Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
«____» _____ 2019 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Численный вероятностный анализ для задач цифровой экономики

09.04.02 Информационные системы и технологии

09.04.02.01 Информационно-управляющие системы

Руководитель	проф., д-р.физ.-мат. наук	Б. С. Добронец
Студент	КИ17-02-1М 031726411	Н. Н. Васильев
Рецензент	д-р техн. наук, доцент	Л. А. Казаковцев
Нормоконтролер	проф., д-р.физ.-мат. наук	Б. С. Добронец

Красноярск 2019

Продолжение титульного листа магистерской диссертации по теме «Численный вероятностный анализ для задач цифровой экономики».

Нормоконтролер Б. С. Добронец

РЕФЕРАТ

Магистерская диссертация по теме «Численный вероятностный анализ для задач цифровой экономики» содержит 73 страницы текстового документа, 19 иллюстраций, 3 таблицы, 49 использованных источников, 1 приложение.

Объект исследования — временные ряды данных большого объёма.

ЦИФРОВАЯ ЭКОНОМИКА, BIG DATA, ЧИСЛЕННЫЙ ВЕРОЯТНОСТНЫЙ АНАЛИЗ, ВРЕМЕННЫЕ РЯДЫ, МЕТОДЫ АГРЕГАЦИИ ВРЕМЕННЫХ РЯДОВ, ПРОГНОЗИРОВАНИЕ, СПЛАЙН, ТРЕНД, ФУНКЦИЯ ПОТЕРЬ, РЕГРЕССИОННАЯ КРИВАЯ.

Цель исследования — повышение качества прогнозирования временных рядов в условиях неопределённости с использованием методов численного вероятностного анализа.

С этой целью в результате выполнения работы был проведён анализ данной области, выявлена актуальность темы исследования, проведены многочисленные исследования методов агрегации и прогнозирования временных рядов.

На основании результатов, полученных в ходе исследования, был предложена и реализована модель прогнозирования временных рядов данных большого объёма для задач цифровой экономики. Преимущества данной модели в сравнении с уже существующими — простота реализации, наглядность, точность результатов. Данный модель применима в областях, где приходится сталкиваться с обработкой данных большого объёма (экономика, производство, наука).

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
« _____» _____ 2019 г.

ГРАФИК
НАПИСАНИЯ И ОФОРМЛЕНИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
в форме магистерской диссертации

Студент: Васильев Николай Николаевич.

Группа: КИ17-02-1М Направление: 09.04.02.01 Информационно-управляющие системы.

Тема выпускной квалификационной работы: «Численный вероятностный анализ для задач цифровой экономики».

График выполнения выпускной квалификационной работы (ВКР) приведён в таблице 1.

Таблица 1 – График выполнения этапов ВКР

Наименование / содержание этапа	Срок выполнения	Примечания
Анализ предметной области, подбор литературы	До 14 февраля 2018	
Составление плана работы над ВКР	До 14 марта 2018	
Разработка и предоставление на проверку первой главы	До 30 мая 2018	
Разработка и предоставление на проверку второй главы	До 30 сентября 2018	
Работа над экспериментальной частью исследования	До 29 декабря 2018	
Разработка и предоставление на проверку третьей главы	До 31 января 2019	
Доработка ВКР в соответствии с полученными замечаниями	До 20 мая 2019	
Разработка тезисов доклада и подготовка презентации для защиты	До 15 июня 2019	
Согласование с руководителем тезисов доклада и презентации	До 20 июня 2019	
Прохождение нормоконтроля	До 28 июня 2019	
Ознакомление с отзывом и рецензией	До 1 июля 2019	
Завершение ВКР к защите с учётом отзыва и рецензии	До 5 июля 2019	

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой
___ Г. М. Цибульский
« ___» _____ 2019 г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ
в форме магистерской диссертации

Студенту Васильеву Николаю Николаевичу.

Группа КИ17-02-1М. Направление 09.04.02 Информационные системы и технологии.

Тема магистерской диссертации «Численный вероятностный анализ для задач цифровой экономики».

Утверждена приказом по университету №_____от_____.

Руководитель магистерской диссертации Б. С. Добронец, профессор, доктор физико-математических наук, заведующий кафедрой систем искусственного интеллекта ИКИТ СФУ.

Исходные данные для магистерской диссертации: методические указания научного руководителя, статьи, книги, научные журналы, монографии по теме исследования.

Перечень разделов ВКР: введение, теоретические основы прогнозирования временных рядов, методы и подходы численного вероятностного анализа для обработки данных больших объёмов и прогнозирования, экспериментальная часть, заключение, список использованных сокращений, список использованных источников, приложения.

Перечень графического материала: плакаты презентации, выполненной в Microsoft Office PowerPoint 2016.

Руководитель ВКР _____ Б. С. Добронец

Задание принял к исполнению _____ Н. Н. Васильев

« __ » ___ 2019 г.

Содержание

ВВЕДЕНИЕ

Большие данные развиваются как новая область исследований в эпоху 21-го века. Эта область связана с изучением объемных наборов данных с несколькими факторами, размеры которых быстро растут со временем. Эти типы наборов данных могут генерироваться из различных автономных источников, таких как научные эксперименты, инженерные приложения, правительственные записи, финансовая деятельность и т.д. С появлением концепции больших данных возникла потребность в новых моделях прогнозирования временных рядов. Для этой цели в этой научной работе рассматривается новая модель прогнозирования временных рядов больших данных, основанная на использовании методов численного вероятностного анализа.

Тема магистерской диссертации связана с исследованием применения методов численного вероятностного анализа для прогнозирования временных рядов данных большого объёма.

Актуальность темы обуславливается необходимостью повышения качества прогнозирования временных рядов данных большого объёма для задач цифровой экономики.

Цель исследования: повышение качества прогнозирования временных рядов в условиях неопределённости с использованием методов численного вероятностного анализа.

Объектом исследования в данной работе являются временные ряды данных большого объёма.

В результате работы была выявлена основная проблема, связанная с качеством обработки данных большого объёма, а также сложностью определения законов распределения случайной величины, функции плотности вероятности и использования больших данных в целях прогнозирования задач цифровой экономики.

Что же такое цифровая экономика? Если обычная "аналоговая" экономика – это хозяйственная деятельность общества, а также совокупность отношений, складывающихся в системе производства, распределения, обмена и потребления. То использование компьютера, интернета, мобильных телефонов уже можно считать "потреблением", в этом случае цифровую экономику можно представить, как ту часть экономических отношений, которая опосредуется Интернетом, сотовой связью, ИКТ [2].

Главной функцией успешной работы практически всех составляющих цифровой экономики является возможность работать с информацией. Поэтому цифровая экономика тесно связана с Big Data.

В настоящее время собираются огромные массивы данных в различных сферах жизнедеятельности человека, но до тех пор, пока эти данные не обработаны и не структурированы, их нельзя использовать и ценности они не представляют. Проблема анализа больших данных состоит в том, что в них всегда есть погрешности, они могут быть неполными или иметь в себе ошибки. Для того чтобы приводить Big Data к надлежащему виду и проводить над ними операции, нужен мощный инструмент работы с данными [3].

Важное научное и практическое значение совершенствования статистического анализа и прогнозирования временных рядов, актуализации системы методов их анализа в соответствии с современными научными разработками, подтверждают актуальность темы исследования.

Планирование и прогнозирование используется в каждом направлении деятельности человека. Например, планирование грузовых перевозок имеет большое практическое значение для производственно-хозяйственной деятельности и управления в данной отрасли. Грузовые перевозки обеспечивают свыше 80% общей выручки железнодорожного транспорта. В данном направлении большую работу провели А.П. Мотренко и В.В. Стрижов в своем исследовании [1], где подняли проблему построения агрегированных прогнозов объемов железнодорожных грузоперевозок.

Охват использования больших данных продолжает расширяться с каждым днем, и он становится почти повсеместным, поскольку он меняет подход компаний ко всему: от повседневных операций до финансового прогнозирования. Понимая, как эти данные повлияют на организацию, и потратив время на то, чтобы сразу же выяснить, что нужно для их эффективного использования, у пользователя будет больше шансов использовать этот мощный инструмент для точного составления прогноза дальнейших действий [4].

Задачи данного научного исследования:

- проанализировать существующие методы агрегирования и прогнозирования временных рядов данных большого объёма;
- описать существующие методы численного вероятностного анализа;
- разработать модель прогнозирования временных рядов данных большого объёма с применением методов численного вероятностного анализа.

1 Теория прогнозирования временных рядов больших данных

1.1 Теория Big data

Для ученых большие данные - это «сбор данных со сложностью, разнообразием, неоднородностью и высокой потенциальной ценностью, которые трудно обрабатывать и анализировать в разумные сроки», в то время как, например, для политиков большие данные являются «новым типом стратегического ресурса в цифровая эра и ключевой фактор, стимулирующий инновации, которые меняют способ производства и жизни людей в настоящее время» [2].

В настоящее время методы прогнозирования на основе Big Data представляют большой практический интерес и позволяют решать широкий спектр задач.

Технологии Big Data — серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов, и значительного многообразия. Данные технологии применяются для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения информации по многочисленным узлам вычислительной сети [5].

Термин Big Data – это наборы данных, размер которых превосходит возможности типичных баз данных по хранению, управлению и анализу информации. В настоящее время множество компаний следят за развитием технологий Big Data.

Существует широко распространенное мнение о том, что большие данные могут помочь в улучшении прогнозов при условии, что мы можем анализировать и обнаруживать скрытые закономерности, и, что прогнозы можно улучшить с помощью принятия решений на основе данных.

В современных условиях организации создают большое количество неструктурированных данных, таких как текстовые документы, изображения,

видеозаписи, машинные коды, таблицы и т. д. Вся эта информация находится во множестве хранилищ, часто даже за пределами баз данных организации.

Компании могут иметь доступ к огромному массиву собственных данных и не иметь необходимых инструментов, которые могли бы установить взаимосвязи между этими данными и сделать на их основе значимые выводы. Основные характеристики традиционной базы данных и базы Big Data приведены в таблице 1.

Таблица 1 – Характеристики традиционной и Big Data базы данных

Характеристика	Традиционная база данных	База Big Data
Объём информации	От гигабайт до терабайт	От петабайт до эксабайт
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурирована	Полуструктурирована или неструктурированна
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Традиционные методы анализа информации не могут угнаться за огромными объемами без остановки увеличивающихся и обновляющихся данных, что в итоге и открывает дорогу технологиям Big Data.

Можно выделить следующие особенности технологии Big Data [6]:

Навыки

Навыки, необходимые для решения проблемы прогнозирования с помощью больших данных, и наличие персонала, квалифицированного для выполнения этой конкретной задачи, является одной из главных задач. В мире, где ученые, исследователи и статистики имеют большой опыт использования традиционных статистических методов более пятидесяти лет для получения точных прогнозов, наличие больших данных само по себе является сложной задачей.

Сигнал и шум

Это более техническая, но чрезвычайно важная задача в прогнозировании больших данных. С большими наборами данных извлечение сигнала становится более сложным. Большинство традиционных методов прогнозирования прогнозируют как шум, так и сигнал, и, хотя они работают относительно хорошо в случае традиционных наборов данных, увеличение отношения шума к сигналу, наблюдаемое в больших данных, с большей вероятностью искажает точность прогнозов. Это говорит о том, что существует необходимость в использовании и оценке использования методов прогнозирования, которые могут фильтровать шум в больших данных и прогнозировать только сигнал. В качестве примера известно, что одним из таких методов является анализ сингулярного спектра (SSA), который стремится отфильтровать шум из заданного временного ряда, восстановить новый ряд, который является менее шумным, а затем использовать этот недавно восстановленный ряд для прогнозирования будущих точек данных.

Аппаратное и программное обеспечение

Современное статистическое программное обеспечение не способно справиться с прогнозированием больших данных, явно видна необходимость в суперкомпьютерах для обработки прогнозов больших данных. Конечно, существуют методы автоматического прогнозирования, которые могут дать результат в течение нескольких секунд. Однако их надежность перед лицом больших данных еще предстоит проверить. Поэтому разумно согласиться с тем, что вычислительные возможности и структура, лежащая в основе статистического программного обеспечения, потребуют улучшений, чтобы успешно справляться с возросшим объемом вводимых данных.

Архитектура алгоритмов

Методы Data Mining предлагаются в качестве важных методов, которые можно использовать для прогнозирования с помощью больших данных. Однако эти методы были разработаны для обработки данных сравнительно меньших размеров, чем для больших данных. Поэтому алгоритмы Data Mining часто не

могут работать с данными, которые не загружены в его основную память, и, следовательно, требуют перемещения больших данных между местоположениями, что может привести к увеличению затрат на сетевую связь. Архитектура аналитики должна быть переработана таким образом, чтобы она могла обрабатывать как исторические данные, так и данные в реальном времени.

Большие данные

Большие данные сами по себе представляют собой проблему для прогнозирования в силу присущих им характеристик. Большие данные эволюционируют и изменяются в реальном времени, и поэтому важно, чтобы методы, используемые для прогнозирования больших данных, могли преобразовывать неструктурированные данные в структурированные данные, точно фиксировать эти динамические изменения и заранее определять точки изменения.

1.2 Прогнозирование временных рядов данных большого объёма

Прогнозирование – одна из самых востребованных задач бизнес-аналитики. Продажи, поставки, заказы – это процессы, распределенные во времени, следовательно, прогнозирование в области продаж, сбыта и спроса, управления материальными запасами и потоками обычно связано именно с анализом временных рядов.

Наиболее распространённой постановкой задачи прогнозирования является задача прогнозирования временных рядов, т.е. функции, определённой на оси времени [19].

Как люди узнают, что цена товара выросла за определенный период времени? Они могут сделать это, сравнив цены товара за определенный период времени. Ряд наблюдений, упорядоченных относительно последовательных периодов времени, является временным рядом. Временной ряд – последовательность наблюдений за изменениями во времени значений параметров некоторого объекта или процесса.

