ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР им. А. А. ДОРОДНИЦЫНА
РОССИЙСКОЙ АКАДЕМИИ НАУК (ВЦ РАН)

УДК 65.012.226,
ВГК ОКП 506190
№ госрегистрации
Инв. №

УТВЕРЖДАЮ
Директор ВЦ РАН,
академик РАН

_____ Ю.Г. Евтушенко

____._____.2014 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:

ОБОСНОВАНИЕ И ВЫБОР НАПРАВЛЕНИЙ ИССЛЕДОВАНИЙ, ИССЛЕДОВАНИЕ ЭКЗОГЕННЫХ ФАКТОРОВ И ИХ ВЛИЯНИЯ НА ОБЪЕМЫ СПРОСА НА ГРУЗОВЫЕ ЖЕЛЕЗНОДОРОЖНЫЕ ПЕРЕВОЗКИ

(промежуточный)

Этап первый

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 2014

СПИСОК ИСПОЛНИТЕЛЕЙ

Научный руководитель, заведующий отделом ВЦ РАН, член-корреспондент РАН	_____ (подпись, дата)	К.В. Рудаков (разделы 1-4)
Исполнители темы:		
Ведущий научный сотрудник, доктор физ.-мат. наук	_____ (подпись, дата)	Стрижов В.В. (реферат, введение, разделы 1-4)
Студент	_____ (подпись, дата)	Мотренко А.П. (разделы 1-4)
Аспирант	_____ (подпись, дата)	Кузнецов М.П. (разделы 1-4)
Инженер-исследователь	_____ (подпись, дата)	Каширин Д.О. (раздел 4, заключение)
Зав. сектором ВЦ РАН	_____ (подпись, дата)	Хорошевский В.Ф. (раздел 5)
Нормоконтролёр вед.научн.сотр., д.т.н., проф.	_____ (подпись, дата)	А.И. Эрлих

РЕФЕРАТ

Отчёт 126 с., 23 рис., 22 табл., 96 источников.

ГИСТОГРАММНОЕ ПРОГНОЗИРОВАНИЕ, ГРУЗОВЫЕ ЖЕЛЕЗНОДОРОЖНЫЕ ПЕРЕВОЗКИ, ДОСТОВЕРНОСТЬ ЭКСПЕРТНОГО ВЫСКАЗЫВАНИЯ, ВРЕМЕННОЙ РЯД, КОРРЕЛЯЦИОННЫЕ МЕТОДЫ, НЕПАРАМЕТРИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ, ОБНАРУЖЕНИЕ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ, ПРОГНОЗИРОВАНИЕ, РЖД, ТЕСТ ГРЕЙНДЖЕРА, ЭКЗОГЕННЫЙ ФАКТОР, ЭКСПЕРТНЫЕ ОЦЕНКИ

Объект прикладных научных исследований – объемы спроса на грузовые железнодорожные перевозки, количественных характеристиках экзогенных факторов, оценке влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки и структуре 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 первом этапе ПНИ:

1. В результате экспертных оценок качественного характера выявлены экзогенные факторы, влияющие на объёмы спроса на грузовые железнодорожные перевозки.

2. Разработаны методы анализа достоверности экспертных достоверности экспертных оценок.

3. Разработаны методы получения количественных оценок влияния экзогенных факторов на объёмы спроса на грузовые железнодорожные перевозки.

4. Проведено патентное исследование по области охвата ПНИ, подтвердившее патентную чистоту проводимых исследований и ожидаемых окончательных результатов.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

Блок-участок	–	Часть железнодорожного перегона, используемое как самостоятельное средство сигнализации и связи, ограниченное проходными светофорами или проходным светофором и станцией.
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 43-х групп перевозимых грузов.
abs	–	Функция ошибки прогнозирования. При ретроспективном прогнозе вычисляется как абсолютное (по модулю) отклонение прогноза от реальных значений прогнозируемого временного ряда.
ADF-тест	–	Augmented Dickey–Fuller test – расширенный тест Дики-Фуллера для проверки временных рядов на стационарность. Необходимость расширения связана с тем, что процесс может быть авторегрессией не первого, а более высокого порядка
ARIMA	–	AutoRegressive Integrated Moving Average
eEPS	–	extended Event Driven Process Chain – широко используемая в мире нотация описания бизнес-процессов
ARMA	–	AutoRegressive Moving Average.
hist	–	Разработанный в ВЦ РАН метод гистограммного прогнозирования. Метод обеспечивает оптимальность свертки построенной по значениям временного ряда гистограммы и функции потерь.
KPSS-тест	–	Тест Квятковского-Филлипса-Шмидта-Шина (Kwiatkowski–Phillips–Schmidt–Shin). Используется для проверки гипотезы о стационарности отклонений временного ряда от детерминированного временного тренда.
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 железно-дорожные перевозки
РЖД	–	ОАО «Российские железные дороги»

ВВЕДЕНИЕ

Современное состояние научно-технической проблемы. Проблема моделирования ГЖДП с целью повышения их экономической эффективности на протяжении многих лет является предметом научных исследования [1–31]. Сложность проблемы обусловлена наличием целого ряда существенных экзогенных (внешних по отношению собственно к ГЖДП) факторов: недетерминированность потребностей заказчиков в услугах транспортировки и погрузки/разгрузки грузов; возникновение запретов на движение товарных составов из-за необходимости обеспечения высокоскоростного движения; наличие в товарных составах вагонов, принадлежащих различным собственникам, и многие другие.

Один из распространенных подходов к решению задачи планирования транспортных потоков состоит в использовании экспертно-зависимых методов, применяемых для моделирования и прогнозирования грузовых перевозок.

Построение таких моделей заключается в попытках формализации различных ситуаций, возникающих в процессе ГЖДП с целью объединить их в единую систему. Так как множество таких ситуаций весьма разнообразно, а сами ситуации зачастую слабо формализуемы, для решения поставленной задачи в рамках эмпирических методов исследователи стремятся найти компромисс между обобщающей способностью системы и ее способностью автономно функционировать.

Следует отметить, что при разработке подобных моделей планирования и прогнозирования ГЖДП неизбежно приходится учитывать специфические особенности перевозчика (регламенты процессов, нормативную базу и т. п.), для РЖД, судя по доступным для изучения источникам, такие модели пока отсутствуют.

Актуальность и новизна темы. Учитывая сказанное, проблема разработки для РЖД математических методов и алгоритмов прогнозирования объемов спроса на ГЖДП с учётом влияния экзогенных факторов, на решение которой нацелено это ПНИ, представляется весьма актуальной, что подтверждается заинтересованностью в реализации этой темы Индустриального партнёра по разработке – РЖД.

Сложность прогнозирования объёмов спроса на ГЖДП обусловлена высокой волатильностью и нестационарностью исторических статистических данных об объёмах спроса, а также наличием в них значительного количества пропусков.

Перечисленные особенности прогнозируемых временных рядов приводят к падению качества прогнозов при использовании существующих в этой области методов.

При создании моделей, методов и алгоритмов прогнозирования объёмов спроса на ГЖДП в этом ПНИ будут учитываться как предыстория самих грузоперевозок в РЖД, так и предыстория влияющих на объёмы спроса экзогенных факторов. Множество таких факторов, в той или иной степени влияющих на объемы спроса на ГЖДП, весьма велико. Поэтому для выбора экзогенных факторов, ощутимо (с точки зрения экономической эффективности ГЖДП) влияющих на объёмы спроса, будут использованы экспертные оценки, достоверность которых в контексте решаемой проблемы будет исследована средствами специально адаптированных для этого современных методов выявления причинно-следственных связей. Таким образом, в разрабатываемых моделях, методах и алгоритмах прогнозировании объёмов спроса на ГЖДП исторические статистические данные о существенно влияющих на объёмы спроса экзогенных факторов, что, как показывают вычислительные эксперименты, проведённые на этом этапе ПНИ, позволяет заметно повысить качество прогнозов.

Разработка моделей, методов и алгоритмов прогнозирования, более точных по сравнению с существующими и обусловливают новизну проводимого исследования.

Исходные данные и научно-технические заделы. Ранее специалистами ВЦ РАН, МФТИ и Российской открытой академии транспорта был решен ряд задач в области прогнозирования загруженности железнодорожных узлов. Были разработаны:

- методы непараметрического прогнозирования загруженности системы железнодорожных узлов по историческим данным,

- алгоритм hist непараметрического прогнозирования загруженности железнодорожных узлов РЖД, основанный на свертке эмпирической плотности распределения значений временного ряда с функцией потерь.

На базе этих результатов был создан и внедрен ряд прикладных систем, основанных на синтезе мультиалгоритмических конструкций. При этом основное конкурентное преимущество решений состояло в объективно оцениваемом каче-

стве прогнозов – минимальная ошибка на тестовых и реальных данных.

Этот научно-технический задел только за период 2009–13 гг. отражен в более чем восьми научных публикациях.

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 способы, модели, 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 раках 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 поиска регрессионных моделей в неявно заданном множестве»

07-07-00181-а, «Развитие теории индуктивного порождения и выбора моделей» 10-07-00422-а, «Методы оценивания и отбора значений признаков в задачах распознавания и классификации» 11-07-00448-а.

Цели и задачи первого этапа, их место в выполнении проекта в целом. В соответствии с п. 2.1 Технического задания на первом этапа выполняемого проекта должны быть достигнуты следующие цели:

- Проведены патентные исследования технического уровня разработок в области охвата выполняемого проекта.

- Выполнен аналитический обзор современной научно-технической, нормативной, методической литературы, затрагивающей научно-техническую проблему, исследуемую в рамках ПНИ.

- Обоснованы и выбраны направления дальнейших исследований с целью разработки математической модели прогнозирования объемов спроса на ГЖДП, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки.

- Выявлены и исследованы экзогенные факторы и их влияния на прогноз объема спроса на ГЖДП, в том числе:

а) проведён экспертный анализа значимости и характера влияния экзогенных факторов на объемы спроса на ГЖДП;

б) описан метод обнаружения причинно-следственных связей в разнородных временных рядах объемов спроса на ГЖДП и экзогенных факторов;

в) описан метод оценки достоверности экспертных высказываний о влиянии экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки;

г) описаны способы расширения подхода Грейнджера для обнаружения структуры связей между временными рядами экзогенных факторов и объемов спроса на ГЖДП;

д) описан алгоритм построения кросскорреляционных зависимостей временных рядов объемов спроса на ГЖДП и экзогенных факторов и приведены результаты его тестирования.

- Проведено исследование предметной области, в том числе:

а) описаны основные процессы в области организации ГЖДП;

б) выявлены источники данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов в историческом и текущем периоде.

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

- На основе аналитического обзора современной научно-технической, нормативной, методической литературы по научно-технической проблеме, исследуемой в рамках проекта, выбрать базовую модель прогнозирования и определить направления ее дальнейшего развития с целью предусмотренной п. 2.1.5 Технического задания разработки математической модели прогнозирования объемов спроса на ГЖДП, учитывающей влияние экзогенных факторов на объемы спроса на ГЖДП.

- Так как в указанной математической модели прогнозирования объемов спроса на ГЖДП должно учитываться влияние экзогенных факторов на объемы спроса на ГЖДП определяемое экспертным путём, должны быть исследованы и адаптированы к использованию в рамках работ, предусмотренных п. 2.1.5 Технического задания, методы оценки достоверности экспертных высказываний, а так же выявления структуры и характера влияния экзогенных факторов на объёмы спроса на ГЖДП.

Провести исследования предметной области, включая основные процессы в области планирования и управления ГЖДП и возможных источников данных об объемах спроса на ГЖДП и влияющих на объёмы экзогенных факторов в историческом и текущем периоде в целях создания информационной базы для проведения вычислительных экспериментов, предусмотренных п. 2.1.7 Технического задания.

1 Аналитический обзор современной научно-технической, нормативной, методической литературы, затрагивающей исследуемую научно-техническую проблему, исследуемую в рамках ПНИ

В этом разделе представлены результаты аналитического обзора современной научно-технической, нормативной, методической литературы, затрагивающей исследуемую научно-техническую проблему.

Рассмотрен 31 научно-информационный источник, в том числе 15 научно-информационных источников за период 2009–13 гг.

Традиционно [1–5], модели спроса на транспортные услуги строятся в рамках предметной области с упором на опыт и суждения экспертов. Построение таких моделей заключается в попытках формализации различных ситуаций, возникающих в процессе перевозок с целью объединить их в единую систему. Так как множество таких ситуаций весьма разнообразно, а сами ситуации зачастую слабо формализуемы, для решения поставленной задачи в рамках эмпирических методов приходится находить компромисс между обобщающей способностью системы и ее способностью автономно функционировать. Характерный для задач планирования транспортных потоков класс экспертно-зависимых методов, применяемых для моделирования и прогнозирования грузовых перевозок, представлен в [6–9].

Одной из базовых моделей, первоначально разработанной для пассажирских перевозок, является четырехступенчатая транспортная модель [6], которая, по мнению специалистов [6, 7], может плодотворно применяться и при моделировании грузовых перевозок. Тем не менее, при прогнозировании грузоперевозок каждая из четырех ступеней модели может существенно отличаться от случая пассажирских перевозок. Важные различия между грузовыми и пассажирскими перевозками состоят в разнообразии лиц, принимающих решения (грузоотправителей, перевозчиков, посредников, водителей, операторов), в разнообразии грузов для транспортировки (от поставок посылок до грузов весом в несколько сотен тысяч тонн) и ограничениях, связанных с неполной доступностью данных (отчасти объясняемой требованиями конфиденциальности). Ниже перечислены четыре ступени модели в контексте грузоперевозок:

Ступень 1 – Производство и спрос. На этом этапе определяются объемы то-

варов, которые необходимо перевезти из различных пунктов отправления и объемы товаров, которые необходимо доставить в пункты назначения. Объемы товаров измеряются в тоннах или денежных единицах (на промежуточных стадиях моделирования).

Ступень 2 – Распределение. На этом этапе, вычисляются потоки грузоперевозок между пунктами отправления и назначения [10] (формируются элементы матрицы отправления-назначения).

Ступень 3 – Разбиение по видам транспорта. На этом этапе происходит распределение товарных потоков между способами транспортировки: автомобильный, железнодорожный, комбинированные перевозки, внутренние водные пути.

Ступень 4 – Назначение. После преобразования грузопотоков, измеряемых в тоннах, в транспортные единицы, грузопотоки могут быть отнесены к сетям (например, назначаются потоки грузовых и пассажирских автомобили в дорожных сетях).

Помимо этих четырех ступеней, как правило, необходимо выполнить ряд дополнительных преобразований. Например, преобразование торговых потоков в денежных единицах в физические потоки в тоннах для определения производства и спроса. Эта конвертация может быть проделана с использованием коэффициентов отношения стоимость/вес для различных товарных групп. Используемые здесь коэффициенты могут оказать существенное влияние на результаты прогнозирования и поэтому важно собрать актуальные данные по конверсии.

Другой тип преобразования заключается в переходе от потоков в тоннах к единицам транспортных средств, таким как тяжелые грузовые автомобили. Этот тип преобразования может потребоваться между выбором вида транспорта и шагом «назначение». Результат конвертации зависит от большого количества принимаемых решений по частоте отгрузки, объемам партий, возвращению грузов и объемам использования транспортных средств. Эти решения могут быть смоделированы в явном виде в рамках дополнительных логистических модулей (например, в модели SMILE [9]), но часто здесь используются фиксированные коэффициенты для конвертации.

Еще один тип преобразования, – «регионализация», – связан с переходом от грубой к более точной системе зонирования.

На первой ступени (Производство и спрос) можно выделить следующие типы применяемых на практике моделей:

- обнаружение тренда и моделирование временных рядов [11];
- моделирование динамических систем [12];
- модели межотраслевого баланса и связанные с ними модели [13].

Все эти модели основаны на агрегированных данных, и в литературе отсутствуют упоминания моделей, использующих дезагрегированные данные.

В трендовых моделях исторические данные экстраполируются в будущее. Временные ряды используются для разработки моделей различной степени сложности, таких как модели фактора роста и авторегрессионные модели скользящего среднего [8, 14]. Последняя модель использует только информацию об объемах грузопотоков и предназначена для краткосрочного прогнозирования. Так же были разработаны модели для прогнозирования временных рядов с учетом дополнительных переменных [15], например, таких как валовой внутренний продукт.

Пример модели, опирающейся на представление процесса транспортировки грузов в виде динамической системы – модель ASTRA (Assessment of Transport Strategies, оценка транспортных стратегий) [12], разработанная в рамках проекта для Европейской комиссии. В модели ASTRA динамика системы, изменения объемов грузоперевозок и влияния экономики на объем перевозок, а также учет землеустройства и окружающей среды моделируются в явном виде [12]. В макроэкономическом модуле ASTRA прогнозируется рост ВВП. Результаты подаются в модуль региональной экономики, который возвращает спрос на грузоперевозки в виде грузопотоков в тоннах для пар отправление-назначение. Изменения в транспортном спросе в свою очередь могут повлиять на ВВП через стоимость транспортировки. Параметры модели динамической системы, как правило, не получают из статистических оценок, а задают согласно существующим источникам. Эти параметры принимаются в качестве начальных значений и затем настраиваются в результате анализа построенной динамической системы (метод «проб и ошибок»). Моделирование динамической системы может включать шаги «распределение» и «разбиение по видам транспорта». Однако, модели динамических систем, как правило, не содержат достаточной пространственной информации о конфигурации транспортной сети для моделирования потоков между зонами.

В основе моделей межотраслевого баланса лежат макроэкономические модели и таблицы межотраслевого баланса, которые описывают (в денежных единицах), что каждый сектора экономики (например, текстильной промышленности) поставляет в другие секторы, включая конечный спрос со стороны потребителей, импорта и экспорта.

Национальные таблицы межотраслевого баланса разработаны для многих стран. Особая форма таблиц межотраслевого баланса – мультирегиональная или пространственная. Такие таблицы включают в себя не только межотраслевые, но и межрегиональные производственные отношения (торговые потоки). Большинство мультирегиональных таблиц межотраслевого баланса различают лишь небольшое количество крупных регионов внутри страны. В моделях межотраслевого баланса предполагается, что для прогнозирования, межрегиональные таблицы можно сформировать на основе прогнозируемого отраслевого роста. Новая таблица межотраслевого баланса может дать оценку будущих торговых потоков между регионами, используя либо:

- фиксированные технические и торговые коэффициенты: текущие объемы производства и торговли экстраполируются в будущее.
- гибкие технические и торговые коэффициенты: оцениваются функции, в которых доля продукции, произведенной в отрасли *P* региона *J*, которая потребляется в отрасли *S* региона *I*, зависит от общего объема производства в отрасли *P* региона *J* и (обобщенных) транспортных расходов, по отношению к другим регионам. Это делает результат ступени «Распределение» чувствительным к изменениям в транспортных расходах и времени транспортировки (форма индуцированного спроса).

Модели последнего типа основаны на использовании большого числа констант, которые со временем устаревают. В работе [15] отмечается, что для предотвращения снижения предсказательной способности подобные модели требуют экспертной перенастройки с течением времени, поэтому в [15] принят подход, основанный на прогнозировании временных рядов. Предложена модель национальной системы прогнозирования железнодорожных перевозок, основанная на стохастических уравнениях, отражающих следующие взаимосвязи между временными рядами:

- В модель спроса на железнодорожные перевозки грузов включаются объем и средняя длина маршрута перевозки товара.

- Спрос на перевозки на национальном уровне определяется спросом на товары, и может быть оценен с помощью одного из двух подходов: регионального, в рамках которого суммируется спрос на товар во всех регионах; и анализа коэффициентов, которые связывают транспортное потребление фирмы с уровнем производства. Оба подхода рассматривают национальный транспортный спрос в зависимости от объемов товарной продукции и реальной средней удельной стоимости транспортировки.

- Количество и тип вагонов, которые будут использоваться при перевозках, а также ожидаемый общий дохода, зависят от уровня и товарной структуры будущих железнодорожных грузовых перевозок.

- Суммарный вес (в тоннах) товара, перевозимого по железным дорогам в произвольный период времени, зависит от общего уровня выходного товара, структурных изменений (централизации или децентрализации производства по отношению к потреблению), и доли перевозимого товара в общем объеме железнодорожных перевозок.

- В краткосрочной перспективе, железнодорожные тарифы будут иметь незначительное влияние на объемы промышленного производства. Однако они в значительной мере определяют долю железнодорожных перевозок, которая является функцией железнодорожных и нежелезнодорожных тарифов, качества обслуживания, и эластичности интермодальных замещений.

В рамках этой модели национальной системы прогнозирования железнодорожных перевозок измеряется влияние макроэкономической деятельности и межвидовой конкуренции на потребности заказчиков в железнодорожных грузоперевозках. В модели учитываются следующие величины (для заданного товара и периода времени): суммарный объем перевезенного товара, перевезенный за учитываемый период (в млн. тонн); средняя длина маршрута перевозки товара за период (в милях к тоннам); средний доход на тонну перевозимого товара; суммарная стоимость товара, произведенного за период; количество коммерческих зарегистрированных транспортных средств.

В работах [16, 17] рассмотрена задача определении факторов влияющих на

объемы перевозок. В [17] на основании метода MANOVA делается вывод о влиянии таких факторов, как погодные условия или праздничные дни, на объем автомобильного трафика. В [16] описана концепция системы моделирования грузоперевозок, с точки зрения влияния различных финансовых факторов на увеличение доли железнодорожных перевозок и сделана попытка определить полный спектр факторов, которые могут повлиять на доли рынка железнодорожных перевозок, с особым акцентом на определение «рычагов», таких как изменения в налогообложении грузового транспортного средства или дорожные заторы. «Рычаги» также включают прямое субсидирование железнодорожных услуг, поощрение развития железнодорожных связей в промышленных объектах и интермодальных терминалах, инвестиции в улучшение инфраструктуры, и меры, направленные на улучшение качества или производительности ЖД сети. При формировании модели считается, что единственный существенный рынок, на котором ЖД перевозки не могут конкурировать с автомобильными – *грузы, перемещаемые в небольших транспортных средствах*. Доля грузов, перевозимых в небольших транспортных средствах, как правило, гораздо выше для коротких расстояний, и варьируется в зависимости от вида груза. Объемы грузов, перевозимых в небольших транспортных средствах, оцениваются на ранней стадии моделирования (отдельно для различных групп товаров и характерных расстояний) и удаляются из потенциальных рынков ЖД перевозок. Построение таких моделей существенно опирается на опыт и суждение эксперта, поэтому перечисленные процедуры не обладают достаточной гибкостью для соответствия требованиям к современным системам планирования и оперативного управления.

При решении задачи прогнозирования грузоперевозок возникает проблема хранения и обработки большого количества динамически изменяемой информации. Большая часть этой информации является неинформативной. В работе [18] для решения проблемы предлагается построить сжатое признаковое описание данных для выделения информативных шаблонов. Предлагается количественный метод для систематического выделения шаблонов в структуре спроса на трафик в исторических данных. Метод основан на технике кластерного анализа и позволяет учитывать априорные знания, что облегчает интерпретацию и практическое использование результатов. Предлагаемая процедура классификации шаблонов применяется к

данным о почасовых объемах трафика на испанском шоссе за 5 лет. Полученные результаты демонстрируют точность метода при выделении сезонных и суточных характерных шаблонов спроса на перевозки.

При планировании пассажирских перевозок [19] важную роль играет моделирование и прогнозирование задержек поездов. В работе [20] для этого используются аналитический и «порождающий» подходы. Аналитический подход является более общим и не требует детальной информации о расписании движения поездов на исследуемом участке сети. Этот подход используется для получения долгосрочного анализа и планирования загруженности железнодорожной сети. В свою очередь, «порождающий» подход основан на имитационном моделировании конкретного участка сети с известными расписанием и инфраструктурой [21]. Он позволяет моделировать задержки, но вычислительно более сложен.

При рассмотрении более детальных данных возникает необходимость учета взаимодействия поездов. Поезд влияет на движение следующего за ним поезда с помощью системы сигнализации. Путь делится на блок-участки, на каждом из которых может находиться только один поезд. Поезд может беспрепятственно продолжать движение, если впереди свободно два или более блок-участка. Иначе он получает сигнал остановиться и не покидать блок-участок или снизить скорость, чтобы суметь остановиться в пределах следующего участка.

Пропускная способность участка сети определяется как количество поездов, проходящих по участку за час. Пропускная способность считается неувеличиваемой, если при ее увеличении движение поездов становится невозможным из-за задержек. Исследование [20] является частью проекта по улучшению функционирования железнодорожной сети Швеции. В частности, решается задача составления оптимального расписания движения поездов.

При составлении «хорошего» расписания необходимо найти компромисс между увеличением пропускной способности и другими характеристиками, ухудшающимися при повышении пропускной способности. В [21], например, вводится показатель надежности как взвешенная сумма средней задержки и ее среднеквадратичного отклонения, а в [20] показателем хорошо составленного расписания считается его устойчивость к задержкам.

Для построения «хорошего» расписания необходимо проанализировать си-

туации, ведущие к возникновению и распространение задержек, выделить ключевые параметры, определяющих загруженность железнодорожной сети. К таким ключевыми параметрами в рассматриваемых здесь источниках относят вторичную задержку, использованную часть запасного времени и неоднородность. Мера неоднородности при этом определяется как разница во времени следования поездов различных типов.

Аналитический подход используется также в [22, 23]. Задержка поезда в зависимости от времени в [22] моделируется случайным процессом со стационарными независимыми приращениями, а в [23] – как случайный винеровский процесс. Вводится понятие критической дистанции – расстояния между поездами, при котором возникает вторичная задержка, определяется критическая дистанция между поездами для различных систем сигнализации.

Задача моделирования задержек решается и в применении к грузовым перевозкам. В [24, 25] описывается проблема перегруженности узлов при взаимодействии различных ведомств, в частности при доставке груза в порт. Предлагается сосредоточить в «одних руках логистической цепи» управление поставкой груза клиенту, что позволит согласованно развивать железнодорожную и портовую инфраструктуру. В статье приводится SWOT-анализ такого решения.

В [26] в рамках принятой в РЖД Программы совершенствования работы и развития сортировочных станций на 2010–15 гг. предложена система, автоматизирующая управление сортировочными станциями.

В [27] рассмотрена проблема пересадок на железнодорожном транспорте. Предложен итеративный алгоритм, объединяющий глобальную задачу минимизации суммарной задержки всех пассажиров с локальной задачей планирования движения поездов.

В качестве базового алгоритма для сравнения полученных прогнозов в [24–26] используется модель авторегрессионного скользящего среднего ARMA. Для построения прогнозов данных с выраженной периодичностью в [28] используется модель sARIMA. В [19] для решения подобных задач применяются нейронные сети. Также для построения краткосрочных прогнозов объемов перевозок используются марковские цепи [29, 30] и непараметрические методы [31].

Помимо указанных выше математически-статистических методов прогно-

зирования, для прогнозирования объемов грузоперевозок применяются макроэкономические методы прогнозирования [97, 98]. Методы заключаются в анализе балансного метода и поиске конкурентного равновесия в выпукло-вогнутой игре. Так, концепция равновесия по Линдалю предполагает, что вложения экономического агента в финансирование общественного блага определяется предельной полезностью для него общественного блага. Под общественными благами в этой модели понимаются товары и услуги, которыми могут пользоваться одновременно разные потребители. Примерами отраслей, производящих такие товары и услуги, являются инфраструктурные отрасли (естественные монополии). Основным результатом исследования таких моделей является теорема о механизме конкурентного равновесия, который должен допускать дифференциацию цен (перекрестное субсидирование), т.е. потребители различных типов за предоставленное общественное благо должны платить по разным тарифам.

В настоящее время Индустриальный партнер, ОАО «РЖД», использует полуавтоматический метод прогнозирования объемов спроса на грузовые железнодорожные перевозки, подробнее рассмотренный в разделе 2.1. Принцип этого метода заключается в построении авторегрессионной модели с оценкой коэффициентов методом наименьших квадратов. Также отдельным подразделением Индустриального партнера производится построение прогнозов объемов спроса на грузовые железнодорожные перевозки с помощью макроэкономических методов. После получения обоих прогнозов производится экспертная коррекция прогноза, построенного на авторегрессионной модели, с учетом результатов макроэкономического прогноза. В данном проекте основное исследование посвящено развитию именно авторегрессионного метода прогнозирования.

Т.к. в текущем процессе прогнозирования Индустриального партнера присутствуют оба метода построения прогноза, сравнение с макроэкономическими методами прогнозирования в рамках проекта не проводится. Разрабатываемый в рамках данного проекта инструментарий прогнозирования, основанный на статистико-математических методах, предлагается использовать как замену полуавтоматического метода прогнозирования объемов спроса на грузовые железнодорожные перевозки. Соответственно разрабатываемые в рамках проекта методы предлагается использовать в процессах Индустриального партнера вместе с макроэкономиче-

скими методами.

Как показывает анализ приведенной литературы, предложенные в ней подходы и методы являются достаточно эффективными и при соответствующей адаптации и доработке смогут повысить качество прогнозирования объемов ГЖДП в РЖД и, как следствие, повысить качество планирования движения железнодорожных составов.

2 Обоснование и выбор направлений дальнейших исследований с целью разработки математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки

2.1 Постановка и анализ проблемы прогнозирования объемов спроса на ГЖДП с учетом влияния экзогенных факторов

Пусть заданы два индексно упорядоченных множества $\mathcal{N}$ и $\mathcal{K}$ имён эндогенных и экзогенных факторов, соответственно.

Далее считается, что все эндогенные и экзогенные факторы как-то изменяются во времени и измеримы, то есть каждому фактору в каждый момент времени можно сопоставить некоторое числовое значение – натуральное (из множества $\mathbb{N}$ с включением нуля), целое (из множества $\mathbb{Z}$) или рациональное (из множества $\mathbb{Q}$).

Для обозначения временных рядов эндогенных и экзогенных факторов будут соответственно использоваться записи:

$$\mathbf{s}_n(\mathrm{T}),\ n \in \mathcal{N} \tag{2.1}$$

и

$$\mathbf{x}_k(\mathrm{T}),\ k \in \mathcal{K}, \tag{2.2}$$

где T – длина некоторого дискретизированного промежутка времени [0, T), состоящего из T равновеликих открытых справа полуинтервалов, а записи $n \in \mathcal{N}$ и $k \in \mathcal{K}$ указывают на имена эндогенных и экзогенных факторов с индексами n и k, соответственно.

Временные ряды $\mathbf{s}_n(\mathrm{T})$ и $\mathbf{x}_k(\mathrm{T})$ представляют собой наборы:

$$\mathbf{s}_n(\mathrm{T}) = \left\| s_{nj} \right\|_{j=1}^{\mathrm{T}} \quad \text{и} \quad \mathbf{x}_k(\mathrm{T}) = \left\| x_{kj} \right\|_{j=1}^{\mathrm{T}}, \tag{2.3}$$

где s_{nj}, и x_{kj}, $j = 1, ..., \mathrm{T}$ – значения n-го эндогенного и k-го экзогенного фактора в начале j-го, открытого справа полуинтервала времени с продолжительностью, равной шагу дискретизации промежутка времени [0, T).

Под эндогенными факторами $\mathcal{N}$ в этом отчёте понимаются грузы – объекты ГЖДП, а под экзогенными $\mathcal{K}$ – внешние факторы, влияющие на объёмы спроса на перевозку грузов. Подробное описание эндогенных и экзогенных факторов приве-

дено в подразделе 3.1.

Учитывая, что как сами экзогенные факторы, так и гипотезы о характере их влияния на объёмы ГЖДП выявляются экспертным путём, рассматриваемая проблема прогнозирования включает в себя три подпроблемы, требующих решения в контексте поставленной проблемы:

- разработка методов прогнозирования спроса на объёмы ГЖДП с учётом специфики РЖД;
- разработка методов выявления, анализа влияния и прогнозирования экзогенных факторов, влияющих на объёмы ГЖДП;
- разработки математической модели и методов прогнозирования объемов спроса на ГЖДП, учитывающих влияние экзогенных факторов.

Именно в таком контексте в этом разделе проведено обоснование и выбор направлений дальнейших исследований.

На текущий момент Индустриальный партнер в лице ОАО «РЖД» использует модель прогнозирования на основе однофакторных линейных регрессионных моделей [101], в котов которых зависимыми показателями являются прогнозируемые объемы погрузки грузов, а независимыми показателями - соответствующие внешние факторы:

$$\hat{x}_k(T) = a_k + b_k s_k(T),$$

где $\hat{x}_k(T)$ – оценка объема погрузки k в прогнозному году T, $s_k(T)$ – значение соответствующего грузам группы k внешнего фактора в году T, a_k и b_k – параметры линейной регрессии для грузов группы k.

При определении параметров регрессии Индустриальный партнер производит анализ ретроспективного периода [101]. Известные значения показателей прошлых лет отбираются экспертным путем с учетом предполагаемой тенденции их динамики в рассматриваемой перспективе. При этом количество лет ретроспективного периода не должно быть меньше количества лет прогнозного периода.

