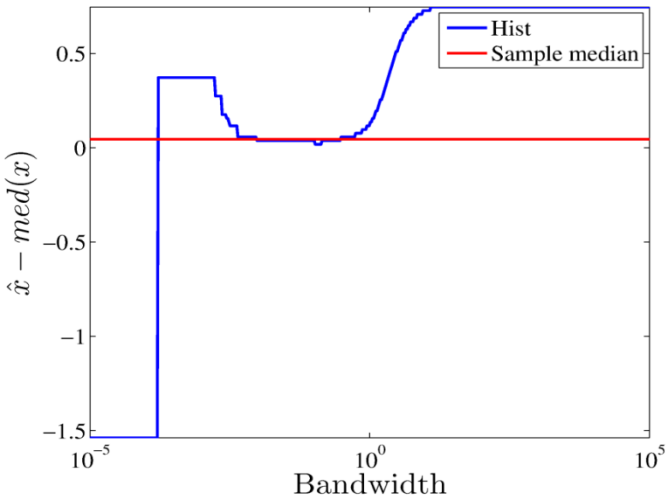

Рисунок 1.4 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 100$ и количестве точек свертки $n = 300$ для абсолютной функции потерь.

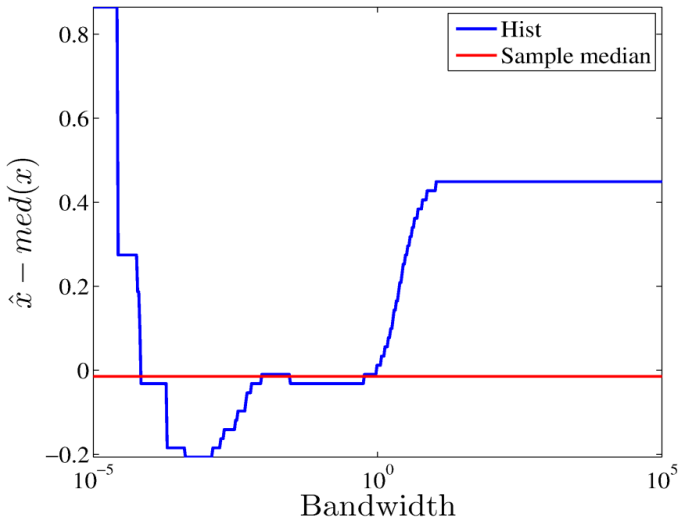

Рисунок 1.5 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 1000$ и количестве точек свертки $n = 300$ для абсолютной функции потерь.

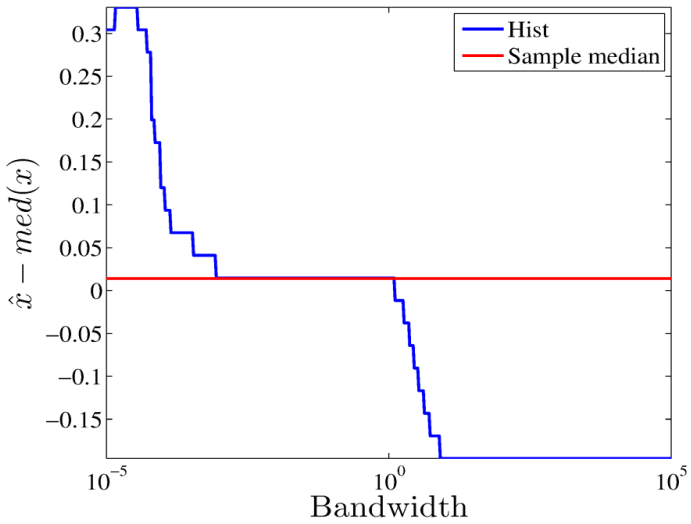

Рисунок 1.6 – Зависимость качества прогнозирования от ширины окна при длине выборки $T = 10000$ и количестве точек свертки $n = 300$ для абсолютной функции потерь.

1.3.3 Зависимость прогноза от числа точек свертки n

При увеличении длины выборки T наблюдается сходимость отклонения прогнозов от оптимального прогноза. При этом при любом значении длины выборки T прогнозы колеблются вокруг некоторого предельного значения. Амплитуда этих ко-

лебаний уменьшается с ростом количества точек свертки. При увеличении длины выборки скорость уменьшения амплитуды колебаний (т.е. скорость сходимости прогнозов алгоритма hist_K) не меняется. Скорость сходимости прогнозов не зависит от длины выборки. В этом разделе рассматривается, как ведут себя прогнозы алгоритма hist при использовании выборок разной длины. Результаты экспериментов на выборках стандартного нормального распределения длины T = 100, 1000, 10000 для квадратичной и абсолютной функций потерь изображены на рисунках 1.7–1.12. Использовалось гауссово ядро с фиксированной шириной окна $h = 0{,}1$ и различным количеством точек свертки.

Точность прогнозов увеличивается с ростом длины выборки T – разница между предельным значением прогнозов, полученных алгоритмом hist и выборочными статистиками уменьшается. На длинных выборках прогнозы алгоритма hist_K сходятся к выборочному среднему при квадратичной функции потерь и к выборочной медиане при абсолютной функции потерь. Скорость приближения предельного значения прогнозов к выборочной медиане ниже, чем к выборочному среднему.

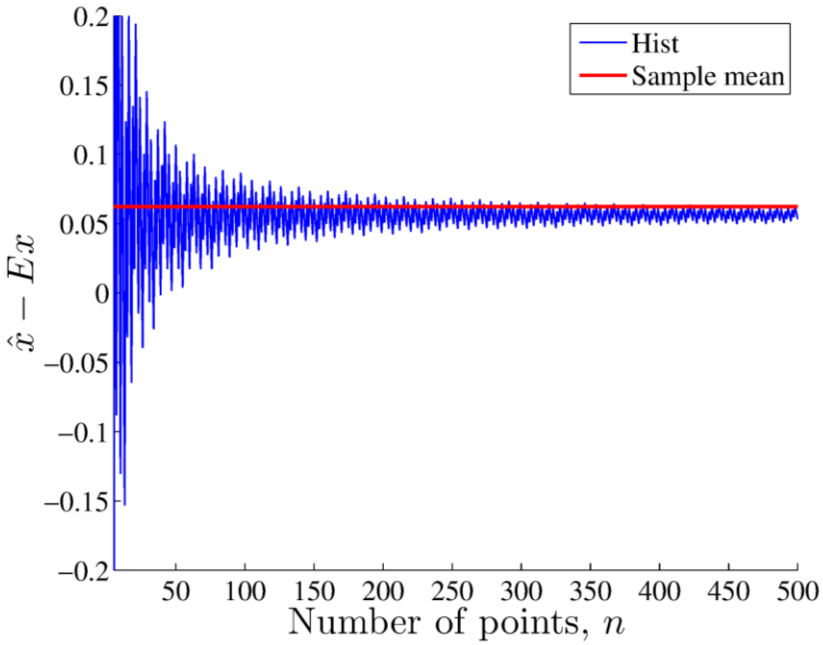

Рисунок 1.7 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 100$ и ширине окна $h = 0{,}1$ для квадратичной функции потерь.

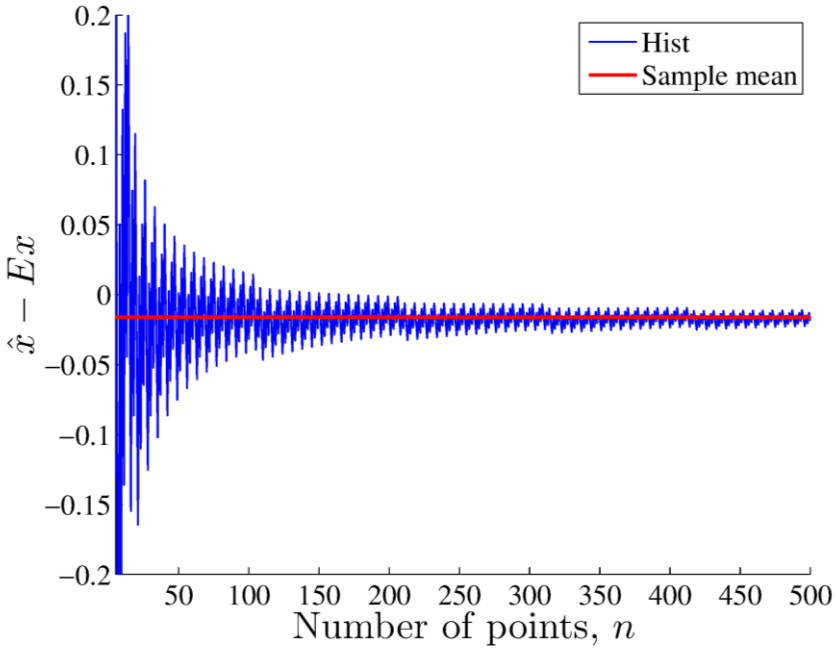

Рисунок 1.8 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 1000$ и ширине окна, $h = 0{,}1$ для квадратичной функции потерь.

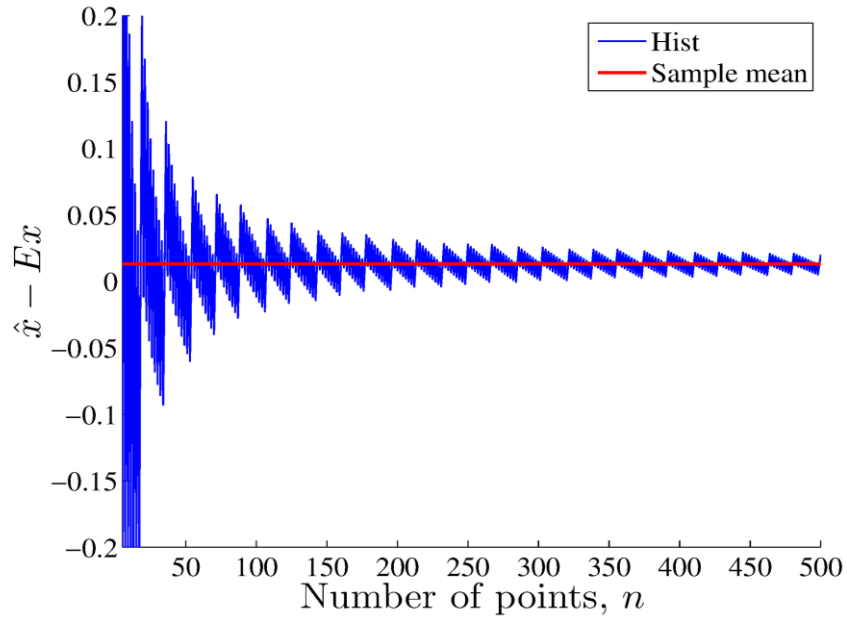

Рисунок 1.9 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 10000$ и ширине окна, $h = 0{,}1$ для квадратичной функции потерь.

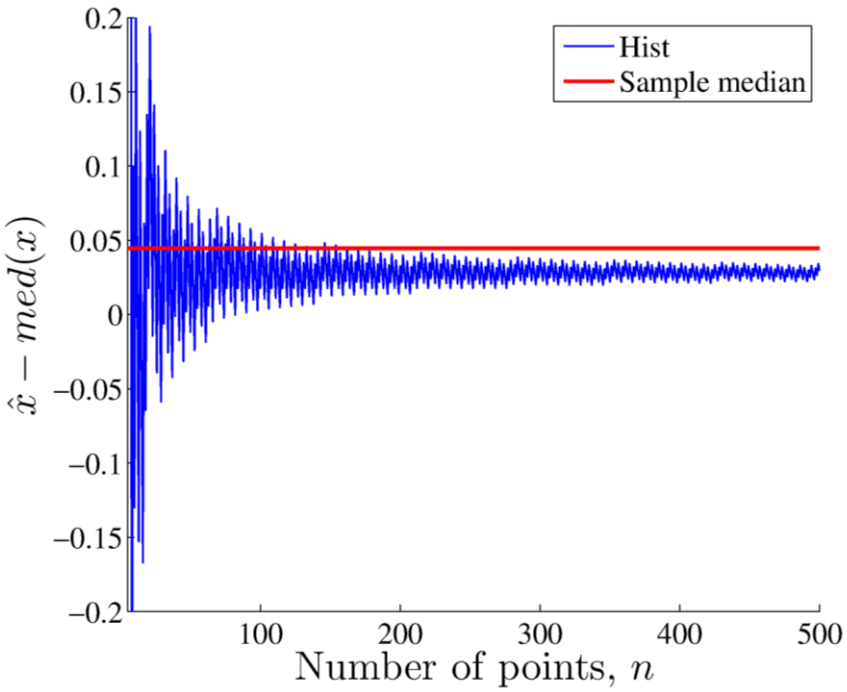

Рисунок 1.10 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 100$ и ширине окна $h = 0{,}1$ для абсолютной функции потерь

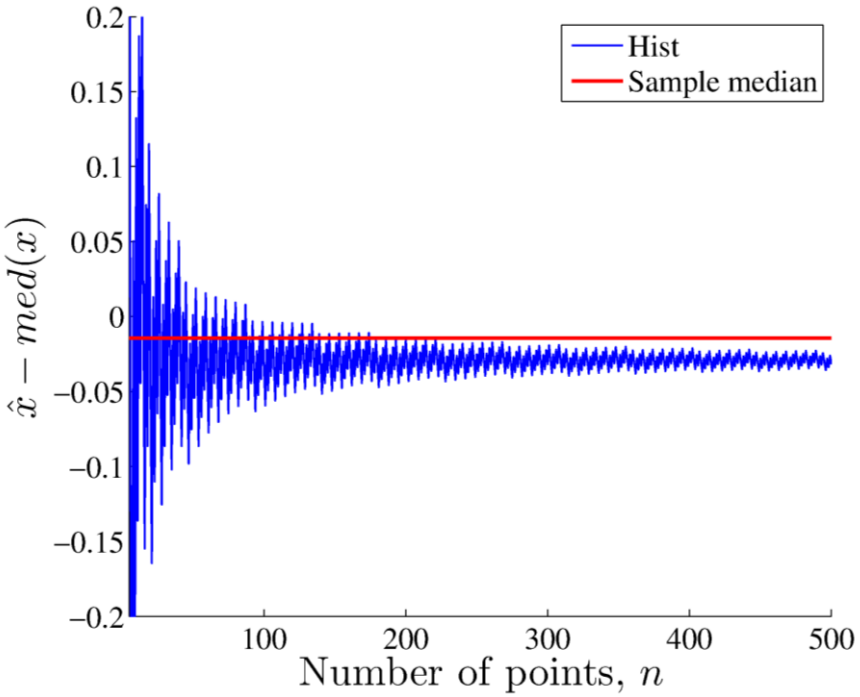

Рисунок 1.11 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 1000$ и ширине окна $h = 0{,}1$ для абсолютной функции потерь.

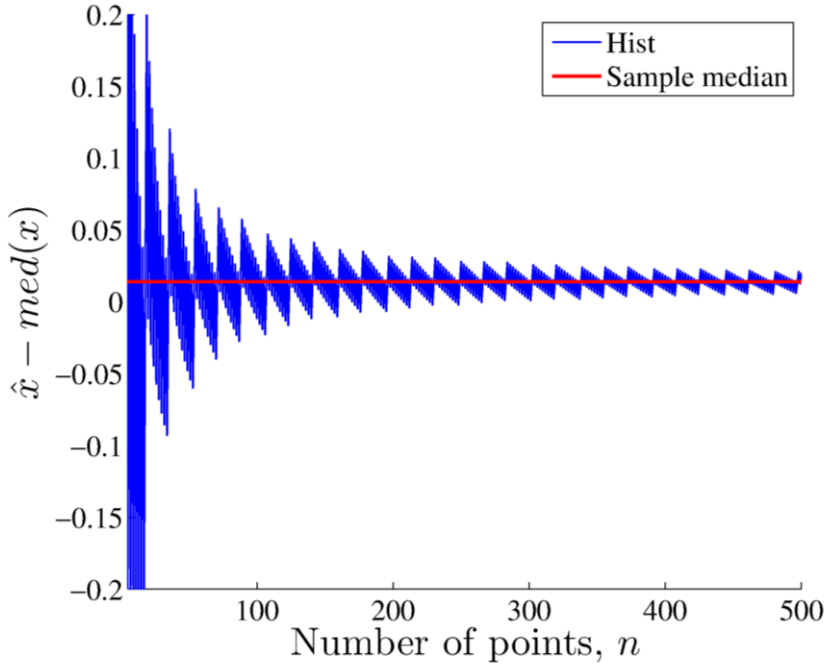

Рисунок 1.12 – Зависимость качества прогнозирования от количества точек свертки при длине выборки $T = 10000$ и ширине окна $h = 0{,}1$ для абсолютной функции потерь

1.3.4 Зависимость прогноза от формы ядра и ширины окна

Проведены исследования поведения прогнозов алгоритма hist при использовании различных ядерных функций (гауссовой, прямоугольной, треугольного ядер и ядра Епанечникова) и разной ширины окна h. Рисунки 1.13–1.20 иллюстрируют проведенные эксперименты для выборки длиной $T = 500$ точек. На рисунках 1.13–1.16 и 1.17–1.20 представлены графики отклонения прогнозов от оптимального для квадратичной (рисунки 1.13– .16) и абсолютной (рисунки 1.17–1.20) функций потерь при использовании различных ядер для стандартного нормального распределения. Для каждого ядра и функции потерь на одном графике представлены кривые поведения прогнозов для ширины окна h = 0,001, 0,01, 0,1, 1, 10 и соответствующая функции потерь выборочная статистика. На графиках по оси абсцисс отложено количество точек свертки n, по оси ординат – отклонение прогноза от оптимального значения. Черной пунктирной линией на рисунках 1.13–1.16 изображено выборочное среднее, на рисунках 1.17–1.20 – выборочная медиана.

На графиках видно, что поведение прогнозов алгоритма hist_K с использованием ядерных оценок плотности распределения слабо зависит от выбора вида ядра. Для всех ядер при ширине окна $h = 0{,}001$ отсутствует сходимость прогнозов при увеличении количества точек свертки. Медленная сходимость прогнозов наблюдается для $h = 0{,}01$ в случае прямоугольного ядра и квадратичной функции потерь, а также прямоугольного ядра и абсолютной функции потерь. Для более широких окон при использовании любых ядер сходимость прогнозов к предельному значению есть.

1.3.5 Определение оптимального количества столбцов n

С увеличением количества столбцов гистограммы увеличивается точность оценки (11) ожидания потерь $L(c)$ и прогноз алгоритма hist стремится к оптимальному по (6) ответу c^*, минимизирующему математическое ожидание потерь. Однако неограниченное увеличение числа n на практике не имеет смысла, так как выборки, по которым строятся гистограммы, имеют конечную длину. В связи с этим прогнозы,

полученные с помощью алгоритма hist, будут с увеличением количества n столбцов гистограммы сходиться не к истинному оптимальному ответу c^*, а к его оценке, сделанной по имеющейся выборке.