Другими словами, расположение данных в соответствии с их временем появления является временным рядом. Это хронологическое расположение данных. Здесь время — это просто способ связать все явление с подходящими ориентирами. Время может быть часами, днями, месяцами или годами.

Временной ряд изображает отношения между двумя переменными. Время является одной из этих переменных, а вторая является любой количественной переменной. Нет необходимости, чтобы отношение всегда показывало приращение изменения переменной по отношению ко времени. Отношение тоже не всегда уменьшается [20].

Прогнозирование данных временных рядов является неотъемлемым компонентом управления, планирования и принятия решений. Следуя тенденции больших данных, временные ряды больших данных доступны из многих разнородных источников во всем большем количестве областей применения. Высокодинамичный и часто колеблющийся характер этих областей в сочетании с логистическими проблемами сбора таких данных из различных источников ставит новые задачи в области прогнозирования.

Традиционные подходы в значительной степени полагаются на обширные и полные исторические данные для построения моделей временных рядов и, таким образом, более не применимы, если временные ряды являются короткими или, что еще более важно, периодическими. Кроме того, большое количество временных рядов необходимо прогнозировать на разных уровнях агрегации с предпочтительно низкой задержкой, в то время как точность прогноза должна оставаться высокой. Это почти невозможно, при сохранении традиционной ориентации на создание одной модели прогноза для каждого отдельного временного ряда [21, 22].

Преимущества использования временных рядов [23]:

- наиболее важным применением изучения временных рядов является то, что это помогает прогнозировать будущее поведение переменной на основе прошлого опыта;

- это полезно для бизнес-планирования, поскольку помогает сравнивать фактическую текущую производительность с ожидаемой;

- из временных рядов изучается поведение явления или рассматриваемой переменной в прошлом;

- возможность сравнивать изменения значений разных переменных в разное время или в разных местах и т.д.

Задачи анализа и прогноза реальных временных рядов возникают во всех аспектах человеческой деятельности (в экономике, медицине, энергетике, промышленности и т.д.) Современные методы статистического прогнозирования позволяют с высокой точностью прогнозировать практически все возможные показатели. При анализе временных рядов можно выделить две основные цели [24]:

- определение природы временного ряда;

- прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям).

В последние два десятилетия были разработаны много методов прогнозирования. Несмотря на наличие широкого спектра методов и алгоритмов, многие проблемы в задах прогнозирования ещё далеки от своего разрешения [27, 28].

1.3 Компоненты временного ряда

Различные причины или силы, которые влияют на значения наблюдения во временном ряду, являются компонентами временного ряда. Четыре категории компонентов временного ряда

- тренд

- сезонные изменения

- циклические Вариации

- случайные или нерегулярные движения

Сезонные и циклические вариации - это периодические изменения или краткосрочные колебания.

Тогда детерминированная составляющая может быть записана в виде:

$$di=ti+si+ci, \quad (1)$$

где ti—тренд, si – сезонная компонента, ci – циклическая компонента.

Тренд

Тренд показывает общую тенденцию увеличения или уменьшения данных в течение длительного периода времени. Тренд - это плавная, общая, долгосрочная, средняя тенденция. Не всегда необходимо, чтобы увеличение или уменьшение было в одном и том же направлении в течение данного периода времени.

Наблюдается, что тенденции могут увеличиваться, уменьшаться или быть устойчивыми в разные периоды времени. Но общая тенденция должна быть восходящей, нисходящей или стабильной. Население, сельскохозяйственное производство, производимые товары, число рождений и смертей, число предприятий или предприятий, число школ или колледжей - вот некоторые из примеров, демонстрирующих тенденции движения.

Линейный и нелинейный тренд

Линейный тренд описывает равномерное изменение показателя во времени, нелинейный же наоборот показывает резкие изменения показателей.

Периодические Колебания

Есть некоторые компоненты во временном ряду, которые имеют тенденцию повторяться в течение определенного периода времени. Они действуют регулярно и спазматическим образом.

Сезонные изменения

Это ритмические силы, которые действуют регулярно и периодически на протяжении периода менее года. Они имеют одинаковую или почти одинаковую картину в течение 12 месяцев. Это изменение будет

присутствовать во временных рядах, если данные записываются ежечасно, ежедневно, еженедельно, ежеквартально или ежемесячно.

Эти вариации вступают в игру либо из-за природных сил, либо из-за созданных человеком условий. Различные сезоны или климатические условия играют важную роль в сезонных колебаниях. Например, производство сельскохозяйственных культур зависит от времени года, продажи зонтов и плащей в сезон дождей, а также продажи электрических вентиляторов и кондиционеров в летний сезон.

Циклические вариации

Изменения во временном ряду, которые действуют в течение более одного года, являются циклическими изменениями. Это колебательное движение имеет период колебаний более года. Один полный период - это цикл. Это циклическое движение иногда называют «деловым циклом».

Это четырехфазный цикл, состоящий из фаз процветания, рецессии, депрессии и восстановления. Циклические колебания могут быть регулярными, а не периодическими. Подъемы и спады в бизнесе зависят от общего характера экономических сил и взаимодействия между ними.

Случайные или нерегулярные движения

Есть еще один фактор, который вызывает изменение исследуемой переменной. Они не являются регулярными вариациями и являются чисто случайными или нерегулярными. Эти колебания являются непредвиденными, неуправляемыми, непредсказуемыми и непредсказуемыми. Этими силами являются землетрясения, войны, наводнения, голод и любые другие стихийные бедствия.

1.4 Предварительный анализ временных рядов

Предварительный анализ временных рядов экономических показателей заключается в основном в выявлении и устранении аномальных значений уровней, а также в определении наличия тренда.

Проблема обнаружения аномалий для временных рядов обычно формулируется как нахождение точек данных выбросов относительно некоторого стандартного или обычного сигнала. Хотя существует множество типов аномалий, основным считаются, такие как: неожиданные всплески, спады, изменения трендов и изменения уровня.

Причинами появления таких аномальных наблюдений могут быть какие-либо технические ошибки, или ошибки первого рода, к которым относятся:

- ошибки, возникающие при передаче информации;

- ошибки, возникающие при агрегировании и дезагрегировании показателей.

Кроме того, аномальные уровни во временных рядах могут возникать из-за воздействия факторов, имеющих объективный характер, но проявляющихся эпизодически или очень редко – ошибки второго рода, они устранению не подлежат.

Критерий Ирвина предполагает использование следующей формулы:

$$\mu_{расч} = (|y_к - y_к\ пред|)/\sigma, \quad (2)$$

где $y_к$ – сомнительное значение, $y_{к\ пред}$ – предыдущее значение в вариационном ряду, считая от верха ряда или последующее (считая от низа ряда), σ – генеральное среднеквадратическое отклонение (СКО) нормально распределённой случайной величины. Если расчётное значение больше табличного – таблица 2, сомнительное значение считают аномальным (грубой ошибкой).

Таблица 2 – Пример табличных значений критерия Ирвина

Число наблюдений	Уровень значимости = 0.05	Уровень значимости = 0.01
2	2.8	3.7
3	2.2	2.9
10	1.5	2.0
20	1.3	1.8
30	1.2	1.7
50	1.1	1.6
100	1.0	1.5
400	0.9	1.3
1000	0.8	1.2

Проверка ряда на наличие тренда может выполняться несколькими методами.

- Метод средних. Изучаемый ряд динамики разбивается на несколько интервалов (обычно на два), для каждого из которых определяется средняя величина. Выдвигается гипотеза о существенном различии средних. Если эта гипотеза принимается, то признается наличие тренда. В более мощном критерии Кокса и Стюарта весь анализируемый ряд динамики разбивают на три группы и сравнивают между собой уровни первой и последней групп.

- Метод серий. По этому способу каждый конкретный уровень временного ряда считается принадлежащим к одному из двух типов: например, если уровень ряда меньше медианного значения, то считается, что он имеет тип А, в противном случае – тип В.

1.5 Агрегирование данных

Прежде чем использовать информацию в целях прогнозирования, её необходимо дополнительно готовить к анализу в системах обработки Больших Данных; приходится принимать решения о пригодности данных, принципах агрегации и необходимом уровне качества.

Агрегация данных — это тип процесса извлечения данных и информации, при котором данные ищутся, собираются и представляются в обобщенном формате на основе отчетов для достижения конкретных бизнес-целей или процессов и / или проведения человеческого анализа.

Целью агрегации временных рядов является объединение набора периодов в группы таким образом, чтобы члены группы - исходные периоды - были максимально похожими. Затем группа представлена одним периодом. Группировка временных рядов основана в большинстве методов на измерении расстояния между атрибутами каждого члена группы. Для точной группировки необработанные исходные данные должны быть предварительно обработаны в нужном формате. На основании этого применяется метод агрегации для создания групп. С точки зрения достижения осуществимого проекта системы, различные варианты интеграции экстремальных периодов могут быть включены впоследствии. Наконец, агрегированные временные ряды должны быть сокращены таким образом, чтобы их средние значения соответствовали средним значениям исходного временного ряда.

Агрегирование данных может быть выполнено вручную или через специализированное программное обеспечение.

Использование таких процедур агрегирования, как усреднение, исключение экстремальных значений (эмиссия), процедура сглаживания может привести к потере важной информации. Поэтому выбор метода агрегирования является важной задачей, поскольку без предварительного исследования легко получить дополнительную неопределённость, которой нет в исходной постановке. Для агрегирования данных используются различные математические модели. В тех случаях, когда данные могут быть представлены частотными распределениями рассматриваемых характеристик или признаков, возможно использование кусочно-полиномиальных моделей.

Частным примером кусочно-полиномиальных моделей является гистограмма, которая представляет собой кусочно-постоянную функцию агрегирования. Гистограмма, с точки зрения процесса агрегирования, во многих

случаях представляет собой альтернативу операциям усреднения или построения интервальных данных. В отличие от указанных операций применение гистограмм позволяет повысить точность вычисления за счёт использования информации о частотном распределении данных вместо замены набора данных одним значением, например, значением выборочного среднего или моды.

Помимо гистограмм и частотных полигонов в качестве математических моделей агрегатов можно использовать сплайны, как кусочно-полиномиальные функции.

Сплайн представляет собой достаточно гладкую кусочно - полиномиальную функцию. Этот подход полезен по следующим причинам. Поскольку сплайн является кусочно-полиномиальной функцией, то его можно рассматривать как функцию агрегирования данных. Функция агрегирования выполняет численные обработки наборов данных и возвращает сплайн значения. Сплайны полезны для анализа неопределённости в данных из-за того, что они адекватно представляют частотное распределение данных. Эти математические модели по своей сущности есть не что иное как преобразование данных в агрегаты, для соответствующей процедуры агрегации.

Также в прогнозировании используется гистограммный временной ряд, который агрегирует данные обычных временных рядов, представляющих собой Big Data. Он описывает ситуации, когда в течение каждого момента времени известны гистограммы, аппроксимирующие функции плотности некоторых случайных величин. Подобные ситуации возникают, когда необходима агрегация большого числа данных в некоторые моменты времени. Во многих случаях гистограммы более информативны, чем, например, среднее значение. Причины для использования гистограмм могут быть сформулированы следующим образом:

- можно использовать их для любой исходной плотности вероятности;
- они могут описывать данные с достаточной степенью точности;

- простая и гибкая структура упрощает их использование.

В символическом анализе данных и Data Mining, гистограммы используются для исследования множества различных процессов и применяются для описания изменчивости количественных признаков.

Актуальность сплайн агрегации заключается в более информативных моделях для представления и анализа изменчивости данных в задачах восстановления зависимостей, чем у других формы агрегирования. Такие процедуры агрегирования помогают сократить объем вычислений при обработке данных и являются важной основой для извлечения полезных знаний из больших объёмов данных. Обладая более высоком порядком сходимости, чем гистограммы и кусочно-линейные функции, сплайны повышают точность вычислений. Разработанные методы снижают уровень информационной неопределённости и существенно сокращают время обработки данных и выполнение численных процедур. Новыми результатами являются применение кусочнополиномиальных моделей к агрегации данных и метод построения регрессионных зависимостей на основе численного вероятностного анализа.

1.6 Функция потерь

Все алгоритмы в машинном обучении основаны на минимизации или максимизации функции, которая называется «целевой функцией». Группа функций, которые минимизируются, называются «функциями потерь». Функция потерь — это мера того, насколько хороша модель прогнозирования с точки зрения способности прогнозировать ожидаемый результат. Наиболее часто используемый метод определения минимальной точки функции - «градиентный спуск».

Не существует единой функции потери, которая работает для всех видов данных. Это зависит от ряда факторов, включая наличие выбросов, выбор алгоритма машинного обучения, временную эффективность градиентного спуска, простоту поиска производных и достоверность прогнозов.

На выбор функции потерь влияют особенности решаемой задачи. Общего правила выбора функции потерь не существует. Чаще всего используются следующие функции потерь – рисунок 1:

- среднеквадратичная (MSE) – A;
- средняя абсолютная ошибка (MAE) – B;
- сглаженная средняя абсолютная ошибка (Huber loss) – C;
- квантильная потеря (Quantile Loss) – D;

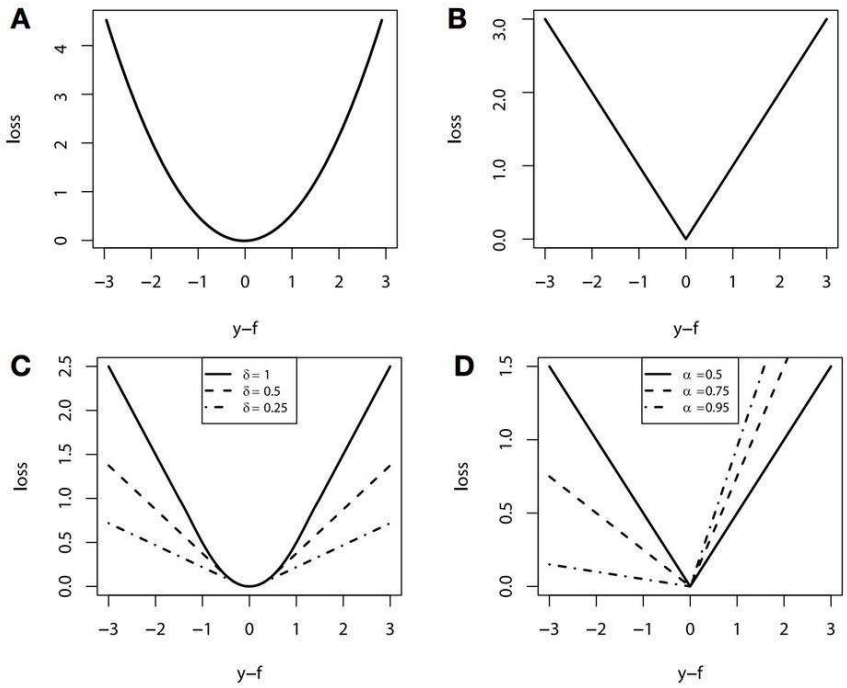

Рисунок 1 – Графики функции потерь

А. П. Мотренко и В. В. Стрижов в своем исследовании [1] по прогнозам объемов железнодорожных грузоперевозок приводят сравнение применения функций потерь при использовании алгоритмов ARIMA+hist (модели авторегрессионного скользящего среднего + гистограммного метода).