Партнером отмечено [101], что получаемые данной однофакторной моделью оценки прогноза могут являться существенно завышенными или заниженными. В этом случае, для расчета прогноза по годам партнером было предложено использовать авторегрессионную зависимость, учитывающую значение объемов предыдущего временного периода:

$$\hat{x}_k(T) = \hat{x}_{k-1}(T)\frac{a_k + b_k s_k(T)}{a_k + b_k s_k(T-1)}.$$

Партнером указано, что описанные модели прогнозирования погрузки следует регулярно актуализировать, по мере того как появляются новые фактические данные об объемах погрузки и значениях внешних факторов, а также скорректированные прогнозы внешних факторов на рассматриваемую перспективу.

Опираясь на данную партнером информацию о виде прогностической модели, в рамках проекта был сформирован набор методов прогнозирования временных рядов, каждый из которых является фактическим обобщением метода, используемого Индустриальным партнером. Сравнительный анализ этих методов указан в разделе 2.2.2. На основе сравнительного анализа был выбран базовый алгоритм прогнозирования ARMA – авторегрессионный метод прогнозирования с автоматической настройкой параметров. Этот метод, во-первых, является обобщением метода Индустриального партнера, что позволяет проводить дальнейший сравнительный анализ, опираясь на полученные методом ARMA результаты прогнозирования. Во-вторых, метод ARMA удовлетворяет набору критериев выбора, приведенных в разделе 2.2.1, и продемонстрировал наилучшее качество ретроспективного прогноза среди всего множества методов (см. раздел 2.2.2).

Исходя из вышесказанного, выбранный базовый алгоритм ARMA позволяет устранить две основные слабые стороны текущего метода Индустриального партнера. Во-первых, автоматический метод построения авторегрессии позволяет достигать приемлемой точности прогноза, в то время как текущий метод в некоторых случаях может получать существенно завышенные или заниженные оценки. Во-вторых, автоматизация оценки параметров авторегрессионной модели является критичным условием для обеспечения высокой детализации прогноза при минимизации стоимости построения прогнозов.

2.2 Методы прогнозирования спроса на объёмы ГЖДП

В этом подразделе проведён сравнительный анализ методов прогнозирования временных рядов грузоперевозок и дано обоснование выбранных методов прогнозирования.

Сравнительный анализ выполнен в два этапа: на первом этапе в результате сравнительного анализа наиболее распространенных алгоритмов прогнозирования выбран базовый. Основной особенностью базового алгоритма является то, что он является фактическим обобщением полуавтоматического метода прогнозирования, используемого Индустриальным партнером. На втором этапе проведено сравнение базового алгоритма с разработанным в ВЦ РАН алгоритмом hist.

Сложность прогнозирования исторических временных рядов грузоперевозок обусловлена слудующими особенностями используемых данных. Во-первых, данные грузоперевозок отличаются высокой волатильностью. Значительный разброс в наблюдаемых значениях прогнозируемых временных рядов повышает погрешность прогнозирования, особенно при высокой частоте периодизации прогноза. Кроме того нестационарность, наблюдаемая для многих рассматриваемых временных рядов, и, как следствие, несостоятельность получаемых оценок параметров распределения данных, могут приводить к получению неадекватных прогнозов. Дополнительную сложность создают пропуски в данных, вызванные отсутствием наблюдений для некоторых отсчетов времени. В связи с перечисленными проблемами в текущем разделе выдвигается ряд требований, которым должен удовлетворять алгоритм прогнозирования для решения поставленной задачи.

2.2.1 Критерии выбора алгоритма прогнозирования

Алгоритм прогнозирования объемов грузоперевозок должен удовлетворять следующим базовым критериям.

- Критерий точности: ошибка прогнозирования при ретроспективном прогнозе должна быть минимальна.
- Критерий согласованности: при наличии иерархической структуры временных рядов отношение, задающее структуру, должно сохраняться для прогнозов, полученных на различных уровнях иерархии.
- Критерий учета экзогенных факторов: алгоритм прогнозирования должен использовать экзогенные факторы, влияющие на целевую переменную.
- Получаемый прогноз должен удовлетворять физическим ограничениям, связанным со спецификой задачи: прогнозы должны быть неотрицательны и не превышать пропускной способности сети.

Кроме того, прогностическая система должна формировать прогноз в соответствии со следующими требованиями:

- в заданном периоде: на месяц посуточно и подекадно, на квартал помесячно, на период от года помесячно и поквартально.
- с разложением по:

а) группам грузов: каждой группе поставлен в соответствии определённый идентификатор;

б) родам подвижного состава: полувагоны, крытые вагоны, цистерны, платформы, прочие;

в) комбинированному разложению, учитывающему перечисленные факторы.

2.2.2 Сравнительный анализ методов непараметрического прогнозирования

Для построения прогностической модели в этом проекте будут применяться непараметрические методы прогнозирования. В этом пункте с целью выбора наиболее подходящего для рассматриваемой предметной области метода проведен сравнительный анализ современных методов непараметрического прогнозирования. При этом в качестве критериев сравнения рассматривается сумма квадратов невязок (MSE) и сумма абсолютных значений невязок (abs).

Помимо указанных критериев особое внимание уделено гибкости методов с точки зрения учета экзогенных факторов и пределам применимости методов относительно рассматриваемых данных.

При выборе базового алгоритма прогнозирования, отвечающего поставленным требованиям, и учитывающего влияние экзогенных факторов, были исследованы основные методы непараметрической регрессии, такие как ядерное сглаживание, сглаживание сплайнами, авторегрессия, скользящее среднее и др. [32–36]. Они заключаются в присвоении имеющимся значениям временного ряда некоторых весов и комбинации взвешенных значений для получения прогноза.

Ядерное сглаживание (Kernel Smoothing**)** [33]. Метод решает задачу непараметрического прогнозирования временного ряда на несколько отсчетов времени, решается и задача восстановления регрессии. Для достижения наилучшего качества прогноза используется динамический выбор внутренних параметров

модели: одной из пяти ядерных функции и так называемой «ширины окна» h. Выбор h осуществляется в результате достаточно сложной в вычислительном отношении минимизации некоторой, вообще говоря, не унимодальной функции зависящей от ядерной функции.

Качество прогноза, достигаемое методом ядерного сглаживания, дает теоретически эффективные оценки. В частности, доказан ряд фактов [33] о достижении минимума критерия MSE при небольшом разбросе значений критерия. Однако для достижения хороших результатов необходим правильный выбор ядерной функции, описывающей близость временных отсчетов. Необходимость выбора ядерной функции делает метод вычислительно неоптимальным, однако позволяет учесть экзогенные факторы в качестве аргументов функции. Помимо выбора функции, необходимо производить выбор ширины окна на исторических данных: при недостаточной ширине алгоритм частично учитывает шум, а при завышенной ширине алгоритм начинает чрезмерно сглаживать ряд, теряя информацию о виде прогнозируемого ряда. Перечисленные свойства указывают на то, что алгоритм ядерного сглаживания теоретически оптимален в смысле критериев сравнения, однако является вычислительно неэффективным.

Экспоненциальное сглаживание (Exponential Smoothing) [33] – один из самых простых методов в использовании. Прогнозирование осуществляется посредством индуктивного построения прогнозируемых значений как взвешенных средних от значений обучающей выборки, вес осреднения и являются параметром сглаживания.

Проведённые исследования выявили следующие надостатки: метод рассчитан на прогнозирование с небольшим горизонтом и при этом не позволяет учитывать тренд и сезонные изменения. Учет влияния этих и других экзогенных факторов требует существенной дополнительной модификации и усложенения. Помимо этого, метод экспоненциального сглаживания не позволяет получать статистически эффективные оценки прогноза. Тем не менее, этот метод является выжным для рассмотрения, поскольку требует мало вычислительных затрат и позволяет получать быстрое значение прогноза.

Метод «Гусеница» (SSA) [32]. При прогнозировании временных рядов этим методом обучающая выборка, используя встроенный метод главных компо-

нентов, раскладывается в сумму компонентов, каждый компонент интерпретируется (тренд, периодика, шум), и по каждому из них строится свой прогноз. Прогноз временного ряда строится по определённым правилам как суперпозиция прогнозов для его компонентов.

Проведённые исследования выявили быстрый рост вычислительной сложности метода с длиной учитываемой истории прогнозируемого ряда. На практике это ограничивает применение метода прогнозированием коротких временных рядов с ярко выраженной периодичнотью или мультипериодичностью. Тем не менее, данный метод позволяет учитывать экзогенные факторы путем увеличения авторегрессионной матрицы и, как следствие, вычислительной сложности задачи. Также проведенные исследования показали низкую эффективность метода в смысле критериев сравнения. Предположительно, низкая эффективность связана с выбором недостаточного количества компонентов для достижения приемлемой вычислительной эффективности.

Авторегрессионное интегрированное скользящее среднее (ARIMA) [33]. Модель ARIMA является обобщением модели авторегрессионного скользящего среднего (ARMA). В свою очередь, модель ARMA является синтезом авторегрессионной модели и модели скользящего среднего и работает в предположении, что ряд стационарен. Поскольку исследуемые временные ряды грузоперевозок, вообще говоря, не являются стационарными, предлагается применять процедуру взятия последовательной разности.

Приведенные исследования выявили высокую эффективность метода в смысле критериев сравнения (в особенности, критерия abs). Кроме того, данный метод обладает достаточной гибкостью для включения произвольного количества экзогенных переменных с помощью расширения модели до т.н. модели векторной авторегрессии. При этом вычислительная сложность прогнозирования растет линейно, а сложность оценки параметров – субквадратично по количеству экзогенных переменных, что делает метод ARIMA подходящим базовым методом решения задачи прогнозирования.

Локальные методы прогнозирования временных рядов (Local Forecast-

ing) [34]. Этот класс методов основан на поиске похожих паттернов во временном ряде для прогнозирования конечного участка паттерна. К параметрам алгоритма относится число «ближайших соседей» – наиболее близких паттернов к исследуемому, функция расстояния между паттернами и длина рассматриваемого вектора предыстории. Прогноз временного ряда строится как агрегированное значение конечных элементов наиболее похожих паттернов в предыстории.

Данный метод показал лучшие результаты прогнозирования одномерного временного в смысле критерия сравнения с SSA. Однако метод не проявляет достаточную гибкость в смысле включения в рассмотрение экзогенных факторов: появляется необходимость решать задачу выбора наилучшей метрики, что является вычислительно трудоемким.

Прогнозирование временного ряда с помощью приближения производными рядами (Subseries Superposition Producing) [33]. Метод строит прогноз временного ряда на основании пронозов для производных от него рядов. Например, производными являюся ряды, значения которых в каждой точке равны средним значениям исходного ряда в некоторой временной окрестности точки. Для построения производных временных рядов задается алфивит базисных функций, и строится оптимальная суперпозиция базисных функций. На основании спрогнозированных значений суперпозиций строится прогноз исходного временного ряда. Этот метод показал низкую вычислительную эффективность и низкую эффективность в смысле критериев сравнения. Тем не менее, метод является перспективным в смысле включения в модель экзогенных факторов, и его идея может быть применена в дальнейшем развитии.

Выбор моделей и признаков в задачах прогнозирования (Model Selection) [36]. Этот алгоритм прогнозирования основан на расширении пространства признаков с помощью порождающих функций и последующем выборе линейных моделей прогнозирования. В качестве модели, описывающей отношение между зависимой переменной и свободными переменными, используется полином Колмогорова-Габора с заведомо избыточным числом элементов, из которых осущест-

вляется выбор оптимальной модели. Для выбора информативных параметров используется метод Белсли.

При исследовании данный метод показан наихудшую эффективность на краткосрочном прогнозировании. Однако, предложенная идея выбора модели может быть использована в дальнейшем при исключении неинформативных экзогенных факторов.

2.2.3 Сравнение методов прогнозирования

Описанные в п. 2.2.1 алгоритмы были протестированы на 121 реальном временном ряде. Рассматривалось краткосрочное и долгосрочное прогнозирование. Длина обучения составляла в среднем 400 отсчетов. При краткосрочном прогнозировании длина прогноза составляла примерно 1% от длины обучения. Алгоритмы сравнивались по средней ошибке MSE. На каждом из рядов лучший из алгоритмов получал 6 баллов, следующий – 5, и так далее; худший алгоритм получал 0 баллов. В таблице 2.1 приведена сумма баллов, а также среднее время работы алгоритма на ряде.

Таблица 2.1 – Сравнение работы алгоритмов на краткосрочном прогнозировании

Название алгоритма	Сумма баллов	Среднее время работы на ряде
SSA	198	менее секунды
Model Selection	92	менее секунды

Продолжение таблицы 2.1

Название алгоритма	Сумма баллов	Среднее время работы на ряде
Local Forecasting	463	менее секунды
ARMA	462	менее секунды

Subseries Superposition Producing	295	35 секунд
Kernel Smoothing	503	120 секунд
Exponential Smoothing	528	менее секунды

Видно, что по сумме баллов и времени работы вперед выходят четыре алгоритма с примерно одинаковыми результатами: Local Forecasting, ARIMA, Kernel Smoothing, Exponential Smoothing.

Также четыре алгоритма были протестированы на тех же временных рядах для долгосрочного прогнозирования: длина прогноза составляла примерно 10% обучения. Каждому алгоритму были присвоены баллы от 3 до 0. Результаты представлены в таблице 2.2.

Таблица 2.2 – Сравнение работы алгоритмов на долгосрочном прогнозировании

Название алгоритма	Сумма баллов	Среднее время работы на ряде
SSA	85	менее секунды
Local Forecasting	221	менее секунды
ARMA	205	менее секунды
Exponential Smoothing	215	менее секунды

На долгосрочном прогнозировании алгоритмы Local Forecasting, ARMA и Exponential Smoothing работают значительно лучше, чем SSA. При этом алгоритм ARMA является масштабируемым на дополнительные экзогенные переменные при небольшом увеличении вычислительной сложности.

По совокупности всех критериев, указанных в подразделе 2.1 (в частности, точность, эффективность и масштабируемость, а также его сходство с алгоритмом, используемым Индустриальным партнером) алгоритм ARMA был выбран в качестве базового метода прогнозирования грузоперевозок.

В качестве основного исследуемого метода, сравниваемого с методом ARMA прогнозирования данных грузоперевозок, был выбран разработанный ранее в ВЦ РАН алгоритм hist. Этот алгоритм основан на модификации алгоритма квантильной регрессии и вычисления эмпирической функции распределения [37, 38]. В частности, в предположении о локальной стационарности временного ряд (выполнено условие Дики-Фуллера [39]), строится гистограмма распределения его значений и вычисляется свертка гистограммы с экспертно заданной функцией потерь

для каждого возможного прогнозируемого значения. При этом рассматривается несимметричная функция потерь. Оптимальным прогнозом является то значение центра сегмента гистограммы, которое доставляет минимальное значение свертке. Для разработанного алгоритма показано, что средняя ошибка прогноза не падает с увеличением длины предыстории. Также проверена его применимость на практике к прогнозированию нестационарных временных рядов. Показано, что алгоритм применим к прогнозированию нестационарных временных рядов, хотя ошибка прогнозирования для нестационарных рядов в среднем выше, чем для стационарных.

Сравнение прогнозов на неделю и на месяц с помощью алгоритмов ARMA и hist при различных функциях ошибки приведены в таблице 2.3. В этой таблице для временных рядов о прибытии вагонов на различные станции приводятся результаты проверки рядов на стационарность (слобец «N/S»: 0 – ряд стационарен, 1 – нет) значения усредненных ошибок алгоритмов при прогнозировании на неделю и на месяц. Для алгоритма hist рассмотрены варианты свертки гистограммы с различными функциями ошибок: сумма квадратов невязок (SSE), сумма абсолюных значений невязок (abs). Кроме того, рассмотрен вариант экспертно заданной трапециевидной функции потерь (trap).

Таблица 2.3 – Сравнение методов ARMA и hist для прогнозирования грузоперевозок

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
1	0	NaN	0,05	0,05	0,05	NaN	0,05	0,05	0,05
2	1	NaN	34,63	33,23	33,46	NaN	34,05	33,02	33,28
3	0	NaN	0,33	0,29	0,29	NaN	0,34	0,30	0,30
4	0	NaN	20,88	19,87	20,02	NaN	20,75	20,11	19,89
5	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
6	0	NaN	0,52	0,38	0,38	NaN	0,53	0,38	0,38
7	0	NaN	0,58	0,38	0,38	NaN	0,58	0,38	0,38
8	0	NaN	0,63	0,42	0,42	NaN	0,64	0,41	0,41
9	1	12,92	8,27	5,88	12,76	15,43	8,94	6,04	13,50
10	0	NaN	1,23	1,01	6,62	2,26	1,26	1,02	7,07
11	0	NaN	0,40	0,26	0,26	NaN	0,40	0,26	0,26
12	0	NaN	3,25	3,51	6,88	NaN	3,26	3,57	7,04
13	0	NaN	0,87	0,79	0,79	NaN	0,88	0,79	0,79
14	0	4,33	2,43	2,31	2,64	4,80	2,42	2,34	2,43

Продолжение таблицы 2.3

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
15	0	NaN	0,69	0,49	0,49	1,22	0,71	0,51	0,51
16	0	NaN	0,20	0,15	0,15	NaN	0,20	0,15	0,15
17	0	NaN	0,54	0,41	0,41	NaN	0,52	0,39	0,39

18	0	15,00	8,33	7,47	10,19	16,22	8,45	7,56	10,23
19	1	30,10	19,45	16,89	17,77	34,86	20,53	17,77	18,71
20	0	2,12	1,42	1,17	7,01	2,37	1,43	1,17	7,44
21	0	NaN	0,19	0,12	0,12	NaN	0,20	0,12	0,12
22	0	NaN	0,44	0,33	0,33	NaN	0,45	0,34	0,34
23	0	2,85	1,66	1,33	1,33	2,87	1,67	1,33	1,30
24	0	2,24	1,22	1,11	1,21	2,39	1,21	1,11	1,21
25	0	2,33	1,35	1,18	1,26	2,60	1,36	1,18	1,28
26	0	1,61	0,97	0,65	0,65	1,79	0,98	0,66	0,66
27	0	1,25	0,78	0,53	0,53	1,39	0,79	0,55	0,55
28	0	2,74	1,56	1,09	1,09	2,83	1,56	1,08	1,08
29	0	NaN	0,32	0,23	0,23	NaN	0,33	0,23	0,23
30	0	6,66	3,90	3,57	9,54	7,48	3,99	3,63	9,77
31	0	2,92	1,56	1,51	1,99	3,19	1,55	1,52	2,04
32	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
33	0	6,38	3,60	3,01	3,92	6,96	3,61	2,90	4,00
34	0	1,85	0,94	0,89	0,87	1,95	0,92	0,88	0,86
35	0	0,99	0,67	0,47	0,47	1,04	0,67	0,47	0,47
36	0	1,77	1,07	0,94	0,95	1,83	1,04	0,93	0,93
37	0	1,25	0,71	0,55	0,55	1,34	0,72	0,56	0,56
38	0	NaN	0,55	0,51	0,51	NaN	0,52	0,50	0,50
39	0	NaN	5,63	5,16	10,09	NaN	5,93	5,30	10,45

Сравнение проводилось для рядов, содержащих не менее 1/5 ненулевых значений от количества всех значений ряда. При большем количестве ненулевых значений ARMA работает некорректно (в таблице 2.3 результат прогнозирования в этих случаях обозначен как NaN).

Проведенные исследования показали, что критерию согласованности с физическими ограничениями отвечает большинство расмотренных алгоритмов. В то же время, критерию согласованности прогнозов не соответствует ни один из них, что приводит к необходимости разработки процедуры согласования. Критерию минимальной ошибки при ретроспективном прогнозировании лучше всего соответствует алгоритм hist.

2.3 Методы выявления, учёта влияния и прогнозирования экзогенных факторов

2.3.1 Выявление экзогенных факторов

Методы получения и обработки экспертных оценок и высказываний [40–45] с середины 1960-х гг. (метод PATTERN – Planning Assistance Through Technical Evaluation Relevance) успешно используются для поддержки принятия решений, касающихся управления сложными слабоструктурированными объектами.

Как известно, существует два вида экспертиз: управлнческие и оценочные. В контексте рассматриваемой проблемы выявления выявления экзогенных факторов, влияющих на объёмы спроса на ГЖДП, речь может идти только об оценочных экспертзах.

Существуют следующие основные виды оценочных экспертиз:

- количественные оценки объектов экспертизы (назначение объектам экспертизы количественных показателей, параметров, характеристик и т. п.);
- качественная кластеризация – отнесение объектов экспертизы к тому или иному виду или классу по некоторому выбранному качественному критерию;
- качественное ранжирование объектов экспертизы по тому или иному качественному критерию;
- рейтингование – определение сравнительной предпочтительности объектов экспертизы в тех или иных количественных шкалах;
- иные подходы числовой или нечисловой природы, используемые для оценивания объектов экспертизы.

При экспертном выявлении экзогенных факторов, влияющих на объёмы спроса на ГЖДП, в силу пока слабой структурированности этой проблемы, речь может идти только о качественной кластеризации экзогненных факторов путём попарных сравнений, в результате которой каждому эндогенному фактору $n \in \mathcal{N}$, характеризующему спрос на объёмы ГЖДП, будет поставлен в соответствие некоторый набор $\mathcal{K}_n$ экзогенных факторов, которые по мнению экспертов так или иначе влияют на эндогенный фактор n.

Следует отметить, что в результате проведения такой экспертизы наборы $\mathcal{K}_n$ для некоторых n могут оказаться пустыми, и, наоборот, некоторый экзогенный фактор $k \in \mathcal{K}$ может не попасть ни в один из наборов $\mathcal{K}_n$.

Последующий анализ достоверности экспертных оценок и характера влияния статистических свойств k-го, $k \in \mathcal{K}_n$ экзогенного фактора на прогнозы n-го эндогенного для экспериментального использования на втором этапе проекта (п. 2.1.6 Технического задания) будет проведён специально адаптированным для этого методом выявления причинно-следственных связей между временными рядами $\mathbf{x}_k(\mathrm{T})$ и $\mathbf{s}_n(\mathrm{T})$. Метод и результаты его тестирования описаны в подразделе 3.6.

2.3.2 Учёт влияния экзогенных факторов

Процедура выявления причинно-следственных связей может быть рассмотрена как построение логических заключений на основе некоторого определения понятия причинно-следственной связи. Существуют три способа выявления причинно-следственных связей:

- индуктивный,
- дедуктивный,
- абдуктивный.

Индуктивный способ выявления связей состоит в построении моделей причинно-следственных отношений, лежащих в основе рассматриваемого явления.

Дедуктивный способ выявление связей объясняет и предсказывает явления на основе измеряемых реальных данных и модели причинно-следственных связей.

Абдуктивное способ выявление связей заключается в построении эмпирических выводов о причинах явления на основе модели причинно-следственных отношений.

В работе [46] выделены следующие основные компоненты системы, выявляющей причинно-следственные отношения:

- Данные об исследуемом явлении.
- Математическое определение понятия причинно-следственной связи (causation theory) – описание условий, выполнение которых необходимо для принятия решения о наличии причинно-следственной связи между рассматриваемыми объектами.
- Способ обучения. При фиксированном определении понятия причинно-следственной связи способ обучения определяет потенциальные связи между объектами, которые ложатся в основу модели рассматриваемого явления. (Процесс обучения рассматривается как процедура поиска модели).
- Результатом обучения становится модель причинно-следственных связей явления (causal model). Возможности дальнейшего применения полученной модели зависят от способа определения понятия причинно-следственной связи, использованного при ее построении. Необходимо отметить, что хотя некоторые модели являются более общими, чем другие, ни одна из них не является абсолютно адекватной

для всех задач связанных с обнаружением причинно-следственных связей и всех возможных областей.

Таким образом, проблема выявления влияния экзогенных факторов на эндогенные заключается в выборе способа наиболее адекватного решаемой задаче определения понятия причинно-следственной связи.

Цель выявления причинно-следственной связи в этом ПНИ – повышение качества прогнозирования эндогенных факторов посредством включения в прогностическую модель экзогенных факторов. По этой причине в качестве базового метода был выбран тест Грейнджера, проверяющий временные ряды на наличие связи в смыле улучшения прогнозируемости временного ряда эндогенных факторов с учетом временных рядов экзогенных факторов.

При выборе базового метода был рассмотрен ряд распространённых подходов к выявлению причинно-следственных связей между статистическими временными рядами. Большой популярностью в современной литературе, посвященной этой проблеме, пользуется моделирование структурных уравнений (structural equation modelling) [13, 14]. В рамках описанного там подхода начальная модель строится, исходя из априорных предположениях о структуре причинно-следственных связей между измеренными временными рядами и рядами, значения которых по каким-либо причинам на начальном этапе неизвестны, затем модель определённым образом модифицируется в соответствии с реальными данными. Очевидным недостатком этого метода является необходимость в первичных предположениях о структуре связи.

В работе [15] предложен другой подход: сначала посредством статистического сравнени исследуемых временных рядов делается предварительный вывод о наличии причинно-следственной связи между ними, а затем этот вывод проверяется на временных рядах того же характера, но измеренных в других условиях. Этот подход использован в подразделе 3.4. для расширения метода Грейнджера.

Метод сходящегося перекрестного отображения (convergent cross mapping, CCM) [16, 17], позволяет определить, принадлежат ли временные ряды одной динамической системе. Этот метод был разработан для выявления причинно-следственных связей в случаях, когда тест Грейнджера неприменим или не может обнаружить связи. Метод основан на преобразовании пространства состояний системы, и срав-

нении ближайших соседей одной и той же точки в преобразованных системах и заключается в проверке сходимости коэффициента корреляции между спрогнозированными и исходными значениями исследуемого ряда при увеличении объема выборки. Этод метод использован в пункте 3.5.1 как альтернативный базовому, так как является комплементарным к нему в смыле области применимости.

Отметим, что рассматриваемые методы используются для получения бинарной оценки степени влияния экзогенных факторов, то есть лишь для ответа на вопрос о наличии влияния. При этом ранжирование факторов по степени влияния не производится, поскольку предложенные в проекте методы прогнозирования автоматически вычисляет степень влияния на основе оценок весов факторов.

2.4 Направления дальнейших исследований и разработок

По результатам исследований, проведенных в подразделе 2.2, в качестве основы для разработки метода непараметрического прогнозирования объемов спроса на ГЖДП, учитывающего влияние экзогенных факторов, а также специфику бизнес-процессов и нормативов РЖД выбран разработанный в ВЦ РАН алгоритм гистограммного прогнозирования hist.

Основным преимуществом алгоритма hist и его вариаций по сравнению с методами, рассмотренными в предыдущем подразделе, является его применимость для прогнозирования стационарных временных рядов с большим количеством одинаковых значений, в том числе и нулевых.

Однако для промышленного использования алгоритма hist в прогнозировании объемов спроса на ГЖДП, учитывающего влияние экзогенных факторов на объемы спроса, необходимы доработки, направленные на повышение его эффективности.

Одно из направлений доработки связано с расширением функциональности алгоритма в области построения и обработки гистограм. При этом будут использованы хорошо зарекомендовавшие себя подходы и методы, изложенные в [47–52].

На втором этапе ПНИ согласно Техническому заданию будет проведены исследование и разработка математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки. При этом в исследованиях и разработке должно учитываться влияние экзогенных факторов на объемы спроса.

Для корректного учета экзогенных факторов и повышения стабильности гисторамного метода прогнозирования в рамках п. 2.1.5.5 Технического задания будут разработаны методы пространственного и временного агрегирования значений рядов [47]. В разработке будет использовано понятие «центральной» гистограммы, минимизирующей расстояние до гистограмм других временных рядов, подлежащих агрегированию [48]. Кроме того, в рамках п. 2.1.5.5 Технического задания будут рассмотрены ранговые гистограммы [49] и их применение для оценки ансамблевых прогнозов, в частности, для определения надежности ансамблевых прогнозов.

Для повышения вычислительной эффективности, будет разработаны метод оптимизации длины предыстории. С этой целью будет проведено исследование зависимости средней ошибки прогнозирования от длины предыстории. Будет вычислено расстояние Кульбака-Лейблера [50] между парой распределений, построенных по наборам точек, отличающихся на 10 точек. При выборе меры близости временных рядов для последующего решения задачи агрегированного согласования, будет проведено исследование и анализ применимости различных коэффициентов, характеризующих меру близости двух распределений друг к другу. Такие коэффициенты в различных источниках называются расстоянием между распределениями [50], мерами разделяющей информации [51], мерами статистического расстояния [52]. В работе [52] приведено сравнение многих известных расстояний с точки зрения скорости сходимости эмпирического распределения к истинному, а также качественного поведения функции расстояния при сходимости. В работе [50] описан метод порождения коэффициентов «непохожести» двух распределений и показано, что многие известные функции расстояния могут быть получены этим методом.

Так как математическая модель прогнозирования объемов спроса на грузовые железнодорожные перевозки должна учитывать различные экзогенные факторы, описанные в подразделе 2.1, а также специфику РЖД (регламенгты процессов и нормативы в области грузовых железнодорожных перевозок и др.), в рамках п. 2.1.5.3 Технического задания будет разработана модификация алгоритма hist, которая позволит учитывать влияние экзогенных факторов. Для получения более устойчивых оценок совместных и условных гистограмм предполагается использовать современные методы [53–59] уточнения многомерных гистограмм по наборам

маргинальных гистограмм.

Для дальнейшего повышения качества прогнозирования, в рамках п. 2.1.5.5 Технического задания будет также добавлена функциональность, которая позволит учитывать иерархическую структуру временных рядов, основанную на разбиении на географические районы и типы товаров или грузов. Такое разбиение приводит к необходимости согласования прогнозов по структуре иерархии, чтобы гарантировать совпадение сумм прогнозов рядов нижних уровней с прогнозами рядов верхних уровней. В современной литературе [1)60, 61] для соглазования прогнозов применяется подход Bottom-up: значение суммарного прогноза определяется путем прогнозирования временного ряда, полученного путем суммирования значений всех рядов для различных грузов.

В рамках п. 2.1.5 Технического задания будет разработано обобщение алгоритма согласования из [1)60] с сохранением его преимуществ и устранением недостатков. Для согласования прогнозов будет использован алгоритм GTOP (Game-theoretically Optimal reconciliation), основанный на идеях из [1)62]. Алгоритм GTOP не требует оценки погрешности независимых прогнозов временных рядов, не требует несмещенности независимых прогнозов и имеет теоретическое обоснование улучшения качества прогнозов после проведения согласования. Задача согласования прогнозов рассматривается как поиск равновесия Нэша [1)63] в антагонистической игре игрока, выбирающего согласованные прогнозы, с природой, которая выбирает действительные значения временных рядов, и сводится к решению оптимизационной задачи с ограничениями типа равенства и неравенства. Вид равновесия Нэша задает параметры оптимизационной задачи.

3 Выявление и исследование экзогенных факторов, а также исследования их влияния на объемы спроса на грузовые железнодорожные перевозки

В этом разделе представлены результаты исследования экзогенных факторов и их влияния на объёмы спроса на ГЖДП.

3.1 Проведение экспертного анализа значимости и характера влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки

Для выявления экзогенных факторов было привлечены эксперты из Российской открытой академии транспорта Московского государственного университета путей сообщения.

Работа с экспертами проводилась в два этапа:

Этап 1. Индивидуальные опросы экспертов на предмет выявления собственно экзогенных факторов, влияющих на объемы спроса на ГЖДП (задача классификации), и качественных оценок значимости и характера влияния этих экзогенных факторов на эндогенные факторы (задачи кластеризации посредством попарных сравнений).

Этап 2. Формирование консолидированных групповых оценок методом «круглого стола».

В таблице 3.1 приведены зкзогенные факторы, которые с точки зрения экспертов оказывают влияние на объем грузоперевозок. Для каждого фактора перечислены группы груза, соответствующие этим факторам, и высказывания экспертов о характере влияния. Экспертные высказывания носят качественный характер и представлены в номинальных (оказывает данный фактор влияние на загруженность железнодорожного узла или нет) и ранговых шкалах (степень влияния – высокая, низкая, фактор не оказывает влияния).