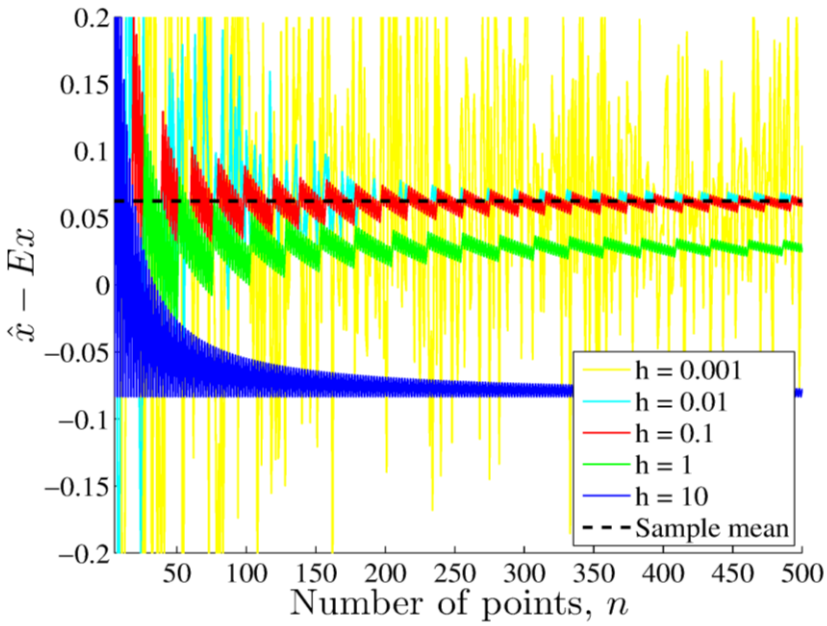

Рисунок 1.13 – Зависимость прогнозов от количества точек свертки при различных значениях ширины окна при выборе гауссова ядра для квадратичной функции потерь

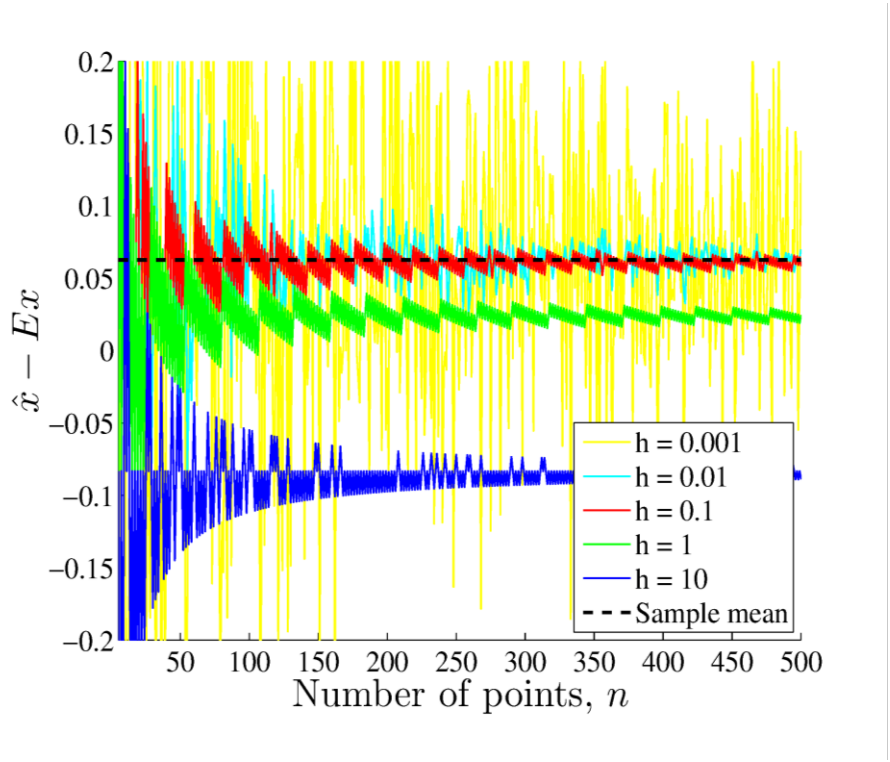

Рисунок 1.14 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе прямоугольного ядра для квадратичной функции потерь

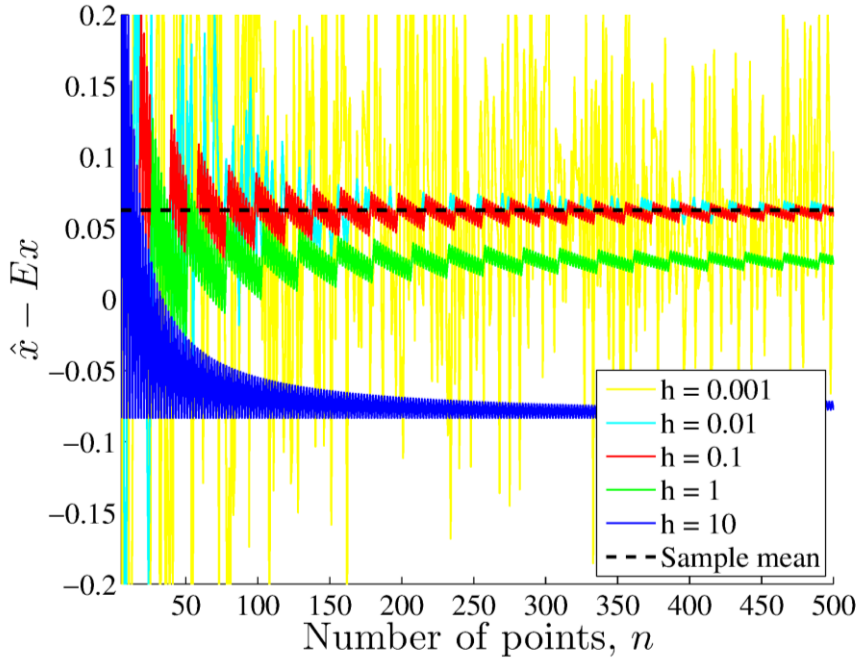

Рисунок 1.15 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе треугольного ядра для квадратичной функции потерь

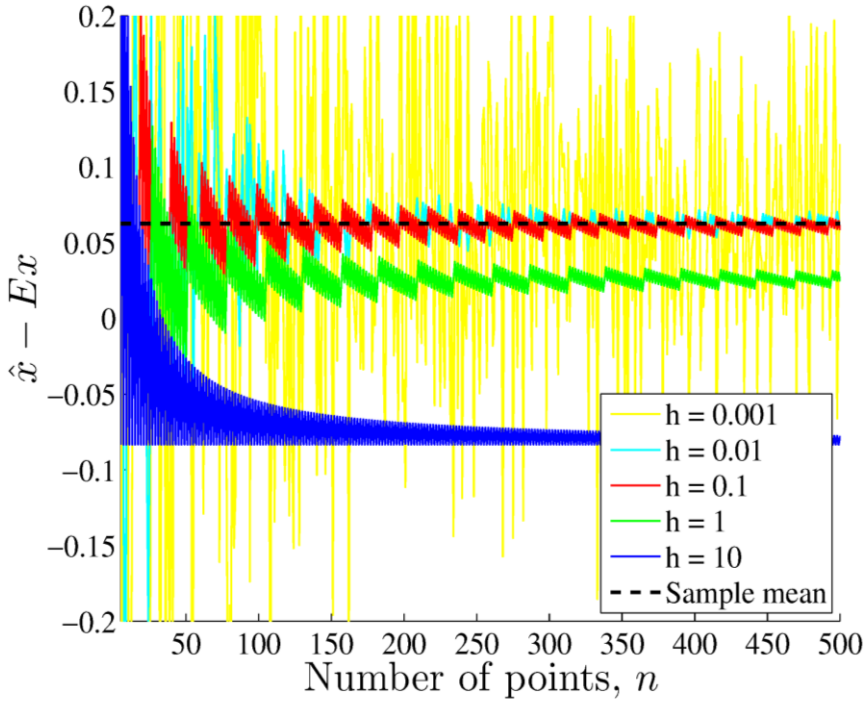

Рисунок 1.16 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе Епанечникова для квадратичной функции потерь

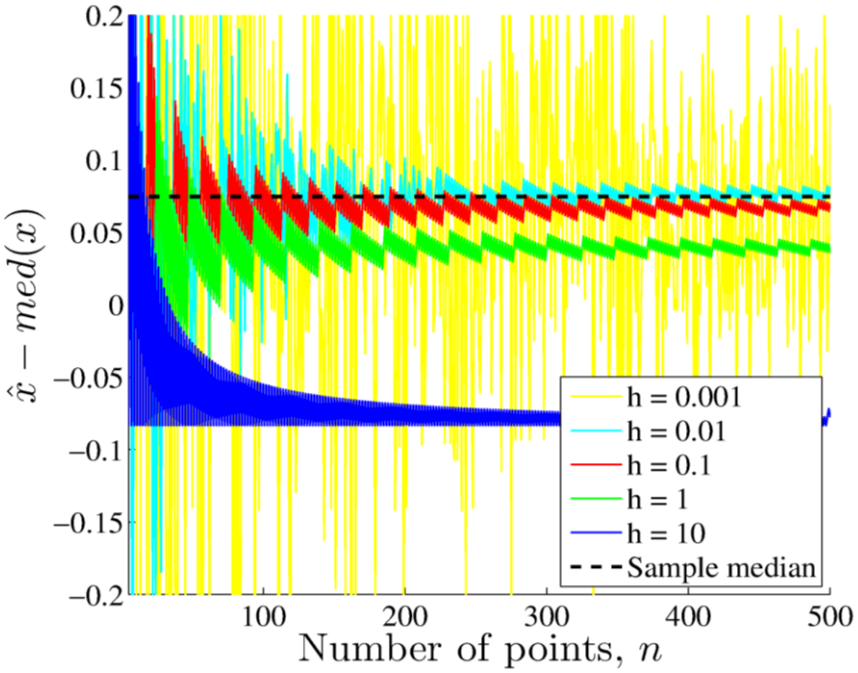

Рисунок 1.17 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе гауссова ядра для абсолютной функции потерь

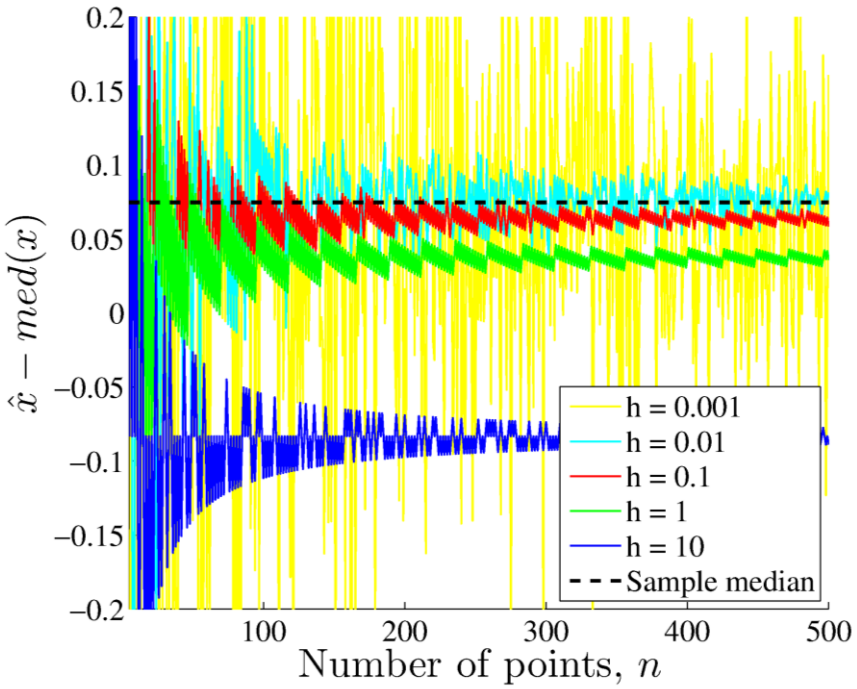

Рисунок 1.18 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе прямоугольного ядра для абсолютной функции потерь

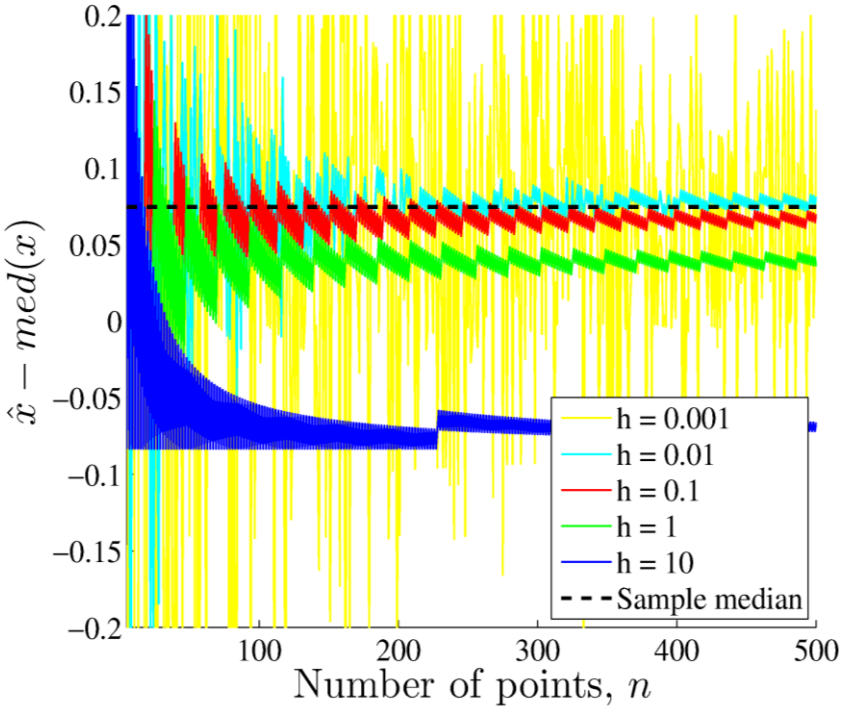

Рисунок 1.19 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе треугольного ядра для абсолютной функции потерь

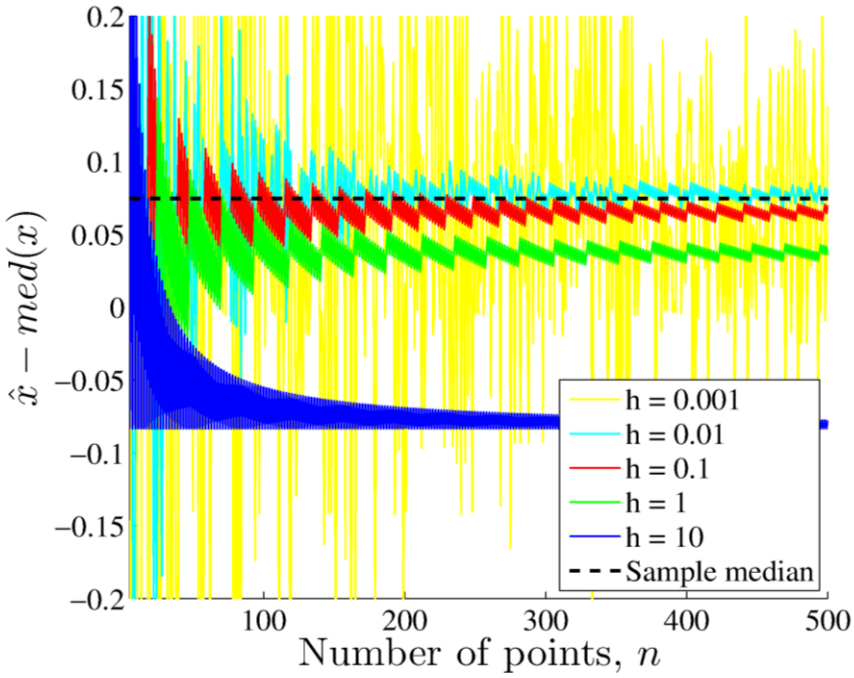

Рисунок 1.20 – Зависимость прогнозов от количества точек свертки при различных значених ширины окна при выборе Епанечникова для абсолютной функции потерь

Из рисунков 1.21 и 1.22 видно, что на практике действительно наблюдаются все те же особенности, о которых сказано выше. Эксперимент был проведён на выборке длины $T = 500$. По оси абсцисс отложено количество столбцов гистограммы, по оси ординат на рисунке 1.21 – отклонение прогноза от истинного математического ожидания, на рисунке 1.22 – отклонение прогноза от истинной медианы распределения. Синий график соответствует прогнозам, полученным алгоритмом hist, красной линией на левом графике отложено отклонение среднего значения выборки (оптимальная оценка в случае неизвестного распределения) от истинного математического ожидания, на правом – модуль отклонения медианы выборки от медианы распределения.

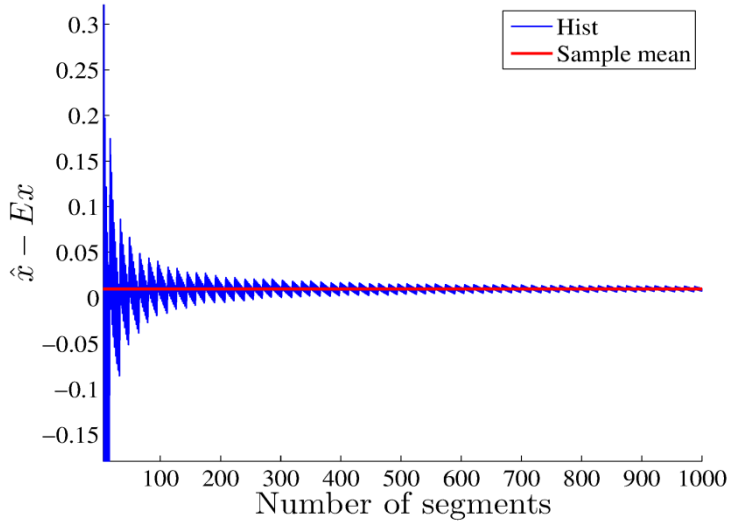

Рисунок 1.21 – Демонстрация особенностей сходимости
для квадратичной функции потерь

На графиках видна стабилизация амплитуды колебания ответов алгоритма hist вокруг предельного значения при большом количестве столбцов в гистограмме. Согласно рисунку 1.21, для квадратичной функции потерь прогнозы, полученные с помощью алгоритма hist сходятся к среднему значению выборки. Аналогично для абсолютной функции потерь (рисунок 1.22) имеет место сходимость прогнозов алгоритма в медиане выборки, а не к медиане распределения. Более того, ввиду конечности выборки, амплитуда колебаний ответов алгоритма вокруг их предельного значения не может бесконечно уменьшаться. Начиная с некоторого момента, прогнозы алгоритма колеблются вокруг предельного значения с постоянной амплитудой.

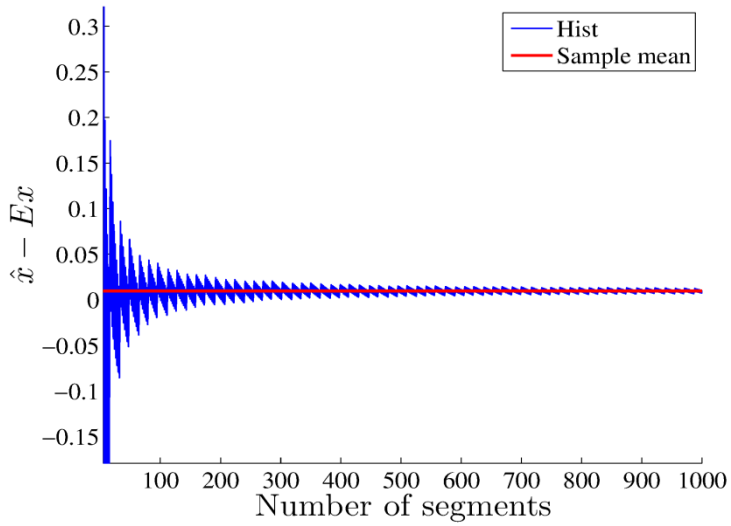

Рисунок 1.22 – Демонстрация особенностей сходимости
для абсолютной функции потерь

Сделанные наблюдения позволяют сделать вывод о том, что не имеет смысла бесконечно увеличивать n. Количество столбцов достаточно выбрать таким, чтобы прогнозы алгоритма стабилизировались вокруг предельного значения. Статистически оптимальным является число столбцов гистограммы, равное $\lceil 3\sqrt[3]{T} \rceil$ [61]. При оптимальном n гистограмма $\hat{p}(u)$ должна наиболее точно приближать плотность $p(u)$ распределения данных. Для определения качества приближения плотности с помощью гистограммы вычислялось расстояние между функциями $p(u)$ и $\hat{p}(u)$ с помощью дивергенции Кульбака-Лейблера [62]:

$$KL(\hat{p}\mathrm{P}p) = \sum_{i=1}^{n} \int_{u_{i-1}}^{u_i} h_i \log \frac{h_i}{p(u)}\, du =$$

$$= b\sum_{i=1}^{n} h_i \log h_i - \sum_{i=1}^{n} h_i \int_{u_{i-1}}^{u_i} \log p(u)\, du.$$

и расстояния Хеллингера [63]:

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

На рисунках 1.23 и 1.24 изображены графики дивергенции Кульбака-Лейблера [62] и расстояния Хеллингера [63] между построенными по выборке гистограммами и истинным стандартным нормальным распределением в зависимости от количества столбцов в гистограмме. Оба графика имеют хорошо выраженный минимум, соответствующий гистограмме с 17 столбцами. С ростом числа столбцов оба расстояния растут примерно линейно с увеличивающейся амплитудой осцилляций – чем больше количество столбцов в гистограмме, тем более она неустойчива.

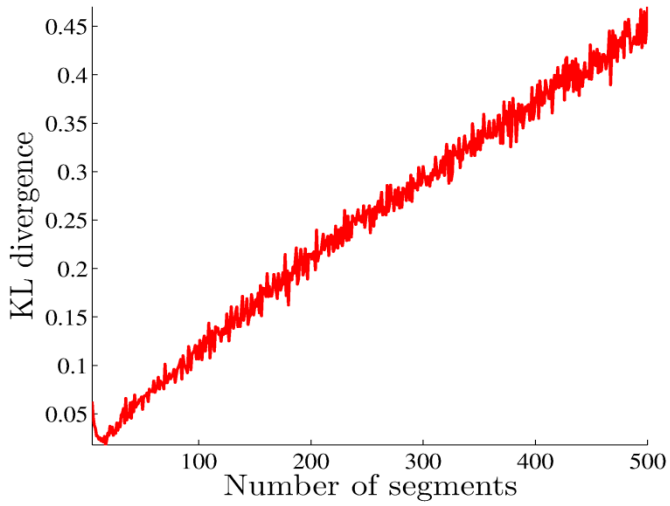

Рисунок 1.23 – Дивергенция Кульбака-Лейблера

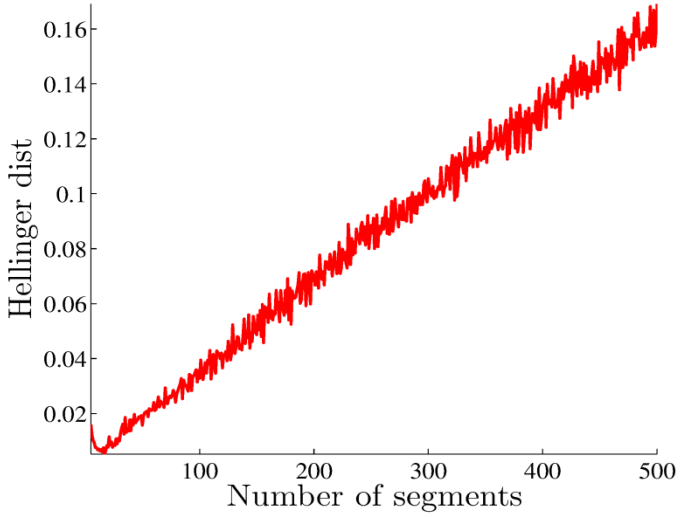

Рисунок 1.24 – Расстояние Хеллингера

1.3.6 Результаты исследования свойств алгоритма

Зависимость качества прогнозов от количества точек свертки n при фиксированной ширине окна и фиксированном ядре имеет такой же характер, как зависимость от количества столбцов в гистограмме – при сходимости наблюдаются осцилляции. Предельное значение прогнозов не совпадает с выборочными статистиками.