20% последних точек каждого временного ряда использовались как контрольные. Для каждой контрольной точки по доступной истории

временного ряда (все точки от первой до предшествующей рассматриваемой контрольной) обучалась выбранная для временного ряда модель ARIMA, затем для обученной модели вычислялся ряд регрессионных остатков. По ряду регрессионных остатков обучался алгоритм hist с заданной функцией потерь и заданным количеством столбцов в гистограмме. Прогноз для контрольной точки складывался из прогноза ARIMA и hist.

Эксперимент проводился для функций потерь: квадратичная, абсолютная, ассиметричная, и вариантов алгоритма hist с 20, 50, 300, 500 столбцами в гистограмме. Средние потери для каждой функции потерь приведены для всех вариантов алгоритма в таблице 3.

Таблица 3 – Пример средних потерь для каждой функции потерь

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

Как видно из таблицы 3, при использовании асимметричной функции потерь двухэтапный алгоритм прогнозирования ARIMA+hist позволяет получать среднюю ошибку прогноза существенно ниже, чем прогнозирование с помощью модели ARIMA. При этом для симметричных функций потерь использование двухэтапного алгоритма прогнозирования не приводит к значительным изменениям по сравнению с прогнозом модели ARIMA.

1.7 Модели прогнозирования временных рядов

Главный инструмент прогнозирования в современной бизнес-аналитике – прогностические модели.

Обобщенная модель прогноза. Набор входных переменных xi (i=1,…,n) – исходные данные для прогноза. Набор выходных переменных yj (j=1,…,m) – набор прогнозируемых величин, n>m. Когда решается задача прогнозирования временного ряда, описывающего динамику изменения некоторого бизнес-процесса, входные значения – наблюдения за развитием процесса в прошлом, а выходные – прогнозируемые значения процесса в будущем. При этом временные интервала прошлых наблюдений и временные интервалы, по которым требуется получить прогноз, должны соответствовать друг другу.

«Наивная» модель прогнозирования. Предполагает, что последний период прогнозируемого временного ряда лучше всего описывает будущее этого ряда. Простейшая модель

$$y(t+1)=x(t), \quad (3)$$

где x(t) – последнее наблюдаемое значение, y(t+1) –прогноз. Чтобы модель учитывала наличие возможных трендов, ее можно несколько усложнить, например преобразовав к виду

$$y(t+1)=x(t)+[x(t)-x(t-1)] \quad (4)$$

или

$$y(t+1)=x(t)[x(t)/x(t-1)] \quad (5)$$

При необходимости учета сезонных колебаний модель модифицируется следующим образом:

$$y(t+1)=x(t-s), \quad (6)$$

где s – показатель, учитывающий сезонные изменения прогнозируемого временного ряда.

Экстраполяция. Если значения функции f(x) известны в некотором интервале [x0, xn], то целью экстраполяции является определение наиболее вероятного значения в точке xn+1. Экстраполяция применима только в тех случаях, когда функция f(x), а соответственно и описываемый ей временной ряд, достаточно стабильна и не подвержена резким изменениям. Наиболее популярный метод экстраполяции в настоящее время – экспоненциальное сглаживание. Основной его принцип заключается в том, чтобы учесть в прогнозе все наблюдения, но с экспоненциально убывающими весами. Метод позволяет принять во внимание сезонные колебания ряда и предсказать поведение трендовой составляющей. Например, в случае ряда с «нулевым» трендом, можно выбрать следующую модель экспоненциального сглаживания

$$y(t+1)= \lambda\, y(t)+(1-\lambda)\, x(t), \quad (7)$$

где x(t) – последнее наблюдаемое значение, y(t) – прогноз на момент времени t, y(t+1) –прогноз на момент времени t+1. $0< \lambda$

Прогнозирование методом среднего и скользящего среднего. Наиболее простая модель этой группы – обычное усреднение набора наблюдений прогнозируемого ряда

$$y(t+1)=(x(t)+x(t-1)+x(t-2)+\ldots+x(1))/t \quad (8)$$

При усреднении сглаживаются резкие изменения и выбросы данных, что делает результаты прогноза более устойчивыми к изменчивости ряда, но в целом эта модель прогноза так же примитивна как «наивная». В формуле прогноза на основе среднего предполагается, что ряд усредняется по всем наблюдениям, но старые значения временного ряда могли формироваться на

основе иных закономерностей и утратить актуальность. Чтобы повысить точность прогноза, можно использовать «скользящее среднее»

$$y(t+1)=(x(t)+x(t-1)+x(t-2)+\ldots+x(t-T))/(T+1), \quad (9)$$

т.е. модель «видит» прошлое на T отсчетов времени и прогноз строится только на этих наблюдениях. Иногда метод скользящего среднего оказывается даже эффективнее чем методы, основанные на долговременных наблюдениях.

Регрессионные модели. Один из методов прогнозирования временных рядов – определение факторов, которые влияют на каждое значение временного ряда. Для этого выделяется каждая компонента временного ряда, вычисляется ее вклад в общую составляющую, а затем на его основе прогнозируются будущие значения временного ряда. Данный метод получил название декомпозиции временного ряда. Исходный временной ряд представляется как композиция тренда, сезонной и циклической компоненты. Для построения прогноза выполняется выделение этих компонент из ряда, т.е. разложение ряда по компонентам.

Рассмотрим прогнозирование методом декомпозиции с помощью тренда. Если тренд линейный, что типично для многих реальных временных рядов, то он представляет собой прямую линию, описываемую уравнением

$$y=a+b\,t, \quad (10)$$

где y—значение ряда, a и b – коэффициенты, определяющие расположение и наклон линии тренда, t – время.

Если уравнение линии тренда известно, то с его помощью можно рассчитать значение тренда в любой момент времени

$$yt+k=a+b(t'+k), \quad (11)$$

где t’ – начало прогноза, k – горизонт прогноза. При использовании сезонности для прогнозирования методом декомпозиции сначала из временного ряда убирается тренд и сглаживается возможная циклическая компонента.

Тогда можно считать, что оставшиеся данные будут обусловлены в основном сезонными колебаниями. На основе этих данных вычисляются так называемые сезонные индексы, которые характеризуют изменения временного ряда во времени.

Например, временной ряд содержит наблюдения по месяцам в течение года. Сезонный индекс, равный 1, будет установлен для месяца, ожидаемое значение в котором составляет 1/12 от общей суммы по месяцам. Если для некоторого месяца устанавливается индекс 1.2, то ожидаемое значение для этого месяца составляет 1/12+20%, а если 0.8 – то 1/12- 20% и т.д. Ясно, что сумма сезонных индексов за год должны равняться 12.

Использовать сезонность для прогнозирования можно тогда, когда сезонные колебания имеют хорошую повторяемость.

Модель ARIMA

Модели ARIMA предоставляют другой подход к прогнозированию временных рядов. Экспоненциальное сглаживание и модели ARIMA являются двумя наиболее широко используемыми подходами к прогнозированию временных рядов и обеспечивают дополнительные подходы к проблеме. В то время как модели экспоненциального сглаживания основаны на описании тренда и сезонности в данных, модели ARIMA стремятся описать автокорреляции в данных.

В основе статистического подхода лежит модель стационарного стохастического процесса ARMA(p, q), предложенная Боксом и Дженкинсом в 1976, для которой должны быть выполнены ограничения: математическое ожидание равно нулю и дисперсия постоянна.

Так как на практике большинство временных рядов являются нестационарными, то для моделирования и прогнозирования таких ВР была предложена и получила распространение расширенная модель стохастического процесса ARIMA(p, q,d) и ее частные случаи. Популярность модели класса ARIMA обусловлена несколькими факторами: она позволила в свое время упростить процесс разработки модели временного ряда, получать модели

широкого класса временных рядов (стационарных и нестационарных) с приемлемыми показателями точности, снизить требования к уровню квалификации пользователя и реализована в распространенных статистических пакетах.

Бокс и Дженкинс [18] предложили выделить класс нестационарных рядов, которые взятием последовательных разностей можно привести к стационарному виду типа ARMA. Если ряд после взятия d последовательных разностей сводится к стационарному, то для прогнозирования его уровней можно применить комбинированную модель авторегрессии и скользящего среднего, обозначаемую как ARIMA(p,d,q). Сокращение I в данной аббревиатуре означает «интегрированный» [17].

Методология Бокса-Дженкинса подбора ARIMA-модели для конкретного ряда наблюдений состоит из четырех этапов:

– идентификация модели – процесс выбора модели, в наилучшей степени, соответствующей рассматриваемому реальному процессу;

– оценивание модели – использование регрессионных методов для получения оценок параметров, включенных в модель;

– тестирование модели – проверка основных предпосылок использования регрессионного анализа, проверка адекватности модели с использованием тестов на нормальность остатков (тест Жарка-Бера), на автокорреляцию остатков (тест Дарбина-Уотсона), на постоянство дисперсий случайных остатков (критерии Кохрана и Голдфалда-Квандта), на качество спецификации модели (F-тест);

– использование модели для прогнозирования.

Идентификация модели

В первую очередь необходимо выяснить, обладает ли изучаемый ряд свойством стационарности. Оценка стационарности исходного временного ряда осуществляется с использованием формальных тестов, например при помощи расширенного критерия Дикки-Фуллера. Кроме этого, при идентификации смешанной модели проводится анализ коррелограммы ряда, для чего строится

график выборочной автокорреляционной функции (ACF). Коррелограмма стационарного временного ряда быстро убывает с ростом порядка k за пределами нескольких первых значений. Если график убывает достаточно медленно, есть основания считать ряд нестационарным; если же не убывает, то исследуемый ряд определенно не стационарен.

Оценивание модели

После структурной идентификации, проведенной на предыдущем шаге, осуществляется параметрическая идентификация модели. Как правило, применение метода наименьших квадратов для этих целей в моделях ARIMA (p,d,q) дает хороший результат.

Оптимальный выбор модели подразумевает, что случайный остаток e_t, t = 1,2,..., n по своим свойствам достаточно близок к белому шуму. Математическое ожидание остатков должно быть равно нулю, дисперсия постоянна на любом участке измерения, а между уровнями ряда случайных остатков не должно быть автокорреляционной зависимости. Фактическая ошибка модели должна быть достаточно случайна, чтобы ее невозможно было уточнить какой-либо иной моделью.

Желательно также, чтобы дисперсия ошибки σ_e^2 была существенно меньше дисперсии самого процесса, т.е. $\sigma_e^2 << \sigma_y^2$. В этом случае модель, описывающая процесс Y_t , снимает значительную часть неопределенности в его изменчивости, что позволяет с большей обоснованностью предсказывать его значения.

Тестирование модели

Проверка основных предпосылкок использования регрессионного анализа: случайный характер остатков модели, равенство нулю математического ожидания остатков, отсутствие автокорреляционной зависимости в остатках, гомоскедастичность дисперсии остатков, подчинение остатков нормальному закону распределения. При выполнении этих

предпосылок оценки коэффициентов регрессии будут обладать свойствами несмещенности, эффективности и состоятельности.

Тест Жарка-Бера представляет собой процедуру определения отклонения от нормальности и основан на значениях выборочного эксцесса и асимметрии. Тестовая статистика JB вычисляется по формуле

$$JB = (n / 6)(A^2 + Ex^2 / 4) \quad (12)$$

где n – количество наблюдений; A – асимметрия; Ex – эксцесс. Статистика JB имеет асимптотическое $\chi 2$ – распределение с двумя степенями свободы и используется для проверки нулевой гипотезы о том, что данные принадлежат нормальному распределению. Нулевая гипотеза, подлежащая проверке, в свою очередь, является составной – о равенстве нулю асимметрии и эксцесса. Как видно из определения статистики JB, любое отклонение от этих значений увеличивает ее значение.

Также применяется F-тест. Общая процедура проверки гипотезы о постоянстве математического ожидания может быть организована так. Если количество наблюдений достаточно велико, то интервал наблюдений делится на $N > 2$ частей, не обязательно одинаковой размерности. Для каждой из частей определяются оценки среднего и дисперсии. Проверяется гипотеза о равенстве оценок средних значений ряда, рассчитанных на этих частях. Для этого может быть использован критерий Фишера.

Прогнозирование на основании модели

В случае если выбранная модель признана адекватной, с ее помощью можно составлять прогнозы на некоторое количество периодов вперед. При поступлении с течением времени в распоряжение исследователя новых данных об уровнях временного ряда, уже сформированная модель ARIMA (p,d,q) может быть применена для составления нового прогноза с иным началом отсчета

времени. Однако если характер поведения данных со временем меняется, то необходима переоценка параметров модели или, в некоторых случаях, поиск модели иной спецификации.

Модель HIST

Модель гистограммного прогнозирования hist разработана ФИЦ ИУ РАН [1] и основана на механизмах математической статистики, в частности функции плотности распределения.