Таблица 3.1 – Перечень экзогенных факторов, выделенных экспертами

Индекс в $\mathcal{K}$	Вид фактора, влияющего на объем грузоперевозок	Группы грузов и отрасли, на которые оказывается влияние	Степень и характер влияния
1	Мировые и внутренние цены на соответствующие товары.	Нефть и нефтепродукты, черные металлы, цветные металлы, удобрения, уголь.	На экспортные перевозки влияние сильное. Связь бывает как прямой, так и обратной.
2	Курс рубля к доллару США.	Грузы, отправляемые на экспорт (нефть и нефтепродукты, металлы, уголь).	Степень влияния для экспортных перевозок зачастую высокая.
3	Сезонность производства природно-климатического характера.	Зерно, овощи, бахчевые культуры.	Степень влияния высокая. Динамика перевозки связана со сбором урожая. Повышенный объем перевозок приходится на июль, август и сентябрь, иногда на октябрь.
4	Сезонность спроса на продукцию.	Строительные грузы (щебень, кирпич, цемент, промышленное сырьё).	Степень влияния высокая. Динамика перевозки связана с сезоном строительных работ. Пик приходится на летние месяцы.
5	Особенности технологического цикла в различных отраслях.	Различные грузы. Например, в соледобыче в период «соленавигации» высокий уровень добычи, а в остальные периоды зависит от величины складских запасов.	Во многих случаях являются ограничителем (как верхней, так и нижней границей) объема погрузки грузов.
6	Сезонность, связанная с навигацией.	Грузы, которые перевозятся водным транспортом.	Степень влияния высокая для производств расположенных в пределах досягаемости судоходных путей.
7	Производственные мощности заводов и других предприятий, производящих товарную продукцию.	Добывающие отрасли (угледобыча, добыча соли, камня гипсового, добыча нефти); нефтеперерабатывающие, нефтехимические, металлургические производства.	Во многих случаях являются верхней границей объема погрузки грузов.

Продолжение таблицы 3.1

8	Рост производственных мощностей предприятий-грузоотпра-вителей и/или спроса на товарную продукцию со стороны предприятий-грузополучателей.	Все грузы.	Степень влияния – высокая в случаях низких мощностей станций погрузки или низких перерабатывающих способностей грузовых фронтов.
9	Запасы готовой продукции на складах предприятий.	Все грузы.	Во многих случаях динамика этого показателя являются мерой спроса на продукцию данного предприятия, а спрос в свою очередь оказывает влияние на объемы производства.
10	Доли погрузки предприятия, приходящиеся на другие виды транспорта.	Все грузы, перевозка которых возможна альтернативными видами транспорта.	На сверхдальних расстояниях перевозок – влияние низкое, для массовых грузов – низкое; для остальных грузов, особенно на короткие и средние расстояния – высокое.
11	Экспортные (импортные) пошлины.	Все экспортируемые или импортируемые грузы.	Степень влияния – высокая как для экспортных и импортных перевозок, так и для внутрироссийских.

Далее, как и в общем описании проблемы (см. подраздел 2.1) индексно-упорядоченное множество имён экзогенных факторов, указанных в этой таблице, как и множество их индесов, будет обозначаться через $\mathcal{K}$.

Всем указанным в таблице 3.1 экзогенным факторам при построении для них временных рядов могут быть поставлены в соответствие некоторые количественные показатели: стоимостные, весовые, объёмные и т. п., а для «сезонных» – количественные идентификаторы (например, весна – 1, лето – 2, осень – 3, зима – 4).

Таким указанным экспертами экзогенным факторам, как, например, «Ограничения экспорта тех или иных родов груза (например, запрет экспорта зерна в 2010 г.) решениями государственных органов» и «Конвенционные запреты и ограничения РЖД» поставить в соответствие количественные показатели и, следовательно, построить временные ряды не представляется возможным. Поэтому учёт

влияния таких факторов в рамках статистического подхода, используемого в этом ПНИ, производиться не будет.

В качестве эндогенных факторов, влияние на которые перечисленных в таблице 3.1 экзогенных факторов следует учитывать при прогнозировании спроса на ГЖДП, эксперты указали 43 группы грузов, представленных в таблице 3.2.

Таблица 3.2 – Выделенные группы грузов

Индекс в $\mathcal{N}$	Группа грузов	Индекс в $\mathcal{N}$	Группа грузов
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	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
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
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
19	Строительные грузы	41	Импортные грузы
20	Промышленное сырье и формовочные материалы	42	Грузы в контейнерах
21	Шлаки гранулированные	43	Остальные и сборные грузы
22	Огнеупоры		

Качественные оценки экспертов о характере влияния экзогенных факторов на эндогенные сведены в таблицу 3.3. Зеленым цветом и значением «+» выделены

клетки, соответствующие парам факторов, относительно связи между которыми была высказана экспертная гипотеза. Пары факторов, к которым экспертные гипотезы относятся неявно, помечены желтым цветом и значением « ? ». Синим цветом и значением « ± » отмечены пары, принадлежность которых к одной из описанных ситуаций возможна, но не может быть установлена без привлечения дополнительной экспертной информации и/или статистической обработки. Для пар факторов, о возможной связи между которыми никаких гипотез высказано не было, клетки цветом не помечены и указано значение « – ».

Таблица 3.3 – Характер влияния экзогенных факторов

N \ *K*	1	2	3	4	5	6	7	8	9	10	11
1	+	+	–	–	±	±	+	+	+	±	±
2	+	–	–	–	±	±	+	+	+	±	±
3	+	+	–	–	±	±	+	+	+	±	±
4	+	–	–	–	±	±	+	+	+	±	±
5	+	–	–	–	±	±	+	+	+	±	±
6	+	+	–	–	±	±	+	+	+	±	±
7	+	+	–	–	±	±	+	+	+	±	±
8	+	+	–	–	±	±	+	+	+	±	±
9	+	+	–	–	±	±	+	+	+	±	±
10	?	?	–	–	±	±	+	+	+	±	±
11	+	?	–	–	±	±	+	+	+	±	±
12	?	–	–	–	±	±	+	+	+	±	±
13	?	–	–	–	±	±	+	+	+	±	±
14	?	–	–	–	±	±	?	+	+	±	±
15	?	–	–	–	±	±	?	+	+	±	±
16	+	–	–	–	±	±	+	+	+	±	±
17	?	–	–	–	±	±	?	+	+	±	±

Продолжение таблицы 3.3

$\mathcal{N}$ \ $\mathcal{K}$	1	2	3	4	5	6	7	8	9	10	11
18	–	–	–	–	±	±	–	+	+	±	±
19	–	–	–	–	±	±	–	+	+	±	±
20	–	–	–	+	±	±	–	+	+	±	±
21	–	–	–	?	±	±	–	+	+	±	±
22	–	–	–	–	±	±	+	+	+	±	±
23	–	–	–	+	±	±	–	+	+	±	±
23–27	–	–	–	–	±	±	–			±	±
28	–	–	+	–	+	±	–	+	+	±	±
29	–	–	+	–	+	±	–	+	+	±	±
30	–	–	+	–	?	±	–	+	+	±	±
31	–	–	+	–	?	±	–	+	+	±	±
32	–	–	+	–	?	±	–	+	+	±	±
33	–	–	+	–	?	±	–	+	+	±	±
34	–	–	+	–	?	±	–	+	+	±	±
35	–	–	+	–	?	±	–	+	+	±	±
36–39	–	–	–	–	±	±	–			±	±
40	–	–	–	–	±	+	–	+	+	+	±

3.2 Разработка и обоснование метода выявления причинно-следственных связей в разнородных временных рядах объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов

В этом подразделе представлены результаты разработки метода выявления причинно-следственных связей в разнородных временных рядах объемов спроса на ГЖДП и экзогенных факторов.

Для выявления причинно-следственных связей в разнородных временных

рядах объемов спроса на ГЖДП и экзогенных факторов адаптированы математические методы выявления причинно-следственных связей [64, 65].

3.2.1 Метод выявления причинно-следственных связей в разнородных временных рядах

Подход Грейнджера, принятый в работах [1)64, 1)65], основан на сравнении прогноза исследуемого временного ряда исключительно по его истории с прогнозом, сделанным с учётом информации о других временных рядах. В случае, если улучшение прогноза подтверждается, говорят [64] о G-зависимости (зависимости по Грейнджеру) исследуемого временного ряда от других временных рядов.

Пусть в прмежутке дискретизированного времени [0, T) заданы два временных ряда:

$$\mathbf{s}(\mathrm{T}) = \left\| s_j \right\|_{j=1}^{\mathrm{T}}, \ s_j \in \mathbb{Q} \tag{3.1}$$

и

$$\mathbf{x}(\mathrm{T}) = \left\| x_j \right\|_{j=1}^{\mathrm{T}}, \ x_j \in \mathbb{Q}, \tag{3.2}$$

где s_j и x_j, – значения соответственно некоторого эндогенного и экзогенного фактора в начале j-го полуинтервала времени j = 1, ..., T.

Требуется определить, является ли временной ряд (3.1) G-зависимым от ряда (3.2).

Процедура применения метода. Пусть **s** и **x** – исследуемые временные ряды, тогда линейная регрессионная модель имеет вид:

$$\hat{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \sum_{t=1}^{\tau} b_t x_{j-t} + \varepsilon_j, \ \text{для} \ j > \tau, \ j = 1, ..., \mathrm{T}, \tag{3.3}$$

$$\hat{x}_j = \sum_{t=1}^{\tau} c_t s_{j-t} + \sum_{t=1}^{\tau} d_t x_{j-t} + \xi_j, \ \text{для} \ j > \tau, \ j = 1, ..., \mathrm{T}, \tag{3.4}$$

где τ – количество предыдущих значений, принимаемых во внимание, a_t, b_t, c_t, d_t, – веса учитываемых объектов, а ε_j и ξ_j – ошибки прогнозирования.

Считается, что временной ряд **x** влечет за собой временной ряд **s**, если модуль ошибки $\varepsilon_j = s_j - \hat{s}_j$ прогнозирования ряда **s** уменьшается при включении в модель значений ряда **x**. Для проверки значимости уменьшения ошибки при добавлении в модель ряда **x** в методе Грейнджера используется статистика Фишера.

Пусть для ряда **s** построена модель (3.3), учитывающая историю ряда **x**, и – линейная регрессионная модель, основанная только на истории **s**:

$$\tilde{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \tilde{\varepsilon}_j, \quad \text{для } \tau < j < \mathrm{T},\ j = 1, ..., \mathrm{T}. \tag{3.5}$$

Тогда применение теста Грейнджера сводится к вычислению статистики Фишера:

$$F = \frac{\mathrm{RSS}_{\tilde{s}} - \mathrm{RSS}_{\hat{s}}}{\mathrm{RSS}_{\hat{s}}} \times \frac{\mathrm{T} - 2\tau}{\tau}, \tag{3.6}$$

где

$$\mathrm{RSS}_{\hat{s}} = \sum_{j=1}^{\mathrm{T}} \varepsilon_j^2 = \sum_{j=1}^{\mathrm{T}} (s_j - \hat{s}_j)^2,$$

$$\mathrm{RSS}_{\tilde{s}} = \sum_{j=1}^{\mathrm{T}} \tilde{\varepsilon}_j^2 = \sum_{j=1}^{\mathrm{T}} (s_j - \tilde{s}_j)^2,$$

и сравнению ее с критическим значением при заданном уровне значимости.

Нулевая гипотеза заключается в предположении, что ряд **x** не оказывает влияния на значения ряда **s**. При нулевой гипотезе F принадлежит распределению Фишера со степенями свободы τ и $\mathrm{T} - 2\tau$. Полученная вероятность отклонить нулевую гипотезу $1 - p(\mathbf{x}, \mathbf{s})$, где $p(\mathbf{x}, \mathbf{s})$ – критическое значение F-статистики, интерпретируется как наличие причинно-следственной связи между рядами **x** и **s**.

Исследование данных о влиянии цен основных биржевых инструментов на объемы перевозок соответствующих групп грузов приведено в подразделе 3.6.

Для оценки качества прогнозирования выделяется контрольная выборка длинной m и строится функционал:

$$Q = \sum_{i=1}^{m} (\tilde{s}_i - s_i)^2, \tag{3.7}$$

где $\tilde{s}_i$ – спрогнозированное значение элемента выборки, s_i – его истинное значение.

Учитывая, что тест Грейнджера применим к стационарным рядам, то есть обладающим не зависящими от времени математическим ожиданием и дисперсией, прежде всего, необходимо удостовериться, что исследуемые ряды – стационарны [66]. Для оценки этого свойства можно использовать ADF-тест, основанный на нулевой гипотезе об отсутствии стационарности или, дополнительно к ADF, KPSS-тест, предполагающий ее наличие. Если проверка даёт отрицательный результат, ряды необходимо модифицировать. Для этого существуют различные способы:

- дифференцирование ряда, то есть переход от непосредственно значений к их изменениям. Эта операция увеличивает вероятность успеха и может повторяться

несколько раз, однако каждая итерация затрудняет интерпретацию полученных данных, поэтому в этой работе дифференцирование проводится лишь один раз. С данными, для которых однократное дифференцирование не приводит к положительному результату (то есть не удается получить стационарный ряд), алгоритм не работает.

- «метод окна» – то есть использование лишь часть известного ряда. Метод основан на следующей идее: что чем короче ряд тем больше он похож на стационарный.

Кроме стационарности в самом определении G-зависимости, данном Грейнджером, заключено предположение о «разделимости» исследуемых временных рядов, то есть предположение о том, что для ряда **s** можно построить модель (3.5), не учитывая никакой информации о временном ряде **x**. В случае линейной зависимости рядов это условие выполняется, однако в случае сложных динамических систем разделимость, как правило, отсутствует.

Выбор параметров. Ключевым параметром при использовании теста Грейнджера является порядок модели (порядок лагирования), то есть количество предыдущих измерений (значений ряда), учитываемых при прогнозировании очередного значения. Если этот параметр не может быть выбран на основе априорного знания, то могут быть использованы информационные критерии Акаике (Akaike information criterion, AIC) [67] или Байеса (Bayesian information criterion, BIC) [68], позволяющие сравнивать модели с различным количеством параметров.

В этом проекте используется только линейная регрессия, поэтому критерий Акаике может быть представлен в виде:

$$\mathrm{AIK} = 2p + m \times \ln\left(\frac{Q}{m}\right),$$

где m – длина контрольной выборки, p – критическое значение F-статистики, Q – функционал (3.7).

Аналогичный вид при линейной регрессии имеет и байесовский критерий, однако функция штрафа за сложность модели (то есть за ее порядок) в этом случае жестче:

$$\mathrm{BIC} = p \times \ln(m) + m \times \ln\left(\frac{Q}{m}\right).$$

В обоих случаях наилучшей модели соответсвует минимальное значение

критерия. Важно, что модели сравниваются по выборкам одинаковой длины.

Работа алгоритма на модельных данных. В ходе вычислительного эксперимента были использованы синтетические данные, состоящие из пяти рядов $\mathbf{w}_i$, $i = 1, \ldots, 5$, сформированных по следующему алгоритму: в каждом из пяти рядов первые три исходных значения w_{ij}, $j = 0, 1, 2, 3$ – некоторые нормально распределённые случайные величины, по которым значения w_{ij}, $i = 1, 4, 5$, $j > 1$, $i = 2$, $j > 3$ и $i = 3$, $j > 2$, начиная с x_{52}, задаются следующими формулами:

$$w_{1j} = 1{,}6 \times w_{1(j-1)} + 0{,}65 \times w_{2(j-2)}$$
$$w_{2j} = 1{,}5 \times w_{2(j-1)} - 0{,}3 \times w_{2(j-2)} - 0{,}3 \times w_{3(j-4)} + 0{,}6 \times w_{4(j-1)}$$
$$w_{3j} = 1{,}8 \times w_{3(j-1)} - 0{,}7 \times w_{3(j-2)} - 0{,}1 \times w_{5(j-3)}$$
$$w_{4j} = 1{,}5 \times w_{4(j-1)} + 0{,}9 \times w_{3(j-2)} + 0{,}4 \times w_{5(j-2)}$$
$$w_{5j} = 1{,}7 \times w_{5(j-1)} - 0{,}5 \times w_{5(j-2)} - 0{,}2 \times w_{3(j-1)}$$

Все сгенерированные ряды (см. рисунок 3.1) оказались стационарными, дополнительная обработка не потребовалась.

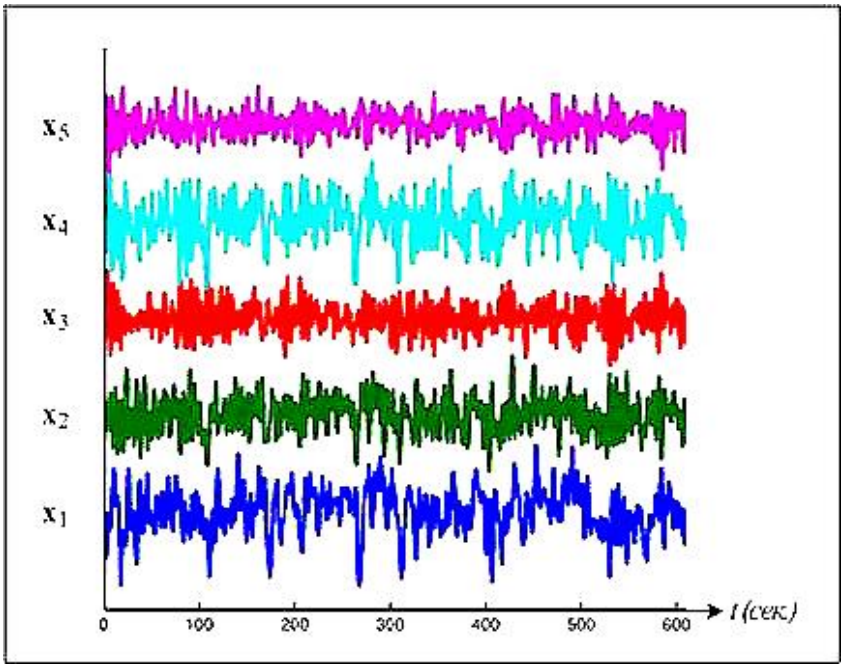


Рисунок 3.1 – Вид исследуемых рядов

Все зависимости, заложенные в исходные данные, в результате вычислительного эксперимента были выявлены (см. рисунки 3.2 и 3.3): ряд $\mathbf{w}_5$ зависит от $\mathbf{w}_1$, изменения рядов $\mathbf{w}_3$ и $\mathbf{w}_4$ влекут за собой изменение $\mathbf{w}_2$, а ряды $\mathbf{w}_3$ и $\mathbf{w}_5$ – взаимнозависимы, критерии AIC и BIC дали оптимальное значение порядка модели, равное 2.

На рисунке 3.2 вершины графа соответствуют номерам рядов, ребро отражает зависимость ряда $\mathbf{w}_{i_2}$ в конце ребра от ряда $\mathbf{w}_{i_1}$ в его начале, дуга (красный цвет) соответствует двусторонней зависимости рядов. На рисунке 3.3 отражена «степень» зависимости: более темные клетки соответствуют более выраженной зависимости рядов, номера которых отложены по оси ординат, от рядов, номера которых отложены по оси абсцисс.

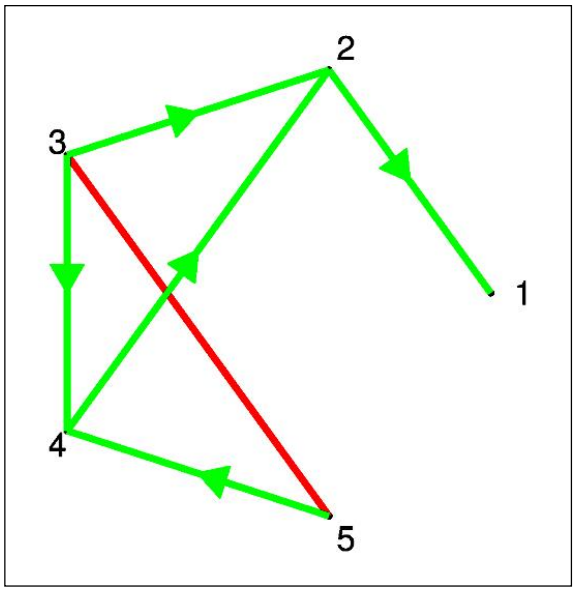


Рисунок 3.2 – Граф зависимостей между рядами

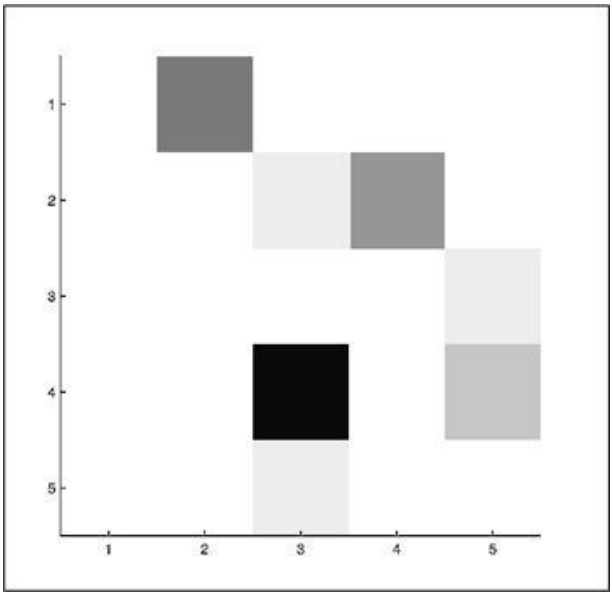


Рисунок 3.3 – Схема «степени» зависимостей между рядами

Работа алгоритма на реальных данных. Вычислительные эксперименты были проведены и на реальных данных – цены на различные виды товаров за каж-

дый месяц с 1960 по 2009 год [69]. В частности, были исследованы два кейса:

Кейс 1 – зависимость цен на сахар в США (на рисунке 3.4 – зелёным цветом) от цен на сахар в Европе (синим цветом).

Кейс 2: - связь между ценами на природный газ в США (на рисунке 3.5 – зелёным цветом) и Европе (синим цветом) и индексом цен на энергию, построенным Всемирным банком (красным цветом).

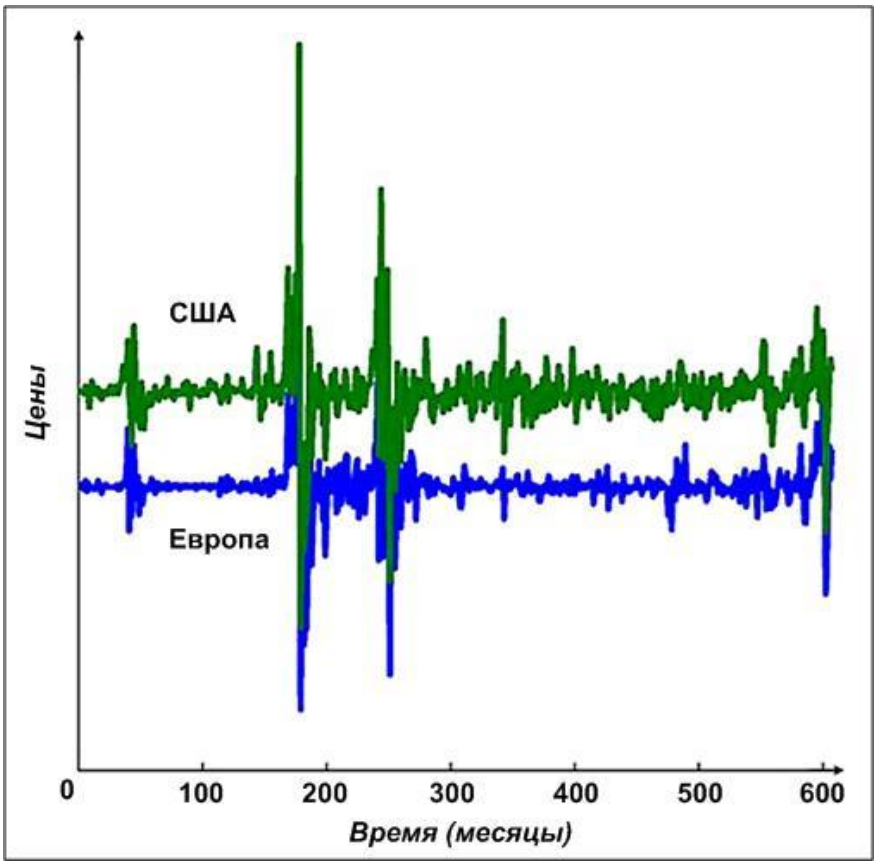


Рисунок 3.4 – Вид временных рядов цен на сахар в Европе и США

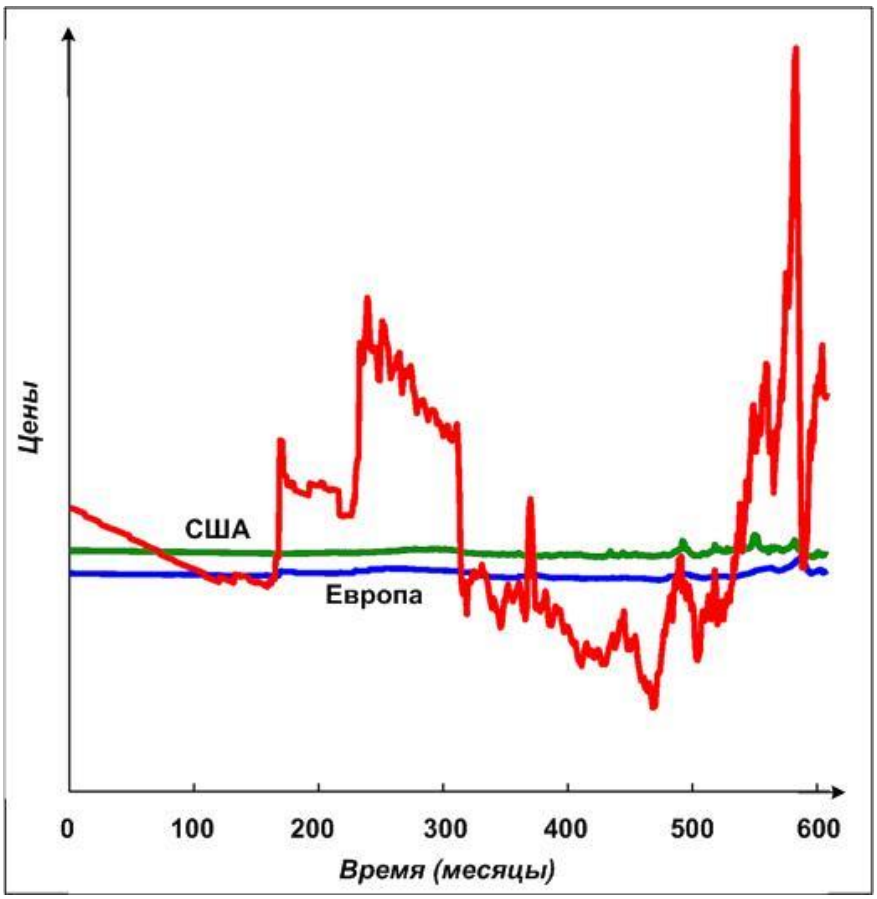


Рисунок 3.5 – Вид временных рядов цен на газ в Европе и США и мировых цен на энергию

В первом кейсе алгоритм выявил зависимость вида «цены на сахар в Европе влияют на цены на сахар в США». В качестве подтверждения правильности этого результата, может служить рисунок 3.6, иллюстрирующий полученные результаты вычислительного эксперимента о зависимости относительных функционалов качества от порядка модели прогноза цен в США с использованием данных о ценах в Европе (на рисунке 3.6 – зелёный цвет) и без учёта европейских цен (синий цвет). Видно, что оценка, полученная с учетом дополнительной информации, более точна.

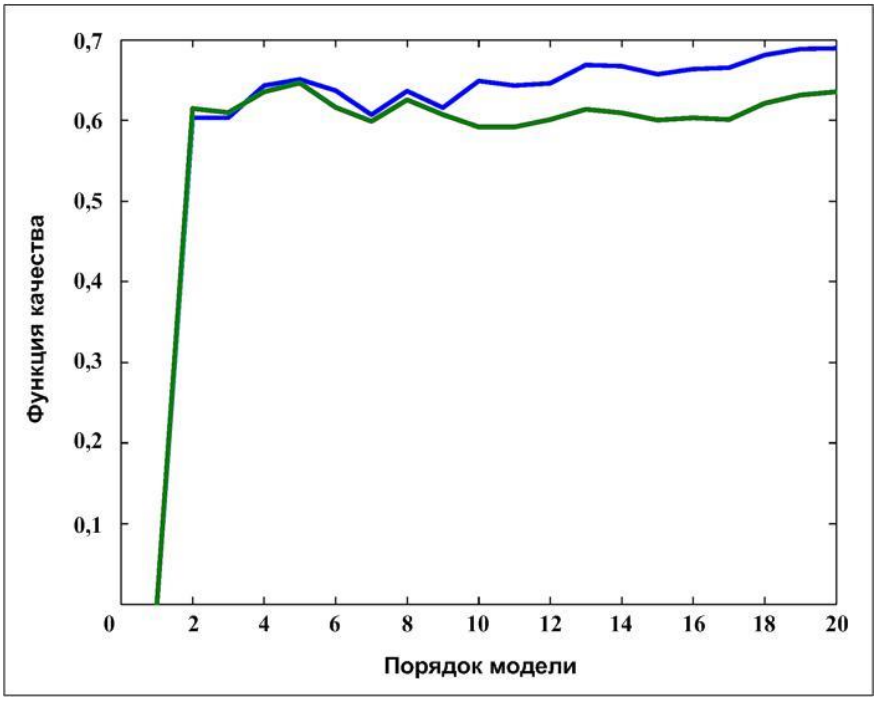


Рисунок 3.6 – Зависимость относительных функционалов качества от порядка модели.

Граф связей, выявленных алгоритмом во втором кейсе, представлен на рисунке 3.7 («1» – цены на природный газ в Европе, «2» – в США, «3» – мировые цены на энергию).

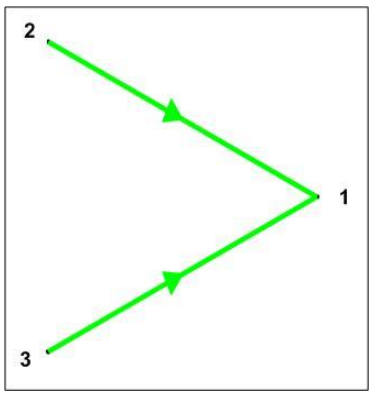


Рисунок 3.7 – Связи между рядами на природный газ в Европе и США с мировыми ценами на энергию

Во втором кейсе временные ряды не являются стационарными, поэтому они в ходе вычислительного эксперимента подверглись модификации. Наиболее действенным способом оказалось дифференцирование. Поэтому на рисунках 3.8 и 3.9 отображены не сами значения исследуемых величин, а их изменения.

На рисунке 3.8 исходный ряд изменений европейских цен на газ представлен зелёным цветом, а их предсказанные значения с учётом динамики цен на газ в США и мировых цен на энергию – синим. Ввиду нестационарности данных, на графиках

отображены не сами значения исследуемых величин, а их изменения.

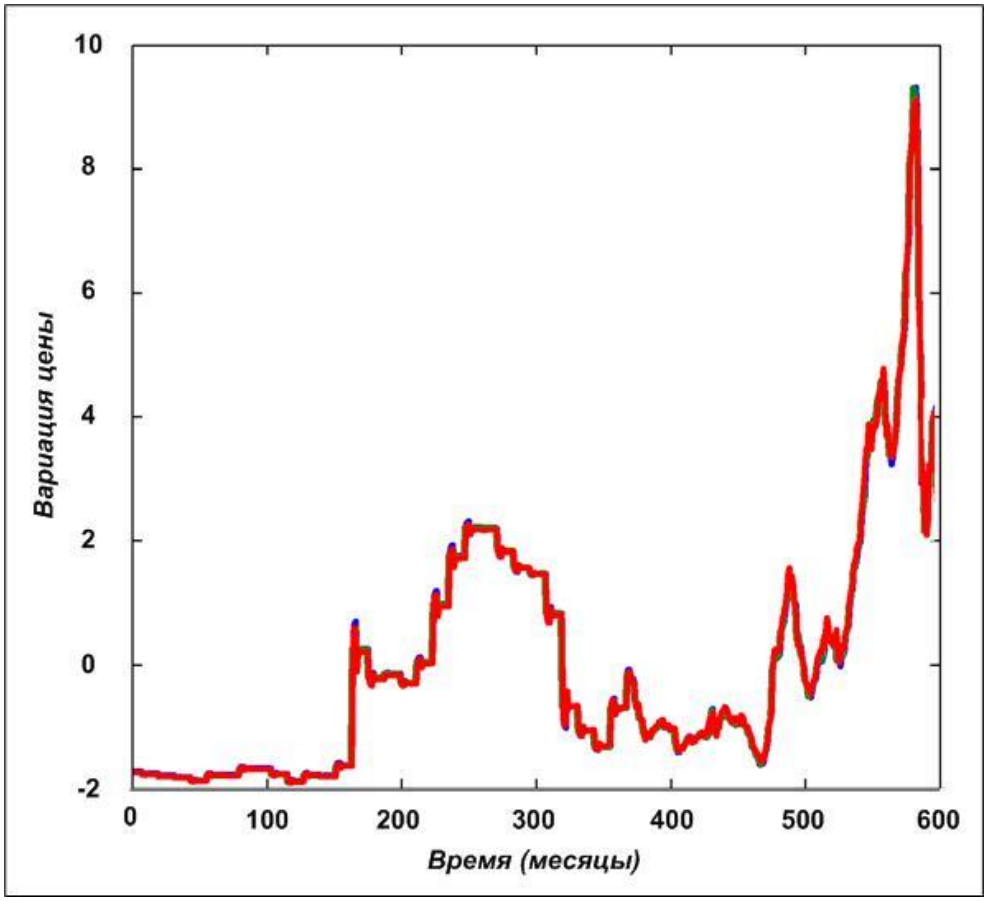


Рисунок 3.8 – Реальная и спрогнозированная динамика европейских цен на природный газ

На рисунке 3.9 в увеличенном для наглядности масштабе (фрагмент только для первых 100 месяцев) показаны: исходный ряд изменений европейских цен на газ (как и на предыдущем рисунке – зелёным цветом), прогноз, построенный только по истории самих европейских цен на газ – красный цвет и прогноз с учётом динамики цен на газ в США и мировых цен на энергию, как и на предыдущем рисунке – синий цвет. Видно, что использование «вспомогательных данных» о ценах в США и мировых ценах на энергию, привело к более точному прогнозу европейских цен.