- Зависимость прогнозов от ширины окна при фиксированном количестве точек свертки не имеет осцилляций. Существует диапазон значений ширины окна

h, в котором качество прогнозов совпадает с качеством, даваемым выборочным средним для квадратичной функции потерь и выборочной медианой для абсолютной функции потерь. В этом диапазоне предельное значение прогнозов при увеличении количества точек свертки совпадает с выборочной статистикой, соответствующей выбранной функции потерь.

- При малой ширине окна h и увеличении количества точек n свертки наблюдаются сильные осцилляции прогнозов и отсутствие сходимости (либо сходимость с очень малой скоростью). Далее идет диапазон значений ширины окна, при которых прогнозы сходятся к выборочной статистике. При дальнейшем увеличении ширины окна сходимость сохраняется, но предельное значение прогнозов все сильнее отличается от выборочной статистики, соответствующей используемой функции потерь. При малой ширине окна наблюдается уменьшение амплитуды осцилляций прогнозов с увеличением количества точек свертки. Качество прогнозов от формы ядра зависит слабо. В целом гауссово ядро обеспечивает более стабильную сходимость прогнозов при увеличении числа точек свертки при любой ширине окна, чем остальные ядра (прямоугольное, треугольное, Епанечникова).
- При увеличении дисперсии σ^2 выборки качество прогнозов падает. Характер осцилляций при увеличении числа точек свертки не меняется. Отклонение предельного значения прогнозов от выборочной статистики при фиксированной ширине окна уменьшается. Это связано с тем, что диапазон значений ширины окна, в котором качество прогнозов алгоритма hist_K совпадает с качеством, достигаемым при использовании выборочных статистик, смещается по оси вправо, то есть оптимальная ширина окна увеличивается с ростом дисперсии. При этом малые значения ширины окна перестают попадать в этот диапазон, что сопровождается увеличением амплитуды колебаний прогнозов при использовании окон малой ширины на выборках с большой дисперсией.

1.4 Разработка и обоснование метода определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки

В этом подразделе описаны результаты исследования методов определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки, выполненного в соответствии с пп. 2.1.5.4, 3.6.4 Технического задания.

Диапазон значений ширины окна, в котором качество прогнозов алгоритма hist совпадает с качеством прогноза при помощи выборочной статистики, расширяется при увеличении длины выборки. При этом с увеличением длины выборки T характер осцилляций прогнозов не меняется, не меняется также их скорость сходимости к предельному значению. Так как реальных задачах длина выборки ограничена, в данном подразделе поставлена задача нахождения оптимальной длины истории T

$$T = \underset{T \in \mathbb{N}}{\operatorname{argmin}} Q(T),$$

достаточной для стабилизации прогноза, но адекватной с точки зрения практической применимости алгоритма. Таким образом, задача определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки сводится к определению функции качества $Q(T)$, определяющей балланс между стабильностью прогноза и длиной истории T.

В работе [64] рассмотрена задача нахождения минимальной длины выборки, необходимой для получения статистически достовеных результатов классификации. Минимальной длина выборки определялась путем тестирования подвыборок рассматриваемой выборки данных на принадлежность одному распределению на основе дивергенции Кульбака-Лейблера между гистограммами этих подвыборок. Предложенный в этом исследовании способ определения оптимальной длины предыстории основан на схожей идее. Рассматривается расстояние между нет необходимости уменьшать расстояние между гистограммой, построенной по выборке размера T, и гистограммой исходной выборки. На рисунке 1.25 показана зависимость качества

гистограммы от длины предыстории T для ряда отправления вагонов с нефтью с ветки 83. По оси абсцисс отложено количество точек T, использованных для построения гистограммы, по оси ординат – дивергенция Кульбака-Лейблера между гистограммой $\hat{p}$, построенной по части точек истории, и гистограммой $\hat{p}^*$, построенной по всем точкам истории. С увеличением длины истории (ширины окна), различие между гистограммами стремится к нулю. Однако, так как гистограмма $\hat{p}^*$ дает лишь приближение истинного распределения данных, при выборе T нет необходимости уменьшать расстояние между гистограммами до нуля. Достаточно выбрать T таким, чтобы это расстояние было мало.

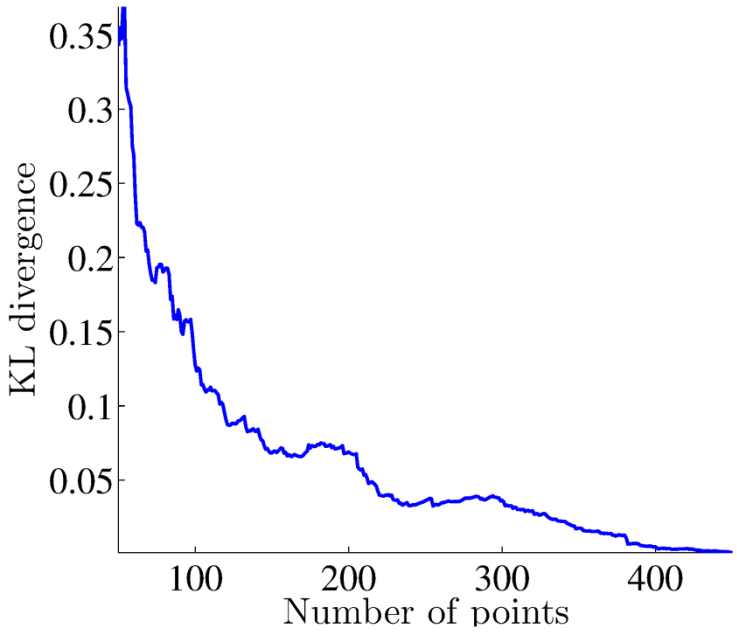

Рисунок 1.25 – Зависимость дивергениции Кульбака-Лейблера от длины предыстории.

Для определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки в данном ПНИ предложено найти баланс между качеством гистограммы и количеством точек, которые необходимо хранить для построения гисторамммы. Качество гистограммы $\hat{p}$, построенной по T точкам, оценивалось ее близостью к распределению

$\hat{p}^*$, восстановленному по всем точкам истории. Таким образом, для оценки оптимального количества точек T минимизируется функцию

$$Q(T) = \mathrm{KL}(\hat{p}||\hat{p}^*) + \alpha T, \tag{14}$$

где γ —вещественный параметр, выбираемый таким образом, чтобы порядки слагаемых совпадали. На рисунке 1.26 изображен график функции $Q(T)$ при $\alpha = 0{,}002$, который имеет глобальный минимум при $T \approx 100$.

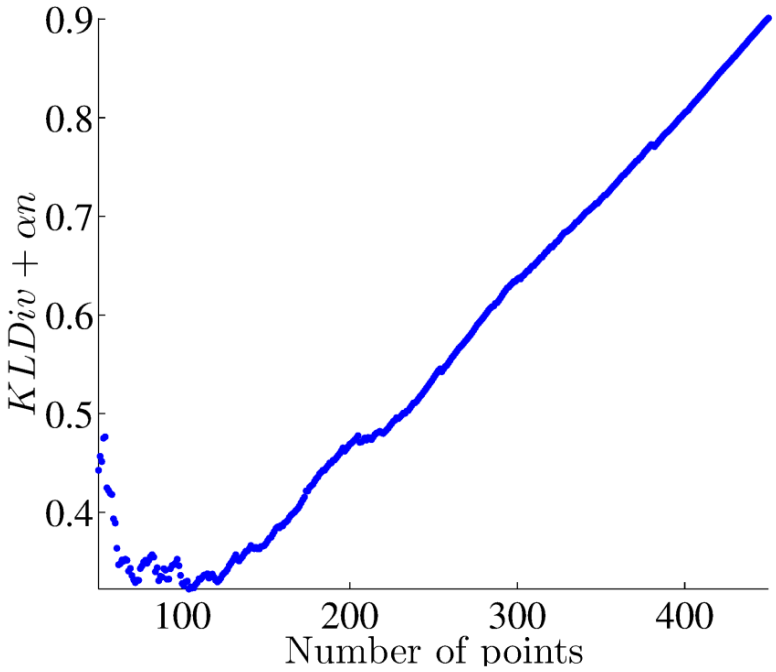

Рисунок 1.26 – Зависимость функции (14) от ширины окна, $\gamma = 0{,}002$

Значение параметра α подбиралось экспериментально. В этом ПНИ рассмотрены временные ряды для перевозок 38 групп грузов. Так как временные ряды различны по характеру (то есть для них характерны разные максимальные значения, разные количества нулевых значений, разная дисперсия значений), для приведения их к одному виду использовалась процедура нормализации, в ходе которой все значения временного ряда делились на максимальное значение данного ряда (если это значение не ноль). После применений этой процедуры, значения всех рядов безразмерны и лежат в отрезке $[0; 1]$. После этого для каждого ряда строилась зависимость

качества прогноза последних 50 точек при вычислении ширины окна путем минимизации (14) от значений параметра α. Затем ошибки, полученные при прогнозировании всех рядов, усреднялись по рядам. Результат избражен на рисунке 1.27 – среднее наилучшее качество прогноза по ветке было достигнуто при $\alpha = 0{,}0017$.

1.5 Разработка и тестирование алгоритмов для выполнения алгебраических операций с гистограммами распределения значений объемов спроса на грузовые железнодорожные перевозки

В этом подразделе описан алгоритм уточнения гистограмм, разработанный на основе исследования методов выполнения алгебраических операций с гистограммами, выполненного в соответствии с пп. 2.1.5.5, 3.6.5 Технического задания. Алгоритм разработан с целью повысить (15) качество прогнозирования алгоритма hist путем учета экзогенных факторов.

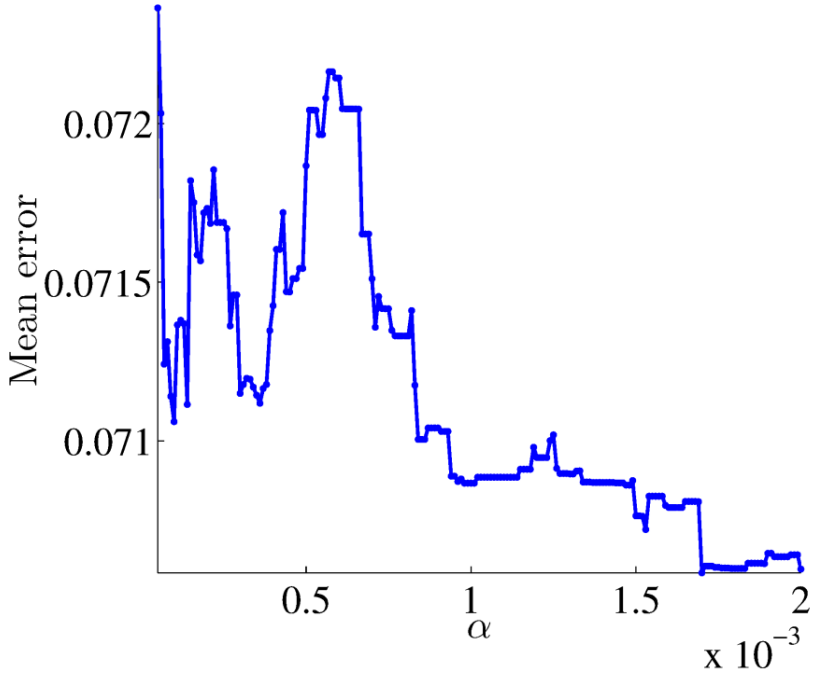

Рисунок 1.27 – Среднее качество прогноза в зависимости от параметра α

При создании моделей, методов и алгоритмов прогнозирования объёмов спроса на ГЖДП в этом ПНИ было принято учитывать как предысторию грузоперевозок в РЖД, так и предысторию влияющих на объёмы спроса экзогенных факторов. Информация о качественном влиянии была получена в рамках первого этапа

данного ПНИ путем анализа экспертных высказываний. Для оценки количественного влияния экзогенных факторов на прогнозируемые временные ряды $\mathbf{x}$ задан набор $\mathbf{Y} = \{\mathbf{y}^1, \dots, \mathbf{y}^N\}$ экзогенных временных рядов $\mathbf{y}^j = \{y_1^j, \dots, y_T^j\}$, испытывающих влияние выделенных в рамках первого этапа данного ПНИ (раздел 3.1 отчета) экзогенных факторов.

При заданной функции потерь $l(\hat{x}, x)$: $\mathbb{R} \otimes \mathbb{R} \mapsto \mathbb{R}$ значение прогноза определяется способом оценки плотности распределения $p(u)$. Для повышения качества прогнозирования в данном ПНИ поставлена задача учета экзогенных временных рядов: на основе измеренных значений эндогенного временного ряда $\mathbf{x}$ и экзогенных временных рядов $\mathbf{y}^j$ построить уточненную гистограмму $\hat{p}(u; v_1, \dots, v_N)$, доставляющую максимальное качество прогнозирования (минимальную ошибку прогноза)

$$l(\hat{x}, x_T) \to \min_{[h_1, \dots, h_n]^T \in [0,1]^n}, \tag{15}$$

где $h_i = \hat{p}(u_i;\ v_1, \dots, v_N)$ при заданых интервалах разбиения $u_0, \dots, u_i, \dots, u_n$.

1.5.1 Задача уточнения прогноза с учетом экзогенных временных рядов

Для учета временных рядов $\mathbf{y}^j$ в гистограммной модели прогнозирования в данном ПНИ предложено уточнить гистограмму $\hat{p}(u)$ временного ряда $\mathbf{x}$ с использованием значений экзогенных временных рядов $\mathbf{y}^j$, $j = 1, \dots, N$. За основу разработанного метода уточнения гистограммы в данном ПНИ взят табличный способ вычисления условных гистограмм. Для оценки гистограммы $\hat{p}_{\mathbf{x}|\mathbf{Y}}(u, v_1, \dots, v_N)$ эндогенного ряда, условной по значениям экзогенных временных рядов $\mathbf{y}^j$ необходимо оценить многомерную гистограмму $\hat{p}_{\mathbf{x},\mathbf{Y}}(u, v_1, \dots, v_N)$, приближающую совместное распределение прогнозируемого временного ряда и экзогенных временных рядов, а затем выбрать срез этой гистограммы на основе конкретных реализаций y_T^j значений экзогенных временных рядов. Таблица 1.1 иллюстрирует процесс оценки условной

гистограммы: в каждой ячейке таблицы содержится значение двумерной гистограммы $\hat{p}_{\mathbf{xy}^j}(u,v)$, в столбце (v_{k-1},v_k) полужирным шрифтом выделены значения условной гистограммы $\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u;y_T)$ эндогенного ряда $\mathbf{x}$, соответствующей значению $y_T^j \in (v_{k-1},v_k)$.

Таблица 1.1 – Связь между совместной и условной гистограммами.

	(v_0,v_1)	…	(v_{k-1},v_k)	…	(v_{K-1},v_K)	$\sum_k$
1	*2*	*3*	*4*	*5*	*6*	*7*
(u_0,u_1)	p_{11}	…	$p_{1k}=\mathbf{h}_1^j\cdot p_{.k}$		p_{1K}	$\mathbf{h}_1^0$
(u_1,u_2)	p_{21}	…	$p_{2k}=\mathbf{h}_2^j\cdot p_{.k}$		p_{2K}	$\mathbf{h}_2^0$
…	…	…	…	…	…	…

Таблица 1.1 – Продолжение

1	*2*	*3*	*4*	*5*	*6*	*7*
(u_{n-1},u_n)	p_{n1}	…	$p_{n2}=\mathbf{h}_i^j\cdot p_{.k}$		p_{nK}	$\mathbf{h}_K^0$
$\sum_i$	p_1^j	…	p_i^j	…	p_K^j	1

При построении многомерных гистограмм ограничением является длина T предыстории рассматриваемых временных рядов. Это ограничение связано с разреженностью таблиц при увеличении их размерности (количества экзогенных временных рядов N). В связи с этим в данном ПНИ предложен альтернативный подход – использовать при прогнозировании эндогенного ряда взвешенную сумму условных гистограмм $\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u,v)$, $j=1,\dots,N$

$$\hat{p}_T(u;v_1,\dots,v_N)=w_0\hat{p}(u)+\sum_{j=1}^{N} w_j\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u,v), \qquad (16)$$

где $\hat{p}_t(u; v_1, \dots, v_N)$ обозначает уточненную гистограмму, построенную с учетом первых $t-1$ значений ряда $\mathbf{x}$ и значений y_t^j экзогенных временных рядов. Таким образом учитывается вся предыстория временного ряда и не происходит потери информации, как в случае с условными гистограммами. Кроме того, данный подход менее требователен к длине истории, так как взвешенная сумма, или смесь гистограмм, не требует построения многомерных гистограмм. Приближая гистограмму взвешенной суммой гистограмм, достаточно вычислить $N+1$ (N двумерных гистограмм и одну одномерную) вместо одной $(N+1)$-мерной, что значительно ослабляет требования к длине предыстории рассматриваемых временных рядов. Кроме того, при использовании смесей гистограмм возможен выбор наиболее информативных экзогенных временных рядов $\mathbf{y}_j$, $j \in \mathcal{J}$ на основе анализа весов w_j соотвествующих компонент $\hat{p}_{\mathbf{x}|\mathbf{y}_j}$ смеси.

Вектор весов $\mathbf{w} = [w_0, \dots, w_N]^T$ компонент смеси (16) максимизирует правдоподобие модели, приближенное с помощью $\hat{p}_T(u; v_1, \dots, v_N)$

$$\mathbf{w} = \underset{\mathbf{w}\in[0,1]^{|\mathcal{J}|}\sum_{j\in\mathcal{J}} w_j=1}{\operatorname{argmax}} \frac{1}{|\mathcal{J}|}\sum_{t=1}^{T} \log\left(\sum_{j\in\mathcal{J}} w_j h_i^j(t)\right), \text{где} h_i^j(t) = \hat{p}_{\mathbf{x}|\mathbf{y}^j}(x_t, y_t^j). \quad (17)$$

Двумерная гистограмма $\hat{p}_{\mathbf{x},\mathbf{y}_j}(u, v)$ представлена таблицей 1.1. Значения p_{ik} соответствуют оценкам вероятности события

$$\{x_{t-1}, y_t^j\} \in (u_{i-1}, u_i) \otimes (v_{k-1}, v_k),$$

означающего совместное попадение значения x_{t-1} временного ряда $\mathbf{x}$ в интервал (u_{i-1}, u_i) и значения y_t^j времнного ряда $\mathbf{y}^j$ – в интервал (v_{k-1}, v_k). Суммирование таблицы по столбцам или строкам дает маргинальные гистограммы $\hat{p}_{\mathbf{x}}(u) \in \{h_1^0, \dots, h_n^0\}$:

$$\hat{p}_{\mathbf{x}}(u) = h_i^0, \text{если } u \in (u_{i-1}, u_i)$$

или $\hat{p}_{\mathbf{y}}(v) \in \{p_1^j, \dots, p_K^j\}$

$$\hat{p}_{\mathbf{y}^j}(v) = p_k^j, \text{если } v \in (v_{k-1}^j, v_k^j),$$

соответственно. Значения h_i^j условных гистограмм $\hat{p}_{\mathbf{x}|\mathbf{y}_j}(u, v)$ задаются выражением:

$$h_i^j = \frac{p_{ik}}{p_k}, \quad p_k = \sum_{i=1}^{n} p_{ik}.$$

Рисунок 1.28 иллюстрирует случай с количеством интервалов гистограммы экзогенного временного ряда $\mathbf{y}_j$ равным $K = 2$. Красным и синим цветом изображены условные гистограммы $\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u, (v_0 + v_1)/2)$ и $\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u, (v_1 + v_2)/2)$. Серым цветом изображена маргинальная гистограмма $\hat{p}_{\mathbf{x}}(u)$.