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

Вход алгоритма: стационарный временной ряд

$$x = \{x_{1\,.\,\ldots\,.} x_T\} \tag{13}$$

и функция потерь

$$\mathrm{l}(\hat{\mathrm{x}}, \mathrm{x}_{\mathrm{T}+1}) \tag{14}$$

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

Порядок вычислений:

Шаг 1: задание количества *n* столбцов гистограммы.

Шаг 2: вычисление ширины столбцов гистограммы

$$x = \frac{\max(\mathrm{x}) - \min(\mathrm{x})}{\mathrm{n}} \tag{15}$$

и координат концов отрезков постоянства $\mathrm{u}_0, \mathrm{u}_{1,\ldots}, \mathrm{u}_\mathrm{n}$ для функции $\hat{\mathrm{p}}(\mathrm{u})$.

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

Шаг 4: вычисление значений свертки

$$\sum_{i=1}^{n} h_i l\left(c, \frac{u_i + u_{i-1}}{2}\right) \tag{16}$$

для всех

$$C \in \left\{\frac{u_1 + u_0}{2}, \dots, \frac{u_n + u_{n-1}}{2}\right\}; \tag{17}$$

выбор C^*, при котором достигается минимальное значение свертки; вычисление соответствующего прогнозируемого значения $\hat{x}$:

$$\hat{x} = c^* \in \left\{\frac{u_1 + u_0}{2}, \dots, \frac{u_n + u_{n-1}}{2}\right\}. \tag{18}$$

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

При заданных выборке данных и функции потерь результат прогнозирования зависит только от количества столбцов гистограммы n. При малых n оценка плотности распределения $\hat{p}(u)$ получается огрубленной, при больших n – более детальной, однако при увеличении возрастает вероятность переобучения прогностической модели.

Алгоритм обеспечивает оптимальность свертки построенной гистограммы и функции потерь, благодаря двухэтапному алгоритму прогнозирования, на первом этапе которого отслеживаются свойства

временного ряда, обусловливающие его нестационарность, такие как тренд и сезонность.

На втором этапе вычисляется поправка, обеспечивающая оптимальность прогноза в случае несимметричной функции потерь. На втором этапе алгоритма в качестве временного ряда выступают регрессионные остатки, однако их плотность распределения не известна. В качестве оценки плотности используется гистограмма значений регрессионных остатков [13].

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

2 Методы и подходы численного вероятностного анализа для обработки данных большого объёма и прогнозирования

2.1 Теория численного вероятностного анализа

Одним из новых подходов в направлении представления, численных методов обработки, моделирования и анализа неопределенных данных является численный вероятностный анализ, как теоретическая и практическая основа информационно-аналитической технологии исследования информационных процессов в условиях элиторной и эпистемической неопределенностей [37].

Численный вероятностный анализ является разделом вычислительной математики, предметом которого выступает решение различных задач со стохастическими неопределенностями в данных в условиях различных видов неопределенности, с применением численных операций над плотностями вероятностей стохастических переменных и их функций. Следующими основаниями определяется потребность в разработке численных операций, определяющих предмет арифметики.

Функция плотности вероятности является одним из способов описания и представления стохастической неопределенности в данных. В классической теории вероятностей рассматриваются аналитические формулы произведения простых арифметических операций нал стохастическими данными. Применение данных процедур в реальности либо невозможно, либо затруднено.

Численный вероятностный анализ представляет собой новый раздел вычислительной математики, предназначенный для решения различных задач со случайными входными данными. Основой численного вероятностного анализа являются понятие вероятностного расширения и численные операции над плотностями вероятности случайных величин [36, 46].

В основе метода лежат численные операции над функциями плотности вероятности случайных значений и вероятностными расширениями. Численный вероятностный анализ направлен прежде всего на разработку методов представления, обработки, численных процедур моделирования и анализа

данных, способствующих снижению уровня неопределенности в зависимости от ее типа, характера, специфических особенностей, объема и источников на всех стадиях информационного процесса, сопровождающего принятие управленческого решения.

ЧВА оперирует с плотностями случайных величин, представленных гистограммами, дискретными и кусочно-полиномиальными функциями [47].

2.2 Гистограммный подход

В тех случаях, когда желательно по данным эксперимента построить оценку плотности вероятностей, экспериментаторы чаще всего прибегают к построению гистограммы [38].

Термин гистограмма был введен знаменитым статистиком Карлом Пирсоном для обозначения "общей формы графического представления". В цитате Оксфордского словаря английского языка из "Philosophical Transactions of the Royal Society of London" упоминается, что слово 'гистограмма' было введено автором лекций по статистике как термин для обозначения общей формы графического представления, т.е. путем маркировки столбцов как областей частотности в соответствии с масштабом их базиса".

В тех случаях, когда желательно по данным эксперимента построить оценку плотности вероятностей, экспериментаторы чаще всего прибегают к построению гистограммы. Идея гистограммного подхода такова: вместе с общими представлениями стохастических величин своими плотностями распределения в виде непрерывных функций, имеется возможность рассматривать случайные переменные, плотность распределения которых имеет вид гистограммы.

Понятия гистограммного и вероятностного расширений вводятся для численного моделирования функциональных зависимостей со случайными аргументами.

Гистограмма — это случайная величина, плотность распределения которой представлена кусочно-постоянной функцией. Гистограмма P

определяется сеткой $\{x_i \backslash i$ 0,..., $n\}$, на каждом отрезке $[x_{i-1}, x_i]$, i = 1,..., n гистограмма принимает постоянное значение p_t.

Процедура ее построения проста и состоит из следующих шагов.

В области возможных значений измеряемой величины X строится сетка, которая определяется

$$\omega = \{x_i | i = 1, 2, 3, \dots, n\} \quad (19)$$

Определяется, сколько выборочных значений m_i от общего числа N оказалось в каждом интервале $(x_{i-1},\ x_i]$.

Над каждым из интервалов строится вертикальный прямоугольник с площадью m_i/N. Высота прямоугольника определяется формулой

$$P_i = m_i/(N(x_i - x_{i-1})) \quad (20)$$

Полученная совокупность прямоугольников и называется гистограммой. Другими словами гистограмма — кусочно-постоянная функция и определяется своей сеткой ω, значениями $\{P_i\}$, принимающая на каждом интервале $(x_{i-1},\ x_i]$ постоянное значение P_i [39].

Ниже на рисунке 2, представлен пример гистограммы.

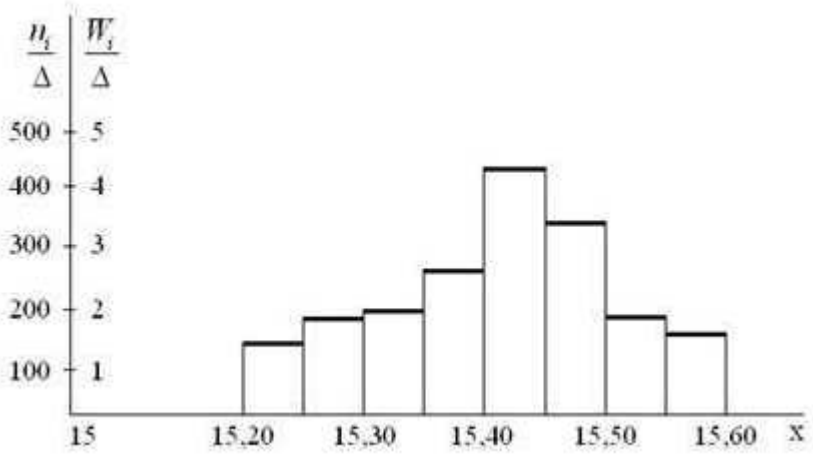

Рисунок 2 - Пример гистограммы

В прикладных задачах зачастую отсутствует возможность получить истинную функцию распределения случайной величины. Тогда следует задавать оценки плотности сверху и снизу. Аппроксимировать такие оценки удобно с применением интервальных гистограмм. Гистограмма является интервальной, если ее функция распределения *P(x)* является кусочноинтервальной функцией.

Наряду с интервальными гистограммами для эпистемической неопределенности применимы гистограммы второго порядка. Гистограмма второго порядка — это гистограмма, каждый столбец которой — гистограмма [43].

Для обработки мнений экспертов и включения результатов в численные процедуры построения вероятностных характеристик показателей привлекательности инвестиционных проектов, предлагается использовать гистограммы второго порядка (ГВП).

Определим гистограмму второго порядка как кусочно-гистограммную функцию. ГВП так же, как и обычная гистограмма, определяется сеткой $\{z_i, i = 0,1,... n\}$ и набором гистограмм $\{P_i, i = 1,2,...,n\}$. На каждом отрезке $[z_{i-1}, z_i]$ ГВП принимает гистограммное значение P_i.

Рассмотрим процедуру построения ГВП. Пусть имеется ряд гистограмм [Yi, i = 1,2,...,N}. Каждой Yj ставится в соответствие вероятность p_i : £ p_t = 1. Для простоты будем считать, что все гистограммы Y_i, заданы на сетке {z_i, i = 0,1,...n} и на отрезке [z_{k-1}, z_k] Y_i принимает значение Y_{ik}. Таким образом, на каждом отрезке [z_{k-1}, z_k] имеем случайную величину Y_k, принимающую значения Y_{ik} с вероятностью $p_{\{}$. Используя эти значения, имеем возможность на каждом отрезке [z_{k-1}, z_k] восстановить гистограмму Pz_k.

На рисунке 3 приведена гистограмма второго порядка, где оттенками серого показано распределение вероятностей. Интервальное распределение (максимальное и минимальное P_t для всех t, где t — случайная величина) изображено граничными линиями. Внутренняя линия определяет «эффективную» плотность вероятности гистограммы второго порядка — математическое ожидание плотностей вероятности P_t в точке x [44, 45].

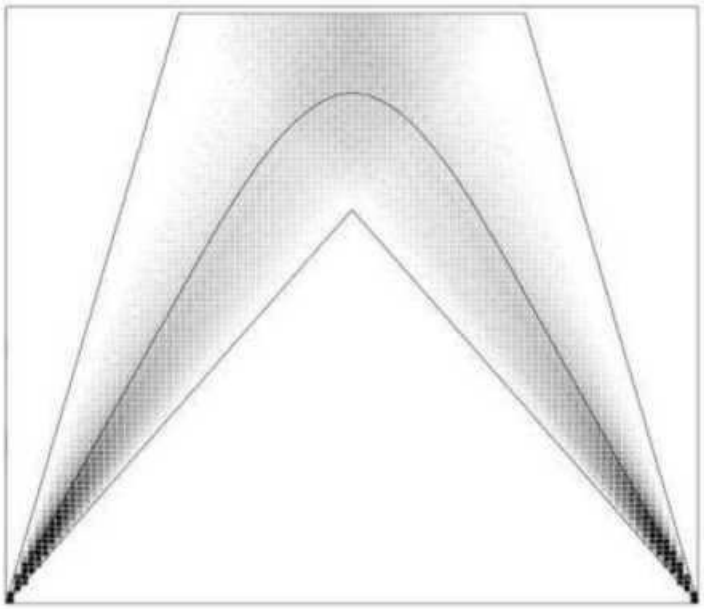

Рисунок 3 — Гистограмма второго порядка

Для осуществления численных операций над «неопределенными» переменными, заданными своими функциями плотности в виде гистограмм

второго порядка, в условиях неопределенности, существует арифметика для ГВП [46].

2.3 Гистограммная арифметика

В рамках численно-вероятностного анализа существуют арифметические операции над гистограммами, такие как сложение, вычитание, умножение, деление и т.д. Для всех этих операций известны аналитические формулы. Например, для нахождения плотности вероятности $p_{x_1+x_2}$ суммы двух непрерывных случайных величин $x_1 + x_2$ используется соотношение [40]

$$p_{x_1+x_2}(x) = \int_{-\infty}^{\infty} p(x-v, v)dv = \int_{-\infty}^{\infty} p(v, x-v)dv \quad (21)$$

Однако такая запись не всегда удобна для численного расчета. Рассмотрим основные принципы разработки гистограммных операций на примере операции сложения. Пусть $z = x_1 + x_2$, и носители $x_1 - [a_1, a_2], x_2 - [b_1, b_2], p(x_1, x_2)$ – плотность распределения вероятностей случайного вектора (x_1, x_2). Заметим, что прямоугольник $[a_1, a_2] \times [b_1, b_2]$ – носитель плотности распределения вероятностей $p(x_1, x_2)$ на рисунке 3 и плотность вероятности z отлична от нуля на интервале $[a_1 + b_1, a_2 + b_2]$. Обозначим $z_i, i = 0, 1, \dots, n$ – точки деления этого интервала на n отрезков. Тогда вероятность попадания величины z в интервал $[z_i, z_{i+1}]$ определяется по формуле [41]:

$$P(z_i < z < z_{i+1}) = \int \int_{\Omega_i} p(x_1, x_2)dx_1 dx_2 \quad (22)$$

где $\Omega_i = \{(x_1, x_2) | z_i \le x_1 + x_2 \le z_{i+1}\}$.

В итоге p_{zi} имеет следующий вид

$$p_{zi} = \int\int_{\Omega_i} p(x_1, x_2) dx_1 dx_2 / (z_{i+1} - z_i). \quad (23)$$

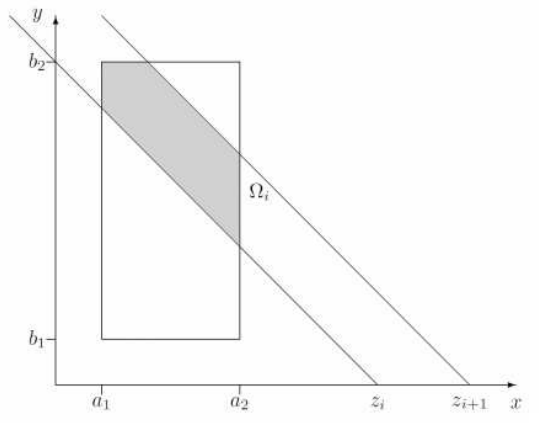

Рисунок 4 – Построение гистограммы суммы двух случайных величин

Численный вероятностный анализ позволяет решать разнообразные задачи обработки и представления данных, содержащих различные виды неопределенности. Разработанная в рамках ЧВА гистограммная арифметика позволяет производить различные операции над плотностями вероятностей.

