

Федеральное государственное учреждение
«Федеральный исследовательский центр
Институт прикладной математики им. М.В. Келдыша
Российской академии наук»

УДК 519.242.33

На правах рукописи

Кислицын Алексей Алексеевич

**МОДЕЛИРОВАНИЕ ИНДИКАТОРОВ РАЗЛАДКИ
В НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДАХ
ЭЛЕКТРОЭНЦЕФАЛОГРАММ**

Диссертация на соискание ученой степени
кандидата физико-математических наук

Специальность

05.13.18 – математическое моделирование, численные методы
и комплексы программ

Научный руководитель:
д.ф.-м.н. Ю.Н. Орлов

Москва, 2021

ОГЛАВЛЕНИЕ

Общие сведения о работе.....	3
Введение.....	11
Глава I. Согласованный уровень стационарности как индикатор разладки.....	16
1.1. Определение согласованного уровня значимости	16
1.2. Примеры применения индикатора СУС.....	22
1.3. Компьютерные программы для моделирования индикаторов разладки	27
1.4. Задача разработки программного комплекса для построения СУС.....	28
Глава II. Эквивалентность по Чернову в задачах математической статистики	30
2.1. Эквивалентные по Чернову операторные функции и полугруппы	30
2.2. Применение эквивалентности по Чернову к уравнению Лиувилля для выборочных распределений	33
2.3. Эквивалентный по Чернову согласованный уровень значимости.....	47
Глава III. Численный алгоритм моделирования разладки в нестационарных временных рядах и программный комплекс.....	51
3.1. Алгоритм построения СУС для анализа нестационарности	51
3.2. Блок-схема программного комплекса	54
3.3. Распараллеливание вычислений	57
3.4. Программирование пользовательского интерфейса	59
Глава IV. СУС как предиктор приступа эпилепсии при анализе ЭЭГ	69
4.1. Гипотеза разладки распределения СУС как предиктора приступа эпилепсии	69
4.2. Статистический анализ рядов данных по отведениям ЭЭГ	72
4.3. Стационарные паттерны функций распределения СУС.....	82
4.4. Верификация индикатора разладки	83
Заключение.....	87
Список литературы.....	89

Общие сведения о работе

Актуальность темы исследования.

Задача разработки высокопроизводительных численных алгоритмов и программных комплексов для обработки больших массивов нестационарных данных, в настоящее время представляет собой все более актуальное направление исследований. В различных сферах расширяется доля использования онлайн вычислений, связанных как с расширением возможностей современных компьютеров, так и с углублением знаний о структуре анализируемых процессов, что требует новых подходов к статистическому анализу неоднородных или плохо структурированных данных. Одной из актуальных задач является проблема статистического распознавания разладки в нестационарных рядах. Для стационарных случайных процессов эта задача имеет решение в виде классических критериев – либо параметрических, когда оцениваются доверительные интервалы принадлежности параметра, характеризующего класс распределений наблюдаемой величины, либо непараметрических, основанных на близости выборочных функций распределения в различных нормах.

Анализ нестационарных процессов может быть проведен только при использовании высокопроизводительных вычислений, что связано с развитием так называемых систем с интенсивным программным обеспечением. Согласно Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года такими системами являются аппаратно-программные комплексы с большим удельным весом программной части. Тема настоящего исследования напрямую связана с обработкой данных, получаемых такими комплексами.