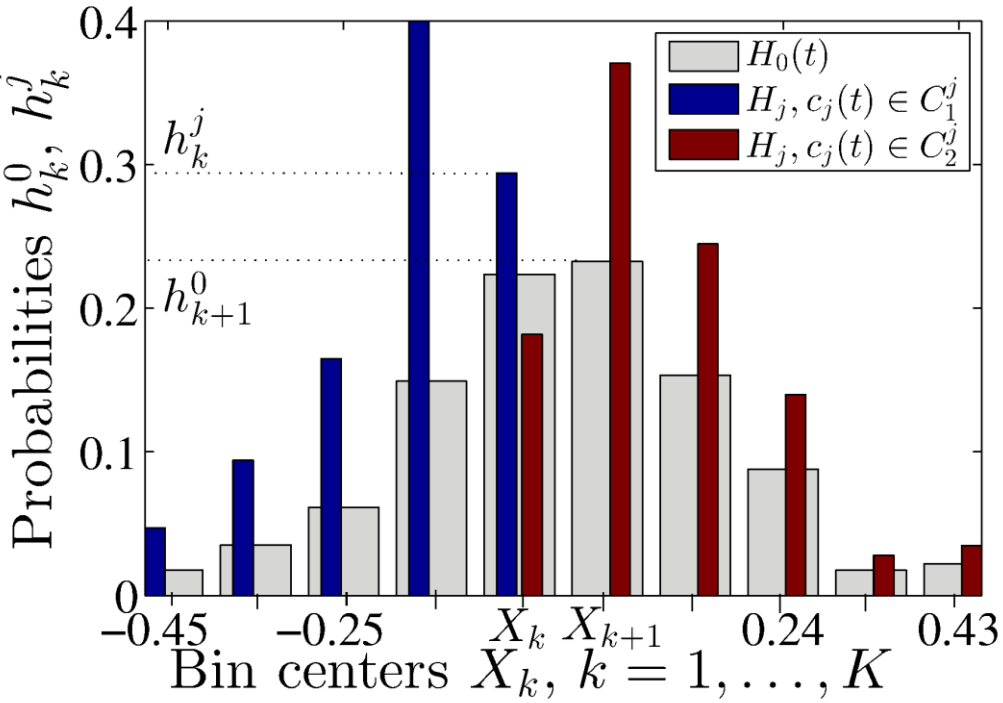

Рисунок 1.28 – Маргинальная и условные гистограммы при $K = 2$

1.5.2 Алгоритм SEM

Настройка весов (17) компонент смеси с параллельным отбором информативных экзогенных временных рядов производилась с помощью алгоритма SEM.

Чтобы для стационарного временного ряда $\mathbf{x} = \{x_1, \ldots, x_T\}$ и набора экзогенных временных рядов $\mathbf{y}^j$ и их производных выбрать с помощью алгоритма SEM

набор $\mathcal{J}$ информативных экзогенных временных рядов и оценить веса w_j компонентов смеси (16), необходимо задать параметр отбора $\alpha \in [0,1)$, количество столбцов эндогенной n и экзогенной K гистограммы, минимальную длину истории $T_{\min}$ и максимальное количество компонент смеси $N_{\max}$. Алгоритм состоит из следующих шагов:

Шаг 1: задание набора информативных экзогенных временных рядов $\mathcal{J} = \{0, \dots, N\}$, начального приближения весов $w_j = 1/|\mathcal{J}|$ и распределения компонент $w_{jt} = 1/|\mathcal{J}|$ для всех $j \in \mathcal{J}, t = T_{\min} + 1, \dots, T$; оценка $\hat{p}_{\mathbf{x}|\mathbf{y}^j}(u, v)$ для всех $j \in \mathcal{J}$; генерация выборки $\tilde{\mathbf{x}} = \{\tilde{x}_t\}_{t=T_{\min}+1}^{T}$ согласно распределению, задаваемому w_{jt}:

$$\tilde{x}_t \sim \sum_{j \subseteq \mathcal{J}} w_{jt} \hat{p}_{\mathbf{x}|\mathbf{y}^j}(u, y_t^j).$$

Шаг 2: вычисление для каждой компоненты $j \in \mathcal{J}$ доли объектов выборки $\tilde{\mathbf{x}}$, описываемой j-той компонентой:

$$T_j = \sum_{t=T_{\min}+1}^{T} \left[\arg\max_{k \in \mathcal{J}} \hat{p}_{\mathbf{x}|\mathbf{y}^k}(\tilde{x}_t, y_t^k) = j \right]$$

и удаление из модели компоненты, описывающие менее $\alpha(T - T_{\min})$ объектов выборки:

$$\mathcal{J} = \mathcal{J} \backslash \{j:\ T_j < \alpha(T - T_{\min})\}.$$

и пересчёт весов w_j для оставшихся компонент $j \in \mathcal{J}$:

$$w_j = T_j / T.$$

Шаг 3: перерасчёт распределения w_{jt}

$$w_{jt} = \frac{w_j \hat{p}_{\mathbf{x}|\mathbf{y}^j}(\tilde{x}_t, y_t^j)}{\sum_{k \in \mathcal{J}} w_k \hat{p}_{\mathbf{x}|\mathbf{y}^k}(\tilde{x}_t, y_t^k)}.$$

Перечисленные шаги повторяются, пока $|\mathcal{J}| > N_{\max}$.

1.5.3 Тестирование алгоритма уточнения гистограммы на основе информации об экзогенных временных рядах

В качестве эндогенных временных рядов, влияние на которые экзогенных факторов следует учитывать при прогнозировании спроса на ГЖДП, был использован набор временных рядов о перевозках групп грузов, представленных в таблице 1.2.

Таблица 1.2 – Расшифровки кодов перевозимых групп грузов.

№	Группа грузов	№	Группа грузов
1	*2*	*3*	*4*
1	Каменный уголь	23	Цемент
2	Кокс	24	Лесные грузы
3	Нефть и нефтепродукты	25	Сахар
4	Торф и торфяная продукция	26	Мясо и масло животное
5	Сланцы горючие (данные отсутствуют)	27	Рыба
6	Флюсы	28	Картофель, овощи и фрукты
7	Руда железная и марганцевая	29	Соль поваренная
8	Руда цветная и серное сырье	30	Остальные продовольственные товары
9	Чёрные металлы	31	Промышленные товары народного потребления
10	Машины и оборудование	32	Хлопок (данные отсутствуют)
11	Металлические конструкции	33	Сахарная свекла и семена
12	Метизы	34	Зерно
13	Лом черных металлов	35	Продукты перемола
14	Сельскохозяйственные машины	36	Комбикорма
15	Автомобили	37	Живность (данные отсутствуют)
16	Цветные металлы, изделия из них и лом цветных металлов	38	Жмыхи
17	Химические и минеральные удобрения	39	Бумага
18	Химикаты и сода	40	Перевалка грузов с водного на ж.д. транспорт (данные отсутствуют)

19	Строительные грузы	41	Импортные грузы (данные отсутствуют)

Таблица 1.2 – Продолжение.

1	*2*	*3*	*4*
20	Промышленное сырье и формовочные материалы	42	Грузы в контейнерах
21	Шлаки гранулированные	43	Остальные и сборные грузы
22	Огнеупоры		

В качестве экзогенных временных рядов в данном ПНИ рассмотрен набор рядов, описывающих внутренние цены на соответствующие товары: сахар, бензин, медь, цинк, золото, никель, пшеницу, мазут, газ, олово, нефть, серебро и свинец за рассматриваемый период времени с различным временным лагом. Кроме того, набор экзогенных временных рядов в данном ПНИ был расширен производными временными рядами – индикаторами возрастания исходных временных рядов.

Выбор значений параметров n, K, $T_{\min}$. Согласно оценкам из [61] было выбрано значение $n = \lceil 3\sqrt[3]{T} \rceil$. При выборе количества столбцов K гистограммы $\hat{p}_{\mathbf{y}^j}(v)$ для каждой из компонент j экспериментально оценивалась вероятность ее включения в модель при различных значениях K. Для проверки гипотезы о независимости вероятности включения компонент в модель от K использован критерий Крускала-Уоллиса [65]. Наблюдаемые значения p-value равнялись около 0,95, что свидетельствует о недостаточности данных для принятия решения о зависимости результатов от K. Так как минимальное количество отчетов $T_{\min}$, необходимое для применения алгоритма, зависит от K линейно, было выбрано минимальное значение $K = 2$. Минимальная длина истории выбрана равной $100K$, из расчета 100 точек на каждый столбец гистограммы, отраженной в таблице 1.1.

Для оценки качества уточнения гистограммы с помощью выбранных информативных временных рядов $\mathbf{y}_j$, $j \in \mathcal{J}$ вычислялось изменение потерь при уточнении гистограммы. На основе 20 прогонов алгоритма SEM для каждого прогнозируемого временного ряда был выбран набор $\mathcal{J}$ из $N_{\max}$ наиболее информативных экзоген-

ных временных рядов и вычислены оценки весов w_j соответствующих выбранным $\mathbf{y}^j$ компонентам смеси в 29 контрольных точках, соответствующих последним точкам $t = 201, \ldots, 229$ истории. Рисунки 1.29 и 1.30 иллюстрируют результаты 20 запусков алгоритма SEM для каждого из эндогенных временных рядов. По осям ординат на графиках рисунков 1.29, 1.30 отложены потенциальные компоненты смесей: лагированные исторические временные ряды $\{x_t\}_{t=1}^{T-\tau-1}$ и экзогенные временные ряды $\{y_t^j\}_{t=1}^{T-\tau}$ (рисунок 1.29) и их производные $\{y_t^j\}_{t=1}^{T-\tau}$ (рисунок 1.30) с лагами $\tau = 1$ и $\tau = 3$, соответственно. По осям абсцисс отложены эндогенные временные ряды (коды перевозимых грузов). Цвет ячейки $(\mathbf{x}, \mathbf{y}^j)$ каждого из рисунков отвечает числу включения соответствующей компоненты в модель (16) для эндогенного временного ряда $\mathbf{x}$. Это число, разделенное на количество прогонов алгоритма, дает оценку вероятности включения j-той компоненты в набор $\mathcal{J}$.

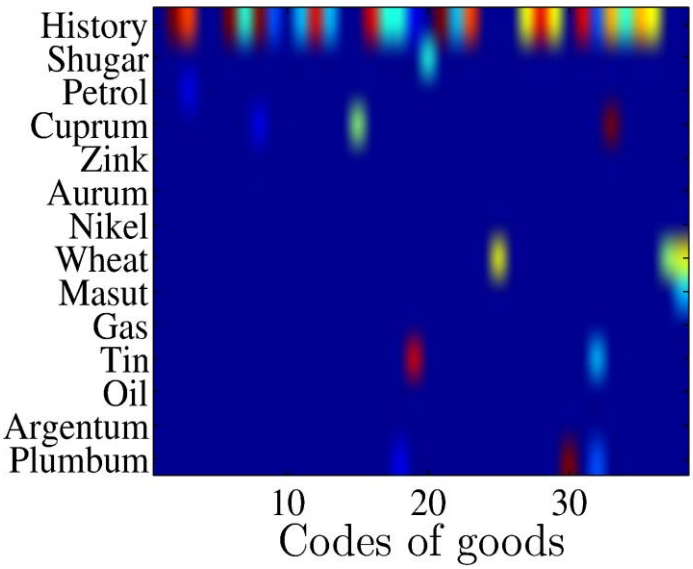

Рисунок 1.29 – Результаты отбора информативных экзогенных временных рядов $\mathbf{c}$ с порядком лагирования $\tau = 1$ для прогнозирования временных рядов объемов грузовых железно-дорожных перевозок

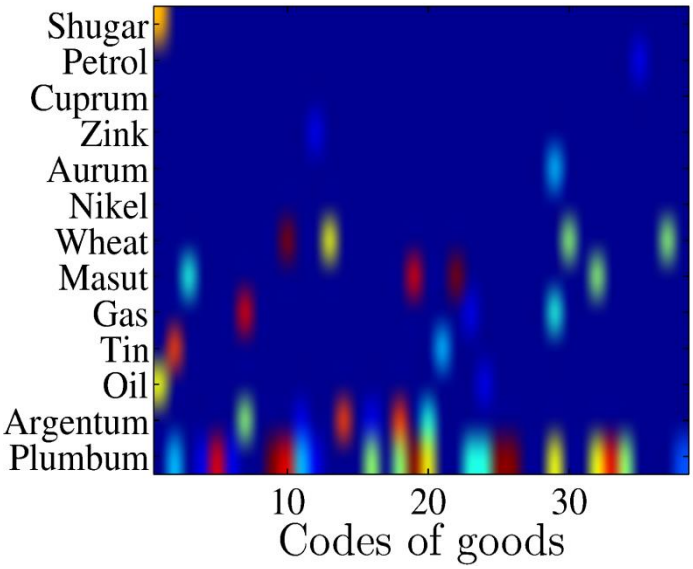

Рисунок 1.30 – Результаты отбора производных экзогенных временных рядов $\mathbf{c}$ с порядком лагирования $\tau = 3$ для прогнозирования временных рядов объемов грузовых железно-дорожных перевозок

На основе выбранных $\mathbf{y}^j$ и соответствующих оценок весов согласно (16) были вычислены уточненные гистограммы $\hat{p}_t(u; v^1, \dots, v^N)$, с помощью которых вычислялись прогноз $\hat{x}_t$ и изменение потерь Δl_t

$$\Delta l_t = l(\hat{x}, x_t) - l(\hat{x}_t, x_t), \quad l(\hat{x}, x) = (x - \hat{x})^2,$$

где $\hat{x}$ – прогноз базовой версии алгоритма hist, $\hat{x}_t$ – прогноз, полученный с помощью уточненной гистограмы. Прогноз выполнялся на 10 точек точек вперед и затем усреднялся. Таким образом была получена выборка $\{\Delta l_t\}$, на основе которой тестировалась гипотеза $\mathbf{E}(\Delta l) = 0$ о равенстве ожидания Δl нулю при альтернативе $\mathbf{E}(\Delta l) > 0$ с помощью t-критерия.

В первом столбце таблицы 1.3 перечислены экзогенные временные ряды, для которых на основе собранной выборки $\{\Delta l_t\}$ было принято решение о статистическом уменьшении потерь при уточнении гистограммы эндогенного временного рядов из второго столбца таблицы (гипотеза $\mathbf{E}(\Delta l) = 0$ отвергалась в пользу альтернативы $\mathbf{E}(\Delta l) > 0$ с p-value < 0,05). Выбранные экзогенные временные ряды приведены в первом столбце таблицы 1.3. Временные ряды со штрихом обозначают про-

изводные временные ряды, число, стоящее напротив названия временного ряда говорит о том, что ряд выбран несколько раз с различным временным лагом: «Мазут'-2» означает, что в модель добавлены 2 экзогенных временных ряда, производных из ряда «Мазут» с различными временными лагами. В столбцах таблицы 1.3 с метками «Δl», «$\#(\Delta l_t > 0)$» and «$\#(\Delta l_t < 0)$» перечислены соответственно среднее значение изменения потерь, доля отсчетов с положительным изменением Δl (чем больше, тем лучше) и доля отсчетов с отрицательным изменением Δl (чем меньше, тем лучше). В последнем столбце содержатся значения p-value, полученные при тестировании гипотезы $\mathbf{E}(\Delta l) = 0$. Меньшие значения p-value соответствуют большей уверенности, что уточнение гистограммы улучшает качество прогнозирования. В таблице 1.3 перечислены только временные ряды с p-value < 0,05.

Таблица 1.3 – Результаты учета экзогенных внешних рядов $\{\mathbf{y}^j, \mathbf{y}^j\}$ в модели hist.

$\mathbf{y}_j$	Типы грузов	Δl	$\#(\Delta l_t > 0)$	$\#(\Delta l_t < 0)$	p-value
1	*2*	*3*	*4*	*5*	*6*
Мазут'-2	Нефть и нефтепродукты	0,16155	0,31034	0,034483	0,0057088

Таблица 1.3 – Продолжение

1	*2*	*3*	*4*	*5*	*6*
Пшеница', Олово', Свинец'	Металлические конструкции	0,33508	0,68966	0,034483	1,96e-06
Газ', Олово', Свинец'	Рыба	0,22634	0,37931	0	0,00016709
Сахар, Свинец'	Другие продоволь-ственные товары	0,27716	0.51724	0,24138	0,029859
Медь, Мазут, Свинец'	Продукты перемола	0,10558	0,27586	0,13793	0,051246
Газ, Свинец'	Комбикорма	0,17925	0,27586	0,10345	0,032905
Нефть'	Жмыхи	0,19045	0,48276	0,13793	0,02642

Пшеница, Мазут	Остальные грузы	0,16718	0,27586	0,034483	0,0043178

1.5.4 Разработка и обоснование методов определения выполнения условия локальной стационарности временного ряда и реализация теста Дики-Фуллера

В этом подразделе представлены результаты исследования свойств алгоритмов прогнозированния в применении к нестационарным временным рядам и методов проверки стациционарности, выполненного в соответствии с пп. 2.1.5.6, 3.6.6 Технического задания.

Описанные в подразделах 1.1–1.5 результаты были получены в предположении (2) о стационарности временных рядов. В этом подразделе рассмотрена задача (1) получения несмещенных прогнозов нестационарных временных рядов. Прогнозирование нестационарных временных рядов включает этапы проверки на стационарность и декомпозиции на стационарную и нестационарные компоненты, описанные в пунктах 1.6.1 и 1.6.2, соответственно.

1.5.5 Анализ качества алгоритмов прогнозирования при наличии нестационарности

Эксперименты на исторических данных. Экспериментальные данные содержат информацию о железнодорожных перевозках 38 номенклатурных типов грузов. Каждый отсчет x_t временного ряда $\mathbf{x}$ которых соответствует одному дню и равен суммарному весу (в тоннах) определенного груза, перевезенному между фиксированными пунктами отправления и назначения. Временные ряды были измерены в период с января 2007 года по май 2008 года, в связи с чем временное агрегирование проводилось только по неделям или месяцам: усреднение данных по кварталам или годам сокращает временные ряды до шести или двух точек, соответственно. При запросе на прогноз по неделям или по месяцам выполнялся подневный прогноз, результаты которого затем усреднялись в соответствии с запросом. Согласно требования к прогнозу, перечисленным во введении к данной работе, прогноз выполняется для объемов перевозок между фиксированными пунктами отправления и назначения для различных типов грузов. При сравнении алгоритмов прогноз вычис-

лялся для всех пар станций, а также для всех пар районов, имеющих ненулевую историю перевозок.

В таблице 1.4 приведены значения выбранных функций потерь MAE (18) и MAPE (19), усредненные по типам перевозимых грузов, вычисленные при прогнозировании по парам станций и парам районов с выбранной детализацией по времени детализации по времени. Полужирным шрифтом выделены лучшие результаты при каждом выборе детализации. Параметры прогнозов с помощью среднего и медианы выбраны экспериментально.

Таблица 1.4 – Ошибки прогнозирования, усредненные по типам грузов, исторические данные.

	По станциям			По районам		
	День	Неделя	Месяц	День	Неделя	Месяц
1	*2*	*3*	*4*	*5*	*6*	*7*
MAE						
ARMA с регуляризацией	1,684	9,323	33,528	4,370	**23,456**	83,595
Среднее по 60 дням	1,927	10,266	35,650	4,892	24,638	**79,417**

Таблица 1.4 – Продолжение.

1	*2*	*3*	*4*	*5*	*6*	*7*
Медиана по 100 дням	**1,276**	**7,813**	**31,995**	**3,974**	24,570	104,070
Нулевой прогноз	1,398	9,444	42,710	4,572	30,970	139,383
MAPE						
ARMA с регуляризацией	0,559	1,218	1,743	0,732	1,281	1,621
Среднее по 60 дням	0,698	1,840	3,288	0,963	2,154	3,410
Медиана по 100 дням	**0,364**	**0,790**	**1,352**	**0,489**	**0,904**	**1,375**
Нулевой прогноз	0,382	0,830	1,428	0,518	0,955	1,466

Согласно таблице 1.4, ошибка прогнозирования, как абсолютная, так и относительная, растет при переходе к менее детальным прогнозам. Оптимальным алгоритмом в среднем оказывается прогноз медианой последних ста отсчетов. Также можно отметить, что относительная ошибка MAPE прогнозирования нулями нена-

много превосходит ошибку при прогнозировании медианой по 100 точкам. Это связано с высокой неравномерностью грузовых перевозок на некоторых направлениях, существенная часть истории некоторых прогнозируемых временных рядов содержит лишь небольшое количества ненулевых отсчетов. На модельных данных, содержащих меньшее количество нулевых значений, такого эффекта не наблюдается. Данный результат свидетельствует о необходимости введения функций потерь, учитывающих специфику решаемой задачи: при использовании стандартных ошибок MAE и MAPE прогноз нулями с математической точки зрения близок к оптимальному, однако на практике такой способ прогнозирования не применим.

На рисунках 1.31 и 1.32 изображены графики зависимости ошибок MAE и MAPE прогнозов соответствующих временных рядов от горизонта прогнозирования. Выраженной зависимости ошибки от горизонта прогнозирования не наблюдается ни для одного из рассмотренных алгоритмов.