2.5 Полиграмма

Полиграмма – кусочно-постоянная непараметрическая оценка плотности f(x) [42]. Построение полиграммы сводится к упорядочению выборки и построению прямоугольников площади K/(N + 1). Полиграмма K-го порядка является асимптотически несмещенной оценкой f(x) и имеет конечные моменты лишь при $r \leq K$.

Для удобства обозначим порядковые статистики, ранги которых кратны K, через $\xi_j = x_{(jk)}$. Введя обозначение m = K/N и введя функцию индикатор интервала с помощью ступенчатой функции:

$c(t) = \{1 : t \geq 0; 0 : t < 0\}$,

результирующую оценку плотности можно записать как:

$$f_N(x) = \frac{1}{m} \sum_{j=1}^{m} \frac{c(x-\xi_i)-c(x-\xi_{i+1})}{\xi_{i+1}-\xi_i}. \quad (24)$$

Теорема. Если неизвестная плотность f(x) > 0, ограниченная вместе со своей первой производной, оценивается полиграммой fN (x) К-го порядка, K = Nα, α < 0.5 fN (x) является состоятельной оценкой f(x).

Для построения полиграммы:

1. Сортируем вариационный ряд в порядке возрастания;

2. Выбираем шаг;

3. Делим вариационный ряд на интервалы и считаем промежуточный элемент между интервалами: $\overset{\wedge}{x_i} = \frac{x_i + x_{i+1}}{2}$

4. Вычисляем частоты hi: $hi = \hat{x} - xi$

5. После этого вычисляем высоту для построения полиграммы

$$Fn(x) = \frac{h}{h_i * N}, \quad (25)$$

где h - шаг, h_i – частоты, N – объем выборки.

На рисунке 5 представлен пример полиграммы.

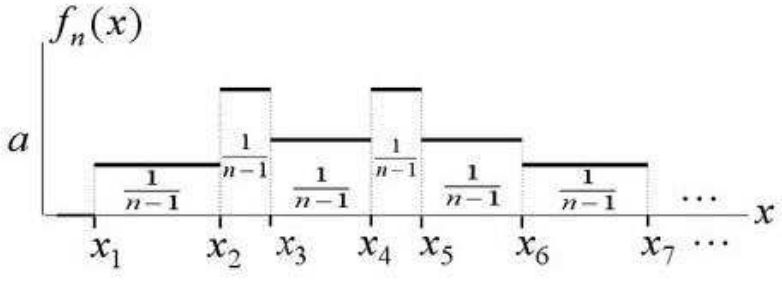

Рисунок 5 – Полиграмма 1 порядка

2.6 Модель частотного полигона

Модель частотного полигона – непрерывная оценка плотности на основе гистограммы, с той или иной формой линейной интерполяции, получающаяся при соединении точек с координатами, т.е. соединяющая середины верхних сторон прямоугольников (кусочно-линейная функция) [43].

Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат w_i. Точки (x_i, w_i) соединяют отрезками прямых и получают частотный полигон. Ниже на рисунке 6, представлен пример частотного полигона.

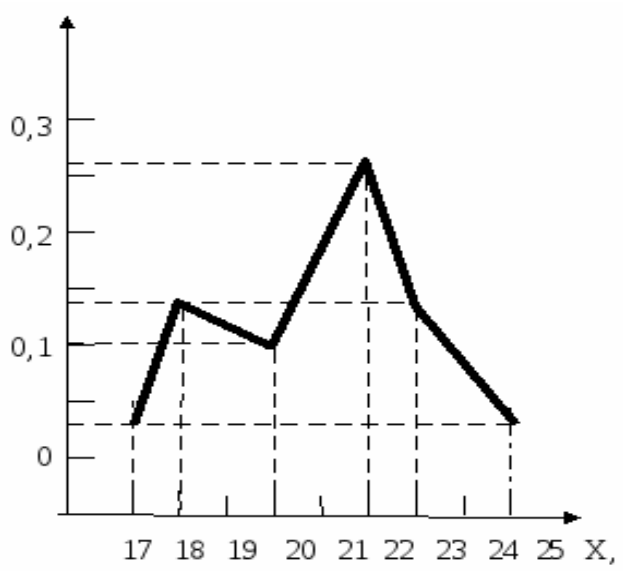

Рисунок 6- Частотный полигон

2.4 Ядерная оценка плотности

Ядерная оценка плотности — это непараметрический способ оценки плотности случайной величины. Ядерная оценка плотности является задачей сглаживания данных, когда делается заключение о совокупности, основываясь на конечных выборках данных. В некоторых областях, таких как обработка сигналов и математическая экономика, метод называется также методом окна Парзена-Розенблатта.

Пусть $(x_1, x_2, \ldots, x_n)$ является одномерной выборкой независимых одинаково распределённых величин, извлечённых из некоторого распределения с неизвестной плотностью f. Наша задача заключается в оценке формы функции f. Её ядерный оценщик плотности равен

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h\,(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\,(\frac{x-x_i}{h}) \tag{26}$$

где K является ядром, т.е. неотрицательной функцией, а $h > 0$ является сглаживающим параметром, называемым *шириной полосы*. Ядро с индексом h называется *взвешенным ядром* и определяется как $K_h(x) = 1/hK(\frac{x}{h})$. Интуитивно стараются выбрать h как можно меньше, насколько данные это позволяют, однако всегда существует выбор между смещением оценщика и его дисперсией. Выбор полосы пропускания обсуждается более подробно ниже.

Существует ряд наиболее часто используемых ядерных функций: однородная, треугольная, бивзвешенная, тривзвешенная, Епанечникова, нормальная и другие. Ядро Епанечникова оптимально в смысле среднеквадратичной ошибки, хотя потеря эффективности для ядер, перечисленных до него, мала. Вследствие удобных математических свойств часто используется нормальное ядро, среднее которого $K(x) = \phi(x)$, где ф является стандартной нормальной функцией плотности.

Построение ядерной оценки плотности находит интерпретацию в областях за пределами оценки плотности. Например, в термодинамике это эквивалентно количеству теплоты, получающеейся, когда ядра оператора теплопроводнсти (фундаментальные решения уравнения теплопроводности) помещаются в каждой точке данных x_i. Похожие методы используются для построения дискретных операторов Лапласа в точках облака для обучения на основе многообразий.

Ядерные оценки плотности тесно связаны с гистограммами, но могут быть наделены свойствами, такими как гладкость или непрерывность, путём выбора подходящего ядра. Чтобы это увидеть, сравним построение гистограммы и ядерной оценки плотности на этих 6 точках, рисунок 7:

1	2	3	4	5	6
-2,1	-1,3	-0,4	1,9	5,1	6,2

Рисунок 7 – Точки для построения гистограммы и ядерной оценки плотности

Для гистограммы горизонтальная ось разделена на подинтервалы, которые покрывают область данных. В этом случае мы имеем 6 отрезков, каждая длины 2. Когда точка данных попадает внутрь отрезка, мы помещаем прямоугольник высоты 1/12. Если в отрезок попадает более одной точки, мы размещаем прямоугольники друг над другом.

Для ядерной оценки плотности мы помещаем нормальное ядро с дисперсией 2,25 (показаны красными пунктирными линиями) для каждой точки данных x_i. Ядра суммируются, давая ядерную оценку плотности (сплошная синяя кривая). Гладкость ядерной оценки плотности очевидна при сравнении с дискретностью гистограммы, так как ядерные оценки плотности сходятся быстрее к истинной лежащей в основе плотности для непрерывных случайных величин.

Сравнение гистограммы (слева) и ядерной оценки плотности (справа), построенных из тех же самых данных. 6 индивидуальных ядер показаны красными пунктирными линиями, ядерная оценка плотности показана синей кривой. Точки данных показаны чёрточками на ленточной диаграмме по горизонтальной оси – рисунок 8.

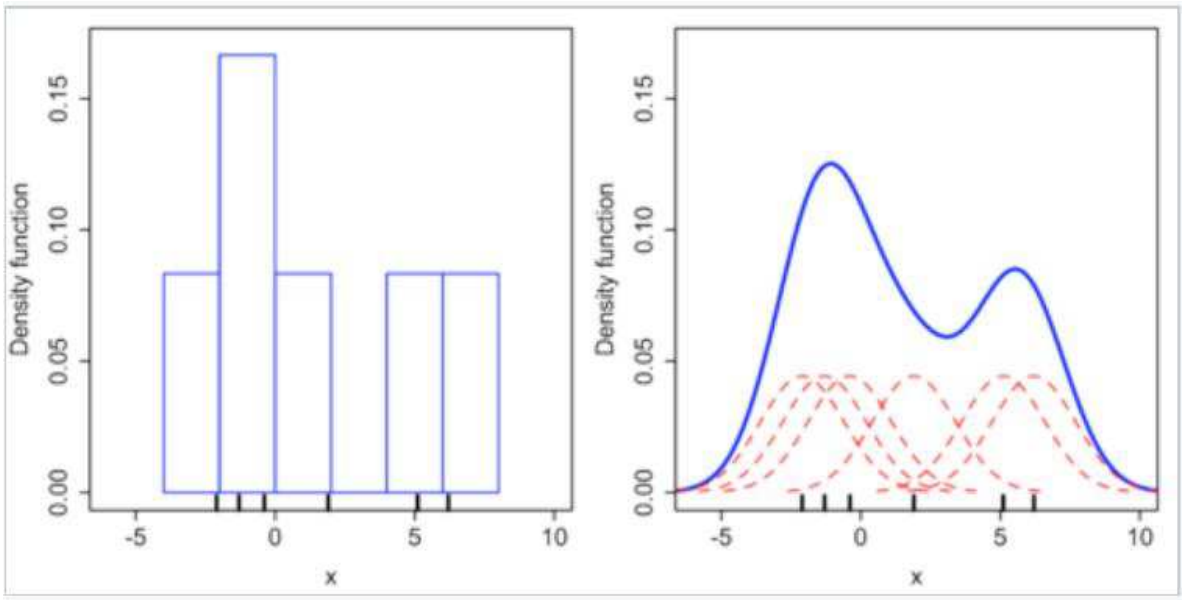

Рисунок 8 – Сравнение гистограммы и ядерной оценки плотности

Оценка плотности ядра может быть сведена к четырем шагам:

На рисунке 9 есть три точки, поэтому n = 3.

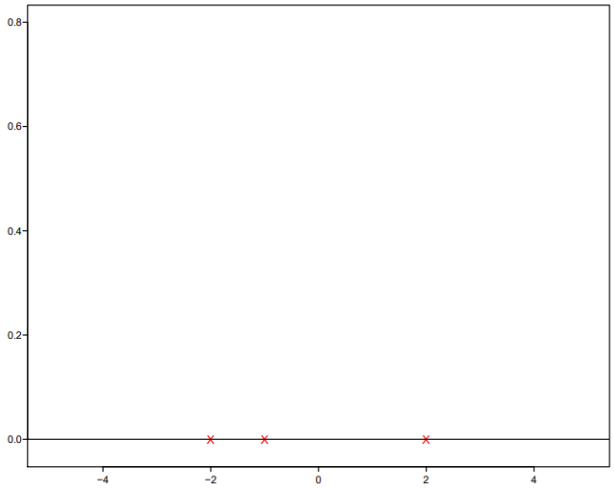

Рисунок 9 – Пример ядерной оценки плотности, 3 точки

Вокруг каждой точки данных рисуется ядро. На рисунке 10 использовано гауссово ядро.

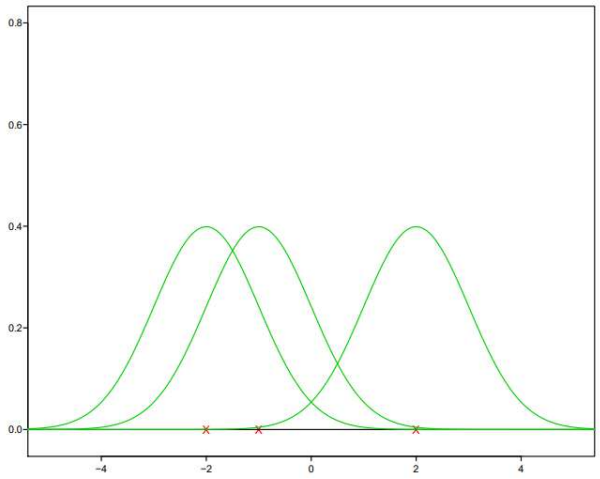

Рисунок 10 – Гауссово ядро

Затем ядра объединяются - синяя пунктирная линия на рисунке 11.

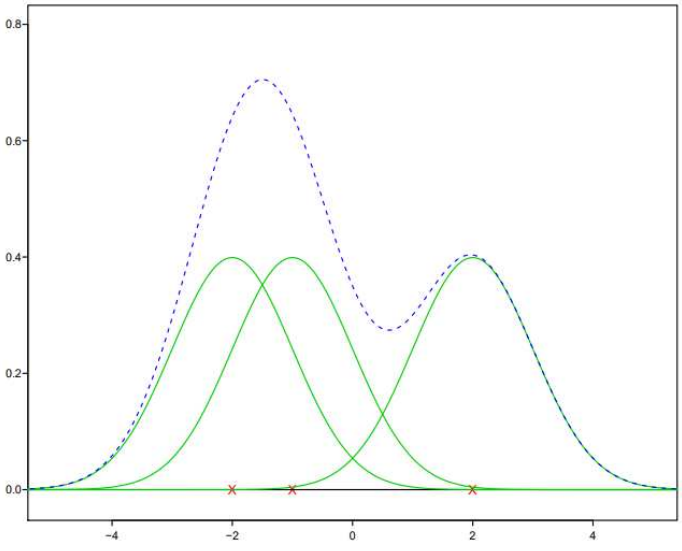

Рисунок 11 – Объединённые ядра

Последний шаг - нормализация распределения. В примере есть три точки, общая площадь под синей пунктирной кривой = 3, следовательно, чтобы восстановить плотность вероятности, мы делим на 3, чтобы получить черную кривую - рисунок 12.