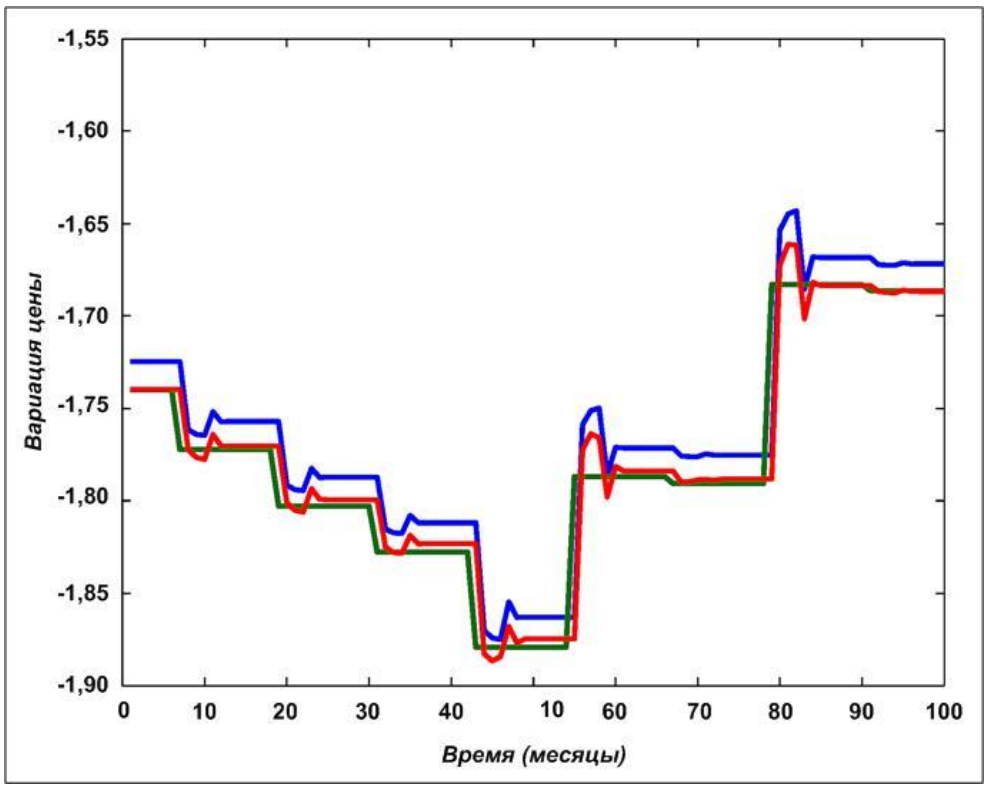


Рисунок 3.9 – Реальная и спрогнозированная динамика европейских цен на природный газ в первые 100 месяцев

Проведённые исследования применимости рассмотренного метода выявления причинно-следственных связей в разнородных временных рядах к заведомо нестационарным временным рядам, дали следующий результат: возможны как точный прогноз, так и полное несоответствие с действительностью. Таким образом, для прогнозирования нестационарных временных рядов необходимо провести модификацию описанного метода. Модифицированный метод построения кросскорреляционных зависимостей временных рядов для выявлени причинно-следственных связей между нестационарными рядами изложен в подразделе 3.5.

3.2.2 Задача оценки достоверности экспертных высказываний

Пусть на отрезке дискретного времени длиной T для некоторого множества точек наблюдения $\mathcal{I}$ (например, набор каких-то железнодорожных станций) заданы два набора временных рядов:

$$\mathbf{S} = \{\mathbf{s}_n^1(\mathrm{T}), \ldots, \mathbf{s}_n^i(\mathrm{T}), \ldots, \mathbf{s}_n^{|\mathrm{I}|}(\mathrm{T})\} \tag{3.8}$$

и

$$\mathbf{X} = \{\mathbf{x}_1(\mathrm{T}), \ldots, \mathbf{x}_k(\mathrm{T}), \ldots, \mathbf{x}_{|\mathrm{K}|}(\mathrm{T})\}, \tag{3.9}$$

где $\mathbf{s}_n^i(\mathrm{T}) = \left\| s_{nj}^i \right\|_{j=1}^{\mathrm{T}}$, $s_{nj}^i \in \mathbb{Q}$ – значение некоторого эндогенного фактора $n \in \mathcal{N}$ в точке $i \in \mathcal{I}$ на j-ом, j = 1, ..., T, полуинтервале времени (например, изменение количества груза n на станции i за j-ый полуинтервал времени), а $\mathbf{x}(\mathrm{T}) = \left\| x_{kj} \right\|_{j=1}^{\mathrm{T}}$, $x_{kj} \in \mathbb{Q}$ –значения экзогенного фактора $k \in \mathcal{K}$ в начале того же полуинтервала времени.

Требуется проверить экспертные высказывания, заданные в таблице 3.3:

$$\mu = \mu(n,k) \in \{«+», «\pm», «?», «–»\}, \tag{3.10}$$

посредством анализа статистической взаимосвязи (3.8) и (3.9) и поставить в соответствие каждому высказыванию значение достоверности.

3.2.3 Метод оценки достоверности экспертных высказываний

Для выделения экзогенных факторов, действительно оказывающих заметное влияние на прогнозируемые временные ряды эндогенных факторов, экспертные оценки подлежат обработке средствами математических методов вывления причинно-следственной связей.

Так как целью данного ПНИ, является в частности, повышение качества прогнозирования влияния экзогенных факторов на эндогенные, был проанализирован ряд наиболее используемых на практике методов выявления причинно-следственных связей между временными рядами. Все они соответствуют различным определениям причинно-следственной связи, отвечающим определенным целям.

Так как ни один из рассмотренных методов ни доказывает наличия причинно-следственной связи между рядами, а лишь проверяет ряды на соответствие некоторому математическому критерию указывающему на отсутствие связи, в качестве базового метода был выбран тест Грейнджера как метод, наиболее адекватно согласующийся с целями оценки достоверности экспертных высказываний.

Как было отмечено в предыдущем подразделе, тест Грейнджера позволяет с вероятностью

$$1 - p(\mathbf{x}_k(\mathrm{T}), \mathbf{s}_n(\mathrm{T})), \tag{3.11}$$

где $\mathbf{x}_k(\mathrm{T})$ и $\mathbf{s}_n(\mathrm{T})$ временные ряды некоторого k-го, $k \in \mathcal{K}$, экзогенного фактора и некоторого n-го, $n \in \mathcal{N}$, эндогенного фактора, T – длина дискретезированного проме-

жутка времени [0, T), на котором заданы значения x_{kj} и s_{nj} соответственно, k-го экзогенного фактора и n-го эндогенного фактора в начальной точке j-го, j = 1, ..., T, полуинтервала времени, а $p(\mathbf{x}_k(T), \mathbf{s}_n(T))$ – критическое значение статистики Фишера (3.6.), позволяет говорить о существовании связи между рядами $\mathbf{x}_k(T)$ и $\mathbf{s}_n(T)$, и, как следствие, о возможном влиянии k-го экзогенного фактора на n-й эндогенный. В качестве дополнительной количественной характеристики, позволяющей повысить или ослабить основанную на (3.11) степень уверенности в наличии такого влияния, может служить получаемый методом сходящегося перекрестного отображения коэфициентом корреляции между перекрестным прогнозом и истинными значениями прогнозируемого временного ряда.

Оценки (3.11), полученные в результате применения описанного метода, служат основой принятия решений о степени достоверности $d_{nk}(T)$ высказываний экспертов (3.10),

$$d_{nk}(T)= (1 - p(\mathbf{x}_k(T), \mathbf{s}_n(T)))\times 100\%, \qquad (3.12)$$

но, строго говоря, только в отношении того промежутка времени [0, T) для которого проводился анализ возможного влияния экзогенных факторов на эндогенные.

Решение о наличии в течение промежутка времени [0, T) влияния экзогенного фактора $k \in \mathcal{K}$ на эндогенный $n \in \mathcal{N}$ принимается при $d_{nk}(T) > 90\%$, при этом в клетку на пересечении n-й строки и k-го столбца таблицы 3.3 ставится «+», в остальные клетки, для которых степень достоверности $d_{nk}(T) \le 90\%$, ставится «–».

В результате функция (3.10) принимает вид:

$$\mu = \mu(n,k) \in \{\text{«+»}, \text{«–»}\}.$$

3.3 Разработка и обоснование методов оценки достоверности экспертных высказываний о влиянии экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки

В этом подразделе описаны способы расширения подхода Грейнджера, которые позволяют расширить описанный в подразделе 3.2 метод выявления причинно-следственных связей в разнородных временных рядах до обнаружения структуры связей между временными рядами экзогенных факторов и объемами спроса.

В рамках метода Грейнджера предполагаетя линейная зависимость между временными рядами. Для выявления связей более сложной структуры требуется

расширение этого метода. Для выявления экзогенных факторов, влияющих на исследуемый временной ряд, и определения структуры связи было предложено использовать информацию о связях между схожими рядами. В проекте рассмотрена задача прогнозирования временного ряда **x** из заданного набора временных рядов, описываемых конечным числом прогностических моделей. Предполагается, что рассматриваемые временные ряды можно разбить на несколько классов, каждый из которых описывается своей прогностической моделью с известной структурой, включающей набор экзогенных переменных, ранее определенный с помощью теста Грейнджера. При известном соответствии моделей и классов временных рядов задача определения структуры модели сводится к задаче классификации временных рядов. Таким образом, задача выявления экзогенных факторов сведена к решению задачи классификации временных рядов [70]. Предложено решение задачи классификации временных рядов с одновременным выбором модели для каждого ряда.

В качестве прогностической модели выбрана криволинейная модель [71, 72]. Такой выбор модели позволяет рассматривать модели различных сруктур, ввести понятие сложности структуры [73], расстояния между рядами [74]. В качестве расстояния между прогностическими моделями рассматривается расстояние между прогнозами временных рядов с помощью этих прогностических моделей. При таком способе сравнения моделей, описание выборки временных рядов и моделей задается четырехиндексной матрицей расстояний между прогнозами временных рядов. Кандидаты на решение задачи выбираются с помощью генетического алгоритма [75, 76]. Каждому решению задачи выбора модели соответствует набор пар «модель f – временной ряд **x**» и двухиндексная матрица парных расстояний между прогнозами $f(\mathbf{x})$ всех временных рядов **x** выбранными моделями f. На ее основе строится кластеризация пар моделей и временных рядов. Предполагается наличие априорной информации о наличии связей между рядами, полученной, например, применением базовой версии теста Грейнджера к исследуемым временным рядам. Для выбора модели задается функционал, характеризующий согласованность результата кластеризации с имеющейся априорной информацией, на основе которого из построенной четырехиндексной матрицы выбирается двухиндексная матрица. Таким образом, предложен алгоритм, позволяющий уточнить информацию о структуре связей для набора временных рядов, полученную применением теста Грейнджера.

Путь задан набор $\Xi = \{\mathbf{x}_i(t)\}_{i=1}^m$ временных рядов

$$\mathbf{x}_i = \{x_i(t)\}_{t=1}^n,$$

где $x_i(t) \in \mathbb{R}$ – отсчет временного ряда $\mathbf{x}_i(t)$ в момент времени t. Для описания временных рядов $\mathbf{x}_i$ фиксируется криволинейную модель:

$$x_i(t) = \sum_{j=1}^k w_j f_j(x_i(t-1)), \quad (3.13)$$

где $f(x_i) = [f_1, \dots, f_k]^T$ – структура модели временного ряда $\mathbf{x}_i$, а коэффициенты w_j – настраиваемые параметры модели. Структурой $f(x)$ модели (3.13) называется набор функций $f_j(x)$ из порождающего множества $G = \{g_i(x)\}_{i=1}^p$:

$$f(x) = \{f_j(x) | f_j = g_{i_j} \in G\}_{j=1}^k.$$

Вектор $\mathbf{w} = [w_1, \dots, w_k]^T$ параметров модели с фиксированной структурой f настраивается на обучающей выборке $D(\mathcal{L})$ и минимизирует функционал качества прогнозирования

$$\mathbf{w} = \underset{w \in \mathbb{R}^k}{\operatorname{argmin}} Q(\tilde{\mathbf{x}}, \mathbf{x} | \mathcal{L}).$$

Функционал Q качества прогнозирования временного ряда **x** был определен следующим образом:

$$Q(\tilde{x}, x | \mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{t \in \mathcal{J}} (\tilde{x}(t) - x(t))^2,$$

где $\tilde{x}(t)$ – спрогнозированное значение отсчета временного ряда, $x(t)$ – его истинное значение. После настройки параметров качество прогнозирования оценивалось на контрольной выборке $D(\mathcal{T})$.

Предполагалось, что ряды из набора Ξ кластеризуются таким образом, что ряды из одного кластера C_k наиболее оптимально (в смысле функционала Q) приближаются моделями f одинаковой или близкой структуры $f \in \mathcal{F}_k$. Наличие подобной кластеризации позволило бы сократить поиск структуры при выборе модели. Выполнялся поиск разбиения множества индексов $\mathcal{I}$ набора временных рядов Ξ на множество индексов кластеров $\mathcal{I}_k$

$$i \in \mathcal{I}_k \Rightarrow \mathbf{x}_i \in C_k,$$

такого, чтобы ряды, близкие согласно заранее заданной функции расстояния $d(\mathbf{x}_i, \mathbf{x}_j)$, лежали в одном кластере C_k:

$$\mathcal{I} = \{1, \dots, m\} = \coprod_{k=1}^r \mathcal{I}_k : \forall i, j \in \mathcal{I}_{k_1}, l \in \mathcal{I}_{k_2} \Rightarrow d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{x}_l). \quad (3.14)$$

Априорная информация о кластеризации временных рядов задана матрицей смежности $\mathcal{K}$, элемент k_{ij} которой определялся следующим образом:

$$k_{ij} = \begin{cases} 1, \text{если временные ряды } \mathbf{x}_i \text{ и } \mathbf{x}_j \text{лежат в одном кластере} \\ -1 \text{ иначе.} \end{cases} \quad (3.15)$$

Был введен функционал, характеризующий согласованность выполненной кластеризации, заданной матрицей смежности $\mathcal{S}$, с экспертными суждениями, заданными матрицей $\mathcal{K}$:

$$A(\mathcal{I}_1, \dots, \mathcal{I}_r) = \sum_{i=1,j=1}^{n} (k_{ij} - s_{ij})^2. \quad (3.16)$$

Согласованность решения тем больше, чем меньше значения функционала (3.16). Обозначив через CQ суммарную ошибку прогнозирования, математическую постановку задачи можно записать в виде:

$$\{\mathcal{I}_1, \dots, \mathcal{I}_r\} =_{\mathcal{I}=\mathcal{I}_1 \amalg \dots \amalg \mathcal{I}_r} CQ(\mathcal{I}_1, \dots, \mathcal{I}_r) \text{при} A(\mathcal{I}_1, \dots, \mathcal{I}_r) \to \min. \quad (3.17)$$

Чтобы учесть при кластеризации расстояние как между рядами, так и между моделями, расстояние между временными рядами вычислялось как расстояние $d(\mathbf{x}_i, \mathbf{x}_j) = \rho(f(\mathbf{x}_i), g(\mathbf{x}_j))$ между прогнозами временных рядов $\mathbf{x}_i$ и $\mathbf{x}_j$ с помощью моделей f и g. Таким обращом для описания набора Ξ временных рядов была сформирована четырехиндексная матрица расстояний между прогнозами временных рядов с помощью моделей:

$$\mathcal{R}(\mathbf{x}_i, \mathbf{x}_j, f, g) = \rho(f(\mathbf{x}_i), g(\mathbf{x}_j)).$$

Принцип работы предлагемого алгоритма отражает диаграмма:

$$\rho((\mathcal{I} \times \mathcal{F})^2) \dashrightarrow \rho(\mathcal{I}^2) \underset{\dashrightarrow}{\mathcal{K}} R.$$

Каждая двухиндексная подматрица $R \in \rho(\mathcal{I}^2)$ матрицы $\mathcal{R} \in \rho((\mathcal{I} \times \mathcal{F})^2)$ с элементами $R_{ij} = \rho(f_i(\mathbf{x}_i), f_j(\mathbf{x}_j))$, $f_i, f_j \in \mathcal{F}$ задает выбор модели (так как каждому ряду $\mathbf{x}_i$ сопоставлена модель f_i из $\mathcal{F}$), и кластеризацию с матрицей смежности $[k_{ij}]$ вида (3.15), определяемой как решение задачи оптимизации

$$\sum_{i,j \in \mathcal{I}} d(\mathbf{x}_i, \mathbf{x}_j) k_{ij} \to \min.$$

Из множества $\rho(\mathcal{I}^2)$ всех таких двухиндексных подматриц выбралась матрица R, оптимальная с точки зрения задачи (3.17). Выполнение условия (3.12) оптимальности выбранной модели и условия согласованности (3.16) кластеризации обеспечивалось выбором подходящей функции расстояния и введением функции компетентности моделей с учетом априорной матрицы смежности. Введение этих дополнительных параметров и подбор функции расстояния позволил улучшить работу алгоритма и осуществить поиск оптимального решения задачи (3.17).

Выбор функции расстояния. Были рассмотрены три варианта функции

расстояния между рядами: функция $d(x_1, x_2)$ показывает близость рядов по значениям, функция $Cor(x_1, x_2)$ показывает, насколько похожи ряды $\mathbf{x}_1$ и $\mathbf{x}_2$ по поведению (возрастание, убывание и пр.), а функция $BC(x_1, x_2)$ учитывает и поведение ряда, и близость по значениям. Значения функций $d(x_1, x_2)$, $Cor(x_1, x_2)$ и $BC(x_1, x_2)$ рассчитываются по следующим формулам:

$$d(x_1, x_2) = \frac{\sqrt{\sum_{i=1}^{m} (\mathbf{x}_1(i) - \mathbf{x}_2(i))^2}}{m},$$

$$Cor(x_1, x_2) = \frac{\sum_{i=1}^{m} (\mathbf{x}_{1(i+1)} - \mathbf{x}_{1i})(\mathbf{x}_{2(i+1)} - \mathbf{x}_{2i})}{\sqrt{\sum_{i=1}^{m-1} (\mathbf{x}_1(i+1) - \mathbf{x}_1(i))^2} \sqrt{\sum_{i=1}^{m-1} (\mathbf{x}_2(i+1) - \mathbf{x}_2(i))^2}}.$$

Функция $Cor(x_1, x_2)$ -коэффициент корреляции Пирсона, показывает схожее поведение значений временных рядов $\mathbf{x}_1$ и $\mathbf{x}_2$

$$BC(x_1, x_2) = \frac{2}{1 + \exp(kCor(x_1, x_2))} d(x_1, x_2).$$

Параметр k в функции $BC(x_1, x_2)$ подбирается вручную.

Функция компетентности для моделей временных рядов. Для того, чтобы улучшить качество кластеризации набора временных рядов и прогностических моделей, необходимо различать временные ряды из одного класса C_1, спрогнозированные прогностическими моделями из того же класса C_1 и временные ряды из одного класса C_1, спрогнозированные прогностическими моделями, лежащими в различных классах. Была введена функция компетентности h, позволяющая увеличить значение расстояния между значениями временного ряда и прогнозом временного ряда моделью другого класса и уменьшить значение расстояния между значениями временного ряда и прогнозом данного временного ряда моделью этого же класса. Каждой прогностической модели f была поставлена поставим в соответствие функция компетентности h вида

$$h(f, \mathbf{x}) = e^{-\beta ||\mathbf{x} - f(\mathbf{x})||},$$

где $\mathbf{x}$ – временной ряд из множества Ξ, $f(\mathbf{x})$– спрогнозированный ряд $\mathbf{x}$ с помощью модели f.

С помощью функции компетентности мы хотим добиться улучшения качества кластеризации набора временных рядов и прогностических моделей. На рисунке 3.10 схематически изображены возможные варианты соответствия моделей и временных рядов классам – матрица ожидаемых расстояний между прогнозами временных рядов из набора Ξ с помощью прогностических моделей из набора $\mathcal{F}$.

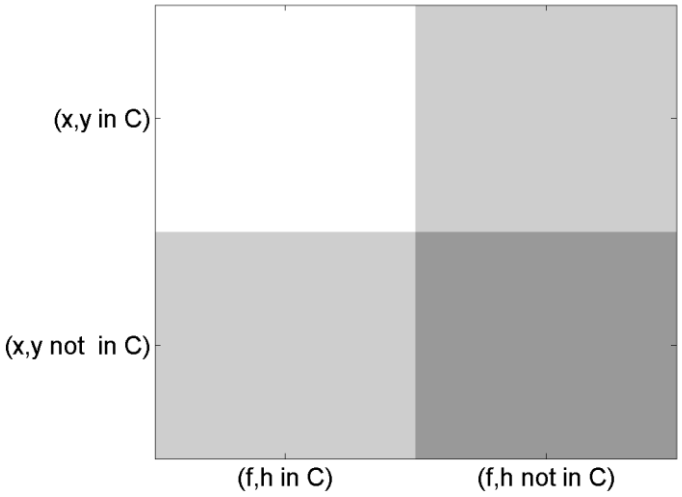


Рисунок 3.10 – Матрица ожидаемые расстояния

Строки и столбцы отражают принадлежность данных моделей и рядов к одному классу. В каждой ячейке этой матрицы показано ожидаемое расстояние между прогнозами временных рядов **x** и **y** с помощью прогностических моделей f и h соответственно. Так, в верхней левой клетке (a_{11}) показывается расстояние между временными рядами из одного класса, спрогнозированными моделью своего же класса. Более светлый цвет означает меньшее расстояние между прогнозами. В первом столбце прогностические модели выбраны из одного класса, во втором столбце для получения прогнозов выбраны модели из разных классов. Соответственно, в первой строке временные ряды **x** и **y** выбраны из одного класса, а во второй строке – из разных. Так, в верхней левой клетке (a_{11}) показывается расстояние между временными рядами из одного класса, спрогнозированными моделью своего же класса. Более светлый цвет говорит о предположительно меньшем расстоянии между прогнозами временных рядов с помощью прогностических моделей.

Выбор модели с помощью генетического алгоритма. Введенный ранее способ вычисления расстояния между рядами и моделями позволил приблизить функционал, характеризующий согласованность выполненной кластеризации с априорной информацией о связях между рядами, заданной матрицей $\mathcal{K} = k_{ij}$ следующим образом:

$$T = \sum_{x_i, x_j \in \Xi} d(x_i, x_j) k_{ij}.$$

Этот функционал был использован при выборе оптимальной матрицы соот-

ветствия (столбцы матрицы соответствуют временным рядам из множества Ξ, а строки матрицы соответствуют прогностическим моделям из множества $\mathcal{F}$) генетическим алгоритмом. Таким образом, при применении алгоритма неявным образом проводился поиск оптимальной прогностической структуры для данного временного ряда, согласованной с априорной информацией.

Перенос структуры связи между парами временных рядов. Была рассмотрена задача повышения качетства прогнозирования сходных временных рядов, путем переноса прогностической структуры. Рассматривались временные ряды $\mathbf{x}_1(t)$, $\mathbf{x}_2(t)$ и $\mathbf{y}(t)$ вида:

$$\mathbf{x}_1 = \{x(t)\}_{t=1}^{m},$$

где $x(t) \in \mathbb{R}$ – отсчет временного ряда $\mathbf{x}_1(t)$ в момент времени t. Была установлена зависимость ряда $\mathbf{x}_1$ от ряда $\mathbf{y}$, задаваемая следующим образом:

$$x_1(t) = \sum_{i=1}^{m} w_{1i} f_i(x_1(t-i), y(t-i)).$$

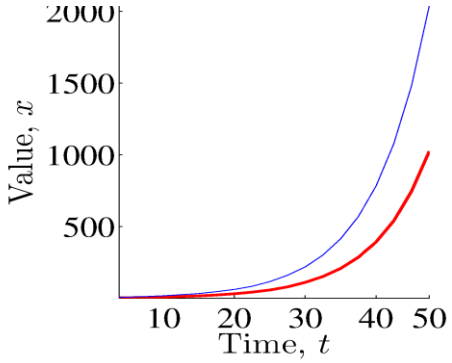
Здесь m – количество предыдущих значений временных рядов $\mathbf{x}_1$ и $\mathbf{y}$, принимаемых во внимание, коэффициенты w_{1i} – веса значений функции f_i, $\mathbf{y}$ – известный временной ряд, т. е. ряд с известным распределением.

Рассматривался временной ряд $\mathbf{x}_2(t)$, близкий к временному ряду $\mathbf{x}_1(t)$ по значению функции расстояния между рядами $r(x_1,x_2)$, подходящую функцию расстояния определим далее. Проверялось влияние временного ряда $\mathbf{y}$ на ряд $\mathbf{x}_2$. Для этого осуществлялся перенос структуры связи f на ряды $\mathbf{y}$ и $\mathbf{x}_2$, представив ряд $\mathbf{x}_2$ в виде:

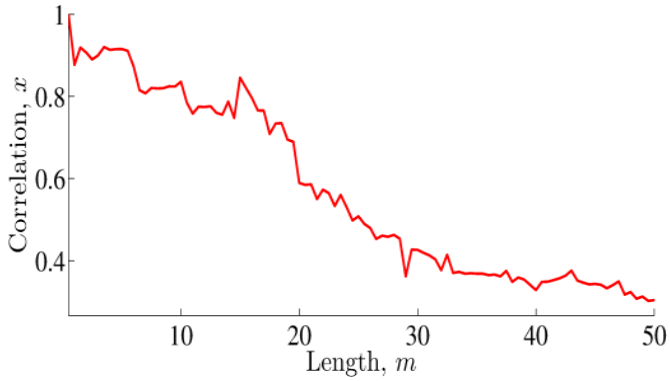
$$x_2(t) = \sum_{i=1}^{n} w_{2i} f_i(x_2(t-i), y(t-i)),$$

где f – перенесенная структура связи между временными рядами $\mathbf{x}_2(t)$ и $\mathbf{y}(t)$, а коэффициенты w_{2i} – новые веса, определяемые по формуле (3.13):

На рисунках 3.11 и 3.12 изображены значения временных рядов $\mathbf{x}_1$ и $\mathbf{x}_2$ (рисунок 3.11(a)), зависимость коэффициента корреляции Пирсона от длины принимаемых во внимание частей временных рядов – функция $Cor(x_1, x_2)$ (рисунок 3.11(b)) и функция $BC(x_1, x_2)$ (рисунок 3.12) от длины временного ряда m.



(a)



(b)

Рисунок 3.11 – (a) Вид исследуемых временных рядов $\mathbf{x}_1$ и $\mathbf{x}_2$, (b) $Cor(x_1, x_2)$

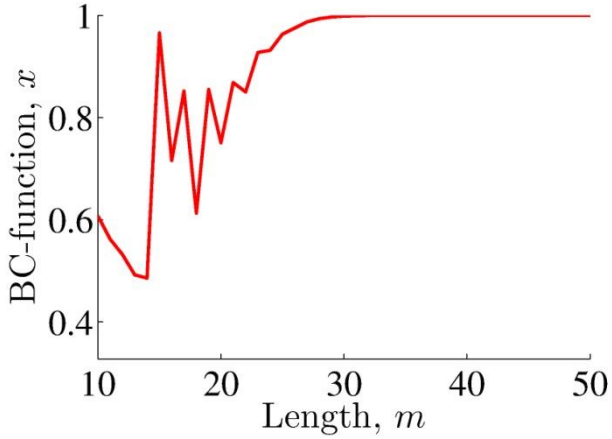


Рисунок 3.12 – BC-функция для временных рядов $\mathbf{x}_1$ и $\mathbf{x}_2$

Рассмотрено несколько вариантов прогнозирования ряда $\mathbf{x}_2$. С помощью линейной регрессии определялись коэффициенты весов для прогнозирования ряда $\mathbf{x}_2$ по его собственным значениям. В этом приближении ститалось, что ряд $\mathbf{x}_2$ зависит от своих предыдущих значений по формуле:

$$x_2(t) = \sum_{i=1}^{n} w_{1i} x_2(t-i).$$

Сложность структуры $f(x, y)$ в этои случае определяется количеством модулей переменных, входящих в функцию f. После выполнения прогноза вычисля-

лось значение функционала качества Q:

$$Q(\tilde{x}, x) = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} (\tilde{x}(i) - x(i))^2.$$

Аналогичные действия выполнялись при нахождении коэффициентов весов для прогнозирования ряда $\mathbf{x}_2$ по значениям временного ряда $\mathbf{x}_1$. В этом случае мы считалось, что значения ряда $\mathbf{x}_2$ задаются формулой:

$$x_2(t) = \sum_{i=1}^{n} w_{1i} x_1(t-i).$$

При проведении экспериментов полагалось, что структура $f(x, y)$ не изменяется и прогноз для ряд $\mathbf{x}_2$ выполниялся в предположении и следующем виде зависимости временного ряда $\mathbf{x}_2$ от временного ряда $\mathbf{y}$ и собственных значений ряда $\mathbf{x}_2$:

$$x_2(t) = \sum_{i=1}^{n} w_{2i} f_i(x_2(t-i), y(t-i)) .$$

Во всех трех случаях после прогнозирования ряда вычислялось среднее значение функционала качества прогнозирования Q и значение сложности структуры для получения зависимости $Q(k)$, где k – сложность структуры $f(x, y)$. Выбиралась структура, доставлющая минимальную ошибку прогнозирования с учетом штрафа за сложность. Включение дополнительной переменной в модель трактовалось, по аналогии с тестом Грейнджера, как решение о наличии связи.

Результаты перебора для структур разной сложности представлены в таблице 3.4.

Таблица 3.4 – Генерирующие и прогностические структуры разных сложностей

Сложность	Генерирующая структура	Прогностическая структура
2	$\mathbf{x}$+$\mathbf{y}^2$	$\mathbf{x}$+$\mathbf{y}^2$
3	log(1+$\mathbf{x}$)+sin$\mathbf{x}$+$\mathbf{y}^2$	$\mathbf{x}$+sin$\mathbf{x}$+$\mathbf{y}^2$
4	log(1+$\mathbf{x}$)+sin$\mathbf{x}$+cos$\mathbf{x}$+$\mathbf{y}^2$	log(1+$\mathbf{x}$)+$\mathbf{x}$+cos$\mathbf{x}$+$\mathbf{y}^2$
5	log(1+$\mathbf{x}$)+sin$\mathbf{x}$+$\mathbf{y}$+$\mathbf{y}^2$+sin$\mathbf{y}$	$\mathbf{x}$+cos$\mathbf{x}$+$\mathbf{y}$+$\mathbf{y}^2$+sin$\mathbf{y}$
6	$\mathbf{x}$+log(1+$\mathbf{x}$)+sin$\mathbf{x}$+$\mathbf{x}^2$+$\mathbf{y}$+$\mathbf{y}^2$	$\mathbf{x}$+log(1+$\mathbf{x}$)+sin$\mathbf{x}$+$\mathbf{x}^2$+sin$\mathbf{y}$+$\mathbf{y}^2$
7	$\mathbf{x}$+log(1+$\mathbf{x}$)+sin$\mathbf{x}$+cos$\mathbf{x}$+$\mathbf{y}$+$\mathbf{y}^2$+sin$\mathbf{y}$	$\mathbf{x}$+log(1+$\mathbf{x}$)+sin$\mathbf{x}$+cos$\mathbf{x}$+$\mathbf{y}$+$\mathbf{y}^2$+sin$\mathbf{y}$

Кластеризация набора временных рядов и моделей. В экспериментах по кластеризации временных рядов и прогностических моделей использовались синтетические временные ряды. Рассматривался набор из пяти временных рядов, x_1, x_2, x_3, x_4, x_5, изображенных на рисунке 3.13 и набор из двух прогностических

моделей $f_1 = sin(\frac{\pi x}{k}), f_2 = (-1)^p px$, где временные ряды x_1, x_2, x_3 (выделены красным цветом) лежат в одном кластере, а ряды x_4, x_5 (выделены синим цветом) – в другом. Каждому ряду из набора соответствует прогностическая модель: ряды x_1, x_2, x_3 прогнозируются моделью вида $f_1 = sin(\frac{\pi x}{k})$, а для рядов x_4, x_5 – моделью $f_2 = (–1)^p px$. На рисунке 3.14 (а–с) изображены временной ряд и его прогноз, полученный с помощью различных моделей из рассматриваемых наборов.