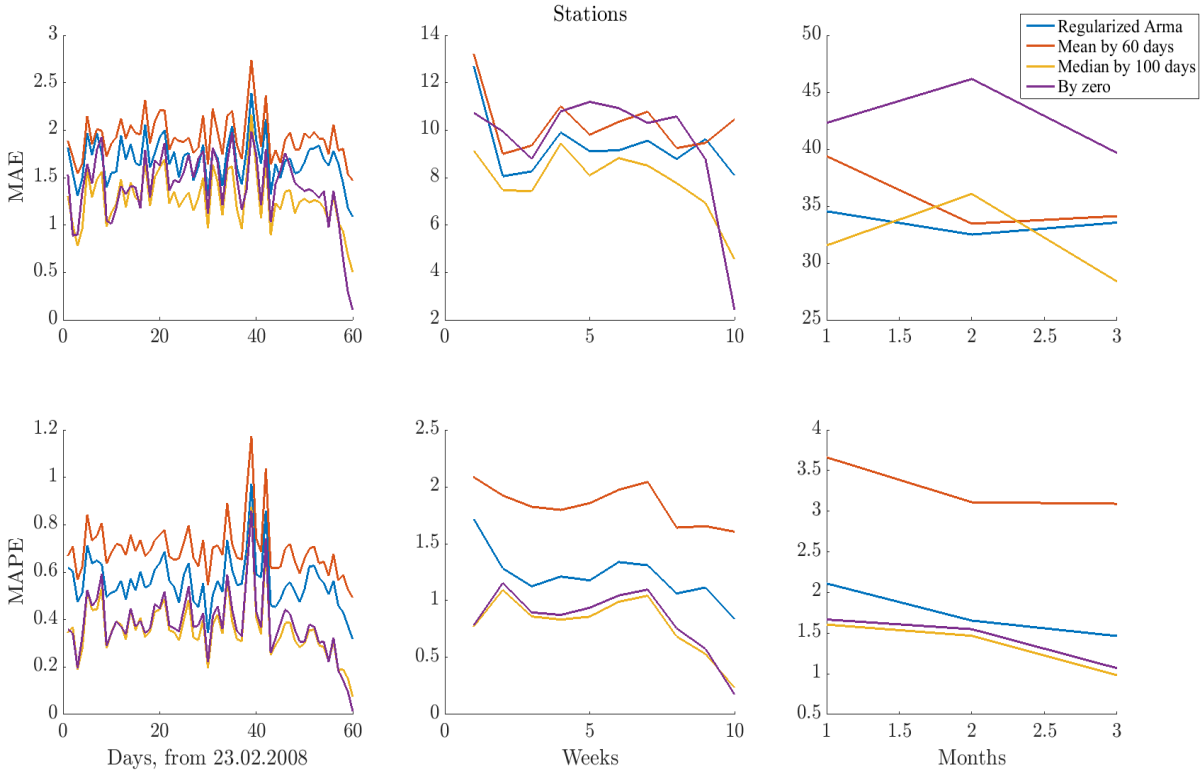

Рисунок 1.31 – Ошибки прогнозирования MAE и MAPE, агрегированные по станциям. Исторические временные ряды

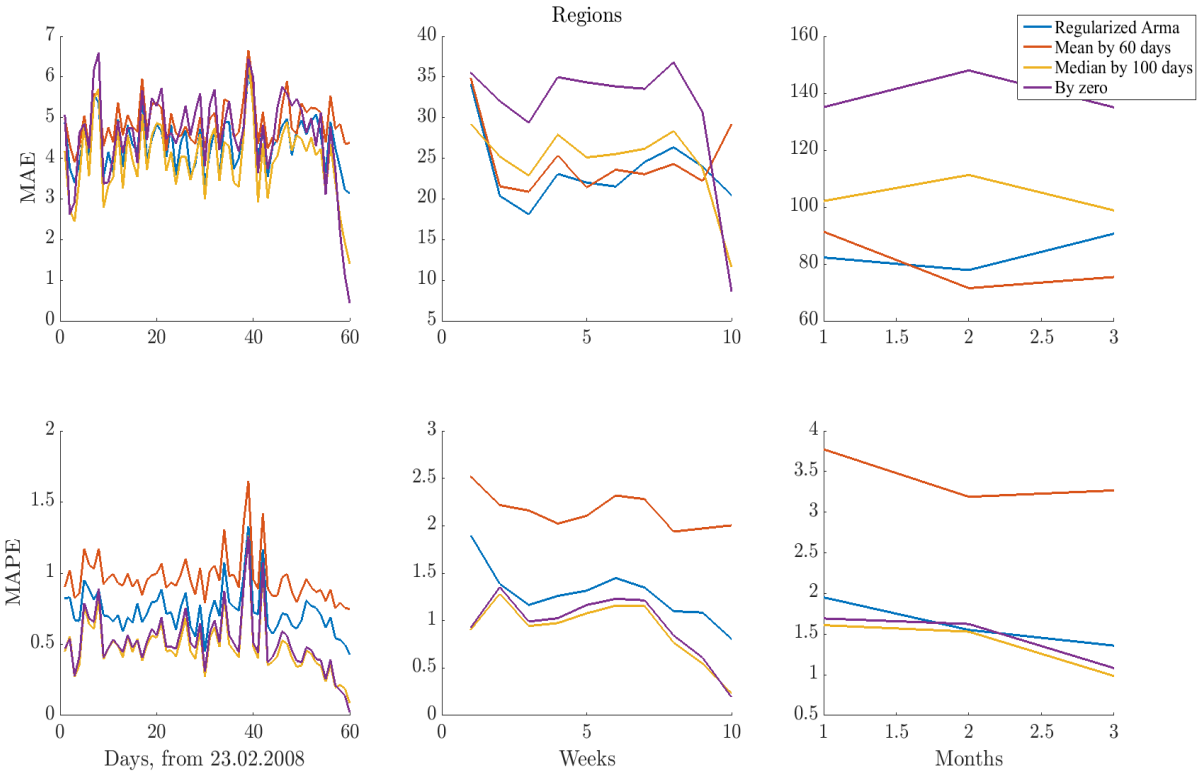

Рисунок 1.32 – Ошибки прогнозирования MAE и MAPE, агрегированные по регионам. Исторические временные ряды

Эксперименты на модельных данных. Для проведения более полных экспериментов в соответствии с пп. 2.3, 3.7 Технического задания был разработан генератор модельных данных, учитывающий особенности структуры грузоперевозок на железнодорожном транспорте, определяемые следующими внешними экспертными данными:

- топологией станций с указанием кодов станций и их принадлежностью к районам;
- информацией о парах станций, используемых для перевозки того или иного товара,
- экзогенными факторами, влияющие на совокупный спрос соответствующего товара;
- историческими данными для экстраполяции структуры перевозок на все области;
- другими экспертными требованиями, такими как устойчивость структуры перевозок во времени.

В таблице 1.5 приведены значения выбранных функций потерь MAE (18) и MAPE (19), усредненные по типам перевозимых грузов, вычисленные при прогнозировании по парам станций и парам районов с выбранной детализацией по времени детализации по времени.

Таблица 1.5 – Ошибки прогнозирования, усредненные по типам грузов, синтетические данные.

	По станциям			По районам		
	День	Неделя	Месяц	День	Неделя	Месяц
1	*2*	*3*	*4*	*5*	*6*	*7*
MAE						
VAR	917,104	**3114,073**	**6986,463**	1946,955	5855,633	17053,388
ARMA с регуляризацией	**913,336**	3381,220	9366,639	**1853,329**	**5168,458**	**14056,409**
Среднее по 5 дням	959,840	4195,548	13987,420	2892,025	18165,117	76272,813
Медиана по 70 дням	984,188	4623,588	16761,591	3168,165	20385,037	84069,813
MAPE						
VAR	**0,100**	**0,057**	**0,025**	0,042	0,021	0,012

Таблица 1.5 – Продолжение.

1	*2*	*3*	*4*	*5*	*6*	*7*
ARMA с регуляризацией	**0,100**	0,062	0,034	**0,040**	**0,019**	**0,010**
Среднее по 5 дням	0,105	0,076	0,051	0,063	0,065	0,055
Медиана по 70 дням	0,107	0,083	0,060	0,068	0,073	0,060

Полужирным шрифтом выделены лучшие результаты при каждом выборе детализации. Как и в случае исторических данных (таблица 1.4), средняя по типам грузов ошибка MAE растет при переходе к менее детальным прогнозам. Напротив, для ошибки MAPE наблюдается обратная зависимость. Оптимальными в этом случае моделями оказались модель векторной авторегрессии при прогнозе по парам станций и авторегрессионного скользящего среднего с регуляризацией параметров модели при прогнозе по парам районов. Оптимальные параметры прогнозирования средним и медианой для синтетических рядов меньше, чем для исторических в связи

с наличием более выраженного тренда. На рисунках 1.33 и 1.34 изображены графики зависимости ошибок MAE и MAPE прогнозов соответствующих временных рядов от горизонта прогнозирования. При росте горизонта прогнозирования для всех алгоритмов наблюдается увеличение ошибки прогнозирования. Этот эффект также связан с наличием тренда.

Поведение прогнозов каждого из алгоритмов продемонстрировано графиками на рисунке 1.35, на котором показаны примеры сгенерированных временных рядов, соответствующих перевозкам грузов с номенклатурными кодами 1 и 3.

Синим цветом на рисунке 1.35 отложены графики сгенерированных рядов. Красным, желтым, фиолетовым и зеленым цветами – прогнозы временных рядов различными алгоритмами. На рисунках видно как при росте горизонта прогнозирования значения прогнозов отклоняются от истинных значений временного ряда. Таким образом, наличие тренда существенно сказывается на точности прогнозов при использовании стандартных алгоритмов прогнозирования, что подтверждает необходимость разработки методов прогнозирования нестационарных временных рядов.

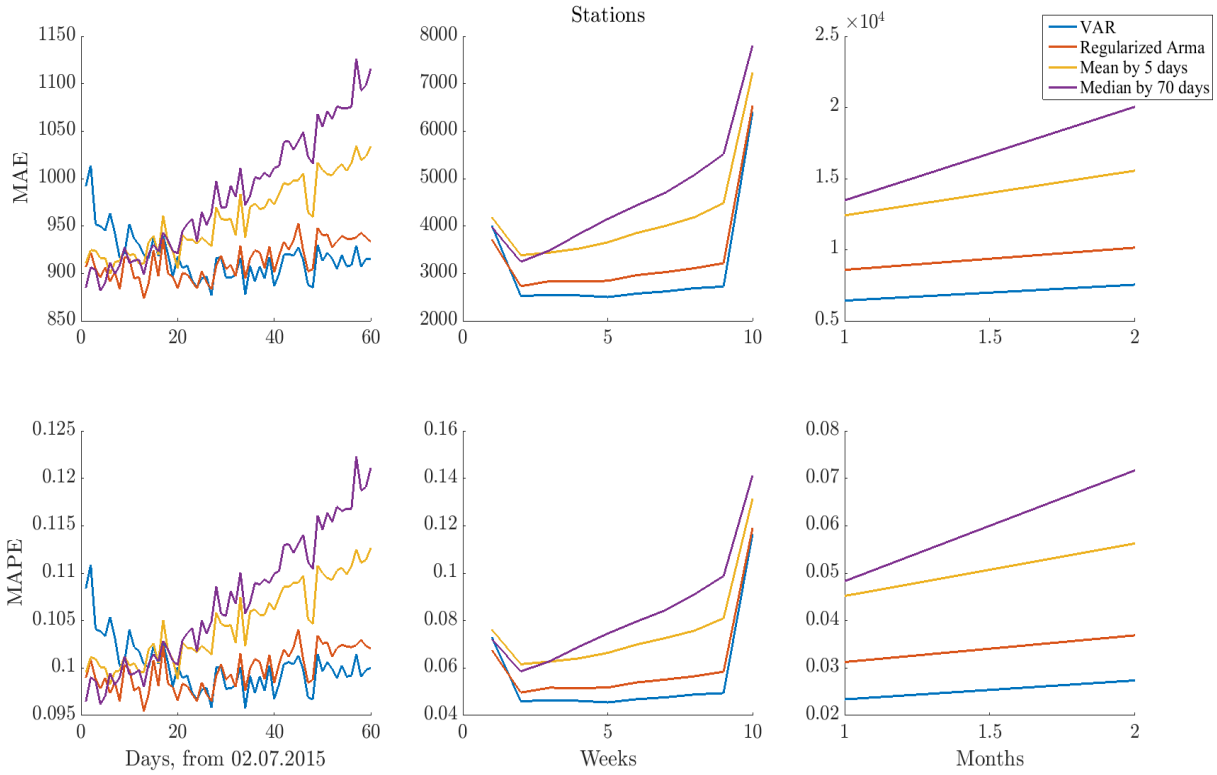

Рисунок 1.31 – Ошибки прогнозирования MAE и MAPE,
агрегированные по станциям. Синтетические временные ряды

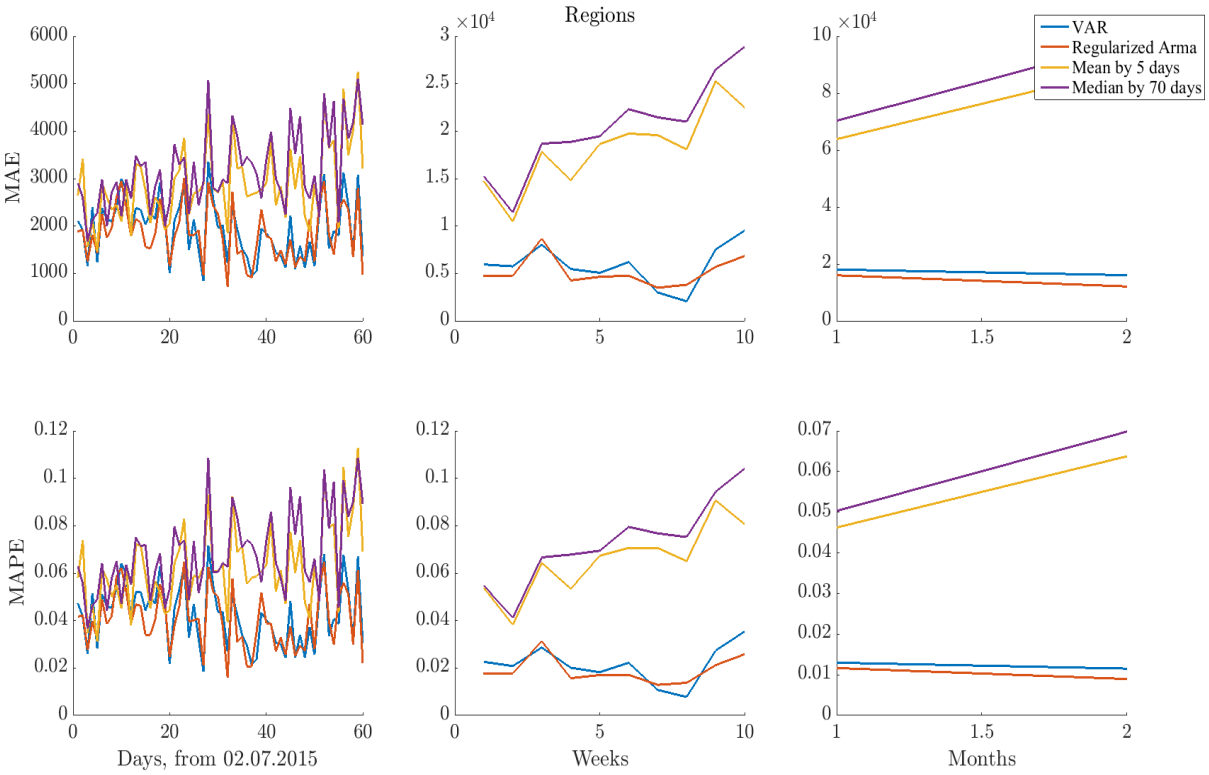

Рисунок 1.34 – Ошибки прогнозирования MAE и MAPE, агрегированные по регионам. Синтетические временные ряды

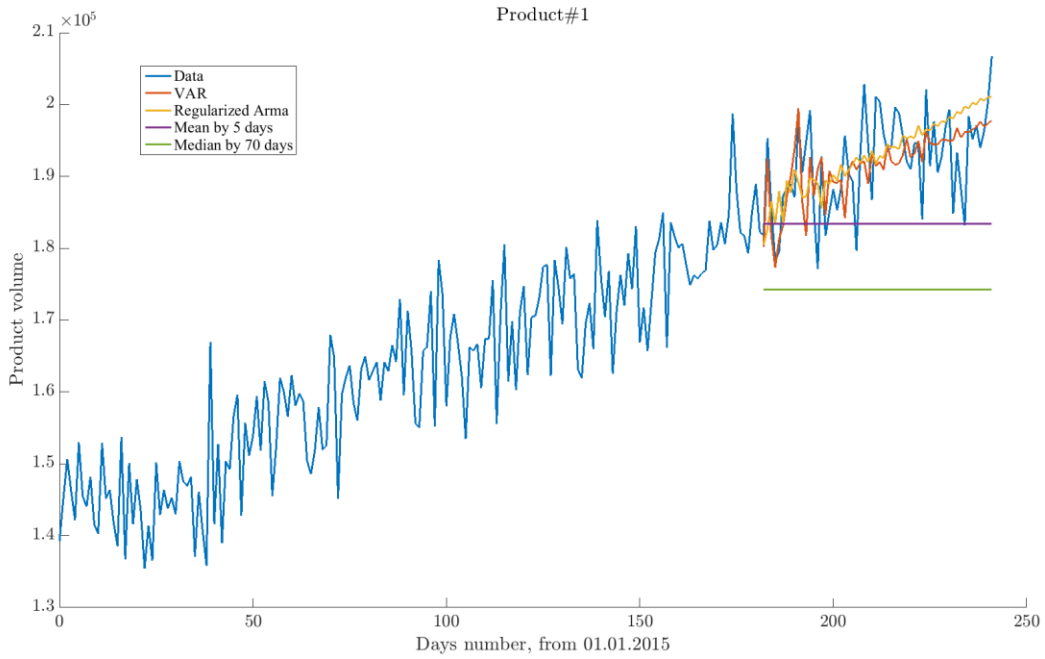

Рисунок 1.35 – Пример синтетического временного ряда и прогнозов, полученных с помощью различных алгоритмов

1.5.6 Реализация теста Дики-Фуллера

При разработке алгоритма hist на временные ряды накладывалось требование стационарности, связанное с предположением что значения временного ряда принадлежат одному распределению с плотностью $p(u)$. Поскольку стационарность временных рядов необходима для использования рассматриваемой прогностической модели, для всех рядов проведен тест Дики-Фуллера на стационарность. Суть теста заключается в проверке гипотезы о равенстве нулю коэффициента b в авторегрессионном уравнении первого порядка

$$x_t - x_{t-1} = bx_{t-1} + \varepsilon_t$$

против альтернативы $b < 0$, так как значения $b > 0$ означают возможность принимать бесконечно большие значения за конечные промежутки. На рисунках 1.36 и 1.37 показаны результаты теста Дики-Фуллера для прогнозируемых временных рядов. Каждая ячейка матриц на рисунках 1.36 и 1.37 соответствует одному временному ряду о прибытии/отправлении вагонов с некоторым типом груза. По оси абсцисс отложены коды веток, по которым отправлялись грузы, по оси ординат – коды типов грузов. Красным обозначены ряды, для которых принято решение о наличии нестационарности. Использовалась реализация теста в среде MatLab.

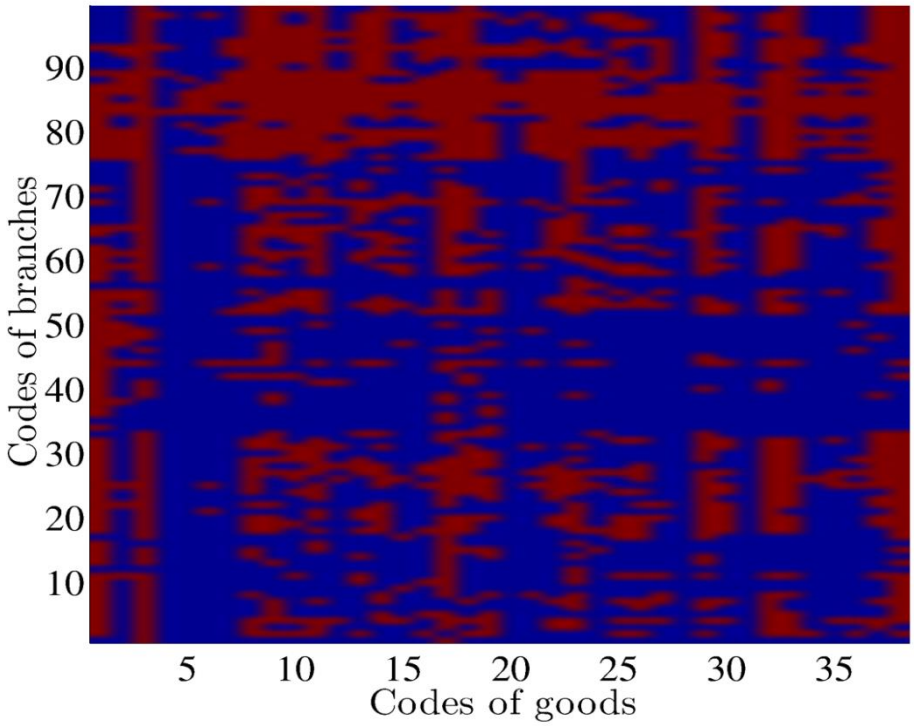

Рисунок 1.36 – Тест Дики-Фуллера для временных рядов прибытия вагонов.