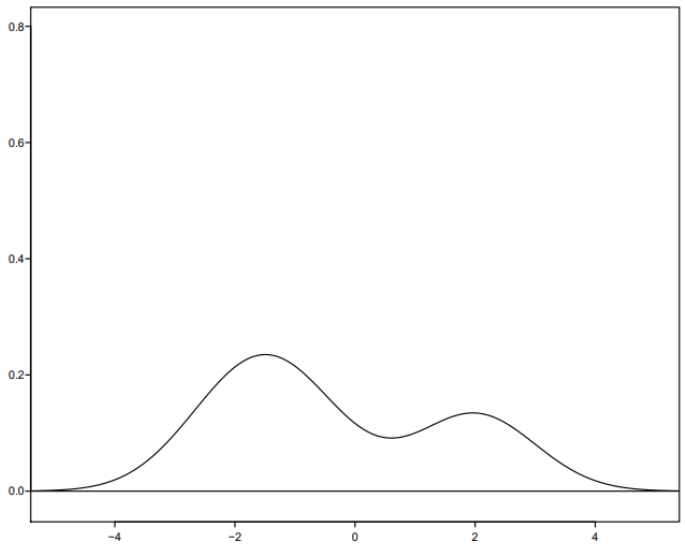

Рисунок 12 – Объединённые ядра

Пусть K - ядро, и предположим, что в нашем образце содержится n значений: x1,. , , xn.

Тогда наша оценка истинной плотности вероятности

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\,(x - x_i) \qquad (27)$$

Мы используем K (x - xi), так как мы рисуем ядро вокруг каждого xi.

2.6 Экстраполяция Ричардсона

В работе [48] рассматривается новый подход к повышению точности в задачах построения аппроксимации функции плотности вероятности и оценке ее погрешности. Подход основан на комбинации ядерных оценок с различными параметрами сглаживания h. Для этих целей используются экстраполяция Ричардсона и правило Рунге.

Использование оценок вторых производных позволяет получить реалистичные оценки математических ожиданий в l_2 норме погрешности

аппроксимации функции плотности вероятности. Знание этих оценок позволяет рассчитать оптимальный параметр сглаживания h [49].

Экстраполяция Ричардсона — очень мощный подход, который может эффективно использоваться для повышения производительности численных расчетов. Он был представлен Ричардсоном в начале XX века, и после этого многие ученые и инженеры многократно использовали для повышения точности численных расчетов математических моделей. В большинстве приложений этот метод до сих пор использовался главным образом в усилиях либо для повышения точности результатов модели. Однако этот подход в виде правила Рунге можно использовать для оценки и проверки величины вычислительных ошибок, попытаться добиться большей точности и повысить эффективность вычислительного процесса.

Экстраполяция Ричардсона является общим подходом для получения результатов высокой точности по формулам низкого порядка. Экстраполяция Ричардсона получила широкое распространение для повышения точности разностных методов решения задач Коши для систем обыкновенных дифференциальных уравнений и краевых задач для дифференциальных уравнений в частных производных [49].

Экстраполяция Ричардсона основана на разложении по степеням h приближенного решения u^h как суммы

$$u^h = u + h^k v + O(h^{k+m}), \tag{28}$$

где u есть искомое точное решение, v - неизвестная функция and h — малый параметр дискретизации, который чаще всего рассматривается, как шаг разностной сетки. Целое значение k характеризует порядок точности приближенного решения, and $m > 0$ характеризует порядок точности приближенного решения членом погрешности $h^k v$. Поскольку u и v не зависят от h, для параметра $h/2$ справедливо разложение:

$$u^{h/2} = u + \left(\frac{h}{2}\right)^k v + O(h^{k+m}). \quad (29)$$

Объединим два разложения таким способом, чтобы исключить ошибку порядка h^k. Умножим (2) на 2^k и вычтем из (1) получаем

$$u = \frac{2^k}{2^k - 1} u^{h/2} - \frac{1}{2^k - 1} u^h + O(h^{k+m}). \quad (30)$$

Таким образом, получено приближение точного решения с более высоким порядком точности.

Поскольку этот подход не требуют значительного увеличения вычислительных затрат, то они будут весьма полезны для практических задач обработки и анализа эмпирических данных.

2.5 Свёртка функций

Концепция свертки является центральной в теории Фурье и анализе линейных систем. В одном измерении свертка между двумя функциями f(x) и h(x) определяется как:

$$g(x) = f(x) \otimes h(x) = \int_{-\infty}^{\infty} f(s)h(x - s)ds, \quad (31)$$

где s - фиктивная переменная интеграции.

Эту операцию можно считать областью перекрытия между функцией f(x) и пространственно-обратной версией функции h(x). Результат свертки двух простых одномерных функций показан на рисунке 13.

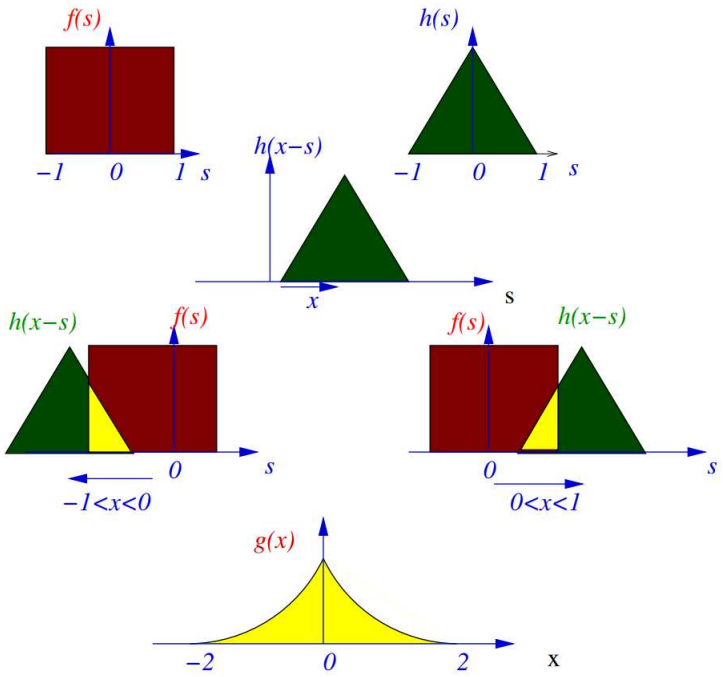

Рисунок 13 – Свёртка двух похожих функций

Теорема о свертке связывает свертку между областью реального пространства с умножением в области Фурье и может быть записана как:

$$G(u) = F(u)\,H(u), \tag{32}$$

где

$$G(u) = F\{g(x)\}$$
$$F(u) = F\{f(x)\}$$
$$H(u) = F\{h(x)\}$$

2.7 Сплайн

В последние годы интенсивно развивается новый раздел современной вычислительной математики – теория сплайнов. Сплайны позволяют эффективно решать задачи обработки экспериментальных зависимостей между параметрами, имеющими достаточно сложную структуру.

Сплайн – кусочно-полиномиальная функция, определенная на отрезке [a, b] и имеющая на этом отрезке некоторое количество непрерывных производных [44].

Наиболее широкое практическое применение, в силу их простоты, нашли кубические сплайны. Основные идеи теории кубических сплайнов сформировались в результате попыток математически описать гибкие рейки из упругого материала (механические сплайны), которыми издавна пользовались чертежники в тех случаях, когда возникала необходимость проведения через заданные точки достаточно гладкой кривой. Известно, что рейка из упругого материала, закрепленная в некоторых точках и находящаяся в положении равновесия, принимает форму, при которой ее энергия является минимальной. Это фундаментальное свойство позволяет эффективно использовать сплайны при решении практических задач обработки экспериментальной информации.

В общем случае для функции $y = f(x)$ требуется найти приближение $y = S(x)$ таким образом, чтобы $f(x_i) = S(x_i)$ в точках $x = x_i$, а в остальных точках отрезка [a;b] значения функций $f(x)$ и $S(x)$ были близкими между собой. При малом числе экспериментальных точек для решения задачи интерполяции можно использовать один из методов построения интерполяционных полиномов. Однако при большом числе узлов интерполяционные полиномы становятся практически непригодными. Это связано с тем, что степень интерполяционного полинома лишь на единицу меньше числа экспериментальных значений функций.

Пример сплайна, наложенного на график частотного полигона изображен на рисунке 14.

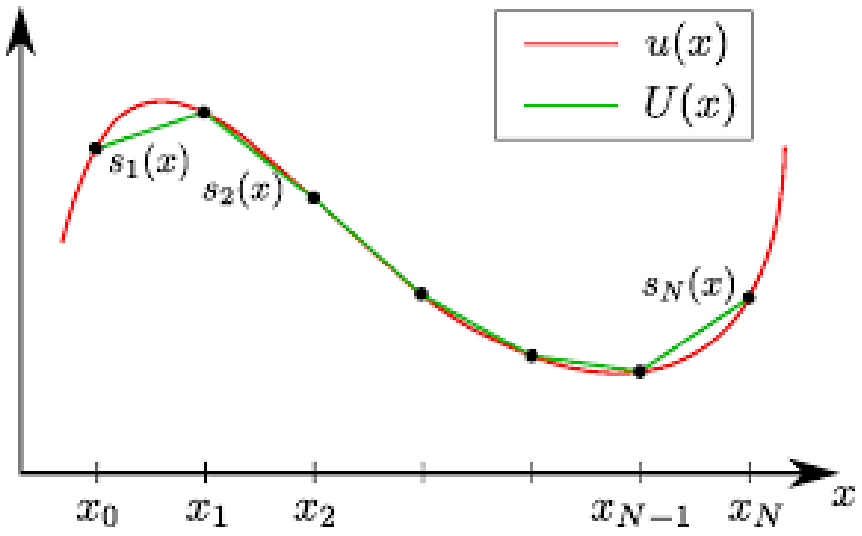

Рисунок 14 – Сплайн

Можно, конечно, отрезок, на котором определена функция, разбить на участки, содержащие малое число экспериментальных точек, и для каждого из них построить интерполяционные полиномы. Однако в этом случае аппроксимирующая функция будет иметь точки, где производная не является непрерывной, т. е. график функции будет содержать точки “излома”.

Кубические сплайны лишены этого недостатка. Исследования показали, что гибкая тонкая линейка между двумя узлами достаточно хорошо описывается кубическим полиномом, и поскольку она не разрушается, то аппроксимирующая функция должна быть, по меньшей мере, непрерывно дифференцируемой [45].

Таким образом, сплайн– это функция, которая на каждом частичном отрезке интерполяции является алгебраическим многочленом, а на всем заданном отрезке непрерывна вместе с несколькими своими производными.

3 Экспериментальная часть

3.1 Разработка функциональной схемы программного модуля

Первым этапом разработки любого программного продукта является разработка функциональной схемы. Она включается в себя функциональные части, такие как: элементы, блоки, функциональные группы и связи межу ними. Графическое построение схемы должно наглядно отражать последовательность функциональных процессов. Была разработана и построена функциональная схема модуля прогнозирования временных рядов данных большого объёма, представленная на рисунке 15.

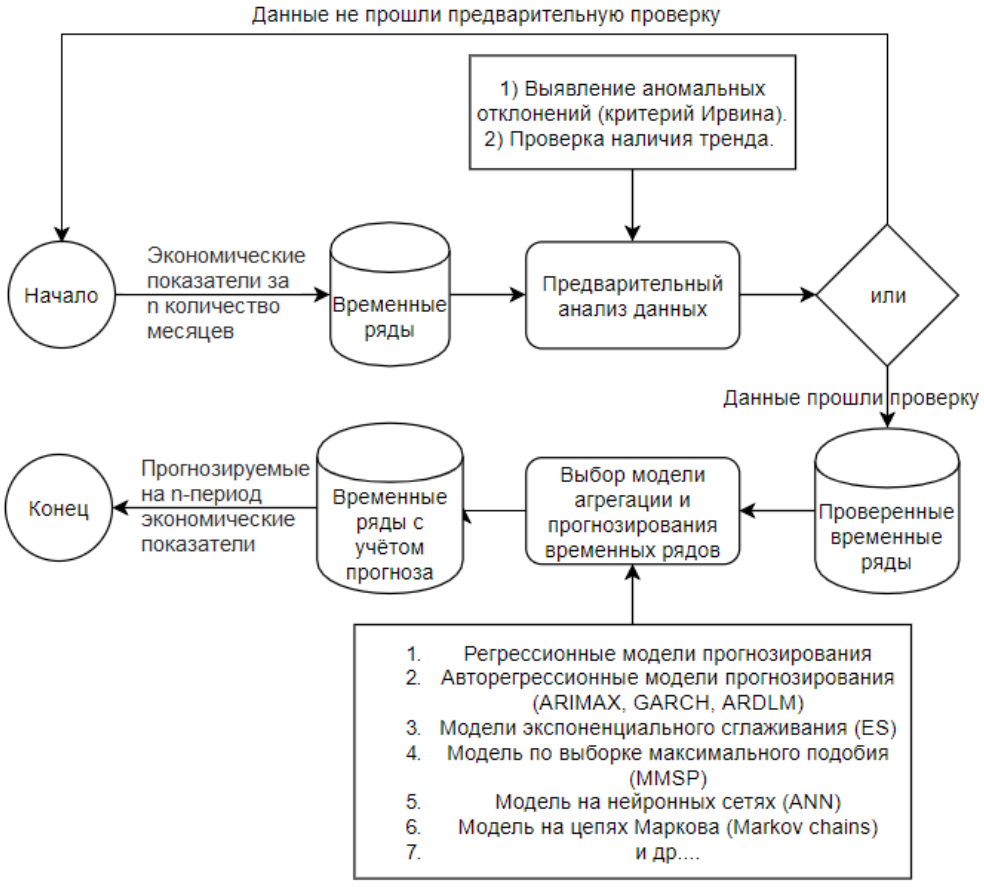

Рисунок 15 – Функциональная схема модуля прогнозирования

На первом этапе берутся данные, с которыми будет производиться работа. Это могут быть реальные экономические показатели, либо

смоделированные данные. Прежде чем прогнозировать реальные данные, лучше все произвести отладку модуля с помощью смоделированных данных, как и будет сделано в этой работе.