Существует много примеров временных рядов, в которых одним из инструментов анализа является модель разладки с учетом нестационарных свойств ряда. Это биржевые ряды цен сделок на финансовые инструменты, а также ряды данных о доходности соответствующих торговых операций, совершенных в рамках определенных стратегий (алгоритмов), где потоки событий составляют до нескольких десятков в секунду по отдельному инструменту. Другим примером являются потоки событий, связанные с анализом трафика беспроводной связи, достигающие десятков тысяч событий в секунду. Менее мощные, но тоже достаточно большие потоки данных требуется обрабатывать при анализе различной биометрической информации о состоянии живых организмов, а также телеметрической информации о состоянии технических систем. Анализ данных электроэнцефалограмм в медицине, статистика показаний сейсмограмм и различных счетчиков радиоактивности, последовательности символов в текстах различной природы

представляют типичный набор задач, в которых требуется отнести результат наблюдения к определенным классам «норма» или «авария». Практически важной задачей является определение промежуточного состояния, которое, собственно, и представляет собой разладку «нормы», чтобы успеть принять меры по недопущению состояния «авария». Модель разладки должна при этом давать вероятностную картину перехода из одного состояния в другое, причем ошибка второго рода (пропуск цели) должна быть минимальна при заданном уровне ошибки первого рода (ложная тревога).

Для стационарных рядов, функции распределения которых известны, вероятности ошибок при принятии или отклонении статистических гипотез могут быть вычислены теоретически. Однако если ряд нестационарный и его выборочные распределения не принадлежат определенному классу, оценить ошибку затруднительно. Тем не менее, решения в таких случаях принимаются на основе классических критериев, только уровень значимости постфактум оказывается хуже, чем утверждается критерием. Следовательно, необходимо скорректировать уровень значимости с учетом нестационарности изучаемой системы.

В настоящей работе предложена модель непараметрического индикатора разладки для рядов, которые не являются стационарными. Выяснилось, что в ряде практически важных случаев распределение этого индикатора оказалось стационарным. Это позволило корректно указать вероятность ошибки первого рода. Разрабатываемый алгоритм и программный комплекс для вычисления этого индикатора направлены на анализ больших потоков данных, что, согласно упомянутой выше Стратегии развития отрасли информационных технологий РФ является в настоящее время актуальным направлением развития практически всех отраслей науки и промышленности.

Степень разработанности научной проблемы.

Для нестационарных временных рядов общего подхода к решению задачи о разладке в настоящее время не существует. Есть результаты, полученные для частных случаев, когда ряды оказываются коинтегрированными. Тогда разладку можно искать с помощью классических стационарных критериев.

Для рядов, которые не сводятся с помощью регрессионного анализа к стационарным, на практике применяются недостаточно обоснованные адаптивные методы. Эти методы используют определенные стационарные алгоритмы, но модельные коэффициенты могут быть пересчитаны при необходимости, то есть они являются некоторыми функциями времени, но зависимость эта не формализована. Недостатком такого подхода для распознавания разладки является то, что ее можно спутать с

изменением параметров модели, тогда как собственно разладкой в этом случае следует считать изменение вида самой модели.

Альтернативным подходом является применение кинетического метода, с помощью которого анализируется эволюция выборочных плотностей функции распределения. Этот подход к анализу нестационарных временных рядов развивается в настоящее время в ИПМ им. М.В. Келдыша РАН. Были предложены новые индикаторы нестационарности, такие как согласованный уровень стационарности и индекс нестационарности. Эти индикаторы использовались главным образом для классификации типов временных рядов, но для анализа разладки они не применялись.

Таким образом, ни на уровне математической модели, ни в виде программного обеспечения индикаторы нестационарности ранее не рассматривались как предикторы разладки.

Цель и задачи диссертационного исследования.

Цель работы заключается в создании программного комплекса для обнаружения разладки в нестационарных временных рядах в реальном времени.

Научно-теоретическая задача состоит в разработке математической модели непараметрического индикатора – предиктора разладки в нестационарном временном ряде (п. 1 паспорта специальности 05.13.18, разработка новых математических методов моделирования объектов и явлений).

В силу большой вычислительной сложности статистической процедуры следует также разработать приближенный метод анализа изучаемых статистик (п. 2 паспорта специальности 05.13.18, развитие качественных и приближенных аналитических методов исследования математических моделей).