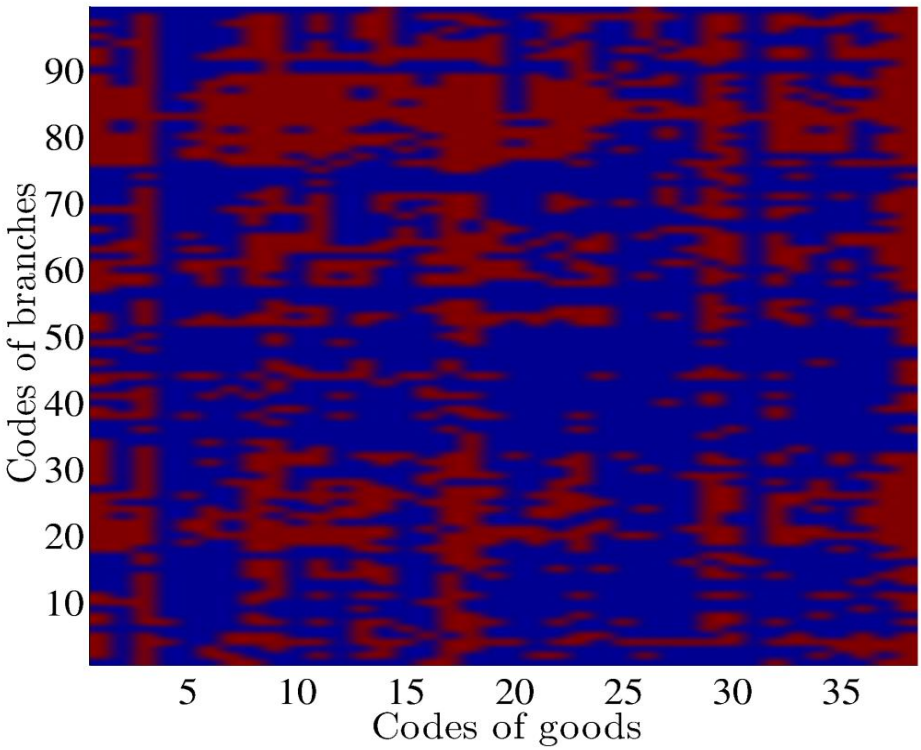

Рисунок 1.37 – Тест Дики-Фуллера для временных рядов отправления (b) вагонов

В случае, когда временной ряд не является стационарным, например имеет восходящий или нисходящий тренд, сезонную компоненту или изменяющуюся с течением времени дисперсию, нельзя предполагать, что значения временного ряда сгенерированы из одного распределения с плотностью $p(u)$. Один из широко используемых методов прогнозирования нестационарных временных рядов, авторегрессионное интегрированное скользящее среднее ARIMA [8], позволяет с хорошим качеством прогнозировать временные ряды с трендом, а также при небольшой модификации и ряды с сезонной компонентой. Однако настройка параметров этого алгоритма осуществляется путем минимизации квадратичной функции потерь. Это приводит к тому, что оптимальный прогноз для модели ARIMA является несмещенным, а регрессионные остатки должны иметь нулевое среднее, быть гомоскедастичными, нормальными и некоррелированными между собой. Ввиду выше сказанного модель ARIMA не подходит для решения задачи прогнозирования в случае несимметричной функции потерь, что отмечается в [9, 10].

В этом ПНИ предложен двухэтапный алгоритм прогнозирования, включающий этап выделения и прогнозирования нестационарной компоненты и прогноза стационарных остатков.

1.5.7 Разработка и обоснование методов прогнозирования нестационарных временных рядов

В этом пункте описан двухэтапный алгоритм $\mathrm{ARIMA} + \mathrm{hist}$ прогнозирования (6) 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 [66], 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 [67] о зависимости смещения прогноза исключительно от функции потерь и дисперсии временного ряда.

Алгоритм $\mathrm{ARIMA} + \mathrm{hist}$, вычислющий прогноз $\hat{x}$ для временного ряда $\mathbf{x} = \{x_1, \dots, x_T\}$ и функции потерь $l(\hat{x}, x_{T+1})$, состоит из следующей последовательности шагов.

Шаг 1: подбор подходящей для временного ряда модели ARIMA по методологии Бокса-Дженкинса [8].

Шаг 2: вычисление прогноза нестационарной компоненты $\hat{x}^{ns}$ на основании выбранной модели ARIMA.

Шаг 3: вычисление регрессионныч остатков $\mathbf{r} = \{r_1, \dots, r_T\}$ для выбранной модели ARIMA.

Шаг 4: задание количества столбцов в гистограмме для алгоритма hist, вычисление прогноза стационарной компоненты $\hat{x}^{s}$ с помощью алгоритма hist.

Шаг 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 $\mathbf{x}$ 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 $\{x_1, \dots, x_T\}$

$$r_t = x_t - \hat{x}_t^{ns}, \quad t = 1, \dots, T$$

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 $\{r_1, \dots, t_T\}$ и решения для этой плотности задачи минимизации ожидаемых потерь (6). Стационарность остатков обеспечивалась выбором подходящей модели ARIMA для прогнозирования нестационарной компоненты.

Выбор модели ARIMA и оценка ее параметров. Временной ряд описывается моделью $\mathrm{ARIMA}(p, d, q)$, если ряд его разностей

$$\nabla^d x_t = (1 - \mathrm{L})^d x_t,$$

где L – оператор дифференцирования временного ряда

$\mathrm{L}^d x_t = x_{t-d}$,

описывается моделью $\mathrm{ARMA}(p, q)$

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

где $\alpha = \mu(1 - \varphi_1 - \cdots - \varphi_p)$, $\varphi_1, \dots, \varphi_p$, $\theta_1, \dots, \theta_q$ – константы, ε_t – гауссов шум с нулевым средним и постоянной дисперсией. При необходимости модель ARIMA можно расширить мультипликативной сезонностью SARIMA. Временной ряд описывается моделью $\mathrm{SARIMA}(p, d, q) \times (P, D, Q)_s$, если

$$\Phi_P(\mathrm{L}^s)\varphi(\mathrm{L})\nabla_s^D \nabla^d x_t = \alpha + \Theta_Q(\mathrm{L}^s)\theta(\mathrm{L})\varepsilon_t,$$

где

$$\Phi_P(\mathrm{L}^s) = 1 - \Phi_1 \mathrm{L}^s - \Phi_2 \mathrm{L}^{2s} - \cdots - \Phi_P \mathrm{L}^{Ps},$$

$$\Theta_Q(\mathrm{L}^s) = 1 + \Theta_1 \mathrm{L}^s + \Theta_2 \mathrm{L}^{2s} + \cdots + \Theta_Q \mathrm{L}^{Qs}.$$

Для оценки параметров модели ARIMA была использована методология Бокса-Дженкинса анализа временных рядов, согласно которой порядок дифферен-

цирования временного ряда d выбирается так, чтобы ряд разностей был стационарным. Параметры p и q выбираются при помощи анализа автокорреляционной функции

$$ACF_\tau = \frac{\sum_{i=1}^{T-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sum_{i=1}^{T}(x_i - \bar{x})^2}, \quad \bar{x} = \frac{1}{T}\sum_{i=1}^{T} x_t$$

(с лагом автокорреляции τ) и частичной автокорреляционной функции

$$PACF_\tau = \begin{cases} \mathrm{corr}(x_{t+1}, x_t), & \tau = 1; \\ \mathrm{corr}(x_{t+\tau} - x_{t+\tau}^{\tau-1}, x_t - x_t^{\tau-1}), & \tau \geq 2, \end{cases}$$

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

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

где $\beta_1, \ldots, \beta_{\tau-1}$ – коэффициенты линейной регрессии. Выбор параметров p и q осуществлялся из следующих соображений:

- в модели $\mathrm{ARIMA}(p, d, 0)$ автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а частичная автокорреляционная функция значимо отличается от нуля при лагах, не больших p;
- в модели $\mathrm{ARIMA}(0, d, q)$ 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).

1.5.8 Анализ качества прогнозов ARIMA+hist

Экспериментальные исследования предложенного алгоритма проводились для трех различных функций потерь:

- квадратичной

$$l(\hat{x}, x_{T+1}) = (x_{T+1} - \hat{x})^2; \tag{18}$$

- абсолютной

$$l(\hat{x}, x_{T+1}) = |x_{T+1} - \hat{x}|; \tag{19}$$

- ассиметричной функции потерь, вычисленной по формуле:

$$l(\hat{x}, x_{T+1}) = \begin{cases} 0{,}5|x_{T+1} - \hat{x}|, & x_{T+1} \leq \hat{x}; \\ 2|x_{T+1} - \hat{x}|, & x_{T+1} > \hat{x}. \end{cases} \tag{20}$$

Все три функции выпуклые, достигают минимума при совпадении прогноза и действительного значения временного ряда. Первые две функции (20) и (21) симметричные, последняя (22) – несимметричная кусочно-линейная функция. Графики этих функций изображены на рисунке 1.38.

Ниже описано применение предложенного алгоритма к прогнозированию временного ряда цен на сахар, изображенного на рисунке 1.39. Для прогнозирования нестационарной части была выбрана модель $\mathrm{ARIMA}(1,0,0)$, остатки которой затем были спрогнозированы с помощью алгоритма hist.

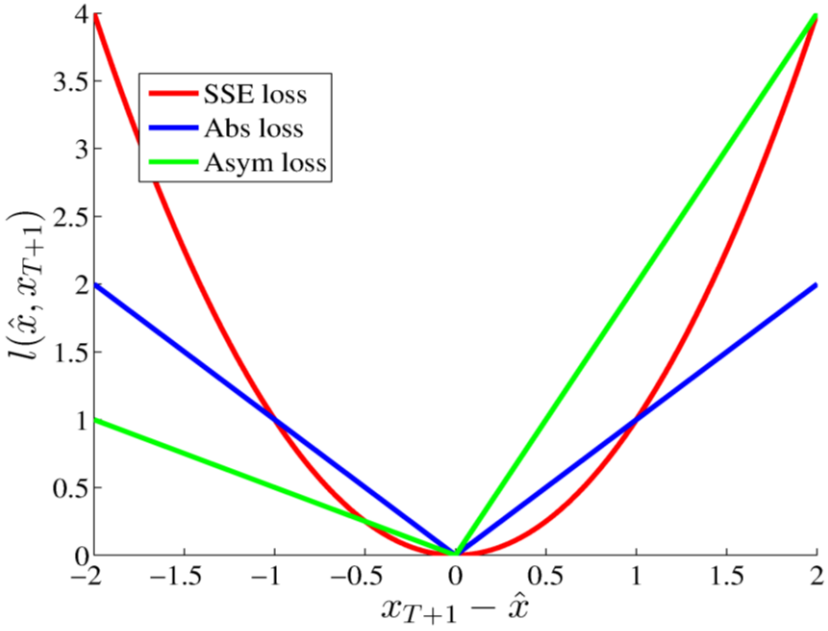

Рисунок 1.38 – Рассматриваемые функции потерь

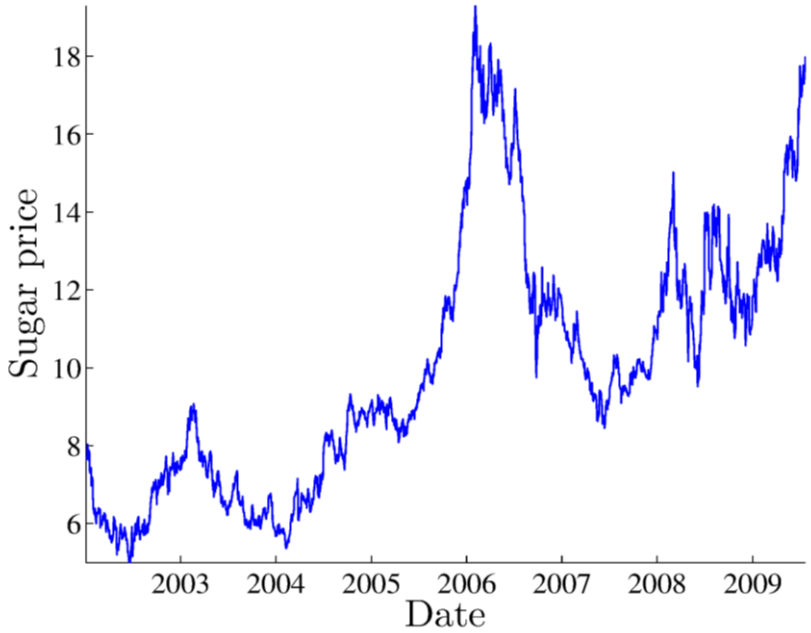

Рисунок 1.39 – Нестационарный временной ряд: цены на сахар.

На рисунках 1.40–1.42 представлены графики зависимости прогноза алгоритма hist от количества столбцов в гистограмме для каждой функции потерь (20)–(22).

На каждом графике по оси абсцисс отложено количество столбцов гистограммы, по оси ординат – прогноз, полученный алгоритмом hist при использовании заданной

функции потерь и гистограммы с заданным количеством столбцов. Видно, что для всех временных рядов и любой функции потерь с увеличением количества столбцов гистограммы полученные прогнозы стабилизируются вокруг предельного значения.

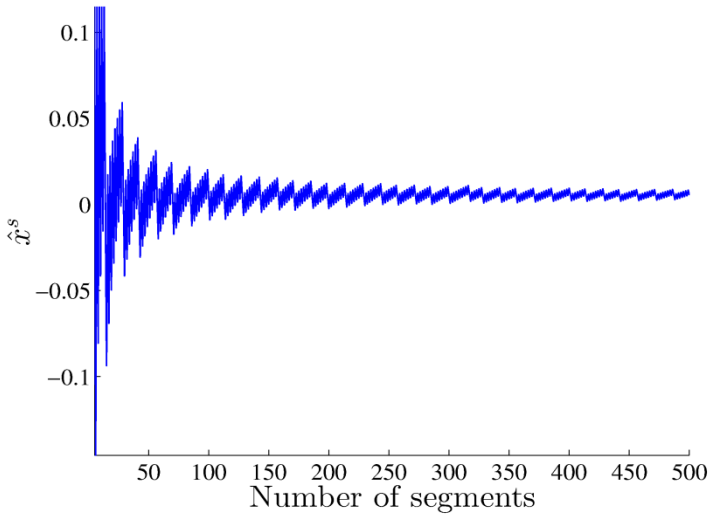

Рисунок 1.40 – Прогнозы алгоритма hist для регрессионных остатков ряда Sugar price для квадратичной функции потерь.

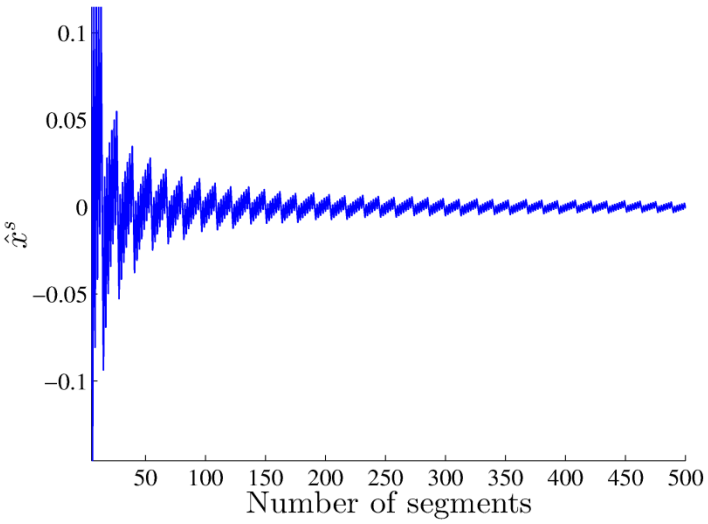

Рисунок 1.41 – Прогнозы алгоритма hist для регрессионных остатков ряда Sugar price для абсолютной функции потерь.

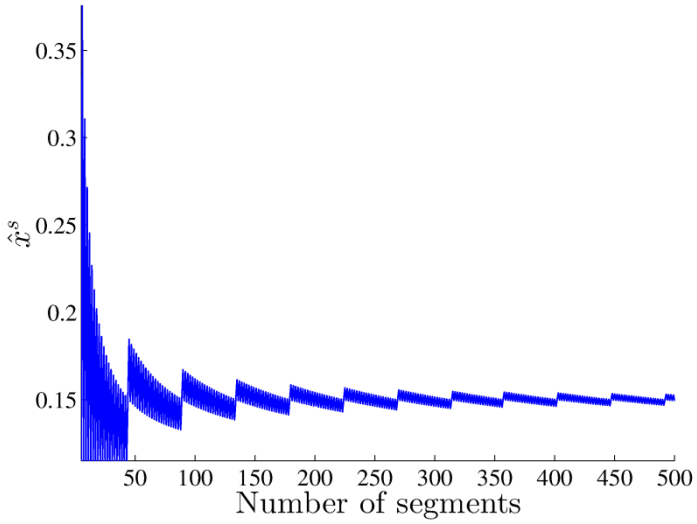

Рисунок 1.42 – Прогнозы алгоритма hist для регрессионных остатков ряда Sugar price для асимметричной функции потерь

Для симметричных функций потерь (20), (21) 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 (22) 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.

Для сравнения качества прогнозов модели ARIMA и связки алгоритмов 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) обучалась выбранная для временного ряда модель ARIMA, затем для обученной модели вычислялся ряд регрессионных остатков. По ряду регрессионных остатков обучался алгоритм 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 (20)–(22) и вариантов алгоритма hist с 20, 50, 300, 500 столбцами в гистограмме. Средние потери для каждой функции потерь приведены для всех вариантов алгоритма в таблице 1.6.

Таблица 1.6 – Средние потери прогнозирования для временного ряда Sugar price.

Алгоритм	Квадратичная функция потерь	Абсолютная функция потерь	Асимметричная функция потерь
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

Как видно из таблицы 1.6, 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.

1.5.9 Быстродействие алгоритма ARIMA+hist

По предварительной грубой оценке сверху вычислительная сложность алгоритма ARIMA+hist прогнозирования объемов спроса на грузовые железнодорожные перевозки не должна превышать полинома степени 3, как по длине прогнозируемого временного ряда, так и по количеству учитываемых временных рядов.

2 Разработка макета модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки на базе математического пакета MatLab

В этом разделе в соответствии с пп. 2.6 и 3.10 Технического задания представлены результаты разработки макета модуля прогнозирования объемов спроса на ГЖДП, выполненной с использованием результатов, проведенных в предыдущем разделе исследований по разработке математической модели прогнозирования объемов спроса на ГЖДП, учитывающей влияние экзогенных факторов на объемы этого спроса, а также специфику бизнес-процессов и нормативов индустриального партнера.

2.1 Назначение и основные функции макета модуля прогнозирования

Основной задачей модуля прогнозирования является построение прогноза объемов спроса на грузовые железнодорожные перевозки по всем парам станций и по парам регионов РЖД в различных временных масштабах (по дням, по неделям и по месяцам).

В рамках выполняемого ПНИ макете модуля прогнозирования реализован двухэтапный алгоритм прогнозирования временных рядов ARIMA+hist, описанный в пункте 1.5.7 предыдущего раздела. Алгоритм позволяет строить прогнозы для нестационарных рядов, а также использует различные функции потерь. При этом в соответствии с п. 4.1.2 Технического задания, этот алгоритм имеет не более чем полиномиальную оценку вычислительной сложности по длине временного ряда и количеству временных рядов и способен обрабатывать не менее 20 различных типов грузов за промежуток времени до двух лет.

Для удовлетворения требованиям, указанным в п. 4.1.1.5 Технического задания, в разработанном модуле прогнозирования реализованы следующие возможности:

- использование в разработанных алгоритмах как исторических данных по объемам спроса на ГЖДП, так и исторических значений экзогенных факторов для определения оптимального вида модели прогнозирования и ее параметров;
- учёт результатов экспертного анализа значимости и характера влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки;

- учёт в структуре входных и выходных потоков данных специфики индустриального партнера, в частности процессов и нормативов в области ГЖДП;
- удовлетворение физическим ограничениям, связанным с прикладной спецификой задачи: получаемые прогнозы являются неотрицательными и не превышают пропускную способность сети РЖД;
- формирование прогнозов для различных временных периодов: посуточно и подекадно, на месяц, на квартал помесячно, на период от года помесячно и поквартально;
- формирование прогнозов с разложением по группам грузов;
- выбор для каждого конкретного варианта прогноза алгоритма, наиболее подходящего согласно заданному критерию точности, с учетом:

a. пространственной иерархии: в целом по сети РЖД, в разрезе групп грузов, в виде «шахматки» станция–станция, в виде «шахматки» регион–регион,

b. временной иерархии: дневной, недельный и месячный горизонты прогнозирования.

Разработанный макет модуля реализует следующие основные функции:

- загрузка данных грузоперевозок, реальных или генерированных;
- запуск алгоритма прогнозирования на загруженных наборах данных;
- построение графиков результатов прогнозирования.

2.2 Функциональная архитектура макета модуля прогнозирования

Макет модуля прогнозирования, выполнен на базе математического пакета MatLab версии R2014b или старше в операционной среде Windows 7 или старше.