На втором этапе данные формируются во временные ряды определенной размерности.

На третьем этапе происходит предварительный анализ данных, где становится ясно, можем ли мы в принципе использовать представленные данные. Для этого с помощью Критерия Ирвина происходит поиск аномальных отклонений значений, а также выполняется проверка на наличие тренда у временного ряда.

На четвёртом этапе, если данные прошли предварительную проверку, происходит выбор метода агрегации и прогнозирования временных рядов. Как уже говорилось, универсального метода не существует и в каждой определённой ситуации лучше подойдет тот или иной метод. Мы будем использовать сплайн-метод численного вероятностного анализа.

На пятом этапе получаются спрогнозированные временные ряды, по которым можно сделать анализ полученных результатов.

3.2 Вычислительный эксперимент

Для вычислительного эксперимента было решено использовать язык разработки Pascal. Интегрированная среда разработки PascalABC.NET ориентирована на создание проектов малой и средней сложности. Она достаточно легковесна и в то же время обеспечивает разработчика всеми необходимыми средствами, такими как встроенный отладчик, средства Intellisense (подсказка по точке, подсказка по параметрам, всплывающая подсказка по имени), переход к определению и реализации подпрограммы, шаблоны кода, автоформатирование кода.

В отличие от многих профессиональных сред, среда разработки PascalABC.NET не имеет громоздкого интерфейса и не создает множество дополнительных вспомогательных файлов на диске при компиляции

программы. Для небольших программ это позволяет соблюсти принцип Одна программа - один файл на диске.

Был смоделирован массив данных размерностью 800*15=12000 значений, функция плотности вероятности которого описывается уравнением:

$$P_4(x) = \begin{cases} \frac{x^3}{6}, & x \in [0,1] \\ -\frac{x^3}{2} + 2x^2 - 2x + \frac{2}{3}, & x \in [1,2] \\ \frac{x^3}{2} - 4x^2 + 10x - \frac{22}{3}, & x \in [2,3] \\ -\frac{x^3}{6} + 2x^2 - 8x + \frac{32}{3}, & x \in [3,4] \end{cases} \quad (33)$$

Функция $f(x)$ - производная функции распределения – характеризует как плотность, с которой распределяются значения случайной величины в данной точке. Эта функция называется плотностью распределения (иначе – «плотность вероятности») непрерывной случайной величины X.

Плотность вероятности — один из способов задания вероятностной меры на евклидовом пространстве R*n*. В случае, когда вероятностная мера является распределением случайной величины, говорят о плотности случайной величины.

Построенный график функции плотности вероятности представлен на рисунке 16.

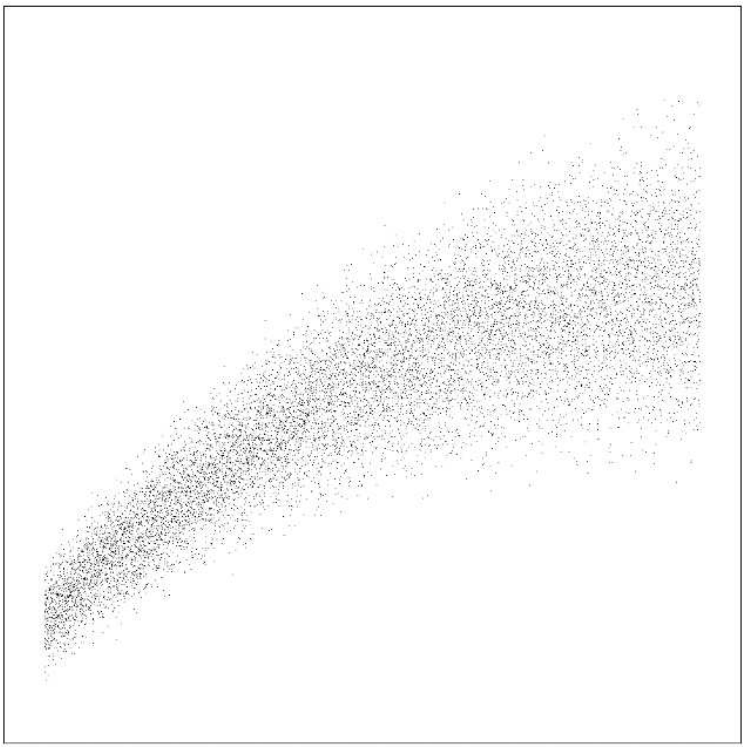

Рисунок 16 – График функции плотности вероятности

На графике представлено сложение четырёх равномерно распределённых случайных величин. На каждом отрезке классический кубический полином – сплайн. На концах производная равна нулю. Внутри кубические полиномы сшиваются так, что сама функция непрерывна и 1-ая, 2-ая производные непрерывны. Смоделируем ситуацию, когда часть данных нам неизвестна, от нас требуется спрогнозировать эти данные – рисунок 17.

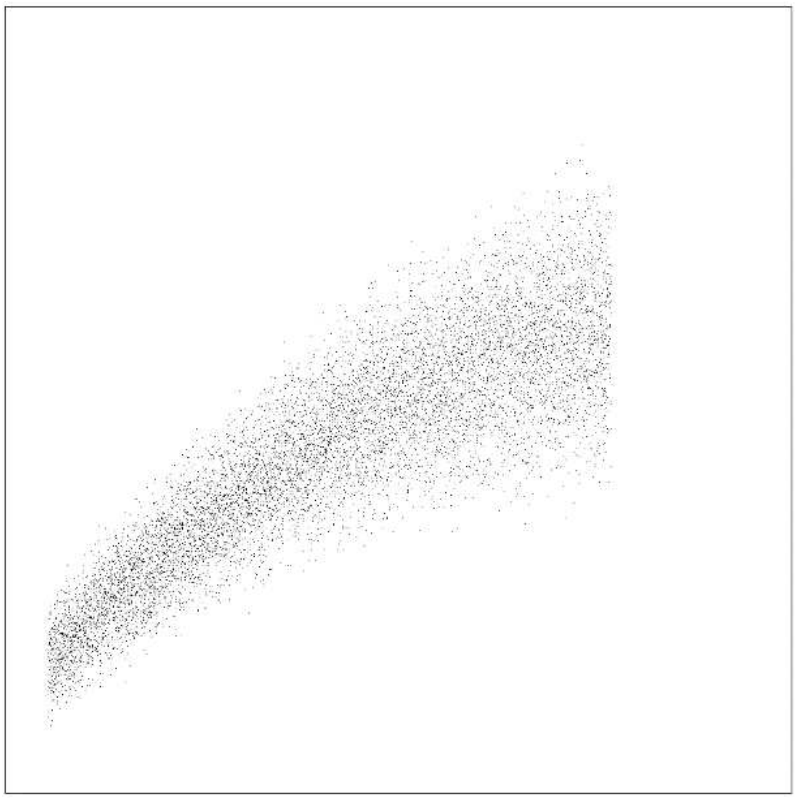

Рисунок 17 – График функции смоделированной ситуации неопределённости

Следующий шаг - строим границы плотности распределения и сравниваем истинную регрессионную кривую (Ирвина Холла =4), гистограммный метод и сплайн – рисунок 18.

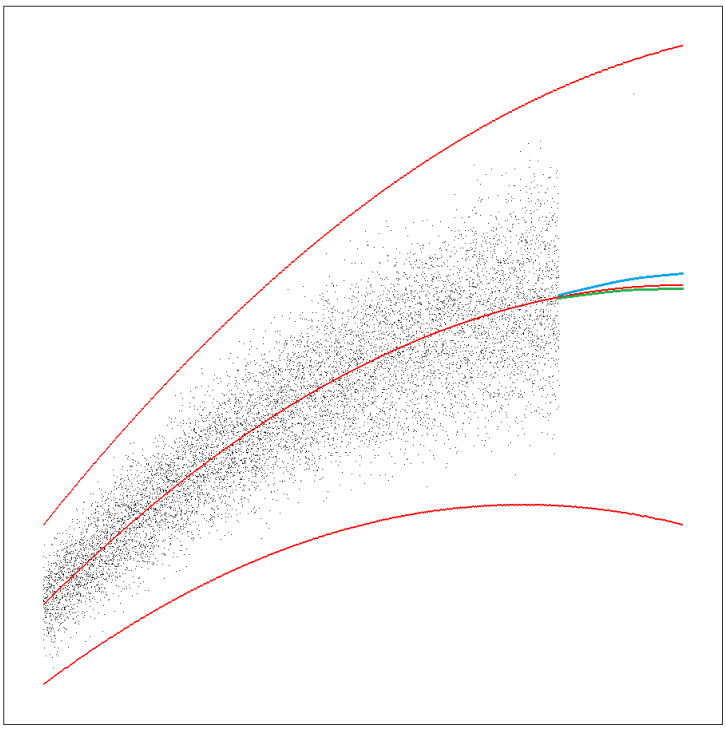

Рисунок 18 – Ирвин Холл 4(красная), гистограмма(синяя), сплайн(зеленая)

Обычно, в таких ситуациях, как, например, в методе Arima+hist, предложенном в работе [1], используют гистограмму. Но мы будем использовать метод ЧВА – сплайн – рисунок 19.

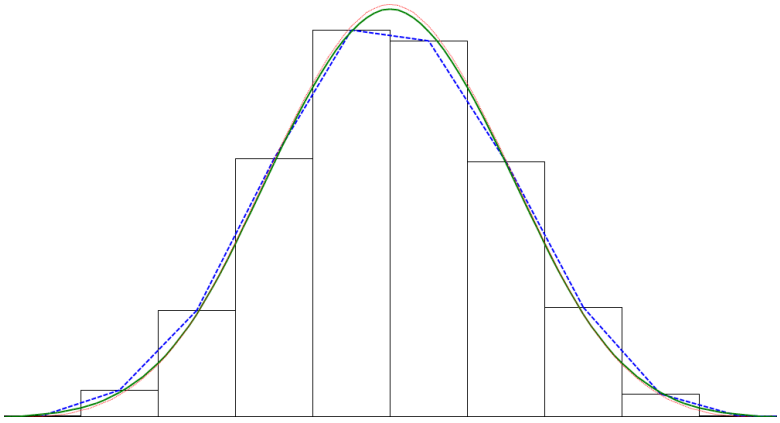

Рисунок 19 – Сравнение прогнозирования

3.3 Результат вычислительного эксперимента

В ходе эксперимента была спрогнозирована неизвестная часть смоделированных данных. По предварительной оценке, метод сплайн численного вероятностного анализа показал точное прогнозирование данных. Метод не потребовал высоких вычислительных мощностей, что делает его доступным во всех сферах применения.

ЗАКЛЮЧЕНИЕ

В ходе выполнения целей и задач данной научно-исследовательской работы был произведен анализ темы исследования, в результате которого были выявлены проблемы, связанные со сложность агрегации временных рядов данных большого объёма, трудоёмкостью классических методов прогнозирования временных рядов, а также ростом количества ошибок при увеличении прогнозируемой выборки.

Главной проблемой анализа больших данных является качество обработки данных большого объёма, а также сложность определения законов распределения случайной величины, функции плотности вероятности и использования больших данных в целях прогнозирования. Планировщики не включают в свои расчётные методы современный математический аппарат.

В итоге, спрогнозированные данные, включаемые в бизнес - план, представляются, как усреднённые, по которым в дальнейшем производятся вычисления в расчётных формулах, и соответственно затрудняет принятие эффективных экономических решений. В качестве аналога существующих методов прогнозирования, был рассмотрены методы численного вероятностного анализа.

В результате выполнения работы были выполнены следующие задачи:

- были проанализированы научные работы по теме исследования;
- описаны теоретические основы прогнозирования временных рядов больших данных, рассмотрены основные методы агрегации данных большого объёма и модели прогнозирования;
- рассмотрена возможность применения методов ЧВА для прогнозирования временных рядов данных большого объёма;
- построена функциональная схема модуля прогнозирования;
- оттестирован модуль прогнозирования временных рядов данных большого объёма с использованием методов ЧВА на примере смоделированных данных.

По результатам исследований была опубликована статья:

1) Васильев Н. Н. Новые подходы к обработке информации в робототехнике // Робототехника и искусственный интеллект: материалы X Всероссийской научно-технической конференции с международным участием (г. Железногорск, 8 декабря 2018 г.) / под науч. ред. В.А. Углева. – Электрон. дан. – Красноярск: ЛИТЕРА-принт, 2018. – 237 с.

СПИСОК СОКРАЩЕНИЙ

ЧВА – Численный вероятностный анализ

СКО – Среднеквадратичное отклонение

MSE – Среднеквадратичная ошибка

MAE – Средняя абсолютная ошибка

ARIMA – Авторегрессионное интегрированное скользящее среднее

ARMA – Модель авторегрессии — скользящего среднего

HIST – Модель гистограммного прогнозирования

ACF – Выборочная автокорреляционная функция

JB – Тест Жарка-Бера

ГВП – Гистограмма второго порядка

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) Стрижов В.В., Мотренко А. П., Кузнецов М.П., Каширин Д.О. Исследование и разработка математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования объемов спроса на грузовые железнодорожные перевозки) – 2015.

2) Shi Y (2014) Big Data: history, current status, and challenges going forward. Bridge 44(4). С. 6–11

3) Доклад о мировом развитии 2016. Цифровые дивиденды. Обзор. Международный банк реконструкции и развития / Мировой банк 2016.

4) Добрынин А. П. и др. Цифровая экономика-различные пути к эффективному применению технологий (BIM, PLM, CAD, IOT, Smart City, BIG DATA и другие) //International Journal of Open Information Technologies. – 2016. – Т. 4. – №. 1. С. 4-11

5) Куприяновский В. П., Намиот Д. Е., Куприяновский П. В. Стандартизация Умных городов, Интернета Вещей и Больших Данных. Соображения по практическому использованию в России //International Journal of Open Information Technologies. – 2016. – Т. 4. – №. 2. С. 34-40

6) Куприяновский В. П. и др. Умные города как «столицы» цифровой экономики //International Journal of Open Information Technologies. – 2016. – Т. 4. – №. 2. С. 41-52.