Практическая задача заключается в создании численного алгоритма вычисления и тестирования индикатора в условиях многомерности потока событий (п. 4 паспорта специальности 05.13.18, реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента).

Объект и предмет исследования.

Объектом исследования является специальная статистика, получаемая по непересекающимся выборкам из нестационарных временных рядов, называемая согласованным уровнем стационарности.

Предмет исследования – свойства выборочных распределений этой статистики с целью выяснения возможности ее использования в качестве предиктора разладки в нестационарных временных рядах. В качестве практического примера ряда, на котором

тестируется соответствующий численный алгоритм, взят многомерный ряд записей электроэнцефалограмм по совокупности отведений для пациентов с приступами эпилепсии.

Теоретическая и методическая основа исследования.

В работе использовались кинетические уравнения, в частности, уравнение Лиувилля и уравнения цепочки Боголюбова для выборочных плотностей функций распределения, методы эквивалентных по Чернову динамических потоков и полугрупп, численные методы решения дифференциальных уравнений в частных производных.

Основу теоретического анализа составляют методы исследования асимптотических свойств стационарной точки уровня значимости, на котором принимается гипотеза об однородности двух выборочных распределений. Исходной статистикой для этого анализа является выборочная функция распределения расстояний между выборочными функциями распределения встык-выборок изучаемого временного ряда.

Информационная база исследования.

Информационная база исследования состоит из учебной литературы по теории вероятностей и математической статистике, а также статей в научных журналах и трудов международных конференций, относящихся к задаче моделирования разладки.

Обоснованность и достоверность результатов исследования.

Достоверность и обоснованность результатов подтверждается использованием строгих математических доказательств и рассуждений и апробированных в научной практике методов численного анализа.

Соответствие диссертации паспорту научной специальности.

Согласно вышеописанным основным целям и задачам данной диссертационной работы, она направлена на разработку математической модели явления разладки, аппроксимирующего ее численного алгоритма и создания соответствующего программного комплекса. Тем самым диссертация соответствует пунктам 1, 2 и 4 паспорта специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ.

Научная новизна результатов исследования.

Новизна работы заключается в том, что впервые теоретический метод построения эквивалентной по Чернову полугруппы был применен в математической статистике для обоснования алгоритма приближенного вычисления индикатора разладки. Также впервые был проведен статистический анализ выборочных функций распределения этого индикатора, что являлось в вычислительном плане серьезным техническим затруднением, связанным с увеличением размерности задачи. Практический результат, полученный при

тестировании предложенного метода на примерах электроэнцефалограмм, состоит в улучшении распознавания приближения приступа эпилепсии. Этот результат получен с использованием нового критерия разладки, в качестве которого предложено использовать статистику, называемую согласованным уровнем стационарности ряда.

Теоретическая ценность и практическая значимость работы.

Теоретическая и практическая ценность работы состоит в том, что в ней построен алгоритм вычисления многомерного индикатора разладки в нестационарном временном ряде, доказаны теоремы о свойствах его аппроксимации, позволяющие значительно сократить время счета, и построен программный комплекс с интерфейсом, реализующий предложенную методику.

Практическая значимость работы подтверждается также применением полученных результатов в проведении научно-исследовательских работ по тематике, утвержденной в ИПМ им. М.В. Келдыша РАН. Результаты исследований были использованы при выполнении проектов ИПМ им. М.В. Келдыша РАН, поддержанных грантами РФФИ (№ 14-21-00025, № 19-71-30004), в совместном проекте ИПМ им. М.В. Келдыша РАН и НМИЦ нейрохирургии им. ак. Н.Н. Бурденко (соглашение № 73 от 05.12.2018), а также в работе по Крупному научному проекту с ННГУ им. Н.И. Лобачевского «Надежный и логически прозрачный искусственный интеллект: технология, верификация и применение при социально-значимых и инфекционных заболеваниях» (договор № 111-20-ЕП от 16.11.2020). Практическая задача, решаемая в рамках перечисленных проектов, относилась к сфере анализа больших данных и была связана с построением предиктора разладки применительно к записям электроэнцефалограмм для пациентов с приступами эпилепсии.