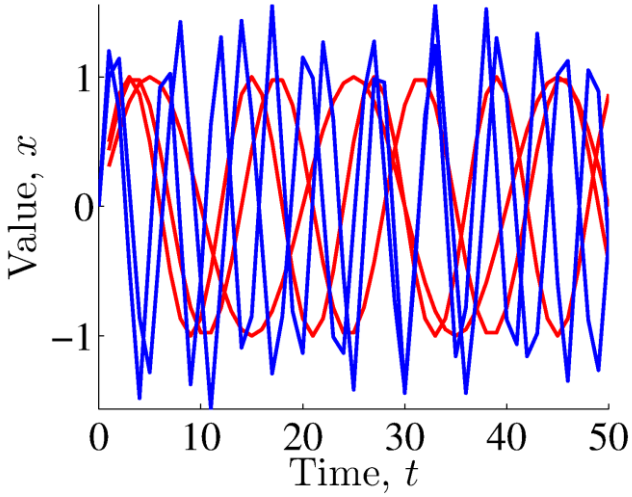
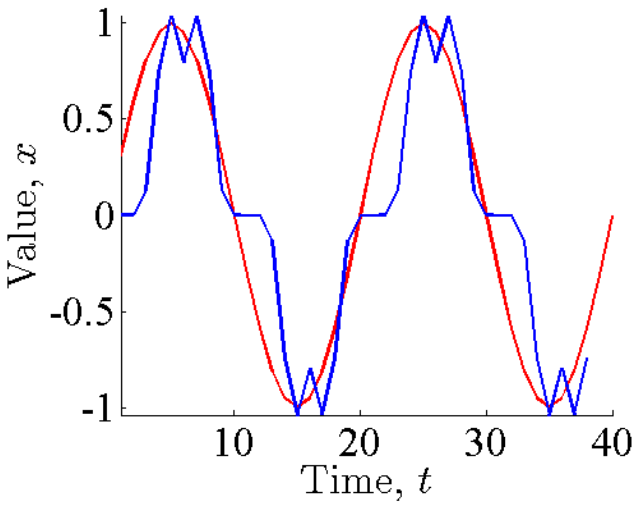
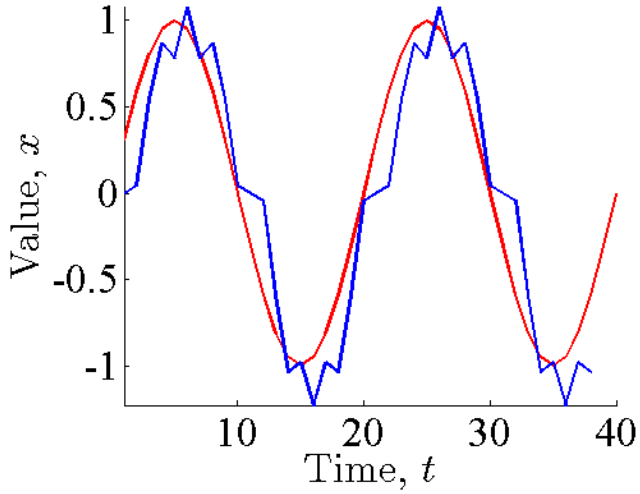


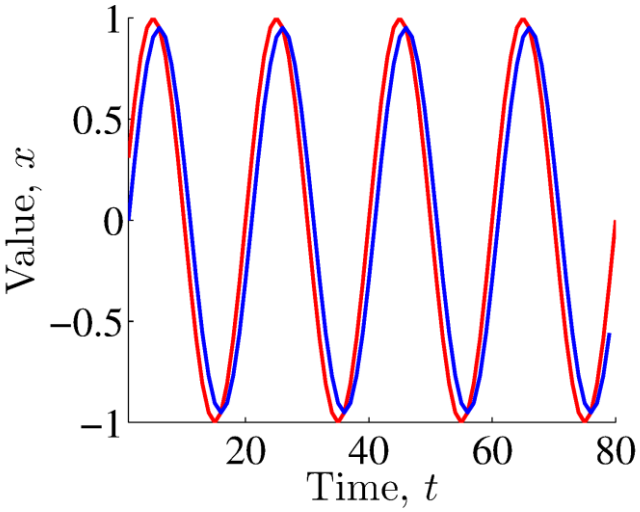
Рисунок 3.13 – Набор временных рядов



(a)



(b)



(c)

Рисунок 3.14 – Соответствие прогнозов временного ряда **x** для моделей различной сложности

В таблице 3.5 приведены значения выбранной алгоритмом двухиндексной матрицы попарных расстояний между прогнозами каждого временного ряда из набора x_1, x_2, x_3, x_4, x_5 с помощью соответствующих моделей. Выделяются два кластера пар «временной ряд – прогностическая модель». Согласно значениям попарных расстояний между прогнозами временных рядов с помощью оптимальных прогностических моделей, ряды и модели правильным образом разделяются на кластеры: временные ряды x_1, x_2, x_3 и прогностическая модель f_1 лежат в одном кластере, а временные ряды x_4, x_5 и прогностическая модель f_2 – в другом.

Таблица 3.5 – Расстояния между прогнозами

Ряд	(x_1, f_1)	(x_2, f_1)	(x_3, f_1)	(x_4, f_2)	(x_5, f_2)

Ряд					
(x_1, f_1)	0	1,5320	2,2903	11,1037	9,8345
(x_2, f_1)	1,5320	0	1,6101	8,0003	7,9271
(x_3, f_1)	2,2903	1,6101	0	6,5115	7,8104
(x_4, f_2)	11,1037	8,0003	6,5115	0	1,3702
(x_5, f_2)	9,8345	7,9271	7,8104	1,3702	0

Влияние функции компетентности на качество кластеризации. Таблица 3.5 показывает попарные расстояния между прогнозами временных рядов x_1, x_2, x_3 и x_4, x_5 с помощью моделей f_1 и f_2 с учетом функций компетентности: прогностической модели f_1 была сопоставлена функцию компетентности $h = e^{-1.6|x-f_1(x)|}$, а прогностической модели f_2 – функция компетентности $h = e^{-0.6|x-f_2(x)|}$. Разрыв диапазонов значений попарных расстояний между $(x_1, f_1), (x_2, f_1), (x_3, f_1)$ и $(x_4, f_2), (x_5, f_2)$ увеличился, то есть устойчивость решения задачи кластеризации была повышена.

Таблица 3.6 – Расстояния между прогнозами на основании функций h_1 и h_2

Ряд / Ряд	(x_1, f_1)	(x_2, f_1)	(x_3, f_1)	(x_4, f_2)	(x_5, f_2)
(x_1, f_1)	0	0,6013	1,3902	104,7137	91,3374
(x_2, f_1)	0,6013	0	0,7889	85,5234	83,9511
(x_3, f_1)	1,3902	0,7889	0	61,2303	72,5400
(x_4, f_2)	104,7137	85,5234	61,2303	0	0,4308
(x_5, f_2)	91,3374	83,9511	72,5400	0,4308	0

Рисунок 3.15 демонстрирует выделение двух кластеров пар временных рядов и прогностических моделей $(x_1, f_1), (x_2, f_1), (x_3, f_1)$ и $(x_4, f_2), (x_5, f_2)$. Более светлые ячейки говорят о большем расстоянии между прогнозами

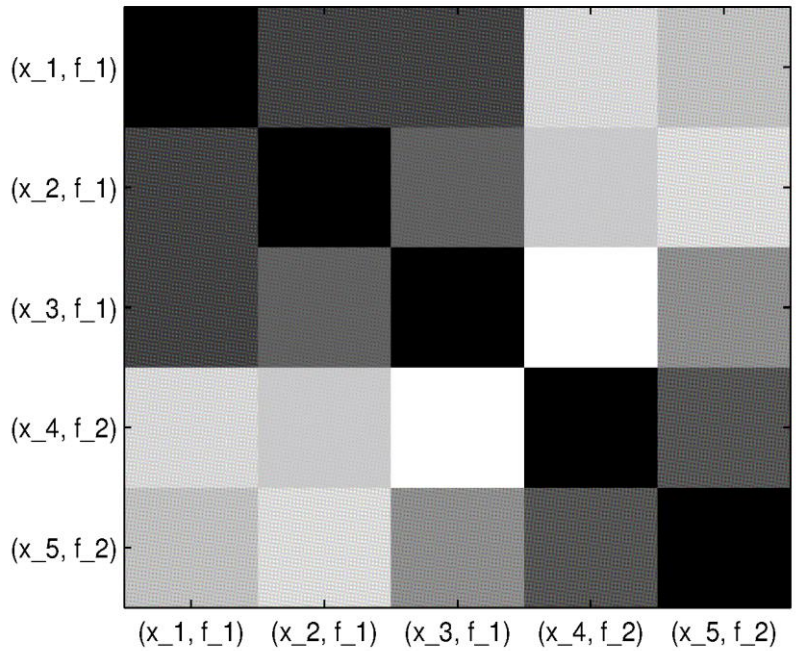


Рисунок 3.15 – Попарные расстояния между прогнозами временных рядов

Таким образом, предложен алгоритм, позволяющий использовать ранее полученные данные о связях между временными рядами и структуре связей для обнаружения новых связей. Алгоритм позволяет распространять информацию о структуре связи на ранее не рассматривавшиеся временные ряды путем кластеризации временных рядов в предположении, что сходные временные ряды прогнозируются моделями одной и той же или близкой структуры.

3.4 Описание и обоснование способов расширения подхода Грейнджера для обнаружения структуры связей между временными рядами экзогенных факторов и объемами спроса на грузовые железнодорожные перевозки

В этом подразделе представлено описание и обоснование расширения подхода Грейнджера для обнаружения структуры связей между временными рядами.

3.4.1 Описание алгоритма и результатов его тестирования на модельных данных

Проведено исследование, посвященное поставленной в подразделе 3.2 задаче обнаружения закономерностей в наборах временных рядов. Рассмотрен метод

выявления причинно-следственных связей, разработанный для временных рядов, описывающих поведение некоторой динамической системы. Под динамической системой понимается множество элементов, для которого задана функциональная зависимость между временем и положением в пространстве каждого элемента этой системы. Состояние динамической системы в любой момент времени описывается множеством вещественных параметров (или векторов), соответствующим определённой точке в пространстве состояний. Закон, который описывает изменение (эволюцию) начального состояния с течением времени, позволяет по начальному состоянию прогнозировать будущее состояние динамической системы. Примерами динамических систем являются механические, физические, химические и биологические объекты, вычислительные процессы. Предложены различные способы описания динамических систем: с помощью дифференциальных уравнений, дискретных отображений, теории графов, теории марковских цепей и т.д. Выбор одного из способов описания задает конкретный вид математической модели соответствующей динамической системы. Понятие связи между временными рядами, описывающими динамические системы, определялось как принадлежность их к одной динамической системе.

Выявление причинности в сложных системах может быть затруднено во многих контекстах, где коэффициент корреляции между переменными меняется во времени и в различные моменты времени может быть положительным, близким к нулю или даже отрицательным в зависимости от состояния системы. Такое поведение, зависящее от состояния системы, является определяющим признаком сложных нелинейные системы [77, 78] и часто всречается в природе. Хотя наличие корреляции между переменными не является ни необходимым, ни достаточно условием причинно-следственной связи, наблюдение за корреляцией лежит в основе многих эвристических методов [77–81].

Рассматриваемый метод разработан для применения к временным рядам, не удовлетворяющих условию отделимости, содержащемуся в опрелелении Грейнджера [82]: информация об экзогенном факторе содержится исключительно в переменной X, и фактор может быть исключен из рассмотрения путем удаления переменной X из модели. Отделимость характерна для чисто стохастических и линейных систем и связана с предположением, что система может быть представлена в виде отдельных

компонент.

Этот подход применим для стохастических систем с сильной связью между переменными, но может приводить к неверным выводам в детерминированных условиях, особенно при исследовании динамических систем со слабой или умеренной связью, так как условие разделимости в таких системах не выполняется. Сложные системы необходимо рассматривать как единое целое (в противоположность традиционным экономическим моделям). Как следует из теоремы Такенса [83, 84], в детерминированных динамических системах, если некий экзогенный фактор, описываемый переменной X, является причиной для Y, информация об этом факторе X избыточно присутствует в Y не может быть полностью удалена из рассмотрения удалением X из модели. Таким образом, нарушается основное условие применимости теста Грейнджера, оставляя вопрос о выявлении причинно-следственной связи в таких системах без ответа.

Метод *сходящегося перекрестного изображения* (convergent cross mapping, CCM) [85, 86] позволяет определить принадлежат ли временные ряды одной динамической системе. Этот метод был разработан для выявления связей в случаях, когда тест Грейнджера неприменим или не может обнаружить связи. Метод CCM основан на преобразовании пространства состояний некоторой динамической системы. В преобразованных пространствах, точками которых являются состояния динамических систем, сравниваются ближайшие соседи, соответствующие одному и тому же моменту времени *t*. На основе полученных ближайших соседей строится искомый прогноз как линейная комбинация этих соседей.

Использование метода сходящегося перекрестного отображения для выявления зависимости между временными рядами. При проверке временных рядов на наличие связи в смысле метода CCM, множество *M* состояний динамической системы

$$\mathbf{x}_j = \varphi(\mathbf{x}_{j-1}), \quad \mathbf{x} \in \mathbb{R}^d \ , \tag{3.18}$$

используются его приближения – множества $M_{\mathbf{x}} = f(M)$ и $M_{\boldsymbol{s}} = f(M)$, где

$$f(x) = (x, \varphi(x), \dots, \varphi^{d-1}(x))$$

может быть использовано для описания системы (3.18). Если ряды **s** и **x** имеют достаточно большую историю, множество $M = \{(s_j, x_j)\}$, $j \in \mathcal{J}$ пар значений временных рядов **s** и **x**, измеренных в *j*-й момент времени, велико. Предполагалась линей-

ная зависимость φ между рядами,

$$\varphi(x_\tau) = x_{j-\tau}, \quad \varphi(s_\tau) = s_{j-\tau},$$

где τ – выбранное в (3.3) и (3.4) значение задержки. Тогда отображения $f_{\mathbf{s}}: M \to M_{\mathbf{s}}$ и $f_{\mathbf{x}}: M \to M_{\mathbf{x}}$, определенные как

$$f_{\mathbf{s}}(s_j, x_j) = (s_j, s_{j-\tau}) \equiv \vec{s}(j),$$
$$f_{\mathbf{x}}(s_j, x_j) = (x_j, x_{j-\tau}) \equiv \vec{x}(j),$$

должны сохранять свойства системы (3.18).

Для точки $\vec{x}(j)$ множества $M_{\mathbf{x}}$ определялись $d+1$ ближайших соседей $\vec{x}(j_1), \vec{x}(j_2), \ldots, \vec{x}(j_d+1)$. При наличии связи между рядами **x** и **s** и индексы

$$j_1, \ldots, j_{d+1} \tag{3.19}$$

ближайших к $\vec{x}(j)$ соседей в $M_{\mathbf{x}}$ являются также индексами близких к $\vec{s}(j)$ соседей в $M_{\mathbf{s}}$. Для проверки этой гипотезы прогноз

$$\hat{s}_j = \sum_{i=1}^{d+1} w_i s_{j_i} \tag{3.20}$$

значения s_j временного ряда **s** вычислялся с помощью $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$

Веса w_i вычислялись экспоненциальным взвешиванием евклидовых расстояний $r(s_{j_i}, s_j)$ от s_j до элементов $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$:

$$w_i = \frac{u_i}{\sum u_k}, \quad u_i = \exp\left(-\frac{r(s_j, s_{j_i})}{r(s_j, s_{j_1})}\right), \quad i = 1, \ldots, d+1,$$
$$r(s, s') = |s - s'|.$$

При увеличении истории $|\mathcal{J}|$ измеряемых временных рядов расстояния между соседними точками множеств $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$ сокращаются. В случае, если ряд **x** действительно оказывал влияние на **s**, при $|\mathcal{J}| \to \infty$ прогноз (3.20) j-го значения ряда **s** становился точнее. Тогда коэффициент корреляции

$$\rho(\hat{s}_j, s_j) = \frac{1}{\sigma_s \sigma_{\hat{s}}} \mathrm{E}(\hat{s}_j - \mathrm{E}\hat{s}_j)(s_j - \mathrm{E}s_j) \tag{3.21}$$

стремился к некоторому ρ_0, не равному нулю. Здесь E – математическое ожидание случайной величины, σ – ее дисперсия. Наличие этой сходимости для исследуемых пар рядов проверялось в вычислительном эксперименте. Чем сильнее ρ_0 отклонялось от нуля, с большей достоверностью принималось решение о зависимости между рядами.

Ниже описана процедура добавления элементов к выборке, предназначенная для выявления сходимости выражения (3.21) к ρ_0. Множество индексов $|\mathcal{J}|$ значений временных рядов разбивается по времени

$$\mathcal{J} = \mathcal{L} \coprod \mathcal{T},$$

в $\mathcal{T}$ содержатся последние исторические индексы временного ряда. Для каждого из элементов с индексами $j \in \mathcal{T}$ был вычислен прогноз $\hat{s}_j$ по формуле (3.20), причем множество M конструировалось без учета элементов $j \in \mathcal{T}$. Для спрогнозированного отрезка временного ряда вычислялся эмпирический коээфициент корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$, для подсчета которого в формуле (3.21) математическое ожидание E и дисперсия σ^2 заменяются на среднее значение ряда, и среднеквадратичное отклонение:

$$\mathrm{E}s \mapsto \bar{\mathbf{s}}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}} s_j,$$

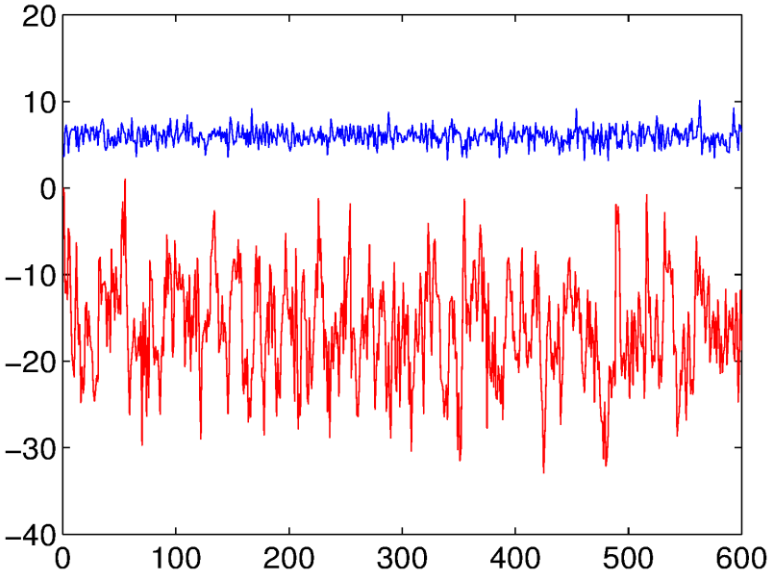
$$\sigma_s^2 \mapsto \frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}} (s_j - \bar{\mathbf{s}}_{\mathcal{T}})^2.$$

Была рассмотрена зависимость коэффициентов корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ исходного ряда $\mathbf{s}_{\mathcal{T}}$ и спрогнозированного ряда $\hat{\mathbf{s}}_{\mathcal{T}}$ и коэффициента корреляции $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ рядов $\hat{\mathbf{x}}_{\mathcal{T}}$ и $\mathbf{x}_{\mathcal{T}}$ от размера выборки. При увеличении объема выборки корреляция (3.21) между спрогнозированными значениями зависимого временнного ряда и его измеренными значениями для зависимых рядов возрастала. В экспериментах увеличивался размер выборки $|\mathcal{J}|$, с сохранением для прогноза последних $0.25|\mathcal{J}|$ значений временного ряда,

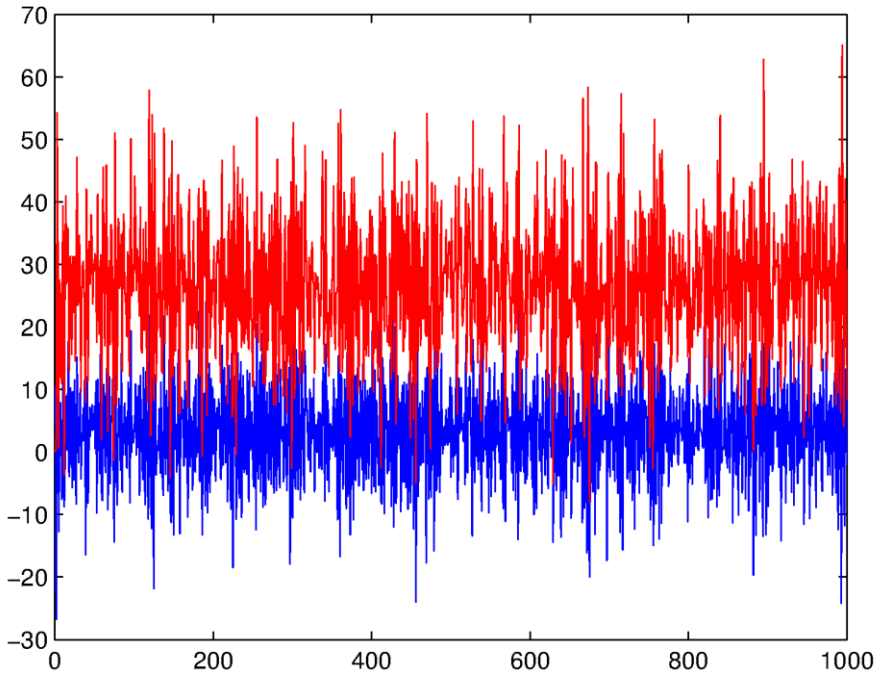
$$\mathcal{T}(|\mathcal{J}|) = \{[0.75|\mathcal{J}|], \dots, |\mathcal{J}|\},$$

где оператор $[a]$ означает округление. При последовательном увеличении $|\mathcal{J}|$ решалась задача (3.7), вычилялись коэффициенты $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$, $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ и корреляция $\rho(\mathbf{s}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ между оставленными для прогноза отрезками рядов.

Были проведены эксперименты на синтетических и реальных данных. Примеры синтетических данных с односторенней и двусторонней связью изображены на рисунке 3.16. В случае (a) ряд **x**, обозначенный синим, генерировался независимо от красного ряда **s**, а ряд **s** зависел от **x**; в случае (b) связь была двусторонней.



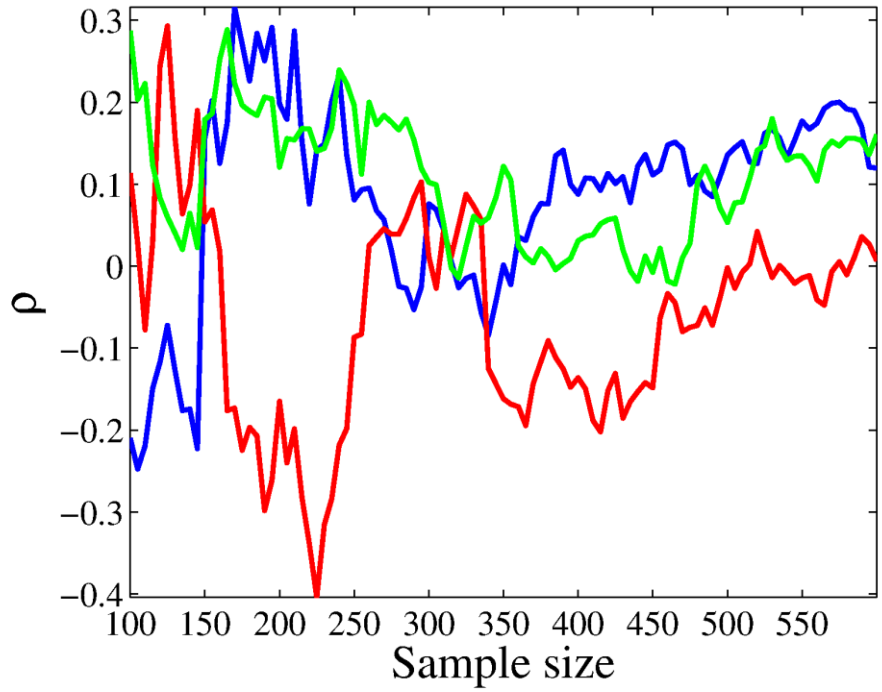
(a)



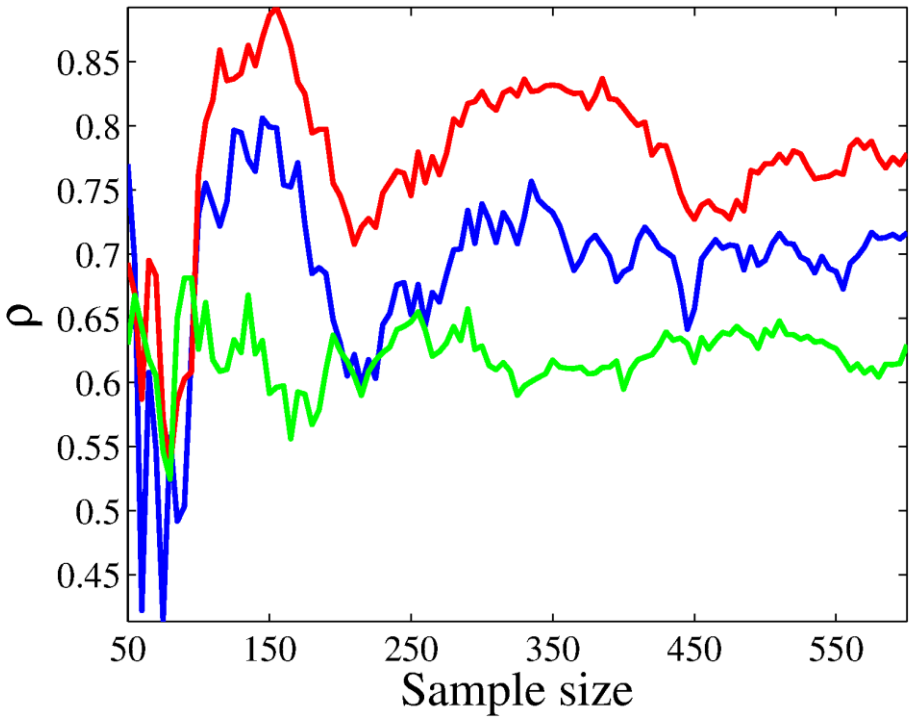
(b)

Рисунок 3.16 – Временные ряды с односторонней (a) и с двусторонней (b) связью

Зависимость коэффициентов корреляции между различными рядами от объема выборки для них изображена на рисунке 3.17: (a) – для рядов с односторонней причинно-следственной связью, (b) – для рядов в двусторонней связью



(a)



(b)

Рисунок 3.17 – Зависимость различных корреляций от объема выборки

Здесь синим цветом изображена зависимость $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ от объема выборки, красным – зависимость $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ от объема выборки. Зеленый цвет соответствует зависимости $\rho(\mathbf{x}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$. В первом случае $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.1$, но наблюдалась сходимость.

Был сделан вывод о наличии слабой связи между рядами. Во втором случае наблюдается сходимость к $\rho_0(\hat{\mathbf{x}}, \mathbf{x}) \approx 0.8, \rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.7$, причем

$$\rho_0(\hat{\mathbf{x}}, \mathbf{x}) > \rho_0(\mathbf{x}, \mathbf{s}), \rho_0(\hat{\mathbf{s}}, \mathbf{s}) > \rho_0(\mathbf{x}, \mathbf{s}),$$

то есть для рядов с двусторонней связью метод ССМ не только выявляет наличие связи, но и справляется лучше чем кросс-корреляция. Метод ССМ показал здесь лучшие результаты, так как он разработан специально для выявления взаимных связей, с которыми тест Грейнджера справляется не всегда.

Эксперименты на реальных данных проиллюстрированы ниже примером об исследовании связи между биржевыми ценами на нефть и загруженностью железнодорожного узла вагонами с нефтью. Ожидался резульат, что цены на нефть влияют на загруженность. В этом случае модуль коэффициента корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ должен был возрастать с объемом выборки и выходить на ненулевую асимптоту. На рисунке 3.18 изображены зависимости коэффициентов корреляции от объема выборки для пары рядов «Нефть и нефтепродукты»–«Цены на нефть», по оси абсцисс отложена длина временного ряда, полученного добавлением в выборку $|\mathcal{J}|$-го элемента.

На графике красного цвета по оси ординат отложены значения коэффициента корреляции спрогнозированых цен на нефть с их истинными значениями, синего – корреляция прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляция истинных значений исследуемых рядов рядов. Хотя коэффициент $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ принимает значения в диапазоне от –0,6 до 0,8, сходимости нет. В вычислительном эксперименте в таких случаях делался вывод об отсутствии связи между временными рядами. Из теории следует, что для получения адекватных результатов с помощью ССМ необходимо, чтобы отображения $f_{\mathbf{s}}, f_{\mathbf{x}}$ были взаимнооднозначными, то есть не наблюдалось самопересечений траекторий $s(j)$ и $x(j)$.

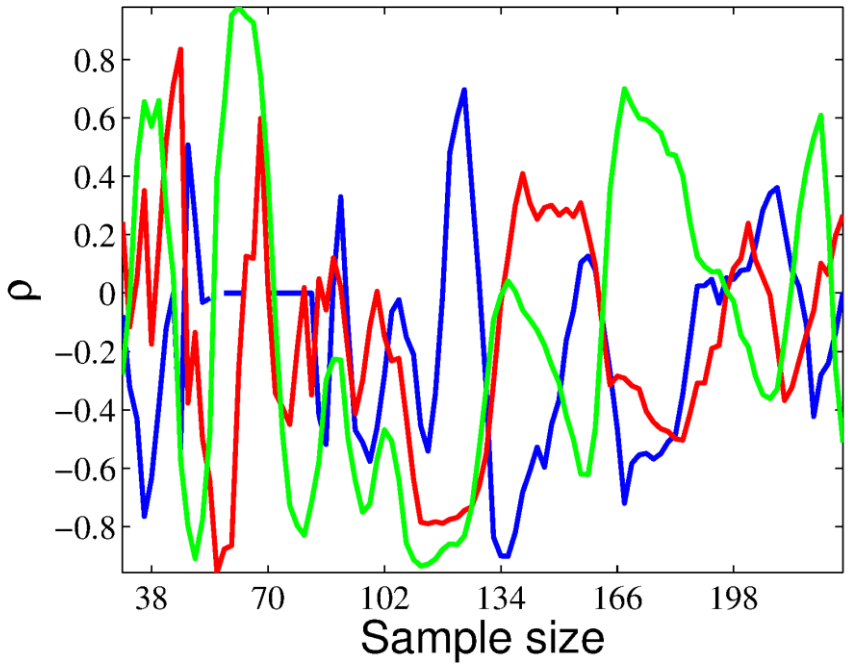
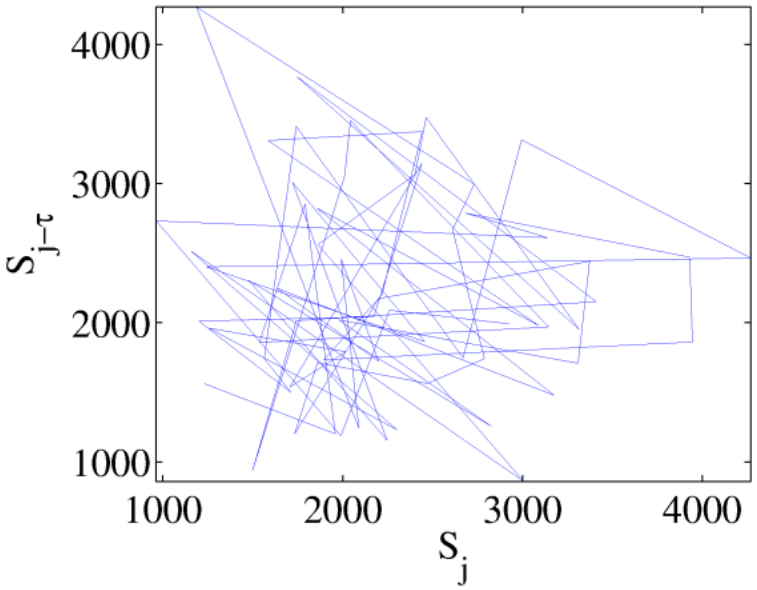


Рисунок 3.18 – Зависимость различных корреляций от объема выборки

Из рисунка 3.19 ясно, что для исследуемых данных это не так: наблюдается значительное число самопересечений. Таким образом, для применения описанного метода к реальным данным требуется дальнейшее исследование условий его применимости и адаптация к исследуемым данным.