7) Добронец Б.С., Попова О.А. Элементы численного вероятностного анализа // Вестник Сибирского государственного аэрокосмического университета. — 2012. — № 2 (42). — С. 19-23.

8) Попова О.А., Велиходский А.С. Информационно-аналитический подход к обработке экономической информации на основе численного вероятностного анализа – диссертация, 2017.

9) СТО 4.2–07–2014 Система менеджмента качества. Общие требования к построению, изложению и оформлению документов учебной деятельности. – Введ. 30.12.2013. – Красноярск: СФУ, 2014. – 60 с.

10) Добронец Б.С., О.А. Попова Гистограммная арифметика для визуально-интерактивного моделирования в задачах принятия экономических решений.

11) Герасимов В.А., Добронец Б.С., Шустров М.Ю. Численные операции гистограммной арифметики и их применения // АиТ. Т. 1991, №2.

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

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

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

15) Добронец Б. С., Интервальная математика. Красноярский государственный университет, 2004.

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

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

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

19) Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. (пер. с англ.) – М.: Мир, 1974.

20) Кильдишев Г.С., Френкель А.А. Анализ временных рядов и прогнозирование. М.: «Статистика», 1973.

21) Орлов Ю.Н., Осминин К.П. // Анализ нестационарных временных рядов, ИМП им. М.В. Келдыша РАН, препринт №36 за 2007 год.

22) Орлов Ю.Н., Осминин К.П. // Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Мат. Мод. 2008.

23) Эфрон Б. // Нетрадиционные методы многомерного статистического анализа. - М.: Финансы и статистика, 1988. - 263 с.

24) Материалы научно-технического совета открытого акционерного общества «Российские железные дороги». – М. : ОАО «РЖД», 2014.

25) Правдин, Н.В. Прогнозирование грузовых потоков / Н.В. Правдин, М.Л. Дыканюк, В.Я. Негрей. – М. : Транспорт, 1987. – 247 с.

26) О совершенствовании среднесрочного и долгосрочного прогнозирования объемов погрузки грузов на сети железных дорог: распоряжение от 23 июля 2012 г. № 1451р.

27) Черных, В.Ю. Прогнозирование нестационарных временных рядов при несимметричных функциях потерь / В.Ю. Черных, М.М. Стенина // Машинное обучение и анализ данных. – 2015. – № 14. – Т. 1. – С. 1893–1909.

28) Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов / Ю.П. Лукашин : Финансы и статистика, 2003. –415 с.

29) Хардле, В. Прикладная непараметрическая регрессия / В. Хардле. – М. : Мир. 1993. – 349 с.

30) Барский, А.Б. Нейросетевые технологии на транспорте / А.Б. Барский // Мир транспорта. – № 2. – 2011. – С. 4–11

31) Кузнецов, А.П. Методологические основы управления грузовыми перевозками транспортной системы дис. д-ра техн. наук / А.П. Кузнецов; Рос. акад. гос. службы при Президенте РФ. – СПб. : РАГС, 2001. – 358 с.

32) Исследование и разработка математических методов и алгоритмов для интеллектуальной системы анализа данных (подсистемы прогнозирования

объемов спроса на грузовые железнодорожные перевозки) : отчет о НИР (промежуточный) / руководитель К.В. Рудаков. – ФИЦ ИУ РАН – М., 2014. – 126с.

33) Канторович, Г.Г. Анализ временных рядов / Г.Г. Канторович // Экономический журнал ВШЭ. – №1. – 2002. – 110 с

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

35) Словари и энциклопедии на Академике [Электронный ресурс]: Режим доступа: http://dic.academic.ru/

36) Герасимов В. А., Добронец Б. С., Шустров М. Ю. Численные операции гистограммной арифметики и их применения // АиТ. Т. 1991, №2. С. 83-88.

37) Добронец Б. С. Попова О. А. Численный вероятностный анализ неопределенных данных. Монография. Красноярск 2014.

38) Добронец Б. С., О. А.Попова Гистограммная арифметика для визуально-интерактивного моделирования в задачах принятия экономических решений.

39) Добронец Б. С., Попова О. А. Представление и обработка неопределенности на основе гистограммных функций распределения и P-boxes // Информатизация и связь. 2014. № 2. С. 23-26.

40) Козиков А. А. О возможностях реализации численного вероятностного анализа в среде R // Международная научная конференция «Молодежь и наука: проспект Свободный». 2016.

41) Крянев А. В., Лукин Г. В. Математические методы обработки неопределенных данных. — 2-е изд., испр. — М.: ФИЗМАТЛИТ, 2006. — 216 с.

42) Математическая энциклопедия: Электронный ресурс. URL: http://dic.academic.

43) Попова О. А. Гистограммный информационно -аналитический подход к представлению и прогнозированию временных рядов // Информатизация и связь. 2014 г. № 2. С. 43-47.

44) Попова О. А. Гистограммы второго порядка для численного моделирования в задачах с информационной неопределённостью Известия ЮФУ. Технические науки. УДК 519.24

45) Попова О. А. Численный вероятностный анализ для агрегации, регрессионного моделирования и анализа данных // Информатизация и связь. 2015 г. № 1. С. 15-21.

46) Moore R. E., Interval Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1966.

47) Васильев Н. Н. Новые подходы к обработке информации в робототехнике // Робототехника и искусственный интеллект: материалы X Всероссийской научно-технической конференции с международным участием (г. Железногорск, 8 декабря 2018 г.) / под науч. ред. В.А. Углева. – Электрон. дан. – Красноярск: ЛИТЕРА-принт, 2018. – 237 с.

48) Попова О. А. Использование экстраполяции Ричардсона для повышения точности обработки и анализа эмпирических данных // Журнал «Измерительная техника», №2, Февраль 2019, 72 с.

49) Scott R. W. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, New York 2015.

ПРИЛОЖЕНИЕ А

Листинг кода

```
Program forecasting;
uses GraphABC;

Var
  h, x, xi, u0, uh, u2h, al : real;
  i,j : integer;

function r4:real;
Var r, t : real;
   i,j : integer;
begin
r:=0;
for i:= 1 to 4 do
r := r +  Random;

r4 := r-2.0;
end;

function pfr4(x:real):real;
Var r, t, x2,x3 : real;
   i,j : integer;
begin

x2:=x*x;
x3:=x2*x;

if x < 0 then r:=0;
if (x >=0) and (x < 1)  then r:= x3/6;
if (x >=1) and (x < 2)  then r:= -x3/2 +x2*2 - 2*x +2/3;
if (x >=2) and (x < 3)  then r:= x3/2  -4*x2 + 10*x -22/3;
if (x >=3) and (x < 4)  then r:= -x3/6 +x2*2 - 8*x +32/3;
if x >= 4 then r:=0;
//функция плотности распределения
pfr4:=r;

end;

function pfal(x,al:real):real;
Var r, t, x2,x3 : real;
   i,j : integer;
begin

pfal:= pfr4(x/al + 2)/al;

end;
```

```
Procedure tt1;
var i: integer;
begin

al:=1;
h:=0.5;

for i:=-8 to 8  do
begin
x:=i*h;
writeln(x:6:3, pfal(x, al): 8:4, pfr4(x/al + 2)/al :8:4);
end;
end;

Procedure tt2;
var i: integer;
begin
xi:=2.0;
uh:=0;
u2h:=0;
h:=0.01;
for i:=1 to 100 do
begin
x:=0.5 + (i-0.5)*h;
al := 0.5+0.5*x;
uh:=uh + pfal(xi, al);
end;

uh:=uh*h;

writeln(uh :8:4, pfr4(2+xi) :8:4);

for i:=1 to 200 do
begin
x:=0.0 + (i-0.5)*h;
al := 0.5+0.5*x;
u2h:=u2h + pfal(xi, al);
end;
u2h:=u2h*h/2;
writeln( u2h:8:4, pfr4(2+xi) :8:4);
u0:=(4*uh - u2h)/3.0;
writeln( u0:8:4,pfr4(2+xi) :8:4);
end;

function par(t : real):real;
begin
  par := 100*(4 - t)*t + 200;
end;
function alpha(t : real):real;
begin
  alpha := 0.5 +0.5*t;
end;
```

```
Procedure Reg01;
Var t,xt : real;
   i, ip,jp,jp0,jpr,jt,c : integer;
begin
h:=2.0/800;

 SetWindowSize(1000,1000);
 for i:=0 to 800 do
 begin
 t:= i*h;
 jp0:= Round(par(t));
 jpr:= Round(200*alpha(t));
  for jt:= 1 to 15 do
 begin
 jp:= Round(par(t)+ 100*r4*alpha(t));
 SetPixel(i+100,jp,RGB(0,0,0));
 end;

 SetPixel(i+100,jp0,RGB(255,0,0));
 SetPixel(i+100,jp0-jpr,RGB(255,0,0));
 SetPixel(i+100,jp0+jpr,RGB(255,0,0));
 SetPixel(i+100,jp0+1,RGB(255,0,0));
 SetPixel(i+100,jp0+1-jpr,RGB(255,0,0));
 SetPixel(i+100,jp0+1+jpr,RGB(255,0,0));

 end;

  MoveTo(50,50);
  Pen.Width := 1;
  setpencolor(clBlack);
  LineTo(50,950);
  LineTo(950, 950);
  LineTo(950, 50);
  LineTo(50, 50);
 Window.Save('forecasting.png');
end;
begin
Reg01;
end.
```

О Т З Ы В

руководителя о магистерской диссертации студента СФУ Васильева Николая Николаевича на тему «Численный вероятностный анализ для задач цифровой экономики», представленную к защите по направлению 09.04.02 — Информационные системы и технологии по программе 09.04.02.01 — Информационно-управляющие системы

Магистерская диссертация на тему: «Численный вероятностный анализ для задач цифровой экономики» выполнена по заданию кафедры «Системы искусственного интеллекта» и посвящена актуальной задаче разработки модуля арифметических операций для работы с дискретными случайными переменными на основе численного вероятностного анализа.

В диссертационной работе рассмотрены новые подходы к реализации прогнозов временных рядов больших данных.

В магистерской диссертации решены следующие задачи:

- изучение публикаций по данной тематике;
- проанализированы способы прогнозирования временных рядов;
- исследованы подходы численного вероятностного анализа для обработки данных большого объёма;
- разработан программный модуль прогнозирования временных рядов;
- для проверки работы модуля было проведено тестирование.

В процессе работы над диссертацией магистрант Васильев Н. Н. принял участие в ряде научно–практических конференций. По результатам исследований опубликована статья.

Материалы работы, можно использовать в научно-исследовательской работе при разработке новых подходов прогнозирования временных рядов данных большого объёма.

Магистерская диссертация оформлена в соответствии с требованиями нормативных документов СФУ.

Выпускная квалификационная работа магистранта удовлетворяет требованиям, предъявляемым к магистерским работам СФУ, и может быть оценена на «отлично», а её автор Васильев Николай Николаевич присуждения квалификации магистр по направлению — Информационные системы и технологии.

Руководитель Б.С. Добронец

профессор, доктор физ.-мат. наук,
профессор кафедры СИИ ИКИТ СФУ

М. П. «30» июня 2019 г.

Федеральное государственное автономное
образовательное учреждение
высшего образования
"СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ"
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой

_____Г.М.Цибульский

«____» _____ 20___г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Численный вероятностный анализ для задач цифровой экономики

09.04.02 Информационные системы и технологии
09.04.02.01 Информационно-управляющие системы

Руководитель	_____ подпись _____ дата	проф., д-р физ.-мат. наук *должность, ученая степень*	Б.С.Добронец
Выпускник	_____ подпись _____ дата	КИ17-02-1М	Н.Н.Васильев
Рецензент	_____ подпись _____ дата	проф., д-р техн. наук *должность, ученая степень*	Л.А.Казаковцев
Нормоконтролер	_____ подпись _____ дата	проф., д-р физ.-мат. наук *должность, ученая степень*	Б.С.Добронец

Красноярск 2019

Продолжение титульного листа магистерской диссертации по теме «Численный вероятностный анализ для задач цифровой экономики».

Нормоконтролер  Б. С. Добронец

РЕЦЕНЗИЯ

на магистерскую диссертацию студента СФУ Васильева Николая Николаевича на тему «Численный вероятностный анализ для задач цифровой экономики», представленную к защите по направлению 09.04.02 Информационные системы и технологии по программе 09.04.02.01 Информационно-управляющие системы

Актуальность обусловлена необходимостью повышения качества прогноза временных рядов для задач цифровой экономики.

Новизна работы заключается в использовании вычислительного вероятностного анализа для повышения точности оценки плотности вероятности в задачах прогноза временных рядов для задач цифровой экономики.

Структура диссертационной работы. Работа состоит из введения, трех глав, заключения.

Результатом данной магистерской диссертации является разработка модуля прогноза временных рядов для задач цифровой экономики.

Апробация. По результатам диссертации опубликована одна печатная работа.

Замечания и предложения по диссертации. Следовало более подробно описать алгоритм построения сплайновой оценки функции плотности вероятности.

Заключение. Изложенный в работе теоретический, графический и демонстрационный материал оформлен достаточно полно, в соответствии с требованиями нормативных документов СФУ. Работа выполнена в полном объеме и на достаточном уровне, в соответствии с поставленной целью. Васильев Николай Николаевич заслуживает присвоения квалификации «магистр» по направлению 09.04.02 Информационные системы и технологии.

Оценка — отлично.

Рецензент
зав. кафедрой Системного анализа
и исследования операций,
ФГБОУ ВО «Сибирский государственный университет науки и технологий
имени академика М.Ф.Решетнева»,
д-р техн. наук, доцент

Лев
Александрович
Казаковцев

«30» июня 2019 г.

Подпись Казаковцева Л.А. удостоверяю
Нач. спец. по нефа.
Начальник УК СибГАУ
г.Красноярск

УПРАВЛЕНИЕ КАДРОВ