Апробация работы.

Материалы диссертации докладывались на научных семинарах отдела вычислительной физики и кинетических уравнений ИПМ им. М.В. Келдыша РАН, на кафедре прикладной информатики и теории вероятностей в Институте прикладной математики и телекоммуникаций РУДН, а также на конференциях:

1. Gennady Bocharov, Alexey Kislitsyn, Rostislav Savinkov, Mario Novkovic, Lucas Onder. Modelling the FRC network of lymph node // International Congress on Systems Immunology, Immunoinformatics & Immune-computation, Taormina, Italy, July 17-18, 2015.
2. Kislitsyn A.A, Orlov Yu.N. On the Distribution of the Stationary Point of Significance Level for Empirical Distribution Function // International Congress on Ultra Modern Telecommunications and Control Systems ICUMT, Moscow, Russia, November 5-8, 2018.

3. Kislitsyn A.A, Orlov Yu.N. Nonstationary stochastic motion modeling by dynamical systems // ECMS Caserta, Italy, June 11-14, 2019.

Публикации.

По материалам диссертации опубликовано 13 работ в рецензируемых научных изданиях, входящих в перечень ВАК, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ. Эти работы включают 4 препринта ИПМ им. М.В. Келдыша РАН, а также 9 статей, входящих в базы данных WoS и SCOPUS, из которых 4 статьи опубликованы в отечественных, а 5 – в иностранных журналах. Имеются 1 патент на изобретение и 1 свидетельство государственной регистрации на программу для ЭВМ.

Личный вклад автора.

Автор диссертации, работая в коллективе соавторов, самостоятельно сформулировал теоретические утверждения, относящиеся к модели разрядки, разработал численный алгоритм и провел серию вычислительных экспериментов. Математическая постановка задач принадлежит научному руководителю. Физическая постановка задачи о предсказании приступа эпилепсии и экспериментальные временные ряды ЭЭГ принадлежат сотрудникам НМИЦ им. акад. Н.Н. Бурденко Козловой А.Б., Корсаковой М.Б. и Машерову Е.Л. В совместных публикациях лично соискателем получены следующие результаты.

В работах [12, 15, 53] соискателем проведены численные расчеты согласованного уровня стационарности и дана оценка точности предиктора.

В работах [33, 54, 69] соискателем проведены работы по созданию трехмерной визуализации результатов расчетов статистических характеристик применительно к задачам математического моделирования в медицине.

В работах [16, 17, 57, 58] соискателю принадлежат доказательства теорем, относящихся к аппроксимации решения уравнения Лиувилля и к аппроксимации стационарной точки эмпирического уровня значимости распределений изучаемых статистик.

В работах [55, 56] соискателем разработан численный алгоритм и построена модель аппроксимации решения уравнения Лиувилля.

Вклад соискателя в патент [34] состоит в проведении верификационных статистических исследований

Вклад соискателя в программный комплекс [18] состоит в разработке архитектуры комплекса, программировании вычислительного алгоритма и интерфейса пользователя.

Основные положения, выносимые на защиту.

1. Разработана математическая модель индикатора разладки нестационарного временного ряда на основе статистики согласованного уровня стационарности в зависимости от двух параметров – длины фрагмента ряда и длины окна встык-выборки (п. 1 паспорта специальности 05.13.18, разработка новых математических методов моделирования объектов и явлений).
2. Разработан метод приближенного вычисления значений индикатора разладки на основе использования формул Фейнмана-Чернова конечнократной аппроксимации полугрупп (п. 2 паспорта специальности 05.13.18, развитие качественных и приближенных аналитических методов исследования математических моделей).
3. Построен численный алгоритм для определения многомерного индикатора разладки в реальном времени и программный комплекс с пользовательским интерфейсом (п. 4 паспорта специальности 05.13.18, реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента).
4. Проведены численные эксперименты по индикации разладки применительно к рядам электроэнцефалограмм для пациентов с приступами эпилепсии и показано, что построенный индикатор имеет стационарное распределение, а его применение характеризуется ошибкой первого рода, равной значению индикатора разладки, и нулевой эмпирической ошибкой второго рода (п. 4 паспорта специальности 05.13.18, реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента).