Макет модуля прогнозирования реализован в виде независимых расчетных модулей, с фиксированными и описанными интерфейсами вызова, форматами входных и выходных данных. Обмен данными между функциональными модулями реализован на базе внутренних вызовов, согласно стандартам используемого математического пакета MatLab.

Основным модулем, осуществляющим прогнозирование на наборе данных, является MatLab-функция SingleDataForecast.m. На вход эта функция получает двухуровневый алгоритм прогнозирования, описанный в п. 1.5.7 и вызывает этот алгоритм с помощью MatLab-функции `feval`:

```
frc = feval(@VarForecast, forecast_query, data, ex_factors, par);
```

где

@VarForecast – указатель на функцию, реализующую двухуровневый алгоритм прогнозирования ARIMA+hist;

forecast_query – запрос на прогнозирование (вырабатывается по входным данным num_days_test и par;

data – исторические (или модельные) данные объёмов грузоперевозок;

ex_factors – исторические (или модельные) данные о значениях экзогенных факторов и характере их влияния на объёмы спроса на грузоперевозки,

par – набор установочных параметров алгоритма прогнозирования (включает в себя длину предыстории и максимальное количество используемых в модели экзогенных факторов);

frc – возвращаемое значение прогноза спроса на грузоперевозки.

Основной модуль прогнозирования в ходе своей работы использует следующие входные данные:

num_days_test – горизонт прогнозирования в днях;

par – параметры функции прогнозирования,

data – массив исторических (или сгенерированных программой ВЦРАН-58.29.29/mdgm-01 модельных) временных рядов объёмов грузоперевозок;

ex_factors – массив исторических (или сгенерированных программой ВЦРАН-58.29.29/mdgm-01 модельных) временных рядов экзогенных факторов, влияющих на объёмы спроса на грузоперевозки.

В качестве выходных аргументов модудь прогнозирования возвращает:

frc – массив прогноза спроса на объёмы грузоперевозок для заданного горизонта прогнозирования,

errors – массив оценок точности прогнозирования.

Техническое описание макета модуля прогнозирования и текст программы на исходном языке приведены в двух отдельных программных документах отчёта за текущий этап:

- ВЦРАН-58.29.29/forecast-01-13-01 «Макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки. Описание программы»;
- ВЦРАН-58.29.29/forecast-01-12-01 «Макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки. Текст программы».

3 Разработка программы и методики тестирования макета модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки

В этом разделе в соответствии с пп. 2.7 и 3.11 Технического задания представлены результаты разработки программы и методики тестирования макета модуля прогнозирования объемов спроса на ГЖДП, описанного в предыдущем разделе.

Целью проведения тестирования является проверка макета модуля прогнозирования на предмет соответствия разработанной программы требованиям ТЗ.

Тестирование и проверка выполняются по следующим позициям:

1) комплектность программной документации,

2) комплектность и состав технических и программных средств,

3) выполнение функций программы,

4) быстродействие программы.

Проверка комплектности программной документации на программное изделие производится визуально представителем службы, ответственной за эксплуатацию. В ходе проверки сопоставляется состав и комплектность программной документации. Проверка считается завершенной в случае соответствия состава и комплектности программной документации, представленной разработчиком, перечню предусмотренной Техническим заданием программной документации.

Проверка комплектности и состава технических и программных средств осуществляется визуальной проверкой наличия всех, предусмотренных технической документацией файлов, необходимых для запуска и исполнения программы.

В папке tests должны находиться следующие файлы:

- SingleDataForecast.m – модуль прогнозирования на одном наборе данных,
- TestAlgoSimulatedData.m – модуль прогнозирования на нескольких наборах данных, в частности, для различных генераций данных.

В папке frc должна находиться функция VarForecast.m, реализующая метод прогнозирования для прогнозирования.

3.1 Проверка выполнения функций

Для тестирования основного модуля прогнозирования TestAlgoSingleData.m в системе MatLab реализован ряд тестов, использующих xUnit Test Framework.

Запуск тестов в среде MatLab выполняется скриптом startUnitTests.m, располагающимся в той же директории, что и функция TestAlgoSingleData.m. При успешном выполнении всех тестов должно появиться сообщение:

«PASSED in XX seconds» (XX – количество секунд).

Реализованы следующие виды тестирования:

1. Тестирование со стандартными параметрами.

а) Для тестирования модуля прогнозирования на наборе данных производится сравнение прогноза, возвращаемого функцией TestAlgoSingleData.m, с ранее полученным прогнозом на том же наборе данных.

б) Набор данных для тестирования генерируется модулем генерации данных (см. раздел 5) со следующими параметрами:

```
sim_par.goods_code = [3, 1]; %              – коды товаров
date_generate_from = '01.01.2015'; %        – начальная дата генерации
date_generate_to = '30.08.2015'; %          – конечная дата генерации
num_days_test = 60; %                       – горизонт прогнозирования
par.max_weights_noise = 0.5; %              – параметр шума
par.num_production_regions = 2; %           – количество регионов производства
                                              товара
par.num_stations_in_branch_pairs = 5; %     – количество пар станций перевозок
                                              товара
par.time_lag = 10; %                        – временная задержка влияния экзо
                                              генных факторов
```

в) Выполнение автоматического тестирования функции прогнозирования VarForecast.m с различными значениями длины предыстории: 5, 10, 20 и 50.

г) Для сравнения полученного прогноза с эталонным прогнозом вычисляется норма разности двух векторов прогноза. Тест считается пройденным успешно, если значение нормы разности не превосходит 10^{-9}.

д) Модуль test_algosingledata_normal.m осуществляет сравнение полученного прогноза с эталонным прогнозом. Внутри себя этот модуль запускает функции

test_var_5, test_var_10, test_var_20, test_var_50 тестирования алгоритма со значениями длины предыстории, соответственно, 5, 10, 20 и 50.

2) Тестирование с вырожденными параметрами.

а) Тестирование модуля прогнозирования на нулевых данных осуществляет модуль test_algosingledata_zero. При этом сравнивается норма двух векторов – вектора прогноза и вектора, состоящего полностью из нулей.

б) Тестирование модуля прогнозирования на нечисловых данных осуществляет модуль test_algosingledata_nan. Тест проверяет совпадение возвращаемого значения со значением NaN (Not a Number).

Результаты тестирования функциональности макета модуля прогнозирования по методике, описанной в предыдущем разделе, заносятся в таблицу 3.1.

Таблица 3.1 – Результаты тестирования функциональности макета модуля прогнозирования.

№№	Описание процедуры	Критерий оценки	Значение успеха	Результат
1	*2*	*3*	*4*	*5*
1	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 5	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = пройден успешно/ не пройден (ненужное зачеркнуть)
2	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 10	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = пройден успешно/ не пройден (ненужное зачеркнуть)
3	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 20	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = пройден успешно/ не пройден (ненужное зачеркнуть)
4	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 50	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = пройден успешно/ не пройден (ненужное зачеркнуть)

Таблица 3.1 – Продолжение.

1	*2*	*3*	*4*	*5*
5	Тестирование модуля прогнозирования на нулевых данных	k – значение нормы разности между полученным прогнозом и нулевым вектором	$k \leq 10^{-9}$	max k = пройден успешно/ не пройден (ненужное зачеркнуть)
6	Тестирование модуля прогнозирования на данных нечислового формата	D – значение прогноза	D is NaN	D is пройден успешно/ не пройден (ненужное зачеркнуть)

3.2 Проверка быстродействия

Согласно п. 4.1.3 Технического задания, алгоритм прогнозирования объемов спроса должен иметь не более чем полиномиальную оценку вычислительной сложности по длине временного ряда и количеству временных рядов. По грубой оценке сверху, приведённой а пункте 1.5.9 этого отчёта, степень полинома не должна превышать 2,5.

Для проверки быстродействия функции TestAlgoSingleData.m в системе реализован скрипт startTimeTests.m. Этот скрипт возвращает два графика:

- график зависимости времени выполнения от длины временного ряда (зависимости от длины предыстории) с отображением дополнительно для оценки соответствия требованиям ТЗ графика (красным пунктиром) линейной аппроксимации данных быстродействия;
- график зависимости времени выполнения от количества учитываемых при прогнозировании временных рядов с отображением дополнительно для оценки соответствия требованиям ТЗ графика (красным пунктиром) аппроксимации данных быстродействия квадратичной параболой.

Техническое описание программы и методики тестирования макета модуля прогнозирования приведено в отдельном программном документе отчёта за текущий этап – ВЦРАН-58.29.29/forecast-01-51-01 «Макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки. Программа и методика тестирования».

4 Тестирование макета модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки

Раздел выполнен в соответствии с пп. 2.8, 3.12 и 6.1.3.6 Технического задания. Целью проведения испытаний была проверка соответствия функциональных и технических характеристик разработанной программы требованиям Технического задания.

Результаты тестирования функциональности макета модуля прогнозирования приведены в таблице 3.1.

Таблица 4.1 – Результаты тестирования функциональности макета модуля прогнозирования.

№№	Описание процедуры	Критерий оценки	Значение успеха	Результат
1	*2*	*3*	*4*	*5*
1	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 5	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = 0 пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
2	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 10	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = 0 пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
3	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 20	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = 0 пройдено успешно/ ~~не пройден~~ (ненужное зачеркнуть)
4	Тестирование модуля прогнозирования на стандартном наборе данных с параметром длины предыстории, равным 50	k – значение нормы разности между полученным и эталонным значением прогноза	$k \leq 10^{-9}$	k = 0 пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
5	Тестирование модуля прогнозирования на нулевых данных	k – значение нормы разности между полученным прогнозом и нулевым вектором	$k \leq 10^{-9}$	max k = 0 пройдено успешно/ ~~не пройден~~ (ненужное зачеркнуть)

Таблица 4.1 – Продолжение.

1	*2*	*3*	*4*	*5*
6	Тестирование модуля прогнозирования на данных нечислового формата	D – значение прогноза	D is NaN	D is NaN пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)

Результаты проверки быстродействия макета модуля прогнозирования, отражены на рисунках 4.1 и 4.2. Синей линией показано достигнутое в ходе тестирования время выполнения, красным пунктиром – аппроксимация прямой и параболой (на рисунках 4.1 и 3.2, соответственно).

Построенные аппроксимации позволяют сделать вывод о том, что вычислительная сложность разработанного алгоритма является субквадратичной по длине предыстории и квадратичной по количеству временных рядов.

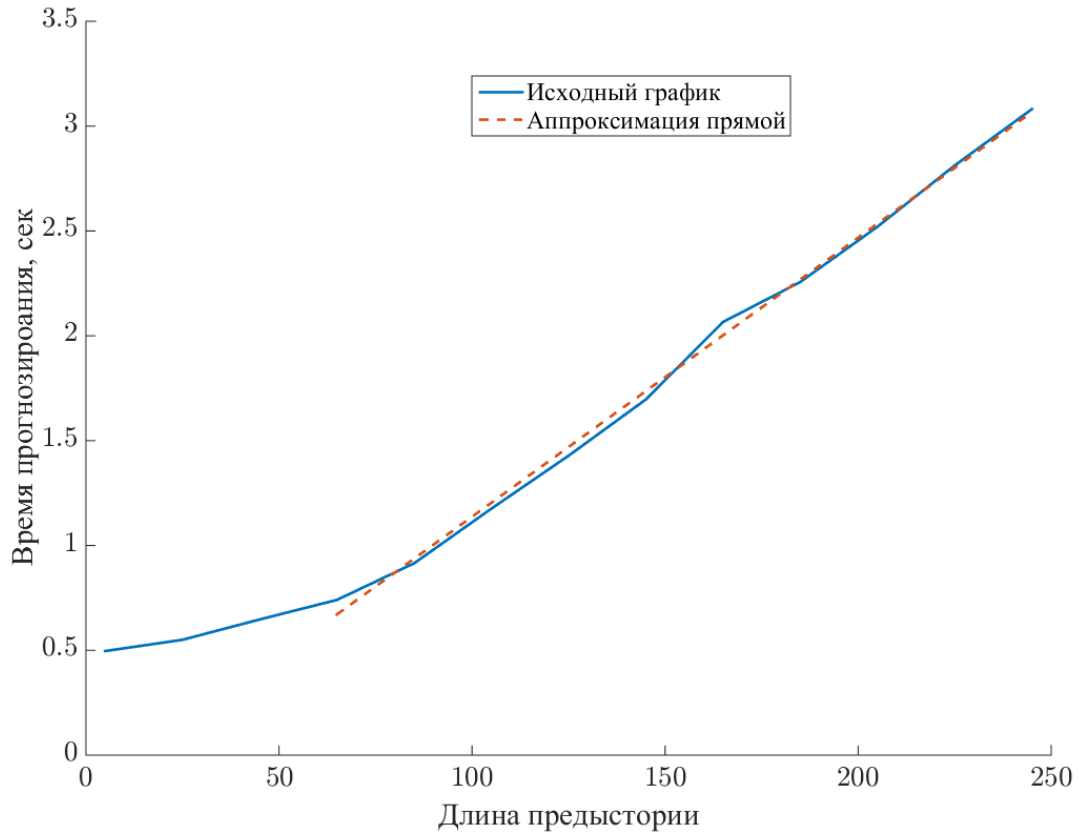

Рисунок 4.1 – Время работы модуля прогнозирования в зависимости от длины предыстории.

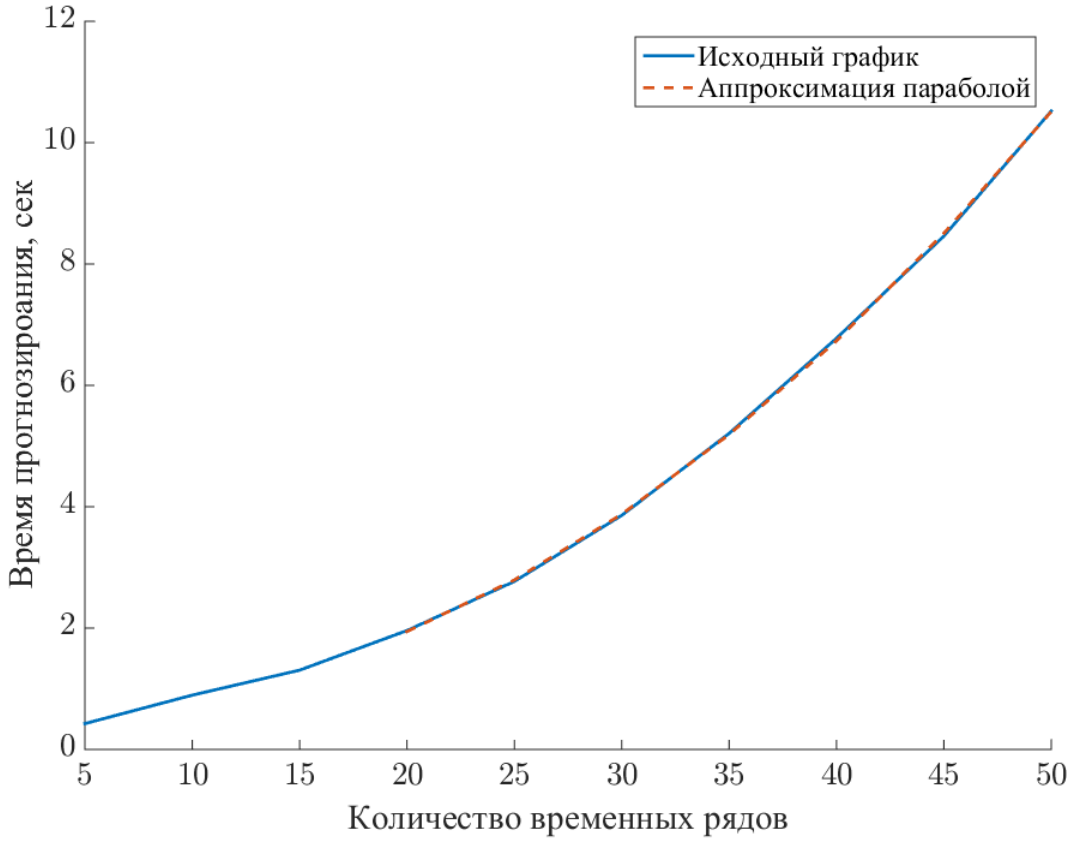

Рисунок 4.2 – Время работы модуля прогнозирования в зависимости от количества временных рядов.

5 Разработка генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов

В этом разделе в соответствии с пп. 2.3 и 3.7 Технического задания представлены результаты разработки генератора модельных исходных данных объемов спроса на ГЖДП и влияющих на них экзогенных факторов.

5.1 Назначение и основные функции генератора модельных исходных данных

Основное назначение программы состоит в генерации в стандартизованном виде модельных данных объемов спроса на ГЖДП и влияющих на них экзогенных факторов, согласующихся с историческими данными о грузоперевозках по парам веток и существующей топологией железнодорожной структурой сети РЖД.

В процессе генерации используются следующие «внешние» данные:

1) топология станций с реальными кодами станций и их принадлежностью к регионам;

2) информация о парах станций, реально используемых для перевозки того или иного товара;

3) экзогенные факторы (цена на основные инструменты и пр.), влияющие согласно экспертным оценкам на совокупный спрос на соответствующие товары;

4) исторические данные о небольшой области РФ для экстраполяции структуры перевозок на все области;

5) другие экспертные требования, например, устойчивость структуры перевозок во времени.

Форматы выходных модельных данных полностью соответствуют форматам используемых для генерации реальных данных.

5.2 Методы генерации модельных данных

Процесс генерации модельных данных включает три последовательных этапа:

- генерация совокупного спроса товара;
- генерация графа грузоперевозок;
- генерация данных для каждого момента времени из заданного диапазона в зависимости от совокупного спроса и графа грузоперевозок.

5.2.1 Генерация совокупного спроса товара

Генерация совокупного спроса на тот или иной вид товара осуществляется в форме одного временного ряда для каждого из заданного набора видов товара. В каждый момент времени значению временного ряда присваивается значение, равное среднему суммарному количеству этого вида товара, перевозимому по всей железнодорожной сети.

При генерации совокупного спроса используются следующие исторические данные:

- среднее значение совокупного спроса,
- исторические временные ряды экзогенных факторов с таблицей показателей их влияния на переменные совокупного спроса на соответствующие товары.
- временной лаг влияния экзогенных факторов.

Совокупный спрос рассчитывается в виде зашумленной линейной комбинации средних значений экзогенных факторов и временного лага их влияния на объёмы спроса на грузоперевозки. Результат генерации:

- сгенерированный модельный временной ряд среднего совокупного количества перевозимого товара;
- массив сгенерированных модельных временных рядов экзогенных факторов, влияющих на объёмы грузоперевозок

5.2.2 Генерация графа грузоперевозок

Генерация графа грузоперевозок осуществляется в виде статичного разреженного взвешенного ориентированного графа грузоперевозок по железнодорожной сети для каждого вида товара. Вершинами этого графа являются все станции железнодорожной сети, а весами ребер – доли совокупного объёма товара рассматриваемого вида, перевозимого между соответствующей парой станций. Отсутствие ребра означает отсутствию перевозок этого вида товара между соответствующей парой станций в любой момент времени.

При построении графа использует следующие реальные данные:

- коды товаров;
- существующая топология станций РЖД;
- количество регионов производства товара,

- количество пар станций перевозок внутри каждой пары регионов.

Результат генерации:

- модельный разреженный орграф сети грузоперевозок.

5.2.3 Генерация модельных данных

Генерация модельных данных объёмов спроса на грузоперевозки каждого вида товара осуществляется распределением сгенерированных данных о совокупном спросе по парам станций из сгенерированного графа грузоперевозок. При этом в целях обеспечения вариаций весов рёбер графа грузоперевозок во времени в каждый момент времени выполняется «зашумление» весов по нормальному закону распределения с соответствующей дисперсией, обеспечивающей общую нормировку весов на единицу.

5.3 Функциональная архитектура генератора модельных исходных данных

Генератор модельных исходных данных выполнен на базе математического пакета MatLab версии R2014b или старше в операционной среде Windows 7 или старше и поддерживает следующие функции:

- загрузка исторических данных объёмов грузоперевозок и реальной топологии сети РЖД для их анализа и извлечения из них информации, необходимой для генерации модельных данных;
- генерация экзогенных факторов на основе исторической информации о влияющих на объёмы ГЖДП экзогенных факторах;
- генерация одномерного временного ряда совокупного спроса для каждого товара с использованием сгенерированных экзогенных факторов;
- генерация разреженного взвешенного графа сети ГЖДП;
- генерация временного ряда объёмов спроса на ГЖДП с учетом совокупного спроса и топологии сети грузоперевозок.

Генератор реализован в виде независимых расчетных модулей с фиксированными и описанными интерфейсами вызова, форматами входных и выходных данных. Обмен данными между функциональными модулями реализован на базе внутренних вызовов, согласно стандартам используемого математического пакета MatLab.

Запускающий модуль программы, осуществляющий управление загрузкой исторических данных грузоперевозок и реальной топологии сети, их анализом и извлечением необходимой для последующей генерации модельных данных – MatLab-функция GenerateData.m.