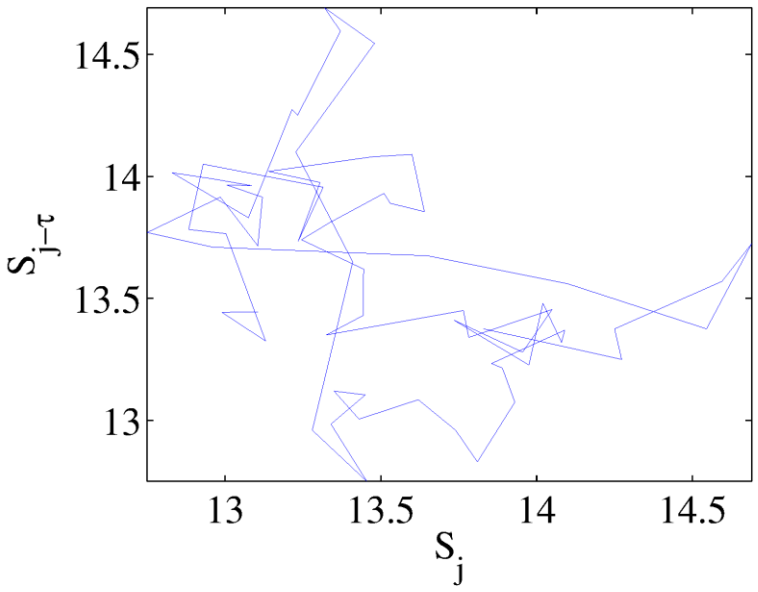


Рисунок 3.19 – Первые сто точек множеств $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$, соединенные по возрастанию индекса j

3.5 Разработка и тестирование алгоритма построения кросс-корреляционных зависимостей временных рядов объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов при прогнозе объема спроса на грузовые железнодорожные перевозки на отдельном железнодорожном узле

В этом подразделе представлено описание алгоритма построения кросскорреляционных зависимостей временных рядов объемов спроса на ГЖДП и экзогенных факторов при прогнозе объема спроса ГЖДП на отдельном железнодорожном узле и результаты его тестирования.

3.5.1 Развитие метода сходящегося перекрестного отображения

Для учета информации о структуре связи между рядами объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов предложена модификация метода сходящегося перекрестного отображения. Предложено рассматривать прогноз, задаваемый не только линейной комбинаций ближайших соседей, как это делается CCM, но также задаваемый более сложной прогностической моделью. Эта модел представляет собой допустимую комбинацию функций из некоторого заданного набора. Качество прогноза характеризуется величиной корреляции (далее *силой связи*) исходного и предсказанного рядов. Исследована зависимость силы

связи рядов от сложности используемой прогностической модели.

В качестве прогностических моделей были использованы линейные комбинации степенных функций из множества $\{s^{j_1}, \dots, s^{j_m}\}$. Для разных наборов показателей степенных функций из множества $\{j_1, \dots, j_m\}$ получаются различные формулы $f_j, j = \overline{1,n}$ для прогностической модели. Таким образом, прогностическая модель выбиралась из множества

$$\mathfrak{F} = \{f_1, \dots, f_n\}$$

допустимых формул. Здесь и далее под сложностью прогностической модели $f_j \in \mathfrak{F}, j = \overline{1,n}$ понимается величина $C_f \in \mathbb{N}$, равная сумме показателей степенных функций в модели.

При проведении эксперимента были использованы степени $k = 0, \dots, 6$. Каждой формуле f_j была поставлена в соответствие своя сложность C_{f_j}, равная сумме показателей степенных функций, используемых в формуле.

Прогноз $\hat{y}(t_i)$ вычислялся с помощью полиномиальной формулы

$$\hat{s}(t_i) = \sum_{k=1}^{E+1} \omega_{i_k} s^k(i_k) \ ,$$

где w_i – веса прогностической модели, $s(i_k)$ – известные значения временного ряда $y(t)$.

Сила связи между рядами была определена как коэффициент корреляции предсказанного $\hat{y}(t)$ и исходного $y(t)$ рядов

$$\rho(\hat{s}, s) = \frac{1}{\sigma_s \sigma_{\hat{s}}} \mathrm{E}(\hat{s} - \mathrm{E}\hat{s})(s - \mathrm{E}s),$$

где E – математическое ожидание случайной величины, σ^2 – его дисперсия. В случае, если прогнозируемый ряд $s(t)$ и используемый при прогнозе ряд $x(t)$ являются дискретной моделью одной и той же динамической системы, ожидалась сходимость силы связи $\rho(\hat{s}, s)$ к некоторой отличной от нуля величине ρ_0 с увеличением размера выборки **D**.

За ошибку $\mathcal{E}(\hat{s}, s)$ прогноза принималась величина отклонения предсказанного $\hat{s}(t)$ от известного результата $s(t)$

$$\mathcal{E}(\hat{s}, s) = \frac{\sqrt{\sum_{i=1}^{m} (\hat{s}(t_i) - s(t_i))^2}}{m} \ .$$

При принятии решения о связи определенной структуры проводился анализ изменения силы связи $\rho(\hat{y}(t_i), y(t_i))$ и ошибки $\mathcal{E}(\hat{y}, y)$ прогноза в зависимости от сложности C_f используемой прогностической модели f и выбиралась модель f оп-

тимальной сложности

$$C_f = \operatorname{argmax}_{C_f} \rho(C_f|\mathbf{D}).$$

В терминах ошибки $\mathcal{E}(\hat{s}, s)$ прогноза задача имеет вид:

$$C_f = \operatorname{argmin}_{C_f} \mathcal{E}(C_f|\mathbf{D}).$$

В качестве экспериментальных данных использовалась синтетически полученная выборка. Рассматривались заведомо временные ряды, полученные с помощью решений системы дифференциальных уравнений Лоренца

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = x(r - z) - y \\ \dot{z} = xy - bz \end{cases}$$

при следующих значениях параметров $\sigma = 10$, $r = 28$, $b = 3$ и начальной точке (0,1,1). Полученное решение системы затем проецировалось в двухмерное пространство (см. [84]). Таким образом были получены временные ряды $x(t)$ и прогнозируемый ряд $y(t)$.

Эксперимент проводился с целью подтверждения или опровержения следующей гипотезы: величина ошибки прогнозирования и сила связи зависят от сложности используемой формулы. Эксперименты показали, что сходимость коэффициента корреляции между спрогнозированным и исходным значением временного ряда происходит независимо от вида проверяемой структуры, однако его предельное значение – определяемая сила связи – падала с увеличением сложности. Полученные результаты свидетельствуют о том, что при принятии решения о связи между временными рядами важно учитывать структуру связи. В противном случае, даже при проверке заведомо зависимых временных рядов на взаимосвязь неадекватной структуры может быть принято решение об отсутствии связи.

3.5.2 Тестирование алгоритма на реальных данных

Экспертами были предоставлены временные ряды, содержащие данные о биржевых ценах на основные инструменты: сахар, бензин, медь, цинк, золото, никель, пшеницу, мазут, газ, олово, нефть, серебро, свинец за 2007–08 гг. Эти временные ряды, нормированные на отрезок [0, 1] изображены на рисунке 3.20(a) и 3.20(b). Как видно из рисунков, в рядах присутствуют пропуски. Кроме того, для многих временных рядов можно выделить наличие тренда, что свидетельствует об их нестационарности.

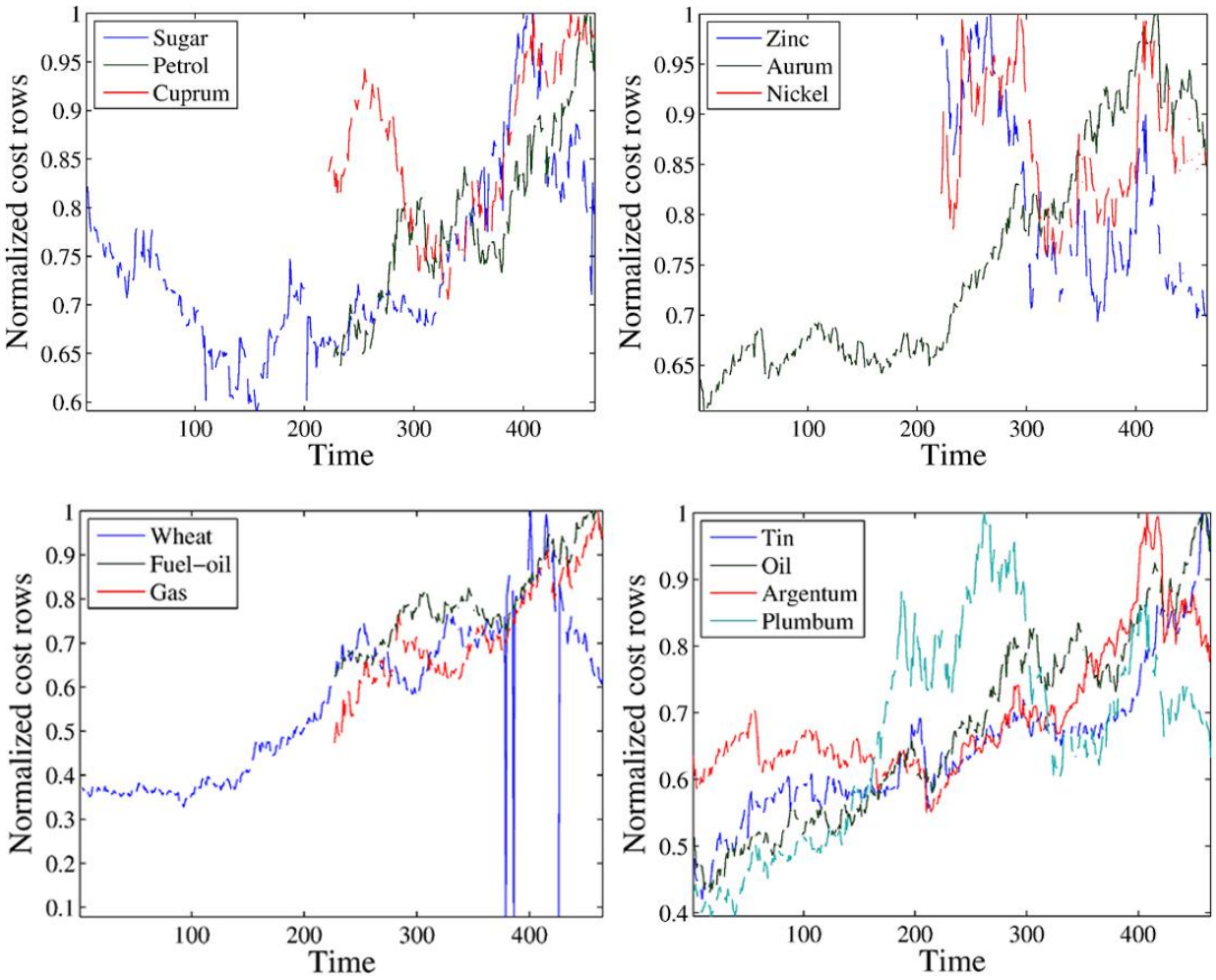


Рисунок. 3.20 – Временные ряды для цен на основные инструменты

3.5.3 Тестирование алгоритма на данных о ГЖДП на отдельном железнодорожном узле

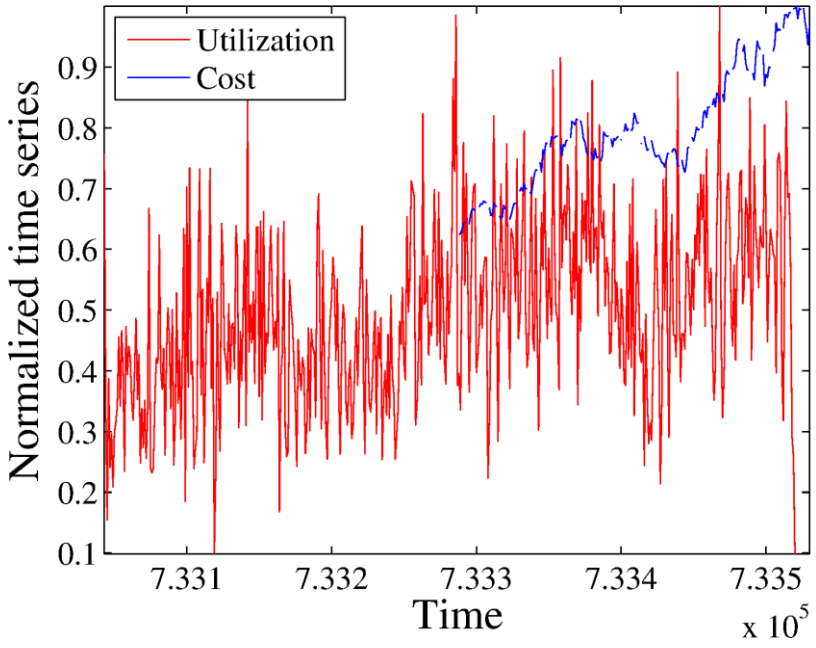
Также были использованы данные о загруженности железнодорожных узлов 83 ветки грузами различных групп каждый день с 01 января 2007 года в течение 473 дней. В таблице 3.7 перечислены рассматренные группы грузов.

Таблица 3.7 – Рассматренные группы грузов

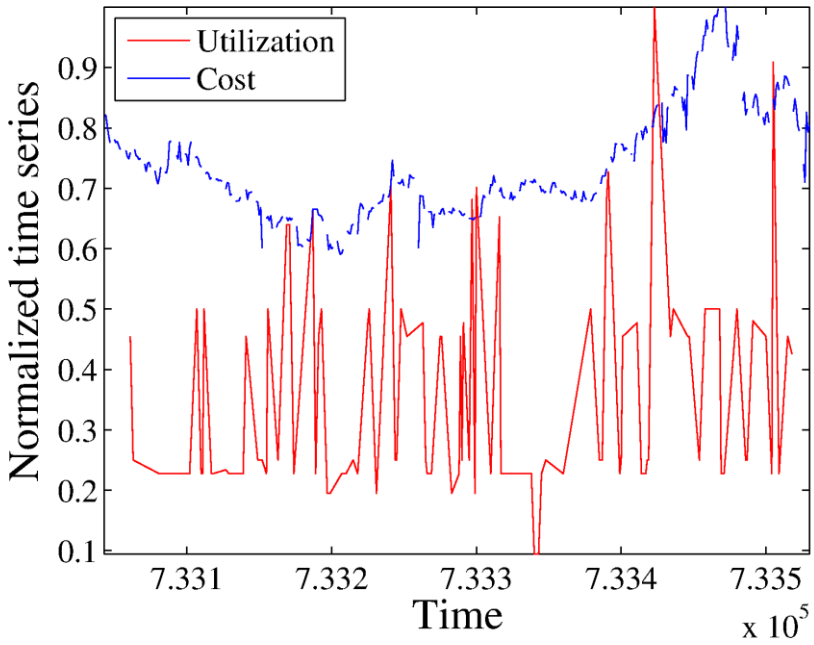
3 – Нефть и нефтепродукты	13 – Лом черных металлов
7 – Руда железная и марганцевая	16 – Цветные металлы, изделия из них и лом цветных металлов
8 – Руда цветная и серное сырье	9 – Черные металлы
11 – Металлические конструкции	25 – Сахар
12 – Метизы	33 – Сахарная свекла и семена
34 – Зерно	35 – Продукты перемола

На основе экспертных высказываний о наличии причинно-следственной связи между временным рядами, представленные в таблице 3.3, рассматривались различные пары (**s, x**) временных рядов.

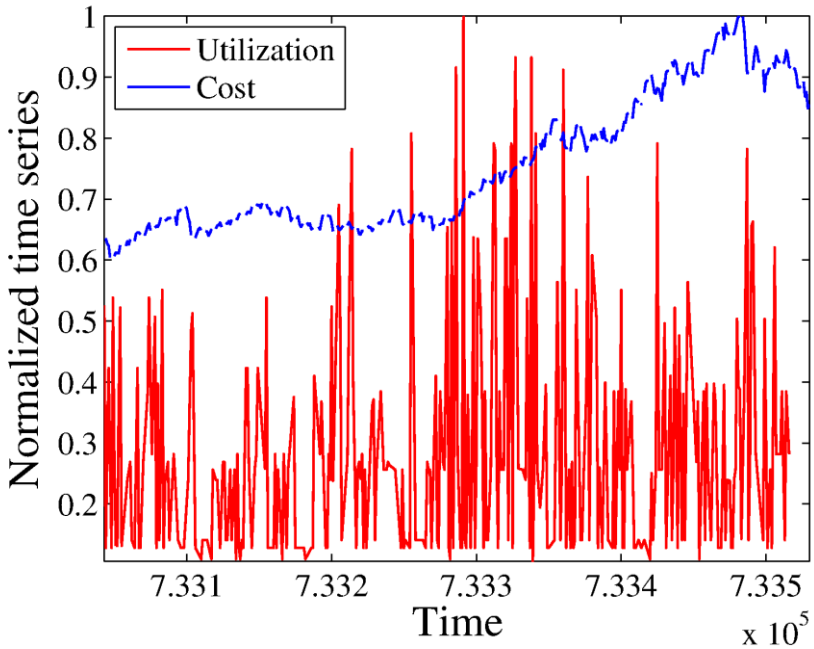
На рисунках 3.21 слева изображены пары временных рядов «Группа груза»–«Цена на некоторый инструмент», каждая точка соответствует одному дню: (a) – «3 – Нефть и нефтепродукты»–«Цены на нефть»; (b) – «25 – Сахар»–«Цены на сахар»; (c) – «16 – Цветные металлы, изделия из них и лом цв. металлов»–«Цены на золото».



(a)



(b)



(c)

Рисунок 3.21 – Пары рядов «Группа груза»–«Цена на некоторый инструмент»

Полный список исследованных в вычислительном эксперименте зависимостей приведен в таблице 3.8.

В этой таблице на первом месте стоит прогнозируемый ряд **s**, на втором – ряд **x**, предположительно оказывающий влияние на **s**. Для каждой пары рядов в таблице 3.8 приведен номер соответствующего этой паре фактора из таблицы 3.3, а также результат теста на зависимость по Грейнджеру и метода сходящегося

перекрестного отображения. В качестве результата теста Грейнджера в таблице приведены значение *p*-value для *F*-статистики (3.6).

Таблица 3.8 – Пары временных рядов рядов, исследованных на зависимость

№№	Пара временных рядов **s** – **x**	№ фактора	Тест Грейнджера		ССМ	
1	«Нефть и нефтепродукты» – «Цены на бензин»	1	$p = 0{,}1131$, $\tau = 6$	–	$\overline{\rho} = -0{,}1296$, $\overline{\rho}_s = 0{,}1630$	–
2	«Нефть и нефтепродукты» – «Цены на мазут»	1	$p = 0{,}0581$, $\tau = 7$	+	$\overline{\rho} = 0{,}1731$, $\overline{\rho}_s = 0{,}1773$	–
3	«Нефть и нефтепродукты» – «Цены на нефть»	1	$p = 0{,}0010$, $\tau = 7$	+	$\overline{\rho} = 0{,}1434$, $\overline{\rho}_s = -0{,}3157$	–
4	«Сахар» – «Цены на сахар»	1	$p = 0{,}0038$, $\tau = 7$	+	$\overline{\rho} = 0{,}1071$, $\overline{\rho}_s = 0{,}0175$	–
5	«Сахарная свекла и семена» – «Цены на сахар»	1	$p = 0{,}0111$, $\tau = 8$	+	N/A	N/A
6	«Зерно» – «Цены на пшеницу»	1	$p = 0{,}1667$, $\tau = 1$	–	N/A	N/A
7	«Продукты перемола» – «Цены на пшеницу»	1	$p = 0{,}5369$, $\tau = 1$	–	N/A	N/A
8	«Сахарная свекла и семена» – «Сезонность производства сахарной свеклы»	3	$p = 0{,}6601$, $\tau = 8$	–	N/A	N/A

Обозначения:
«+»– сделан вывод о наличии связи между временными рядами,
«–» – сделан вывод об отсутствии связи,
«N/A» – не достаточно данных для проведения эксперимента.

В этой таблице на первом месте стоит прогнозируемый ряд **s**, на втором – ряд **x**, предположительно оказывающий влияние на **s**. Для каждой пары рядов в таблицы 3.8 приведен номер соответствующего этой паре фактора из таблицы 3.3, а также результат теста на зависимость по Грейнджеру и метода сходящегося перекрестного отображения. В качестве результата теста Грейнджера в таблице приведены значение *p*-value для *F*-статистики (3.6).

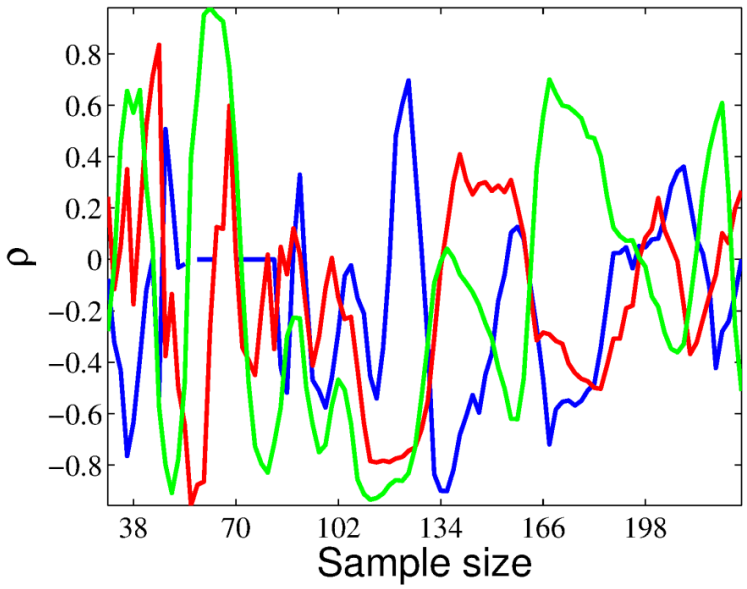
В графе «ССМ» встречается формулировка «Недостаточно данных». Так как метод сходящегося перекрестного отображения основан на проверке сходимости коэффициента корреляции (3.8) при неограниченном увеличении объема выборки, он требует большого объема выборки. Решения принимались при $|\mathcal{J}| > 50$.

Решение о наличии зависимости по Грейнджеру принималось при $p < 0{,}1$.

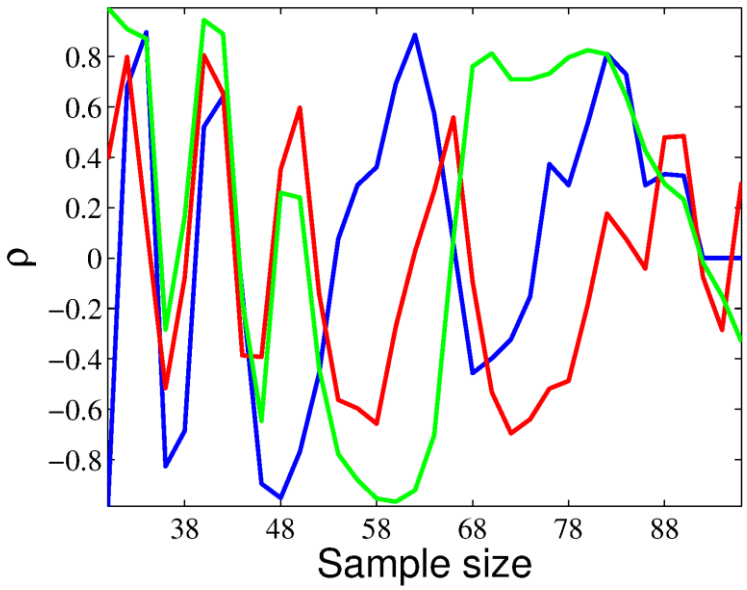
Также для каждой пары исследованных рядов приведено значение параметра задержки τ из модели (3.3)–(3.4). Перед применением теста Грейнджера все ряды были сглажены, после чего протестированы на стационарность с помощью теста Дики-Фуллера, нестационарные ряды были продифференцированы. Дифференцирование проводилось не более одного раза, так как большее количество преобразований могло бы затруднить интерпретацию полученных результатов.

В качестве результатов метода сходящегося перекрестного отображения в таблице приведены средние значения $\bar{\rho}_{\mathbf{s}}$ и $\bar{\rho}$ коэффициентов корреляции $\rho(\hat{\mathbf{s}}_T, \mathbf{s}_T)$ и $\rho(\mathbf{s}_T, \mathbf{x}_T)$. Решение о наличии или отсутствии сходимости коэффициента (3.7) принималось по виду графика зависимости соответствующего коэффициента от длины выборки $|\mathcal{J}|$. Графики для некоторых пар временных рядов изображены на рисунке 3.22: (a) – «Нефть и нефтепродукты» – «Цены на бензин»; (b) – «Нефть и нефтепродукты» – «Цены на мазут»; (c) – «Нефть и нефтепродукты» – «Цены на нефть»; (d) – «Сахар» – «Цены на сахар».

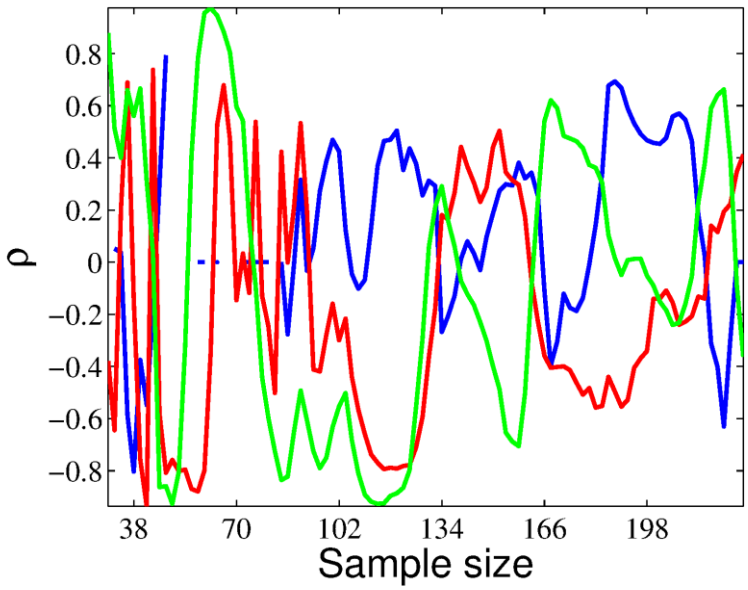
На графике красного цвета по оси ординат отложены значения коэффициента корреляции $\rho(\hat{\mathbf{x}}, \mathbf{x})$ спрогнозированых цен на нефть с их истинными значениями, синего – корреляция $\rho(\hat{\mathbf{s}}, \mathbf{s})$ прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляции истинных значений исследуемых рядов рядов.



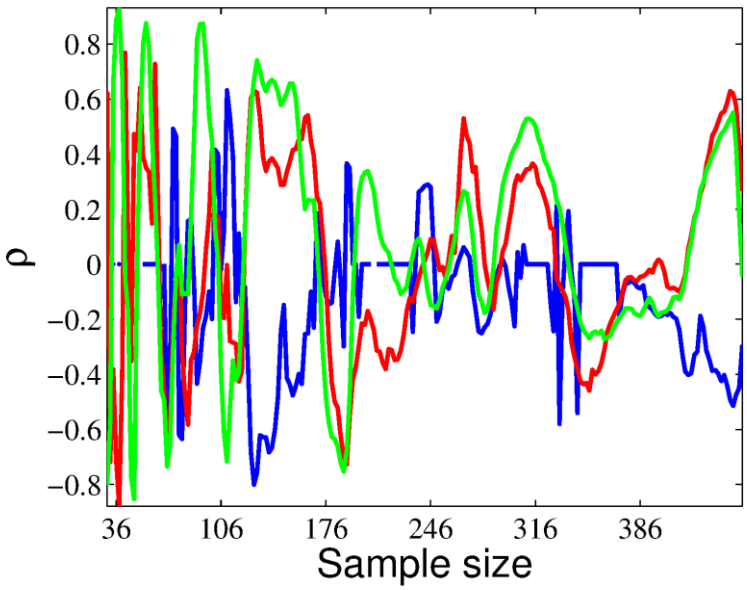
(a)


0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
ρ
38
48
58
68
78
88
Sample size

(b)


0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
ρ
38
70
102
134
166
198
Sample size

(c)



(d)

Рисунок 3.22 – Зависимости различных коэффициентов корреляции от объема выборки.

Для исследования влияния сезонности построен вспомогательный временной ряд **x**, $x_j \in [0, 1]$. Например, свекловичный сахар производится в сентябре–октябре. Для исследования влияния сезонности производства сахара на перевозки сахара будет составлен временной ряд, каждая точка которого соответствует одному дню. Значения ряда, соотвествующие месяцам с января по август включительно, приравниваются нулю, а значения ряда, соответствующие месяцам после октября – единице. Значения в сентябре–октябре вычисляются сглаживанием имеющихся значений.

Результаты, приведенные в таблице 3.8, следует понимать следующим образом. Каждый из рассмотренных методов исследует временные ряды на наличие некоторой связи, причем каждый в своем смысле. Таким образом, ни для одной из пар рядов, рассмотренных в вычислительном эксперименте, не принято решение о наличии связи в смысле метода сходящегося перекрестного отображения, хотя для некоторых из них сделан вывод о наличии зависимости по Грейнджеру. При этом ни наличие зависимости по Грейнджеру, ни зависимость в смысле метода сходящегося перекрестного отображения не доказывает наличие причинно-следственной связи между рядами. Однако определение зависимости по Грейнджеру согласуется с целью уточнить прогноз временного ряда **s**, используя значения временного ряда **x**,

поэтому, если принимается решение о зависимости рядов по Грейнджеру, то считается, что экспертное высказывание $\mu(\mathbf{s}, \mathbf{x})$ подтвердилось, и ему приписывается оценка достоверности $1 - p(\mathbf{s}, \mathbf{x})$.

3.5.4 Схема тестирования системы прогнозирования объемов ГЖДП с учетом влияния экзогенных факторов.

На рисунках 3.23 и 3.24 изображены уровни А-0 и А0 функциональной схемы проектируемой ситемы прогнозирования. На уровне А-0 (рис. 3.23) продемонстрированы входные, управляющие и выходные параметры. На уровне А0 (рис. 3.24) продемонстрированы основные блоки системы. Ниже приводится описание входных, управляющие и выходные параметров в соответствии со схемой А-0 на рис. 3.23.

1) Входные параметры:

Таблица грузоперевозок. Исторические данные об объемах ГЖДП это записи, содержащие дату, коды станции отправления и станции назначения, количество вагонов, которые прошли по маршруту от станции отправления до станции назначения, код груза, род вагонов, суммарный вес груза.

Экзогенные временные ряды. На основе экспертных высказываний о степени и характере влияния внешних факторов на ЖД грузоперевозки был составлен список (см. таблицу 3.1) видов экзогенных факторов, оказывающих наибольшее влияние на объемы ГЖДП. Для каждого из выделенных временных видов был собран набор соответствующих временных рядов, находящихся в свободном доступе: цены на перевозимые товары, курсы валют, погодные данные. Собранные временные ряды составляют базу экзогенных временных рядов, используемых для повышения качества прогнозирования.

Таблица соответствия экзогенных и эндогенных переменных. Для каждого вида экзогенных факторов из таблицы 3.1 эксперты указали тип грузов, на объемы перевозок которых данный фактор оказывает существенное влияние. На основе этих данных построена таблица 3.3 соответствия экзогенных и эндогенных переменных. В таблице 3.3 по горизонтали отложены виды экзогенных факторов из таблицы 3.3, по вертикали - коды перевозимых грузов, расшифровка которых приводится в таблице 3.2. В таблице 3.3 значением «+» выделены ячейки,

соответствующие парам факторов, относительно связи между которыми была высказана экспертная гипотеза. Пары факторов, к которым экспертные высказывания относятся неявно, помечены знаком «?». Знаком « ± » отмечены пары, принадлежность которых к одной из описанных ситуаций возможна, но не может быть установлена без привлечения дополнительной экспертной информации и/или статистической обработки. Пары факторов, о возможной связи между которыми никаких гипотез высказано не было, отмечены знаком «-». При получении запроса на прогнозирование, из таблицы экзогенных временных рядов выбираются временные ряды, соответствующие данному запросу согласно таблице 3.3: выбираются для проверки ряды, отмечтенные « + », « ± », « ?». Информативность выбранных временных рядов проверяется на этапе прогнозирования.

2) Управляющие параметры:

Запрос на прогнозирование с заданной детализацией. При запросе на прогнозирование пользователь задает детализацию прогнозов. Учитывая формат входных данных, возможна детализация по типам грузов (по всем/некоторым), родам вагонов (по всем/некоторым), веткам (по всем веткам, со всех веток на заданную, с заданных веток на заданные, с заданной ветки на все), узлам (по всем/некоторым), времени (по дням, неделям, помесячно).