Структура и объем диссертации.

Диссертация «Моделирование индикаторов разладки в нестационарных рядах электроэнцефалограмм» состоит из введения, четырех глав и заключения. Каждая глава разбита на параграфы, имеющие двойную нумерацию, первая цифра которой указывает на соответствующую главу. Формулы внутри каждого параграфа имеют тройную нумерацию, с указанием на главу и параграф. Рисунки и таблицы имеют двойную нумерацию с указанием на номер главы.

Результаты исследования изложены на 94 страницах и содержат 40 рисунков и 1 таблицу. Библиографический список состоит из 77 наименований, расположенных в алфавитном порядке.

Во введении проводится обзор современных направлений исследований в области построения статистических индикаторов разладки для нестационарных временных рядов. В результате обзора обосновываются и формулируются задачи, решению которых посвящена диссертация.

В первой главе приводятся основные сведения о статистике так называемого согласованного уровня стационарности, даются примеры применения этого индикатора и формулируется концепция индикации разладки на основе анализа его локального изменения. Также в этой главе формулируются задачи по разработке программного комплекса для реализации предложенной идеи.

Во второй главе даются необходимые сведения из теории конечнократных аппроксимаций полугрупп на основе теоремы Чернова. Далее в ней излагаются утверждения по применению метода построения эквивалентных по Чернову полугрупп для аппроксимации решения уравнения Лиувилля относительно выборочных плотностей функций распределения и для аппроксимации стационарной точки уровня значимости выборочной функции распределения расстояний между выборочными функциями распределения анализируемого ряда. Поскольку аппроксимация по Чернову требует значительно меньше вычислительных действий, чем расчет точного значения индикатора в окне произвольной длины и по встык-выборке произвольной длины, то результаты этой главы служат теоретическим обоснованием для построения экономного алгоритма, позволяющего вычислять разладку в скользящем окне в реальном времени.

В третьей главе приводятся алгоритмы решения поставленных задач, которые объединяются в единый программный комплекс. Дается общая блок-схема программного комплекса и описываются его возможности. Также описывается интерфейс пользователя и приводится инструкция с описанием структуры рабочих окон программы.

В четвертой главе приводятся результаты численного эксперимента по индикации разладки в виде отклонения согласованного уровня стационарности от своего ожидаемого значения для прогноза наступления приступа эпилепсии в многомерных временных рядах, полученных из записей электроэнцефалограмм по совокупности отведений.

В заключении подытоживаются основные результаты диссертации и обсуждаются возможные области их применения, указываются ограничения построенной модели и возможности ее совершенствования.

Введение

Распознавание разладки применительно к функционированию некоторой физической или информационной системы представляет собой важную задачу прикладной математической статистики.

Разладкой называется отклонение контролируемых параметров анализируемой системы от нормативных пределов. Эти параметры могут быть собственно теми, которые обеспечивают безопасность эксплуатации изделия, как, например, напряжение в электрической сети, или измерения могут быть косвенными в виде средств неразрушающего контроля, если прямые по ряду причин проведены быть не могут. Таковы, например, методы анализа вероятности физической поломки изделия на основе оценки изменения его предела упругости, когда само изделие не должно ломаться в процессе измерения, поскольку оно встроено в эксплуатационную цепь.

В статистическом смысле классической разладкой называется изменение одного стационарного распределения случайной величины на другое (тоже стационарное), произошедшее в случайный момент времени. В этом случае задачей математической статистики является обнаружение соответствующего момента с наименьшей погрешностью на основе анализа наблюдаемых данных.