Помимо модуля GenerateData.m в состав программы входят четыре функциональных модуля – MatLab-функции:

- GenerateTotalVolume.m – генерация совокупного спроса на товар;
- GenerateTransportationGraph – генерация разреженного взвешенного графа сети грузоперевозок;
- GenerateExFactors.m – генерация экзогенных факторов на основе исторической информации о факторах;
- GenerateGraphData.m – генерация одномерного временного ряда совокупного спроса для каждого товара с использованием сгенерированных экзогенных факторов.

Генератор в ходе своей работы использует следующие входные данные, являющиеся параметрами функции генерации данных GenerateData.m:

goods_code – коды товаров,

date_generate_from – начальная дата генерации;

date_generate_to – конечная дата генерации;

par – параметры генерации:

- par.num_generations – количество генераций,
- par.date_generate_from – начальная дата генерации,
- par.date_generate_to – конечная дата генерации,
- par.num_days_test – количество дней генерации,
- par.max_weights_noise – параметр шума структуры грузоперевозок,
- par.num_production_regions – количество регионов производства товара,
- par.num_stations_in_branch_pairs – количество станций поставок товара в регионе,
- par.time_lag – параметр, характеризующий временной лаг влияния экзогенных факторов;

mean_volume – исторические данные о средних значениях совокупного спроса на товар;

prices.mat – исторические временные ряды экзогенных факторов с характером их влияния на объёмы спроса;

stations.mat – список кодов станций РЖД с регионами.

В качестве выходных данных генератор возвращает в формате, совпадающем с использованными реальными данными:

`data` – массив сгенерированных модельных временных рядов грузоперевозок;

`ex_factors` – массив сгенерированных модельных временных рядов экзогенных факторов, влияющих на объёмы грузоперевозок.

Техническое описание генератора модельных исходных данных и текст программы на исходном языке приведены в двух отдельных программных документах отчёта за текущий этап:

- ВЦРАН-58.29.29/mdgm-01-13-01 «Генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов. Описание программы»;
- ВЦРАН-58.29.29/mdgm-01-12-01 «Генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов. Текст программы».

6 Разработка программы и методики тестирования генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов

В этом разделе в соответствии с пп. 2.4 и 3.8 Технического задания представлены результаты разработки программы и методики тестирования генератора модельных исходных данных объемов спроса на ГЖДП и значений влияющих на них экзогенных факторов.

Целью тестирования является проверка работоспособности макета генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки, учитывающего экзогенные факторы и соответствие разработанной программы техническим требованиям к научно-техническим результатам ПНИ по ТЗ.

Тестирование и проверка выполняются по следующим позициям:

1) комплектность программной документации,

2) комплектность и состав технических и программных средств,

3) выполнение функций программы.

Проверка комплектности программной документации на программное изделие производится визуально представителем службы, ответственной за эксплуатацию. В ходе проверки сопоставляется состав и комплектность программной документации. Проверка считается завершенной в случае соответствия состава и комплектности программной документации, представленной разработчиком, перечню предусмотренной Техническим заданием программной документации.

Проверка комплектности и состав технических и программных средств осуществляется визуальной проверкой наличия всех, предусмотренных технической документацией файлов, необходимых для запуска и исполнения программы.

В директории data должны находиться следующие файлы:

- GenerateData.m – функция, осуществляющая генерацию данных;
- GenerateTotalVolume.m – генерация совокупного спроса на товар;
- GenerateTransportationGraph – генерация разреженного взвешенного графа сети грузоперевозок;
- GenerateExFactors.m – генерация экзогенных факторов на основе исторической информации о факторах;
- GenerateGraphData.m – генерация одномерного временного ряда совокуп-

ного спроса для каждого товара с использованием сгенерированных экзогенных факторов.

- stations.mat – mat-файл, содержащий полный список железнодорожных станций. Данные хранятся в массиве cell array;
- ParseStations.m – функция, осуществляющая загрузку файла stations.mat;
- prices.mat – mat-файл, содержащий реальные экзогенные факторы (цены на основные инструменты).

6.1 Проверка выполнения функций

Для тестирования основной функции GenerateData.m, осуществляющей генерацию данных грузоперевозок за заданный временной период в системе Matlab реализован ряд тестов, использующих xUnit Test Framework.

Запуск тестов в среде Matlab выполняется скриптом startUnitTests.m, располагающимся в той же директории, что и функция GenerateData.m. При успешном выполнении всех тестов должно появиться сообщение:

«PASSED in XX seconds» (XX – количество секунд).

Реализованы следующие виды тестирования:

1. Тестирование со стандартными параметрами.

а) Для тестирования модуля генерации на стандартном наборе данных производится сравнение результата – сгенерированных модельных данных, возвращаемых функцией GenerateData.m, с эталонным набором данных, полученных в результате выполненной ранее генерации данных с тем же набором параметров.

б) Для сравнения полученного результата с эталонным набором данных выполняется тест «три сигма». Тест заключается в подсчете статистики – максимума абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда. Если значение получившейся статистики не превосходит 3, то тест считается пройденным успешно.

в) Для тестирования предварительно сгенерировано три эталонных временных ряда, соответствующих трем различным кодам грузов: goods1, goods2 и goods3. Функции, выполняющие сравнение с эталонными временными рядами для этих грузов, имеют соответствующие названия: test_3sigma_goods1, test_3sigma_goods2 и test_3sigma_goods3.

2) Тестирование с вырожденными параметрами.

а) Тестирование входных параметров NaN (Not a Number), а также входных параметров Inf («бесконечность») осуществляет функция test_generatedata_nan_inf.m. Тест проверяет совпадение возвращаемого значения со значением NaN.

б) Тестирование отрицательных входных параметров осуществляет функция test_generatedata_nan_negative.m. Тест проверяет совпадение возвращаемого значения со значением NaN.

в) Тестирование неправильного формата данных осуществляет функция test_generatedata_wrongtype.m. На вход этой функции подаются два вида значений: (1) значение ts_date_from > ts_date_to (дата начала генерации позже даты конца генерации) и (2) дата неправильного формата.

Результаты функционального тестирования генератора модельных отражаются в таблице 6.1.

Таблица 6.1 – Результаты тестирования функциональности генератора модельных исходных данных.

№№	Описание процедуры	Критерий оценки	Значение успеха	Результат
1	*2*	*3*	*4*	*5*
1	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods1	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = пройден успешно/ не пройден (ненужное зачеркнуть)
2	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods2	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = пройден успешно/ не пройден (ненужное зачеркнуть)

Таблица 6.1 – Продолжение.

1	*2*	*3*	*4*	*5*
3	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods3	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = пройден успешно/ не пройден (ненужное зачеркнуть)
4	Тестирование нулевых значений входных параметров	D – выходные данные	D is NaN	D is пройден успешно/ не пройден (ненужное зачеркнуть)
5	Тестирование бесконечных значений входных параметров	D – выходные данные	D is NaN	D is пройден успешно/ не пройден (ненужное зачеркнуть)
6	Тестирование отрицательных значений входных параметров	D – выходные данные	D is NaN	D is пройден успешно/ не пройден (ненужное зачеркнуть)
7	Тестирование неправильного формата данных	D – выходные данные	D is NaN	D is пройден успешно/ не пройден (ненужное зачеркнуть)

Техническое описание программы и методики тестирования модуля генерации модельных данных приведены в отдельном программном документе отчёта за текущий этап – «Генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов. Программа и методика тестирования».

7 Тестирование генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов по ПМИ генератора модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов

Раздел выполнен в соответствии с п. 2.5, 3.9 и 6.1.3.4 Технического задания. Целью проведения испытаний является проверка соответствия характеристик разработанной программы функциональным требованиям, изложенным в Техническом задании.

Результаты тестирования генератора модельных исходных данных по методике, описанной в предыдущем разделе, представлены в таблице 4.1.

Таблица 7.1 – Результаты тестирования генератора модельных исходных данных.

№№	Описание процедуры	Критерий оценки	Значение успеха	Результат
1	*2*	*3*	*4*	*5*
1	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods1	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = 1,41 пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
2	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods2	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = 1,19, пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
3	Тестирование на стандартном наборе параметров: сравнение с эталоном. Тестирование для груза goods3	k – максимум абсолютной разности между двумя временными рядами, нормированной на среднеквадратичное отклонение эталонного временного ряда	$k \leq 3$	k = 1,56, пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)

Таблица 7.1 – Продолжение.

1	*2*	*3*	*4*	*5*
4	Тестирование нулевых значений входных параметров	D – выходные данные	D is NaN	D is NaN пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
5	Тестирование бесконечных значений входных параметров	D – выходные данные	D is NaN	D is NaN пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
6	Тестирование отрицательных значений входных параметров	D – выходные данные	D is NaN	D is NaN пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)
7	Тестирование неправильного формата данных	D – выходные данные	D is NaN	D is NaN пройден успешно/ ~~не пройден~~ (ненужное зачеркнуть)

ЗАКЛЮЧЕНИЕ

1. Разработана непараметрическая модель прогнозирования объемов спроса на грузовые железнодорожные перевозки на железнодорожных узлах РЖД. Предложена модификация модели с ядерной оценкой плотности прогнозируемого временного ряда. Исследованы свойства непараметрической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, а также зависимость качества модифицированной модели от параметров ядерной оценки.

2. Разработан и протестирован алгоритм построения гистограммы распределения значений объема спроса и вычисления свертки гистограммы с экспертно заданной функцией потерь, исследованы свойства алгоритма. Для обеспечения оптимального качества прогнозирования в условиях ограниченных объемов даннных, разработан метод определения оптимальной длины предыстории временных рядов экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки.

3. В связи с решением учитывать при прогнозировании экзогенные временные ряды, разработан и протестирован алгоритм взвешенного суммирования гистограмм распределения значений объемов спроса на грузовые железнодорожные перевозки.

4. Проведен анализ качества алгоритмов прогнозирования в применении к нестационарным временным рядам. Описана процедура проверки условия локальной стационарности временного ряда с помощью теста Дики-Фуллера. Разработан и обоснован метод прогнозирования нестационарных временных рядов при ассиметричных функциях потерь.

5. Разработан и протестирован генератор модельных исходных данных объемов спроса на грузовые железнодорожные перевозки и влияющих на эти обёмы экзогенных факторов.

6. Разработан и протестирован макет модуля прогнозирования объемов спроса на грузовые железнодорожные перевозки.

Таким образом, задачи этапа 2, предусмотренные Техническим заданием и Календарным планом проекта решены в полном объёме.

http://www.ccas.ru/confer/conf15-r.htm

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ-ДАТА, 2001.

2. Албегов М. М., Бурса Б. И., Истомина Р.П., Медведев В. Г и др. Краткосрочное прогнозирование регионального развития в условиях неполной информации. Едиториал УРСС, 2001.

3. Терёшина Н.П., Галабурда В.Г., Токарев В.А. и др. Экономика железнодорожного транспорта: Учебник для вузов ж.-д. транспорта. М.: УМЦ ЖДТ, 2008.

4. Гасников А.В. Заметка об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях // Математическое моделирование, arXiv:1410.3123, 2015.

5. 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, сс. 1–51. ВЦ РАН Москва, 2014.

6. Шананин А. А. Вычислимая модель железнодорожных грузоперевозок с учетом коммуникационных ограничений // Тезисы докладов 10-й Международной конференции Интеллектуализация обработки информации. Греция, о. Крит, 2014. СС. 192–192.

7. Постановление Правительства РФ от 18.05.2001 № 384 “О программе структурной реформы на железнодорожном транспорте”.

8. Box G. E. P., Jenkins G. M., Reinsel G. C. Time Series Analysis: Forecasting and Control. Englewood Cliffs, 3rd edition, 1994.

9. Patton A. J. and Timmermann A. Properties of optimal forecasts under asymmetric loss and nonlinearity // Journal of Econometrics, 2007. Vol. 140(2), pp. 884–918.

10. Berk R. Asymmetric loss functions for forecasting in criminal justice settings // Journal of Quantitative Criminology, 2011. Vol. 27(1), pp. 107–123.

11. Cipra T. Asymmetric recursive methods for time series // Applications of Mathematics, 1994. Vol. 39(3), pp. 203–214.

12. Koenker R. and Zhijie Xiao. Quantile autoregression // Journal of the American Statistical Association, 2006. Vol. 101(475), pp. 980–990.

13. Koenker R. Quantile regression. Cambridge university press, 2005.

14. Biau G., Bleakley K., Györfi L., Ottucsák G. Nonparametric sequential prediction of time series // Journal of Nonparametric Statistics, 2010. Vol. 22(3), pp. 297–317.

15. Крянев А. В., Лукин Г. В. Математические методы обработки неопределенных данных. Физматлит М., 2003.

16. Kooi R. P. The optimization of queries in relational databases. PhD thesis, Case Western Reserve University, 1980.

17. Piatetsky-Shapiro G., Connell C.. Accurate estimation of the number of tuples satisfying a condition // ACM SIGMOD Record, 1984. Vol. 14, pp. 256–276.

18. Muralikrishna M., DeWitt D. J.. Equi-depth multidimensional histograms // ACM SIGMOD Record, 1988. Vol. 17, pp. 28–36.

19. Berleant D., Cheng H. A. Software tool for automatically verified operations on intervals and probability distributions // Reliable Computing, 1998. Vol. 4, pp. 71–82.

20. Ioannidis Y. E., Poosala V. Balancing histogram optimality and practicality for query result size estimation // ACM SIGMOD Record, 1995. Vol. 24, pp. 233–244.

21. Chen L., Dobra A. Histograms as statistical estimators for aggregate queries // Information Systems, 2013. Vol. 38, pp. 213–230.

22. Poosala V., Haas P. J., Ioannidis Y. E. and E. J. Shekita. Improved histograms for selectivity estimation of range predicates // ACM SIGMOD Record, 1996. Vol. 25, pp. 294–305.

23. König A. C., Weikum G. Combining histograms and parametric curve fitting for feedback-driven query result-size estimation. In Proceedings of the 25th International // Conference on Very Large Data Bases, 1999. Pp. 423–434.

24. Gunopulos D., Kollios G., Tsotras V. J., and Domeniconi C. Approximating multidimensional aggregate range queries over real attributes // ACM SIGMOD Record, 2000. Vol. 29, pp. 463–474.

25. Bruno N., Chaudhuri S., and Gravano L.. Stholes: a multidimensional workload-aware histogram // ACM SIGMOD Record, 2001. Vol. 30, pp. 211–222.

26. Furtado P., Madeira H. Summary grids: building accurate multidimensional histograms database systems for advanced applications // 6th International Conference on Database Systems for Advanced Applications, 1999. Pp. 187–194.

27. Jagadish H., Koudas N., Muthukrishnan S. et al. Optimal histograms with quality guarantees // VLDB ’98 Proceedings of the 24rd International Conference on Very Large Data Bases, 1998. Vol. 98, pp. 275–286.

28. Guha S., Indyk P., Muthukrishnan S., and Strauss M. Histogramming data streams with fast per-item // Processing Automata, Languages and Programming, 2002. Vol. 2380, pp. 681–692.

29. Muthukrishnan S., Poosala V., Suel T. On rectangular partitionings in two dimensions: algorithms, complexity and application // Lecture Notes in Computer Science, 1999. Vol. 1540, pp 236–256.

30. Thaper S., Guha N., Indyk P. and Koudas N. Dynamic multidimensional histograms // Proceedings of the 2002 ACM SIGMOD international conference on Management of data, ACM, 2002. Pp. 428–439.

31. Whittle P. Prediction and Regulation by Linear Least-Square Methods. University of Minnesota Press, 1983.

32. Roopaei M., Zolghadri M., Emadi A. Economical forecasting by exogenous variables // IEEE International Conference on Fuzzy Systems, 2008. Pp. 1491–1495.

33. De Gooijer J. G., Hyndman R. J. 25 years of time series forecasting. International // Journal of Forecasting, 2006. Vol. 22, pp. 443 –473.

34. Peña D., Sánchez I. Multifold predictive validation in armax time series models // Journal of the American Statistical Association, 2005. Vol. 100, pp. 135–146.

35. Syrovátka P., Grega L. Analysis of methodological approaches to evaluation of complementary and substitution relationship in consumer demand for food // Agricultural economics, 2002. Vol. 48(10), pp. 456–462.

36. Kumar M., Patel N. R., Woo J. Clustering seasonality patterns in the presence of errors // KDD, ACM, 2002. Pp. 557–563.

37. Kumar M., Patel N. R. Clustering data with measurement errors. Computational Statistics & Data Analysis, 2007. Vol. 51(12), pp. 6084–6101.

38. Стенина М.М., Стрижов В.В. Согласование агрегированных и детализированных прогнозов при решении задач непараметрического прогнозирования //Системы и средства информатики, 2014. Т. 24, № 2, СС. 21–34.

39. Van Erven T., Cugliari J. Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. Unpublished, available at: https://hal.inria.fr/hal-00920559, 2013 (accessed October, 2014).

40. Raftery A.E. A model for high-order markov chains // Journal of the Royal Statistical Society: Series B, 1985. Vol. 47(3), pp. 528–539.

41. Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition // Proceedings of the IEEE, 1989. Vol. 77(2), pp. 257 – 286, 1989.

42. Wong C.S. and Li W.K. On a mixture autoregressive conditional heteroscedastic model // Journal of the American Statistical Association, 2001. Vol. 96, pp. 982–995.

43. Berchtold A. Mixture transition distribution (mtd) modeling of heteroscedastic time series // Computational Statistics & Data Analysis, 2003. Vol. 41, pp. 399–411.

44. Bartolucci F., Farcomeni A. A note on the mixture transition distribution and hidden markov models // Journal of Time Series Analysis, 2010. Vol. 31(2), pp. 132–138.

45. Berchtold A., Raftery A. E. The mixture transition distribution model for highorder markov chains and non-gaussian time series. Statistical Science, 2002. Vol. 17(3), pp. 328–356.

46. Aitnouri E., Wang S., Ziou D., Vaillancourt J., and Gagnon L.. Estimation of multi-modal histogram's pdf using a mixture model // Neural, Perallel and Scientific Computations, 1999. N. 7, pp. 103–118.

47. Everingham M., Thomas B. Supervised segmentation and tracking of non-rigid objects using a "mixture of histograms" model // Proceedings of 2001 International Conference on Image Processing, 2001. Pp. 62–65.

48. Yu-Ren Lai, Kuo-Liang Chung, Guei-Yin Lin, and Chyou-Hwa Chen. Gaussian mixture modeling of histograms for contrast enhancement // Expert Systems with Applications, 2012. Vol. 39, pp. 6720–6728.

49. Schopf J. M. A practical methodology for defining histograms for predictions and scheduling // In PARCO, 1999. Pp. 664–671.

50. Arroyo J., Gonzalez-Rivera G., Mat´e C., and San Roque A. M.. Smoothing methods for histogram-valued time series: an application to value at risk // Statistical Analysis and Data Mining, 2011. Vol. 4(2), pp. 216–228.

51. Williamson R. C., Downs T. Probabilistic arithmetic. numerical methods for calculating convolutions and dependency bounds // International Journal of Approximate Reasoning, 1990. N. 4, pp. 89–158.

52. Герасимов В. А., Добронец Б. С., Шустров М. Ю. Численные операции гистограммной арифметики и их применения // Автоматика и телемеханика, 1991. № 2, СС. 83–88.

53. Добронец Б. Интервальная математика. Красноярский государственный университет, 2004.

54. Добронец Б. С., Попова О. А. Численные операции над случайными величинами и их приложения // Journal of Siberian Federal University. Mathematics and Physics, 2004. № 2, СС. 229–239.

55. Civanlar M., Trussell H. Constructing membership functions using statistical data // Fuzzy Sets and Systems, 1986. Vol. 18, pp. 1–13.

56. Dubois D., Prade H. Fuzzy sets and statistical data // European Journal of Operational Research, 1986. Vol. 25, pp. 345–356.

57. Grzegorzewski P. Testing fuzzy hypotheses with vague data // Statistical Modeling, Analysis and Management of Fuzzy Data, 2000. Vol. 87, pp. 213–225.

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

59. Кудрявцев Л.Д. Курс математического анализа. В 3 томах. Т. 2. Ряды. Дифференциальное и интегральное исчисления функций многих переменных. М.: Дрофа, 2004

60. Parzen E. On estimation of a probability density function and mode // Annals of Mathematical Statistics, 1962. Vol. 33(3), pp. 1065–1076.

61. Scott D. W. On optimal and data-based histograms // Biometrika, 1979. Vol. 66(3), pp. 605–610.

62. Kullback S., Leibler R.A. On information and sufficiency // Annals of Mathematical Statistics, 1951. Vol. 22(1), pp. 79–86.

63. Pollard D. E. A user's guide to measure theoretic probability. Cambridge, UK:Cambridge University Press, 2002.

64. Мотренко А.П., Стрижов В.В. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульба-ка-Лейблера // Информатика и ее применения, 2014. Т. 8, № 2, СС. 86–97.

65. Кобзарь А. И. Прикладная математическая статистика. М.: Физматлит, 2006.

66. Granger C. W. J. Prediction with a generalized cost of error function. Operational Research Society, 1969. Vol. 20(2), pp. 199–207.

67. Christoffersen P. F., Diebold F. X. Optimal prediction under asymmetric loss // Econometric theory, 1997. Vol. 13(06), pp. 808–817.