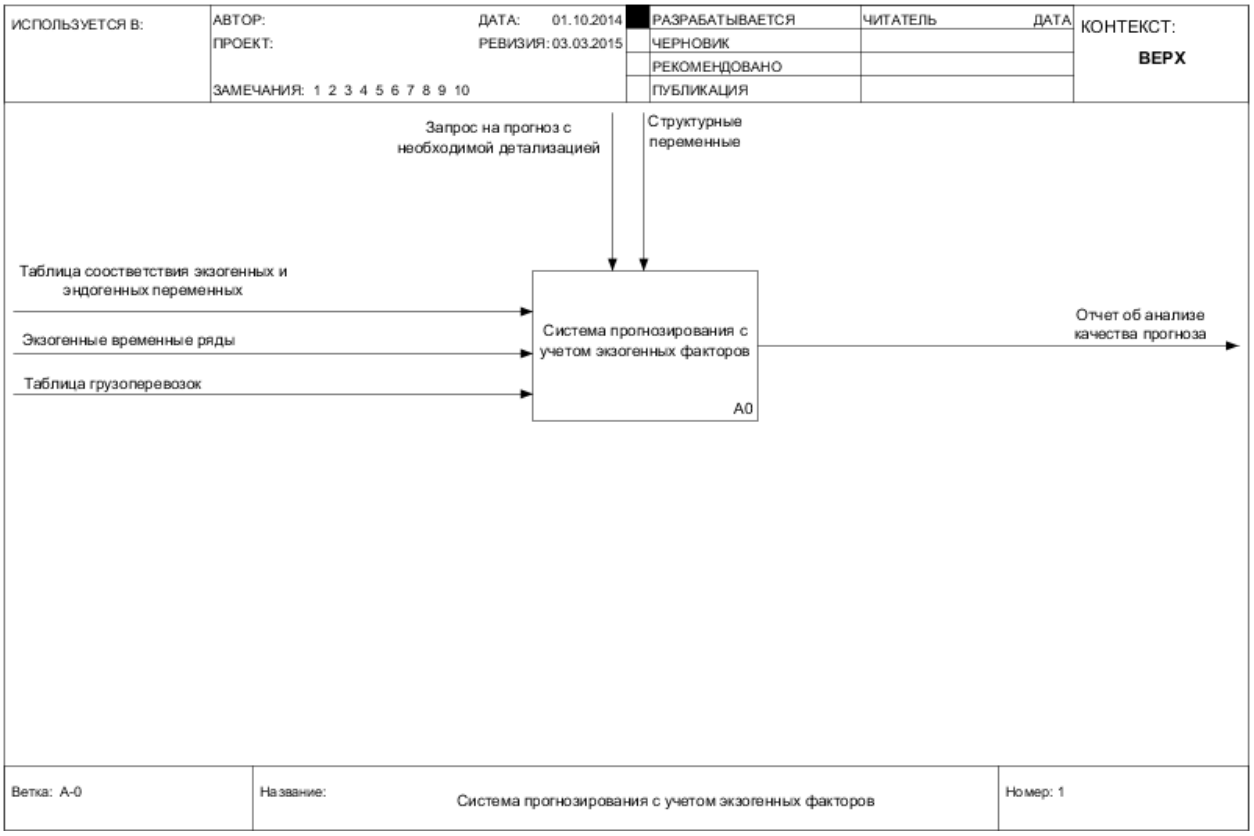


Рис. 3.23 - Уровень A-0 системы прогнозирования.

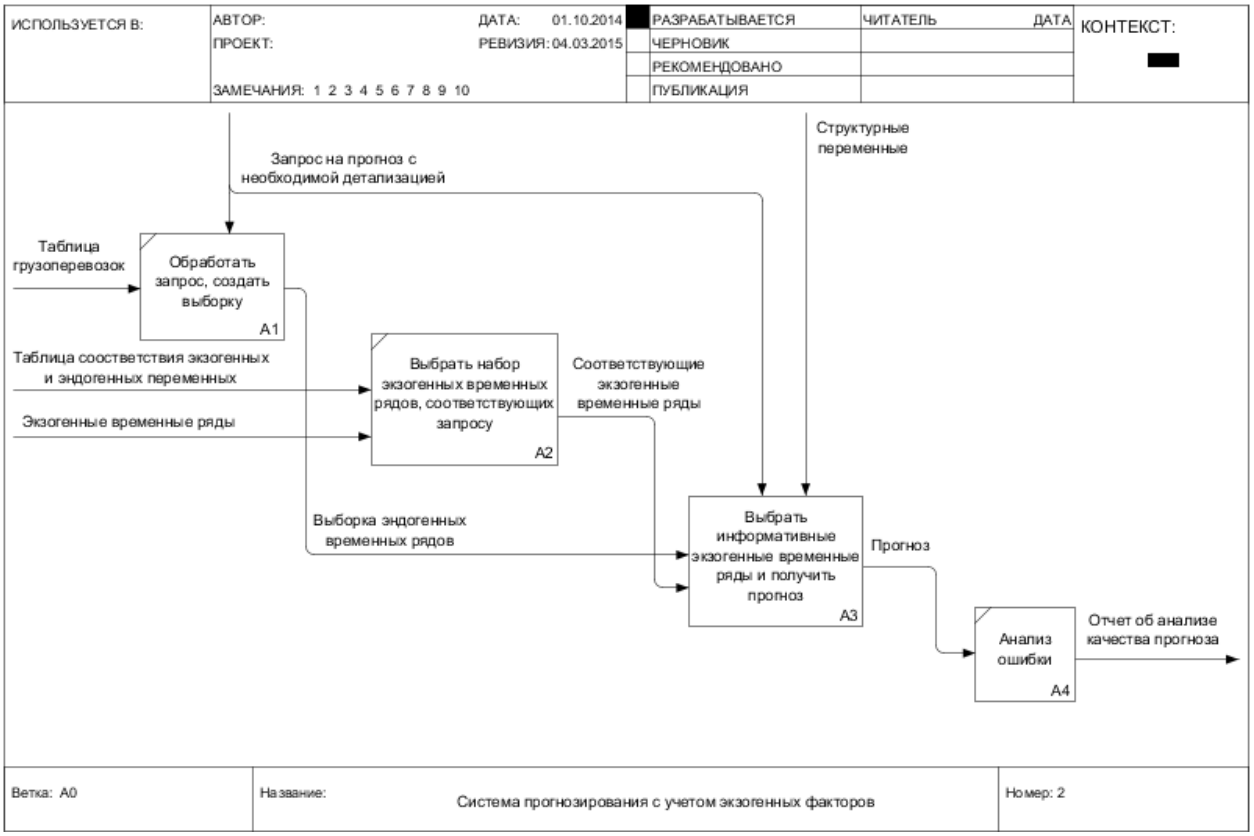


Рис. 3.24 - Уровень A0 системы прогнозирования.

Структурные параметры. Структурные параметры определяют способ про-

гнозирования:

- задают алгоритм прогнозирования и его внутренние параметры,
- определяют, учитывается ли при прогнозе потенциально информативные экзогенные факторы.

В таблице 3.9 перечислены возможные значения структурных параметров. Для каждого структурного параметра приводятся его название (Name), список полей (Fields), их значения по умолчанию (Default) и описание (Description).

Таблица 3.9 - Описание структурных параметров.

Name	Fields		Default	Description
Algorithm	Hist	Для получения прогноза строится свертка гистограммы временного ряда в функцией потерь. Оптимальным прогнозом является то значение центра сегмента гистограммы, которое доставляет минимальное значение свертке		
		K_{min}	5	Минимальное число столбцов гистограммы
		K_{max}	100	Максимальное число столбцов гистограммы
		w_{min}	0.1	Минимальный допустимый вес объекта, учитываемого при построении гистограммы
		$L(x, y)$	$(x - y)^2$	Функция потерь
Algorithm	ARMA			Модель авторегрессионного скользящего среднего
		p	5	Порядок лагирования авторегресионной части модели
		q	5	Порядок лагирования модели скользящего среднего
Exogneity			True	По умолчанию предлагается учитывать выбранные информативные экзогенные временные ряды

Name	Fields	Default	Description
Horizon		10	Количество прогнозов, вычисляемых для анализа качества при ретроспективном прогнозировании

3) Выходные параметры.

Отчет об анализе качества прогноза. Качество прогноза оценивается при ретроспективном прогнозировании. Анализ качества прогноза включает оценку необходимой длины предыстории и оценку достоверности получаемого прогноза при текущем объеме выборки. Проводится сравнение используемого алгоритма прогнозирования с базовыми (ARMA, ARMAX), прогноза с помощью расширенной модели (с учетом экзогенных факторов) и прогноза по историческим значениям.

4 Исследования предметной области, описание основных процессы в области организации грузовых железнодорожных перевозок. Выявление и исследование возможных источников данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов в историческом и текущем периоде

В этом разделе представлены результаты исследований предметной области:

- описаны основные процессы в области организации ГЖДП;
- описаны выявленные источники данных об объемах спроса на ГЖДП в историческом и текущем периоде;
- описаны выявленные источники данных об экзогенных факторах в историческом и текущем периоде.

4.1 Основные процессы в области организации ГЖДП

Основным средством управления и планирования ГЖДП является принятая и внедренная в РЖД система бюджетирования. Основная мотивация РЖД как Индустриального партнера в проведении и софинансировании данного ПНИ – повышение качества бюджетирования доходной части бюджета ГЖДП.

В результате интервьюирования менеджеров и специалистов Индустриального партнёра удалось установить, что критерием качества бюджетирования является результат сравнения целевых показателей бюджета на отчётный период и достигнутых значений этих показателей, рассчитанных по результатам работы индустриального партнера в отчётном периоде. Поэтому Индустриальный партнер заинтересован в снижении различий между запланированными и исполненными бюджетными показателями, то есть в повышении обоснованности и достоверности целевых показателей бюджета.

Для достижения цели ПНИ было проведено исследование предметной области – процессов управления и планирования ГЖДП РЖД, в частности процессов планирования и внесения изменений в доходную часть бюджета ГЖДП.

В процессах планирования Индустриального партнера производится периодическое формирование и утверждение бюджетов разных уровней на различные периоды планирования [99],[100]. В связи с этим было проведено исследование

нормативно-технической и управленческой документации, на основании которого была сформирована общая модель процессов жизненного цикла бюджета, применимая к любому иерархическому уровню организационной структуры РЖД, так же, как и к произвольному периоду планирования. При этом процессы взаимных согласований бюджетов и статей бюджетов между разными уровнями организационной структуры Индустриального партнёра не рассматривались.

Доходная часть бюджета как объект процессов планирования и управления исполнением имеет свой жизненный цикл, состоящий (см. рисунок 4.1) из следующих основных процессов:

- формирование доходной части бюджета,
- исполнение доходной части бюджета,
- мониторинг и контроль исполнения доходной части бюджета,
- корректировка доходной части бюджета.

На рисунке 4.1 изображена укрупнённая схема жизненного цикла доходной части бюджета ГЖДП. Специально обозначено место актуальных прогнозов объёмов ГЖДП. Под актуальным понимается прогноз на оставшуюся часть бюджетного периода, выполненный с учётом текущей статистической информации об объёмах ГЖДП и влияющих на них экзогенных факторов.

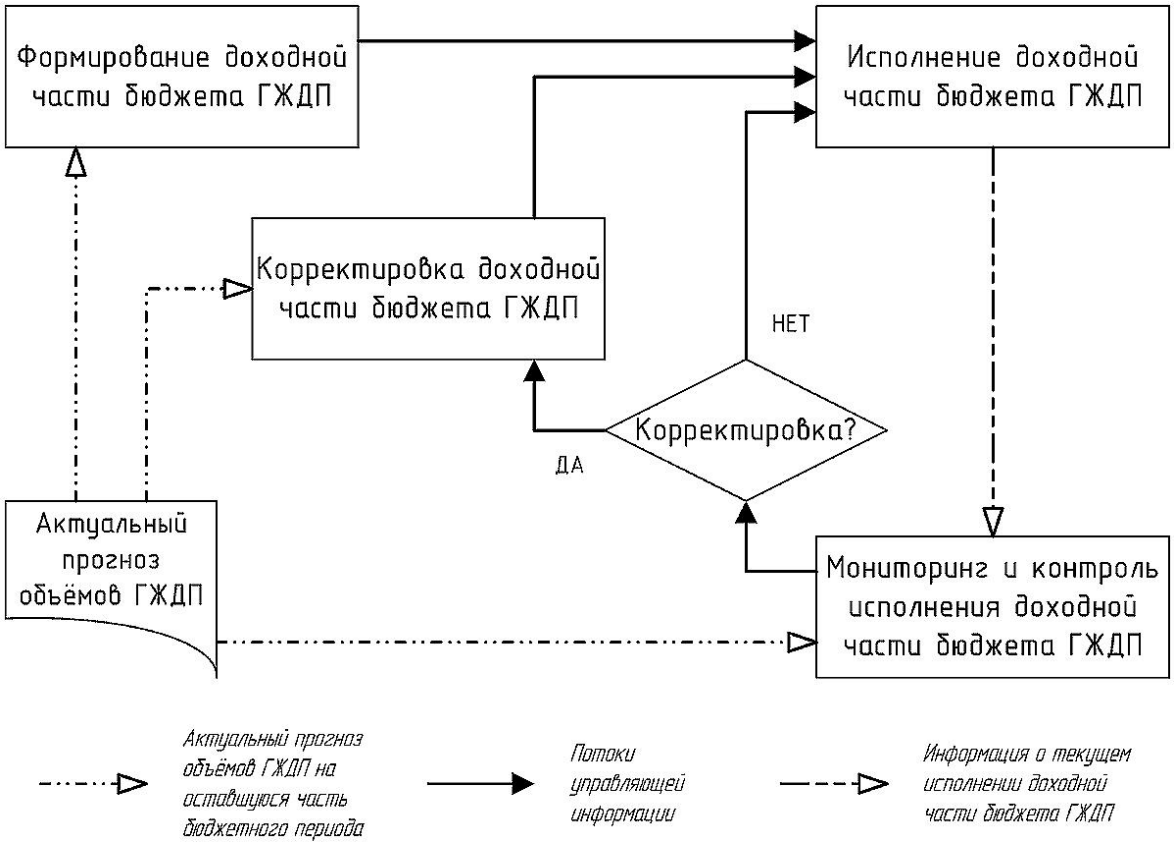


Рисунок 4.1 – Укрупнённая схема жизненного цикла доходной части бюджета ГЖДП

В интересах разработки математических методов и алгоритмов прогнозирования объемов спроса на ГЖДП с учетом влияния экзогенных факторов выполнено исследование двух групп процессов: группа формирования и группа взаимосвязанных процессов исполнения, контроля и пересмотра бюджета.

На рисунке 4.2 представлена модель группы процессов формирования бюджета, с указанием значимых объектов процессов, без привязки к субъектам. Модель исполнена в соответствии с нотацией eEPC.

Доходная статья бюджета рассчитывает на основании значений двух главных показателей ГЖДП:

- Показатель объема погрузки грузов на различных уровнях агрегации узлов железнодорожной цепи (погрузки)

- Показатель средней дистанции отправки грузов между станциями (отправки).

Указанные показатели рассчитываются на основании статистических данных аналогичного прошедшего периода и выражаются в натуральных величинах. Для перевода этих показателей в денежную форму дополнительно производится

расчет средней доходности от погрузок и отправок в аналогичном прошедшем периоде. Таким образом, основная часть доходной статьи бюджета периода рассчитывается как сумма дохода по погрузкам и отправкам, которые в свою очередь рассчитываются как умножение натуральных показателей на средние доходности.

Показатели погрузки и отправки рассчитываются на нескольких уровнях детализации (размерностях):

Погрузка типов грузов: показатель (тонны груза) имеет следующие «привязки»:

- Тип объекта. Тип объекта имеет следующую иерархию:

а) вся железнодорожная сеть РЖД;

б) направление железной дороги;

в) управление железной дороги;

г) железнодорожная станция.

- Тип груза

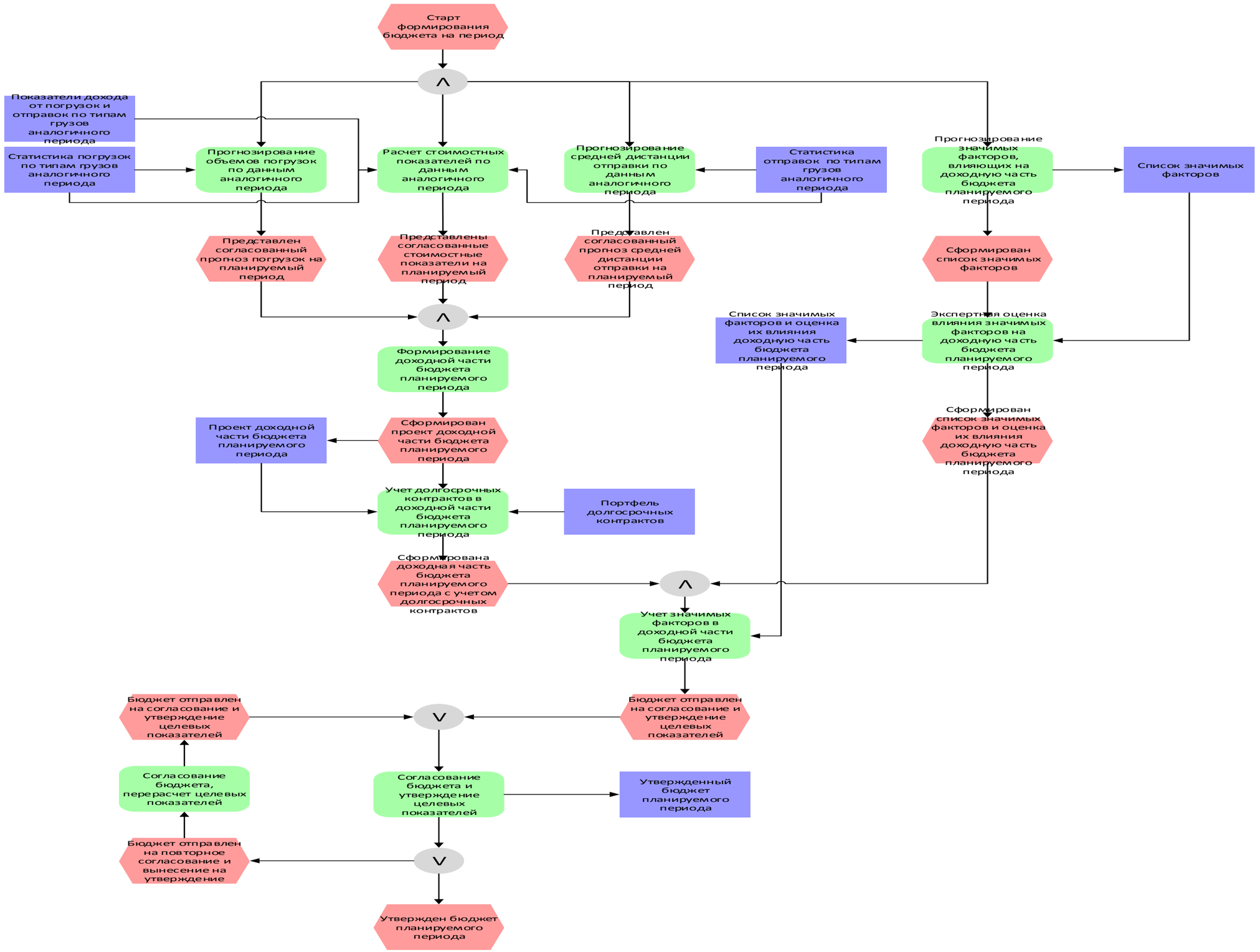


Рисунок 4.2 – Группа процессов формирования бюджета

Средняя дистанция отправки: количественный показатель (км) имеет следующие «привязки»:

- Тип объекта. Тип объекта имеет следующую иерархию

а) Вся железнодорожная сеть РЖД

б) Направление железной дороги

в) Управление железной дороги

- Тип груза

Прогнозирование показателей осуществляется простым методом мультипликации показателей аналогичного прошедшего периода на заданный постоянный коэффициент.

На основании спрогнозированных показателей, с учетом различных ограничений, формируется проект доходной части бюджета.

Для уточнения бюджета используется портфель (пул) долгосрочных контрактов на перевозку грузов, который позволяет скорректировать прогнозируемый спрос задокументированным спросом. На данный момент долгосрочные контракты занимают незначительную долю доходной части бюджета.

Помимо прогнозирования показателей и учета долгосрочных контрактов при подготовке доходной части бюджета производиться выявление и оценка значимых факторов, которые могут повлиять на доходы от грузовых железнодорожных перевозок в планируемом периоде. Для выделения таких значимых факторов, как и для их оценки, привлекаются эксперты, работники плановых отделов индустриального партнера. Выявленные факторы учитываются при корректировке доходной части бюджета согласно ранее подготовленным оценкам экспертов.

После утверждения бюджета производится процесс согласования, доработки и утверждения. На рисунке 4.3 представлена модель группы процессов исполнения, контроля и пересмотра бюджета, с указанием значимых объектов процессов, без привязки к субъектам.

В процессе протекания текущего периода производится детализированный расчет целевых показателей по подпериодам текущего периода (например, по месяцам в квартальном периоде). По окончанию каждого подпериода производится контроль целевых показателей всего периода и их соответствие бюджетным. При наличии значимого расхождения производится корректировка бюджета.

Рисунок 4.3 – Группа процессов исполнения, контроля и пересмотра бюджета

Отдельно проводится процесс мониторинга на наличие значимых факторов, способных существенно изменить планируемые показатели, их оценки и учета.

В текущей ситуации субъектами всех процессов выступают люди, подразделения и дочерние компании. Каждый из процессов проводится трудовым коллективом, автоматизации подвергнута лишь небольшая часть процессов прогнозирования, основанного на простых алгоритмах мультипликации статистических данных.

В результате исследования предметной области выявлены основные процессы управления и планирования грузовых железнодорожных перевозок (процессы – «as is»). Выделены основные процессы, в которых предполагается использование подсистемы прогнозирования, основанной на результатах данного ПНИ.

Будущая подсистема прогнозирования планируется к использованию в обеих группах процессов бюджетирования. В процессах формирования бюджета использование подсистемы прогнозирования при расчете показателей погрузок и отправок может существенно повысить качество итоговых бюджетов. Предлагается применение подсистемы и для прогнозирования и учета значимых факторов при составлении прогнозов. В процессах исполнения, контроля и пересмотра бюджета предлагается использование подсистемы при прогнозировании и учете новых значимых факторов.

В ходе дальнейшей проработки моделей и алгоритмов на втором этапе ПНИ в рамках п. 2.1.5 Технического задания представленные модели процессов будут уточнены и детализированы для учета специфики бизнес-процессов Индустриального партнера.

4.2 Выявление и исследование возможных источников данных об объёмах спроса на ГЖДП и влияющих на них экзогенных факторах в историческом и текущем периоде

4.2.1 Эндогенные факторы

Основным источником данных о спросе на ГЖДП являются учетные системы РЖД. По результатам исследования этих источников разработана реляционная модель данных, количественно описывающих эндогенные факторы. Схема реляционной модели представлена на рисунке 4.4.

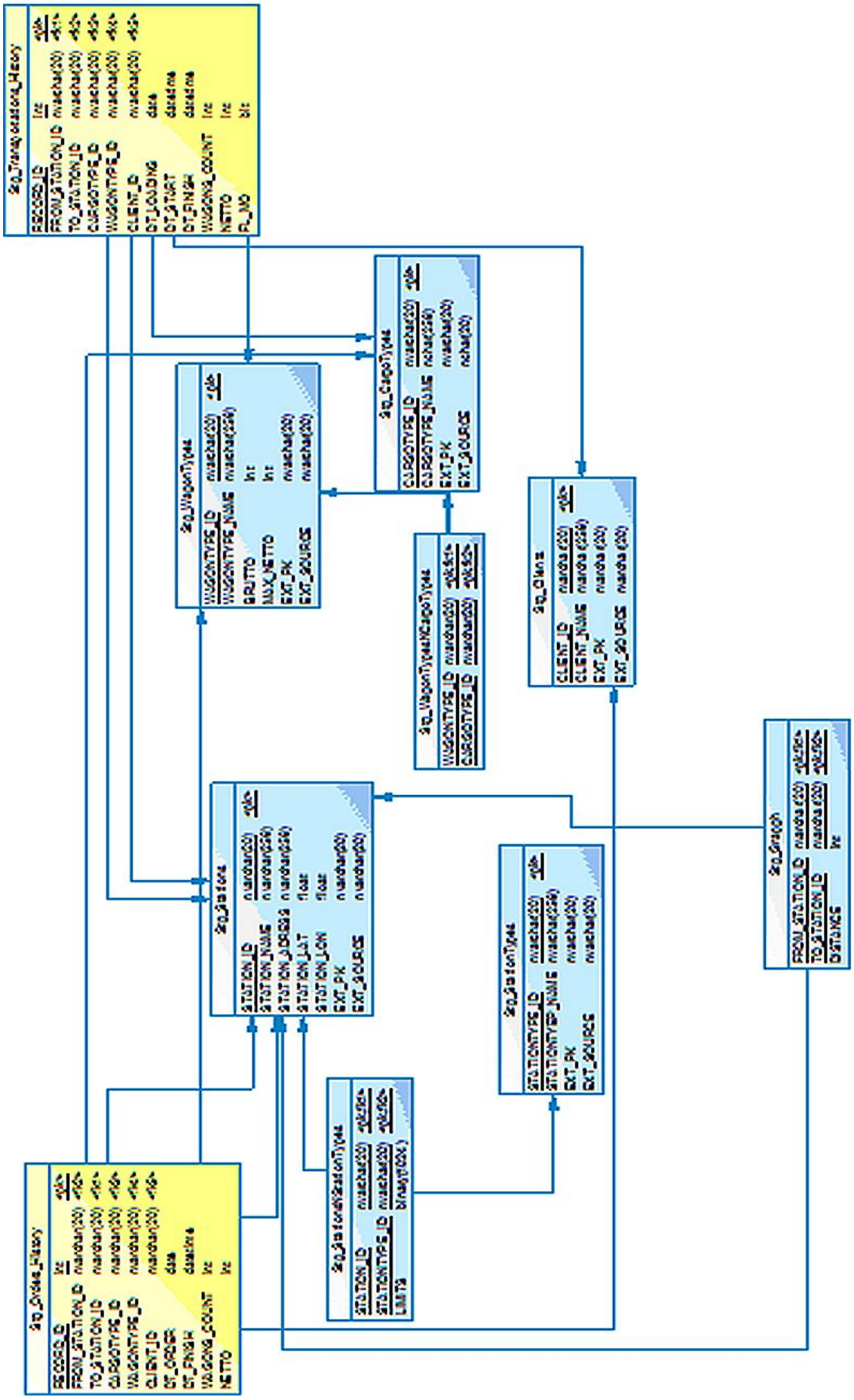


Рисунок 4.4 – Модель исходных данных

Описание представленных на рисунке 4.4 таблиц дано в таблицах 4.1–4.11.

Таблица 4.1 – Список таблиц в модели данных

Наименование таблицы	Назначение таблицы
Stg_CargoTypes	Справочник типов грузов
Stg_Clients	Справочник заказчиков
Stg_Grapgh	Справочник связей между станциями
Stg_Orders_History	История заявок на перемещение грузов
Stg_Stations	Справочник железнодорожных узлов
Stg_StationsNStationTypes	Справочник сопоставления станций и типов станций
Stg_StationTypes	Справочник типов станций
Stg_Transportations_History	История перемещений грузов
Stg_WagonTypes	Справочник типов подвижного состава
Stg_WagonTypesNCargoTypes	Справочник сопоставления типов подвижного состава и типов грузов

Описание полей таблиц, перечисленных в таблице 4.1, представлены в следующих таблицах.

Таблица 4.2 – Stg_CargoTypes

Наименование поля	Комментарий	Тип данных
CARGOTYPE_ID	Код типа груза	nvarchar(20)
CARGOTYPE_NAME	Наименование типа груза	nchar(256)
EXT_PK	Внешний первичный ключ	nvarchar(20)
EXT_SOURCE	Внешний источник данных	nchar(20)

Таблица 4.3 – Stg_Clients

Наименование поля	Комментарий	Тип данных
CLIENT_ID	Код клиента	nvarchar(20)
CLIENT_NAME	Наименование клиента	nvarchar(256)
EXT_PK	Внешний первичный ключ	nvarchar(20)
EXT_SOURCE	Внешний источник данных	nvarchar(20)

Таблица 4.4 – Stg_Grapgh

Наименование поля	Комментарий	Тип данных
FROM_STATION_ID	Код станции	nvarchar(20)
TO_STATION_ID	Код станции	nvarchar(20)
DISTANCE	Расстояние между станциями по длине железнодорожного пути	int

Таблица 4.5 – Stg_Orders_History

Наименование поля	Комментарий	Тип данных
RECORD_ID	Идентификатор записи	int
FROM_STATION_ID	Код станции	nvarchar(20)
TO_STATION_ID	Код станции	nvarchar(20)
CARGOTYPE_ID	Код типа груза	nvarchar(20)
WAGONTYPE_ID	Код типа подвижного состава	nvarchar(20)
CLIENT_ID	Код клиента	nvarchar(20)
DT_ORDER	Дата подачи заявки	date
DT_FINISH	Желаемые дата и время прибытия	datetime
WAGONS_COUNT	Количество вагонов	int
NETTO	Полный вес груза	int

Таблица 4.6 – Stg_Stations

Наименование поля	Комментарий	Тип данных
STATION_ID	Код станции	nvarchar(20)
STATION_NAME	Наименование станции	nvarchar(256)
STATION_ADRESS	Почтовый адрес станции по КЛАДР	nvarchar(256)
STATION_LAT	Широта GPS	float
STATION_LON	Долгота GPS	float
EXT_PK	Внешний первичный ключ	nvarchar(20)
EXT_SOURCE	Внешний источник данных	nvarchar(20)

Таблица 4.7– Stg_StationsNStationTypes

Наименование поля	Комментарий	Тип данных
STATION_ID	Код станции	nvarchar(20)
STATIONTYPE_ID	Код типа станции	nvarchar(20)
LIMITS	Ограничения	binary(1024)

Таблица 4.8 – Stg_StationTypes

Наименование поля	Комментарий	Тип данных
STATIONTYPE_ID	Код типа станции	nvarchar(20)
STATIONTYEP_NAME	Наименование типа станции	nvarchar(256)
EXT_PK	Внешний первичный ключ	nvarchar(20)
EXT_SOURCE	Внешний источник данных	nvarchar(20)

Таблица 4.9 – Stg_Transportations_History

Наименование поля	Комментарий	Тип данных
RECORD_ID	Идентификатор записи	int
FROM_STATION_ID	Код станции	nvarchar(20)
TO_STATION_ID	Код станции	nvarchar(20)
CARGOTYPE_ID	Код типа груза	nvarchar(20)
WAGONTYPE_ID	Код типа подвижного состава	nvarchar(20)
CLIENT_ID	Код клиента	nvarchar(20)
DT_LOADING	Дата погрузки	date
DT_START	Дата и время отправки	datetime
DT_FINISH	Дата и время прибытия	datetime
WAGONS_COUNT	Количество вагонов	int
NETTO	Полный вес груза	int
FL_MO	Признак маршрутной отправки	bit

Таблица 4.10 – Stg_WagonTypes

Наименование поля	Комментарий	Тип данных
WAGONTYPE_ID	Код типа подвижного состава	nvarchar(20)

Продолжение таблицы 4.10

Наименование поля	Комментарий	Тип данных
WAGONTYPE_NAME	Наименование типа подвижного состава	nvarchar(256)
BRUTTO	Вес подвижного состава	int
MAX_NETTO	Ограничение по весу груза	int
EXT_PK	Внешний первичный ключ	nvarchar(20)
EXT_SOURCE	Внешний источник данных	nvarchar(20)

Таблица 4.11 – Stg_WagonTypesNCargoTypes

Наименование поля	Комментарий	Тип данных
WAGONTYPE_ID	Тип подвижного состава	nvarchar(20)
CARGOTYPE_ID	Тип груза	nvarchar(20)

- Коды станций представляют собой шестизначные числа. Станции, в коде которых две первые цифры совпадают, входят в одну железнодорожную ветку. Станций отправления 1566, станций назначения 1902, веток 99.

- Код груза – натуральное число от 1 до 43. Также имеются перевозки, где код груза не указан. Расшифровки кодов грузов приведены в таблице 3.2.

- Род вагона – натуральное число. В имеющихся данных 75 различных типов вагонов.

- Признак маршрутной отправки: если несколько вагонов следуют по одному документу, в разные составы такие вагоны ставить нельзя.

Согласно используемым данным железнодорожный узел рассматривается без детализации по путям и по очередности отправления блоков вагонов.

4.2.2 Экзогенные факторы

Все экзогенные факторы, указанные в таблице 3.1 разделены на несколько основных групп по измеряемым значениям.

1. Биржевые цены на перевозимые грузы на российских и зарубежных биржах.

2. Сезонность и данные о погоде (для учета перевозок товарных грузов,

производство или спрос на которые зависит от сезонно-климатических условий, например, сельскохозяйственная продукция, мазут для ТЭЦ и др.).

3. Макроэкономическая ситуация в России и за рубежом.