Одним из основных методов так называемого однократного обнаружения разладки является метод кумулятивных сумм (см. [2, 10, 37, 63]). Он основывается на методе максимального правдоподобия и состоит в следующем.

Пусть анализируется некоторая функция $f(x)$ по измеряемому множеству данных $f_1, \dots, f_N$. Нулевая гипотеза H_0 отвечает представлению о том, что вся совокупность данных $\{f_m\}_1^N$ соответствует определенной модели A_0 . Альтернативная гипотеза H_1 состоит в том, что существует некоторый номер $k \in (1, N)$ такой, что совокупность данных $\{f_m\}_1^k$ соответствует модели A_0 , а совокупность данных $\{f_m\}_{k+1}^N$ соответствует другой модели A_1 . Модели $A_{0,1}$ в этом подходе параметрические, они отвечают гипотезе об определенном распределении значений f_m , допустим, гауссовом, с заданными параметрами среднего и дисперсии. Допустим, тестируется гипотеза об изменении среднего при постоянной дисперсии. Выбирается индекс $k \in (1, N)$, для которого рассматривается величина скачка в среднем значении, оцениваемом по указанным двум

под-выборкам. Пусть $\Delta = \mu_0 - \mu_1$ есть величина скачка в среднем значении, а σ^2 есть дисперсия наблюдаемых данных. После этого рассматривается сумма

$$B(k) = \frac{\Delta}{\sigma^2} \sum_{n=k}^N \left(f_n - \mu_0 - \frac{\Delta}{2} \right),$$

для которой вычисляется $k^* = \arg \max_k B(k)$. Решающая статистика строится посредством задания критического уровня β , так что если $\max B(k) > \beta$, то справедлива гипотеза H_1 , а если $\max B(k) \leq \beta$, то справедлива гипотеза H_0 .

Отметим недостатки данного подхода. Во-первых, задача в данной постановке параметрическая, для чего существенно знание функционального вида распределения значений наблюдаемой величины. Во-вторых, эта задача решается полным перебором по индексу k , что представляет известное неудобство, особенно для выборок больших объемов. В-третьих, процесс вне точки разладки предполагается стационарным.

Некоторой модификацией такого метода для нестационарного ряда данных является адаптивный двухуровневый подход к обнаружению разладки. Например, в ряде алгоритмов для торговых систем на финансовых рынках условие купли или продажи принимается после так называемого «пробития» траекторией цены некоторого коридора вокруг выборочного среднего, т.е. при наступлении разладки первого уровня. Средняя выборочная цена и ширина коридора являются параметрами задачи. Однако они зависят от длины выборки. Предположим, что существует некоторая модель по определению оптимальной в данных условиях длины выборки. Эта модель действует на промежутке времени, значительно превышающем длину выборки для идентификации разладки, поэтому оптимальная длина окна для идентификации разладки меняется не быстро. Тем не менее, если в результате текущего анализа выясняется, что окно следует поменять, то это означает, что наступила разладка второго уровня, и параметры разладки первого уровня пересчитываются. Однако такие эвристические подходы не имеют строгого математического обоснования.

Среди других стандартных подходов к анализу разладки в стационарных процессах отметим также байесовские подходы [39, 50] и процедуры на основе методов разложения многомерных данных, такие как анализ главных компонент [45, 60, 61, 62] и анализ сингулярного спектра [51, 75, 76, 77]. Однако, как отмечается в работах [45, 47, 48], использование традиционных методов и процедур для обнаружения разладок в системах контроля с интенсивным программным обеспечением не эффективно по причине нестационарности потоков данных.