К первой группе относятся биржевые цены на основные инструменты на российских и зарубежных биржах. Исторические данные по основным инструментам располагаются в открытом доступе, в частности, на ресурсах [87, 88]. Эти источники содержат данные по таким инструментам, как курсы валют (USD–EUR, USD–RUB, EUR–RUB), нефть (Brent), газ (GAZPROM), золото (Gold), серебро (Silver) и другие цветные металлы. В открытом доступе находятся сутотчные данные за несколько лет, за дополнительную плату можно приобрести данные с меньшим временным интервалом (в частности, миллисекундные).

Исторические данные цен на российские биржевые инструменты, относящиеся к исследуемым областям промышленности, располагаются на российском интернет-ресурсе [89]. Ресурс содержит открытые суточные данные по ценам на акции Норильского Никеля (GMK), Лукойла (LKOH), сбербанка (SBER), цены на пшеницу (ZWZ). Кроме того, используются данные российской биржи ММВБ-РТС, представленные на сайте биржи [90].

Предполагается использовать исторические данные Блумберга [91] об основных индексах Dow Jones, S&P, мировых ценах на нефть и газ, мировых ценах на металлы, а также ценах на сельскохозяйственную продукцию: пшеницу, овес, кукурузу, рис, бобы и сою.

Помимо биржевых данных, предполагается использовать данные Центробанка России [92] и Всемирного банка [93]. Используются открытые данные об активах банков, таких как золотовалютные резервы, выданные кредиты и государственные облигации, а также о пассивах – депозитах и капитале.

Второй группой данных являются суточные данные о погоде и сезонности. Исторические данные за несколько лет располагаются на открытых ресурсах [94, 95] и др. Данные содержат информацию о среднесуточной температуре, точке росы, влажности, давлении, силе ветра.

К третьей группе экзогенных показателей относятся основные макроэкономические показатели РФ. К этим показателям относятся валовой внутренний продукт, объемы производства, инвестиции, социальные показатели, инфляция, моне-

тарные показатели, валютный курс, внешняя торговля, бюджетные показатели, государственный долг, банковская система. Месячные данные по этим показателям в единицах прироста за 2011–14 гг. представлены в частности на Интернет-ресурсе [96].

5 Выполнение патентных исследований в соответствии с ГОСТ 15.011-96

В соответствии с пунктом 5.1 Технического задания (Приложение № 1 к Соглашению от 19 июня 2014 г. № 14. 604.21.0041 о предоставлении субсидии) на текущем этапе выполнения ПНИ были проведены предварительные патентные исследования.

Проект, в интересах которого проведены патентные исследования, имеет целью разработку 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».

Объектом патентных исследований являются способы, модели, методы и системы прогнозирования объемов спроса на грузовые железнодорожные перевозки.

Целью патентных исследований является исследование технического уровня разработок в области информационно-аналитической поддержки процессов управления грузовыми железнодорожными перевозками, учитывающих влияние экзогенных факторов на объемы спроса, а также специфику бизнес-процессов и нормативов индустриального партнера (ОАО «РЖД»), а также исследование технического уровня разработок, ориентированных на поддержку процессов анализа ретроспективных данных статистики объемов грузовых железнодорожных перевозок с использованием перспективных математических методов прогнозирования и обработки временных рядов.

Для достижения указанных целей в рамках проведенных патентных исследований были решены следующие задачи:

1) Исследование технического уровня разработок в области создания моделей прогнозирования объемов спроса на грузовые железнодорожные перевозки; методов обнаружения причинно-следственных связей в разнородных временных рядах объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов; алгоритмов построения кросс-корреляционных зависимостей временных рядов объемов спроса на грузовые железнодорожные перевозки и экзогенных фак-

торов.

2) Исследование тенденций развития и прогноз развития технического уровня разработок в области прогнозирования объемов спроса на грузовые железнодорожные перевозки, методов обнаружения причинно-следственных связей в разнородных временных рядах и алгоритмов построения кросс-корреляционных зависимостей временных рядов.

3) Исследование патентно-лицензионной ситуации в области охвата проекта.

Для достижения цели патентных исследований (в соответствии с регламентом) патентно-информационный поиск был выполнен по открытым базам данных официального сайта Российского агентства по патентам и товарным знакам (www.fips.ru); по открытым базам данных Европейского патентного ведомства (ЕПВ) (ep.espacenet.com), а также патентного сайта http://www.freepatentsonline.com/.

С учетом научно-технического уровня исследований и разработок в области охвата проекта, в интересах которого проведено настоящее патентное исследование, приоритетными для патентно-информационного поиска были охранные документы России, а также США, развитых стран Европы и динамически развивающихся стран Юго-Восточной Азии. При этом приоритетный поиск по фирмам не выполнялся, поскольку исследования и разработки в области охвата проекта, как правило, имеют фундаментальный характер и, следовательно, охранные документы самых разных фирм и организаций, специализирующихся в различных предметных областях и секторах рынка, могут быть использованы при создании моделей прогнозирования объемов спроса на грузовые железнодорожные перевозки.

С учетом динамики научно-технического уровня исследований и разработок в области охвата проекта, в интересах которого проведено настоящее патентное исследование, в качестве глубины патентного поиска были выбраны 15 лет (2001-2014 гг.). Кроме того, для фиксации исходных точек изобретательской активности в данной области в результаты выборочно включена информация о более ранних охранных документах.

С учетом структуры и состава рубрик в международных патентных классификаторах, а также целей и задач проекта, в интересах которого проведено настоя-

щее патентное исследование, патентно-информационный поиск проводился в отношении следующих классов МПК 7: G06F9/50, G06F15/00, G06F15/18, G06F17/00, G06F17/20, G06F17/21, G06F17/27, G06F17/28, G06F17/30, G06F19/00, G06K9/00, G06N3/00, G06Q10/04, G06Q50/30, G08G1/00, G08G1/01, G09B19/00, H04L12/70, H04M3/26, B61L27/00.

В соответствии с регламентом патентно-информационный поиск проводился только по зарегистрированным заявкам и патентам на изобретения и полезные модели в открытых базах официального сайта Российского агентства по патентам и товарным знакам (www.fips.ru) и Европейского патентного ведомства (ЕПВ) (ep.espacenet.com), а также патентного сайта http://www.freepatentsonline.com, а научно-техническая, конъюнктурная, нормативная документация и материалы государственной регистрации (отчеты о научно-исследовательских работах), полученные в процессе информационного поиска, в дальнейшем не использовались. Обоснованием принятого регламента патентно-информационного поиска является то, что в соответствии с пунктом 3.1 Технического задания (Приложение № 1 к Соглашению от 19 июня 2014 г. № 14. 604.21.0041 о предоставлении субсидии) эти материалы должны быть представлены в основной части отчета по проекту.

Достаточность выбранных параметров проведенного патентного исследования для обеспечения достоверности полученных результатов подтверждается тем, что глубина патентного поиска в 15 лет позволяет с высокой достоверностью оценить технический уровень разработок, исследовать тенденции развития и дать прогноз развития технического уровня разработок, а также проанализировать патентно-лицензионную ситуацию в области охвата проекта, в интересах которого проведено настоящее патентное исследование; патентная база данных Российского агентства по патентам и товарным знакам (www.fips.ru) содержит информацию обо всех охранных документах, зарегистрированных в России с учетом выбранной глубины патентного поиска; патентные базы данных Европейского патентного ведомства (ЕПВ) (ep.espacenet.com) и патентного сайта http://www.freepatentsonline.com/ содержат информацию обо всех охранных документах, зарегистрированных во всех странах, включая США, страны Европы и в страны Юго-Восточной Азии с учетом выбранной глубины патентного поиска; выбранные для патентного поиска классы международных патентных классифика-

торов полностью покрывают все основные направления исследований по проекту, в интересах которого проведено настоящее патентное исследование.

Проведенный в рамках выполненного патентного исследования анализ выявленных охраняемых технических решений показывает, что основная изобретательская активность в области охвата проекта, в интересах которого проведено настоящее патентное исследование, связана с получением охранных документов на способы, методы и системы прогнозирования на основе анализа временных рядов; в основе используемых в охранных документах решений, как правило, лежат классические методы интеллектуального анализа данных, а технический эффект достигается за счет уменьшения вычислительной сложности предлагаемых алгоритмов и сокращения временных горизонтов прогнозирования; практически отсутствуют охранные документы, защищающие технические решения, ориентированные на обработку экспертных оценок, а выявленные охранные документы обеспечивают достижение технического результата за счет использования известных методов определения коэффициентов ранговой корреляции.

Выводы и обобщения по результатам анализа современного технического уровня разработок в исследуемой области техники представлены в отчете о ПИ в отдельном документе.

По результатам проведенного предварительного патентного исследования рекомендовано на заключительном этапе ПНИ при получении результатов интеллектуальной деятельности, способных к правовой охране в соответствии со ст. 1225 ГК РФ, провести дополнительные патентные исследования в соответствии с ГОСТ Р 15.011-96.

Полный отчет о ПИ представлен отдельным документом.

С учетом вышесказанного, основными направлениями разработок в рамках выполняемого проекта, где должны быть сосредоточены усилия на получении патентоспособных технических решений, соответствующих установленному в рамках выполненного патентного исследования современному техническому уровню и современным тенденциям развития исследуемой области техники, можно считать фундаментальные исследования и прикладные разработки в области интеллектуального анализа данных, включая разработку

- перспективных математических методов прогнозирования на базе ста-

тистически значимых ретроспективных данных;

- методов кластеризации и классификации, а также методов и средств обработки временных рядов с акцентом на эффективных алгоритмах выявлении причинно-следственных связей в разнородных временных рядах;
- методов обработки результатов многопараметрических экспертных оценок.

Научная значимость и прикладная перспективность выполняемых прикладных научных исследований и разработок по проекту подтверждается результатами выполненных патентных исследований, поскольку информационно-патентный поиск не выявил охранных документов, которые могли бы рассматриваться в качестве прямых прототипов и аналогов для методов и средств, разрабатываемых по данному проекту. При этом важно отметить, что охранные документы, которые могут рассматриваться в качестве косвенных аналогов, ориентированы на достижение технических результатов за счет эффективной реализации известных классических подходов к прогнозированию и интеллектуальному анализу данных, в то время как модели, методы и средства, разрабатываемые в рамках проекта, в интересах которого выполнено настоящее патентное исследование, ориентированы на инновационные решения, отсутствующие в охранных документах, выявленных по результатам информационно-патентного поиска.

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

В соответствии с регламентом патентно-информационный поиск был выполнен по открытым базам данных официального сайта Российского агентства по патентам и товарным знакам (www.fips.ru), по открытым базам данных Европейского патентного ведомства (ЕПВ) (ep.espacenet.com), а также патентного сайта http://www.freepatentsonline.com и проводился в классах МПК 7 G06F9/50, G06F15/00, G06F15/18, G06F17/00, G06F17/20, G06F17/21, G06F17/27, G06F17/28, G06F17/30, G06F19/00, G06K9/00, G06N3/00, G06Q10/04, G06Q50/30, G08G1/00,

G08G1/01, G09B19/00, H04L12/70, H04M3/26, B61L27/00 на глубину 15 лет с включением выборочной информация о более ранних патентах для фиксации исходных точек изобретательской активности в области охвата проекта. При этом с учетом научно-технического уровня исследований и разработок в области охвата проекта, в интересах которого выполнялось настоящее патентное исследование, приоритетными для патентно-информационного поиска были охранные документы, зарегистрированные в Российской федерации, а также в США, развитых странах Европы и динамически развивающихся странах Юго-Восточной Азии. Приоритетный поиск по фирмам не выполнялся, поскольку исследования и разработки в области охвата проекта, как правило, имеют фундаментальный характер в силу чего охранные документы самых разных фирм и организаций, специализирующихся в различных предметных областях и секторах рынка, могут быть использованы при создании моделей прогнозирования объемов спроса на грузовые железнодорожные перевозки.

Соответствие заданию на выполнение настоящего патентного исследования и его полнота подтверждается тем, что глубина патентного поиска в 15 лет позволяет с высокой достоверностью оценить технический уровень разработок, исследовать тенденции развития и дать прогноз развития технического уровня разработок, а также проанализировать патентно-лицензионную ситуацию в области охвата проекта, в интересах которого проведено настоящее патентное исследование; патентная база данных Российского агентства по патентам и товарным знакам (www.fips.ru) содержит информацию обо всех охранных документах, зарегистрированных в России с учетом выбранной глубины патентного поиска; патентные базы данных Европейского патентного ведомства (ЕПВ) (ep.espacenet.com) и патентного сайта http://www.freepatentsonline.com/ содержат информацию обо всех охранных документах, зарегистрированных во всех странах, включая США, страны Европы и в страны Юго-Восточной Азии с учетом выбранной глубины патентного поиска; выбранные для патентного поиска классы международных патентных классификаторов полностью покрывают все основные направления исследований по проекту, в интересах которого проведено настоящее патентное исследование.

Таким образом, выполненное патентное исследование полностью соответствует заданию на их проведение и регламенту патентного поиска.

ЗАКЛЮЧЕНИЕ

В результате выполнения работ этапа 1 ПНИ получены следующие результаты.

1 Аналитический обзор современной научно-технической, нормативной, методической литературы, затрагивающей научно-техническую проблему, исследуемую в рамках НИР.

Анализ и обзор Аналитический обзор современной научно-технической, нормативной, методической литературы показал, что предложенные подходы и методы являются достаточно эффективными и при соответствующей адаптации и доработке смогут повысить качество прогнозирования объемов ГЖДП в РЖД и, как следствие, повысить качество планирования движения железнодорожных составов.

2 Обоснованы и выбраны направления дальнейших исследований с целью разработки математической модели прогнозирования объемов спроса на грузовые железнодорожные перевозки, учитывающей влияние экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки. В качестве основы для разработки метода непараметрического прогнозирования объемов спроса на ГЖДП, учитывающего влияние экзогенных факторов, а также специфику бизнес-процессов и нормативов РЖД выбран разработанный в ВЦ РАН алгоритм гистограммного прогнозирования hist.

3 Сформировано описание основных процессов в области организации грузовых железнодорожных перевозок, описание выявленных источников данных объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов в историческом и текущем периоде. Выделены основные процессы, в частности, процессы жизненного цикла бюджета различных периодов индустриального партнера, в которых предполагается использование подсистемы прогнозирования, основанной на результатах выполнения данного ПНИ.

4 По результатам исследований экзогенных факторов и их влияния на прогноз объема спроса на грузовые железнодорожные перевозки получены следующие результаты:

4.1 Перечень основных экзогенных факторов, влияющих на объемы спроса на грузовые железнодорожные перевозки, а также результаты экспертного анализа

значимости и характера влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки, представленные в виде таблицы оценки значимости и характера влияния. Полученные результаты использованы в дальнейшей работе по изучению и тестированию влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки.

4.2 Предложен метод обнаружения причинно-следственных связей в разнородных временных рядах объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов. Выявлено, что в результате применения описанного метода обнаружения причинно-следственных связей в разнородных временных рядах объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов возможно построение как точного прогноза объемов спроса на грузовые железнодорожные перевозки, так и прогноза, несоответствующего действительности.

4.3 Предложен метод оценки достоверности экспертных высказываний о влиянии экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки. Предложенный метод применен при формировании списка основных экзогенных факторов, влияющих на объемы спроса на грузовые железнодорожные перевозки, а также при формировании результатов экспертного анализа значимости и характера влияния экзогенных факторов на объемы спроса на грузовые железнодорожные перевозки.

4.4 Предложены способы расширения подхода Грейнджера для обнаружения структуры связей между временными рядами экзогенных факторов и объемов спроса на грузовые железнодорожные перевозки, а также алгоритм, позволяющий использовать ранее полученные данные о связях между временными рядами и структуре связей для обнаружения новых связей. Алгоритм позволяет распространять информацию о структуре связи на ранее не рассматривавшийся временные ряды путем кластеризации временных рядов в предположении, что сходные временные ряды прогнозируются моделями одной и той же или близкой структуры.

4.5 Результаты тестирования алгоритма построения кросс-корреляционных зависимостей временных рядов объемов спроса на грузовые железнодорожные перевозки и экзогенных факторов, показавшие что при принятии решения о связи между временными рядами важно учитывать структуру связи. В противном случае,

даже при проверке заведомо зависимых временных рядов на взаимосвязь с неадекватной структуры может быть принято решение об отсутствии связи.

Исследование экзогенных факторов и их влияния на прогноз объема спроса на грузовые железнодорожные перевозки позволило сформировать методологический и технологический задел для дальнейшего исследования и разработки математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования объемов спроса на грузовые железнодорожные перевозки).

5 Проведены патентные исследования, показавшие нарастающий интерес к разработкам и исследованиям по текущей теме ПНИ и схожим темам, а также высокую активность в области защиты прав во многих странах мира. Область охвата проекта относится к наукоемким секторам изобретательской активности, где используются результаты фундаментальных исследований по интеллектуальному анализу данных. Проанализировано, что среди выявленных в результате информационно-патентного поиска охранных документов нет патентов и заявок на изобретения, которые могут препятствовать применению результатов ПНИ в Российской Федерации, а также препятствовать получению охранных документов в других странах.

Таким образом, задачи этапа 1 проекта решены в полном объёме.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kullman, B. C. A model of rail/truck competition in the intercity freight market. Massachusetts Inst. of Technology. Dept. of Civil Engineering. Thesis. 1973. Ph. D.

2. Baumol W. J., Vinod H. D. An Inventory Theoretic Model of Freight Transportation Demand. // Management Science, 1970. Vol. 16(7), pp. 413–21.

3. Oum Tae H. A Cross-Sectional Study of Freight Transport Demand and Rail-Truck Competition in Canada. // The Bell J. of Economics, 1979. Vol. 10(3), pp. 463–82.

4. Guglielminetti P., Leyvraz J.-P., Rivier R. Freight transport planning: An optimisation model for transalpine rail network // Swiss Transport Research Conference, 2001.

5. de Jong G., Gunn H. F., Walker W. National and international freight transport models: an overview and ideas for further development. // Transport Reviews, 2004. Vol. 24(1), pp. 103–124

6. Pendyala R. M., Shankar V. N McCullough R. G. Freight travel demand modeling: synthesis of research and development of a framework, Transportation Research Record, 1725, J. of the Transportation Research Board // National Research Council, Washington D.C., 2000, pp 9–16.

7. Garrido R. A. Spatial interaction between trucks flows through the Mexico-Texas border. // Transportation Research Part A: Policy and Practice, 2000. Vol. 33(1), pp. 23–33.

8. Tavasszy L. A., et al. (1998) Scenario-wise analysis of transport and logistic systems with a SMILE. // Paper presented at the 8th WCTR conference, Antwerp, Belgium.

9. ASTRA consortium. Final report: assessment of transport strategies. / University of Karlsruhe, Germany. 2000.

10. Tsekeris T., Stathopoulos A. Treating Uncertain Demand Information in Origin — Destination Matrix Estimation with Traffic Counts // J. of Transportation Engineering, 2008. Vol. 134(8), pp. 327-337.

11. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. –М.: Финансы и статистика. 2003. –415 с.

12. Хардле В. Прикладная непараметрическая регрессия. М.: Мир. 1993. –349 с.

13. Cortez P., et al. Evolving time series forecasting ARMA models. // J. of Heuristics, 2004. Vol. 10(4). pp. 419–429.

14. McFadden D., Talvitie A., and Associates. Forecasting the Values of Exogenous Variables: Socioeconomic Variables. // Demand Model Estimation and Validation.. Chapter 3, Part 4. –University of California, 1977.

15. Rao P. S. Forecasting the demand for railway freight services // J. of Transp. Econ. and Policy, 1978. Vol 12(1), pp.7–26.

16. Hawthorne J., et al. A Rail Freight Forecasting Model for the Strategic Rail Authority. Sinclair Knight Merz and Strategic Rail Authority, UK // Estimation of railway capacity consumption using stochastic differential equations. 2012. –18 p.

17. Shan Z., Qiu R., Yingjie Y. A MANOVA of Historical and Unhistorical Factors for Traffic Volume Forecasting // Proceed. of the 2nd Int. Conf. ICTIS. China: June 29-July 2, 2013. pp. 391–397.

18. Soriguera F. Deriving Traffic Flow Patterns from Historical Data. // J. Transp. Eng., 2012. Vol 138(12), pp. 1430–1441.

19. Duddu V. R., Pulugurtha S. S. Principle of Demographic Gravitation to Estimate Annual Average Daily Traffic: Comparison of Statistical and Neural Network Models // J. of Transportation Engineering, 2013. Vol. 139(6), pp. 585–595.

20. Lindfeldt A. Congested railways: Influence of infrastructure and timetable properties on delay propagation // Royal Institute of Technology. –Stockholm, Sweden. 2012. –44 p.

21. Huang B., et al. A Simulation Study for Minimizing Operating Speed Variation of Multilane Highways by Controlling Access // Proceed. of the 2nd Int. Conf. ICTIS. China: June 29-July 2, 2013, pp. 1393–1401.

22. Medeossi G. Capacity and reliability on railway networks: a simulative approach. PhD thesis. University of Trieste. 2009. –140 p.

23. Stok R. Estimation of railway capacity consumption using stochastic differential equations. PhD thesis. University of Trieste, 2005.

24. Попов П. А. Стохастическое моделирование движения поездов в задачах интервального регулирования. // Наука и транспорт, 2013. Т. 6, № 2, сс. 8–19.

25. Кокурин И. М., Катцын Д. В., Малыгин И. Г. Организация экспорта каменного угля на основе концентрации логистической цепи поставок. Наука и транспорт, 2013. Т. 6, № 2, сс. 15–17.

26. Иванченко В. Н. Современные информационные технологии управления сложными процессами расформирования–формирования поездов. // Наука и транспорт. 2013. Т. 6, № 2, сс. 64–69.

27. Dollevoet T., et al. An iterative optimization framework for delay management and train scheduling. // Econometric Inst. Report EI2012-10. 2012. –23 p.

28. Tong M., Xue H. Highway Traffic Volume Forecasting Based on Seasonal ARIMA Model // J. of Highway and Transp. Res. and Develop., 2008. Vol. 3(2), pp. 109–112.

29. Jiang L., et al. Prediction of Traffic Volume in Bridge Load Random Process Based on Grey Markov Chain // J. of Highway and Transp. Res. and Develop., 2012. Vol. 6(1), pp. 61–65.

30. Mei H., et al. Short-Term Traffic Volume Prediction for Sustainable Transportation in an Urban Area // J. of Comp. in Civil Engineering, 2013. Online Publication: http://ascelibrary.org/doi/pdf/10.1061/(ASCE)CP.1943–5487.0000316.

31. Shi X., Ren Q.. The Advanced Nonparametric Model for Short-Term Traffic Volume Forecasting // Proc. of the 11th Int. Conf. of Chinese Transport. Prof. (ICCTP '11). hina, August 2011. pp. 1442–1453.

32. Draper N., Smith H. Applied Regression Analysis. –N.Y.: Wiley, 1998. –736 p.

33. Магнус Я. Р. Эконометрика. Нач. курс. Учебн. –М.: Дело. 2004. –576 с.

34. Федорова В. П. Локальные методы прогнозирования временных рядов. Дипл. раб. / Научн. рук.: Дьяконов А. Г. –М.: МГУ им. М. В. Ломоносова, 2009.

35. McNames J. Innovations in local modeling for time series prediction. Ph.D. thesis. Stanford University (USA), 1999.

36. Стрижов В. В. Методы выбора регрессионных моделей. –М.: ВЦ РАН, 2010. –60 с.

37. Вальков А. С., Кожанов Е. М., Мотренко А. П., Хусаинов Ф. И. Построение кросскорреляционных зависимостей при прогнозе загруженности железнодорожного узла // Машинное обучение и анализ данных. 2013. Т. 1, № 1, сс. 505–51.

38. Koenker R., Bassett G., Jr. Regression Quantiles // Econometrica, 1978. Vol.46 (1), pp. 33–50.

39. Friedman J., et al. Additive Logistic Regression: a Statistical View of Boosting. // Annals of Statistics, 1998. Vol. 38(2), pp. 337–374.

40. Архангельский Н. Е. и др. Экспертные оценки и методология их использования. –М.: Изд-во МЭСИ, 1974. –125 с.

41. Евланов Л. Г., Кутузов В. А. Экспертные оценки в управлении. –М.: Экономика, 1978. –133 с.

42. Бешелев С. Д., Гурвич Ф. П. Экспертные оценки. –М.: Наука, 1982. –161 с.

43. Ларичев О. И. и др. Выявление экспертных знаний. / Под ред. С. В. Емельянова. –М.: Наука, 1989. –128 с.

44. Литвак Б. Г. Экспертные технологии в управлении. –М.: Изд-во: Дело, 2004. –400 с.

45. Орлов А. И. Организационно-экономическое моделирование. Учебн. в 3 ч. / Ч. 2 : Экспертные оценки. – М.: Изд-во МГТУ им. Н. Э. Баумана. 2011. –486 с.

46. Sangüesa R., Cortés U. New perspectives on Causal Networks: the first CaNew workshop. // Int. J. of Approximate Reasoning, 2000. Vol. 24, pp. 1–10.

47. Kullback S., Information Theory and Statistics. –N.Y.: Wiley, 1959.

48. Chernoff H. A Measure of Asymptotic Eciency for Tests of a Hypothesis Based on the sum of Observations. // Ann. Math. Statist., 1952. Vol. 23(4), pp. 493–655

49. Колмогоров А. Н. О приближении распределений сумм независимых слагаемых неограниченно делимыми распределениями. // Труды Моск. мат. общ., 1963. Т. 12, сс. 437–451.

50. Ali S. M., Silvey S. D. A general class of coecients of divergence of a distribution from another. // J. of Royal Statistical Soiety. Series B (Methodoligical). 1966. Vol. 28(1), pp. 131–142.

51. Csiszár I. Shields P. C. Information theory and statistics: A tutorial. // Foundations and Trend in Communications and Information Theory. –Hanover (USA): now Publishers Inc. 2004. –115p.

52. Gibbs A. L., Su F. E. On Choosing and bounding probability metrics // International Statistical Rev., 2002. Vol. 70(3), pp. 419–435.

53. McFadden D., Talvitie A., and Associates. Forecasting the Values of Exogenous Variables: Socioeconomic Variables. // Demand Model Estimation and Validation.. Chapter 3, Part 4. –University of California, 1977.

54. Muralikrishna M., DeWitt D. J. Equi-depth multidimensional histograms // Proc. of the 1988 ACM SIGMOD int. conf. on Management of Data. Vol. 17, Issue 3, pp. 28–36.

55. Gunopulos D., et al. Approximating multi-dimensional aggregate range queries over real attributes. // Proc. of the 2000 ACM SIGMOD int. conf. on Management of Data. Vol. 29, Issue 2, pp. 463–474.

56. Thaper N., et al. Dynamic multidimensional histograms. // Proc. of the 2002 ACM SIGMOD int. conf. on Management of Data. pp. 428–439.

57. Arroyo J., Maté C. Forecasting histogram time series with k-nearest neighbors methods. // International J. of Forecasting, 2009. Vol. 25(1), pp. 192–207.

58. Arroyo J., et al. Forecasting with interval and histogram data. Some financial applications. // Handbook of Empirical Economics and Finance. –CRC Press, 2010. pp. 247–280.

59. Hamill T. M. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. // J.: Mon. Weather Rev., 2001. Vol. 129, pp. 550–560.

60. Grunfeld Y., Griliches Z. Is aggregation necessarily bad? // The Review of Economics and Statistics, 1960. Vol. 42(1), pp. 1–13.

61. Fogarty D. W., Blackstone J. H., Hoffman T. R. Production and Inventory Management. 2nd ed., –Cincinnati (USA): South-Western Publishing Co., 1991/. –870 p.

62. Van Erven T., Cugliari J. 2013. Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. // URL: https://hal.inria.fr/hal-00920559, (accessed October 21, 2014).

63. Меньшиков И. С. Лекции по теории игр и экономическому моделированию. –М.: Изд-во. Контакт Плюс, 2010. –336 с.

64. Granger C. W. J.. Investigating Causal Relations by Econometric Models and Cross-spectral Methods // Econometrica, 1969.Vol. 37, pp. 424–432.

65. Hamilton J. D.. Time series analysis. Princeton University Press, 1994. –816 p.

66. Kwiatskowski D., et al. Testing the null hypothesis of stationarity against the alternative of a unit root. // J. of Econometrics, 1992. Vol. 54, pp. 159–178.

67. Akaike H. A new look at the statistical model identification. // IEEE Trans. Autom. Control, 1974. Vol. 19, pp. 716–723.

68. Schwarz G. Estimating the dimension of a model. // Annals of Statistics, 1978. Vol. 6(2), pp. 461–464.

69. Gujarati D. Basic Econometrics, 5th ed. –The McGraw-Hill Companies, 2009. –922 p.

70. Maharaj E. A. Cluster of time series. // J. of Classifcation, 2000. Vol. 17(2), pp. 297–314.

71. Стрижов В. В., Крымова Е. А., Выбор моделей в линейном регресссионном анализе // Информационные технологии, 2011. Т. 10, сс. 21–26.

72. Кузнецов М. П., Стрижов В. В., Медведникова М. М. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах. // Научно-технический вестник СПб ГПУ. Информатика. Телекоммуникации. Управление, 2012. № 5, pp. 92–95.

73. Рудой Г. И., Стрижов В. В. Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных. // Информатика и её применения, 2013. Т. 7, № 1, сс. 17–26.

74. Douzal-Chouakria A. Classification trees for time series. // Pattern Recognition, 2012. Vol. 45(3), pp. 1076–1091.

75. Глибовец Н. Н., Медвидь С. А. Генетические алгоритмы и их использование для решения задачи составления расписания // Кибернетика и системный анализ, 2003. № 1. сс. 95–108.

76. Sivanandam S. N., Deepa S. N. Introduction to Genetic Algorithms. –Springer Berlin Heidelberg, 2008. –442 p.

77. Abarbanel H. D. I. Tools for the Analysis of Chaotic Data // Nonlinear Signal and Image Analysis, 2006. V. 808, pp. 1–17.

78. Bischi G. I. (ed.), Chiarella C. (ed.), Gardini L. (ed.). Nonlinear Dynamics in Economics. –Berlin: Springer, 2010. –381 p.

79. Lo T. T., Hsu H. H.. Change in the dominant decadal patterns and the late 1980s abrupt warming in the extratropical Northern Hemisphere. // Atmospheric Science Letter, 2010. Vol. 11(3), pp. 210–215.

80. Chen Y., Bressler S. L., Ding M.. Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. // J. of Neuroscience Methods, 2006. Vol. 150, pp. 228–237.

81. May R. M., Levin S. A., Sugihara G. Ecology for Bankers. // Nature, 2008. Vol. 451, pp. 893–895.

82. Мотренко А. П. Использование теста Грейнджера при прогнозировании временных рядов // Машинное обучение и анализ данных, 2011. Т. 1. № 1, сс. 51–60.

83. Takens F. Detecting strange attractors in turbulence. // In D. A. Rand and L.-S. Young. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, 1981. Vol. 898, pp. 366–381.

84. Deyle E. R., Sugihara G. Generalized theorems for nonlinear state space reconstruction. PLoS ONE 6(3) –2011, e18295.

85. Sugihara G., May R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. // Nature, 1990. Vol. 344, pp. 734–741.

86. Вальков А. С. и др. Построение кросскорреляционных зависимостей при прогнозе загруженности железнодорожного узла // Машинное обучение и анализ данных, 2013. Т. 1, № 1, сс. 505–518.

87. Google Finance. URL: https://www.google.com/finance/.

88. Yahoo Finance. URL: http://finance.yahoo.com/stock-center/.

89. Fusion Media Ltd. URL: http://ru.investing.com/.

90. Московская Биржа. URL: http://rts.micex.ru/.

91. Bloomberg L.P. URL: http://www.bloomberg.com/markets.

92. Центральный банк Российской Федерации. URL: http://www.cbr.ru/.

93. World Bank Group. URL: http://www.worldbank.org/.

94. Гисметео. / НПЦ «Мэп Мейкер». URL: http://www.gismeteo.ru//.

95. Weather Underground. URL: http://www.wunderground.com/.

96. Институт комплексных стратегических исследований (ИКСИ). URL: http://www.icss.ac.ru/macro/.

97. Вычислимые модели и численные методы для анализа тарифной политики железнодорожных грузоперевозок / М. Ващенко, А. Гасников, Е. Молчанов и др. // Сообщения по прикладной математике. — ВЦ РАН Москва, 2014. — С. 1–51.

98. Гасников А.В., “Заметка об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях”, *Матем. моделирование,* 2015

99. Распоряжение от 2 декабря 2013 г. N 2684р «Об утверждении регламента формирования и контроля исполнения консолидированных бюджетов холдинга "РЖД".

100. Распоряжение от 6 марта 2014 г. N 597р «О внесении изменений в распоряжение ОАО "РЖД" от 30 октября 2012 года n 2174р».

101. Распоряжение от 23 июля 2012 г. N 1451р «О совершенствовании среднесрочного и долгосрочного прогнозирования объемов погрузки грузов на сети железных дорог»