В данной диссертации представлена методика идентификации разладки в многомерных нестационарных временных рядах на примере записей электроэнцефалограмм по совокупности отведений. Работа посвящена построению индикатора разладки как предиктора изменения состояния системы. Применительно к практической теме исследования разладкой считается приступ эпилепсии. Инструментом анализа является статистика расстояний между выборочными функциями распределения. В качестве индикатора рассматривается стационарная точка уровня значимости распределения расстояний между выборочными распределениями.

Тема диссертации лежит в направлении исследований по нестационарным случайным процессам, проводимых в отделе кинетических уравнений и вычислительной физики ИПМ им. М.В. Келдыша РАН применительно к различным областям практической деятельности. Метод кинетических уравнений применительно к задачам математической статистики состоит в том, что изменение функции распределения или ее плотности трактуется как результат усреднения большого числа детерминированных факторов. При этом для функций распределения оказывается возможным выписать уравнение, в рамках которого моделируется их эволюция. Построение такого модельного уравнения на основе обработки наблюдаемых данных и нахождение его численного решения на некоторый прогнозный горизонт являются центральными задачами анализа нестационарных временных рядов в рамках этого подхода. Основное отличие этого метода от других состоит в том, что изучается не одна наблюдаемая в эксперименте траектория в виде последовательно регистрируемых значений случайной величины, как это имеет место, например, в моделях регрессионного типа, а строится ансамбль возможных траекторий, выборочные распределения которых имеют те же статистические характеристики, что и наблюдаемые в эксперименте.

Кинетический подход позволяет разработать методику для обнаружения разладки в нестационарных процессах с помощью различных эконометрических и экономико-математических моделей, применяемых при изучении и оптимизации задач маркетинга и менеджмента, управления предприятием, стабильности технологических процессов, в задачах надежности и обеспечения технологической и экологической безопасности, при описании функционирования технических устройств и объектов, разработки организационных схем и т.д.

В настоящее время существует более тысячи статистических тестов или критериев, которые применяются для того, чтобы с некоторой точностью отнести изучаемый случайный процесс к тому или иному классу, т.е. использовать для его описания определенную математическую модель. Доказательные результаты относятся к

стационарным процессам, что позволяет (если процесс действительно таков) по одной выборке корректно оценить вероятность того или иного значения функционала от генеральной совокупности. Однако во многих актуальных практических задачах при большом числе наблюдений за случайным процессом, осуществляемым в скользящем окне, обнаруживается, что если процесс не является стационарным, то число ошибок в принятии той или иной статистической гипотезы оказывается в разы больше, чем уровень значимости, на котором по классическому критерию принималось решение. Тем самым возникает настоятельная потребность снижения ошибки прогнозирования и разработки метода, позволяющего более точно определять уровень доверия.

В прикладных задачах часто используется критерий согласия Колмогорова [19, 59] для определения близости выборочной функции распределения случайной величины ξ к распределению генеральной совокупности. Именно, статистика $D_n = \sup_x |F_n(x) - F(x)|$ супремума модуля разности выборочной и точной интегральных функций распределения стремится по вероятности к нулю с ростом объема выборки n так, что случайная величина $\sqrt{n}D_n$ имеет асимптотическое распределение в виде табулированной K -функции Колмогорова. На основе этого утверждения были получены различные широко применяемые асимптотические критерии [19] о принадлежности двух выборочных распределений одной генеральной совокупности: критерий Колмогорова-Смирнова, Вальда-Волфовица, Вилкоксона, Манна-Уитни, Гнеденко-Королюка и другие критерии, применяемые к оценкам выборочных моментов (Стьюдента, Фишера, Крамера-Уэлча, «омега-квадрат» и др.). Все эти критерии являются стационарными, а их модификация для нестационарного случая представляется затруднительной.

В нестационарном случае все асимптотические критерии, гарантирующие увеличение точности аппроксимации с увеличением объема выборки, не состоятельны. Аналогичные проблемы возникают и при использовании сглаженного скользящего усреднения. Если ряд нестационарный, то средние (скользящие, «растущие» - т.е. взятые по выборке растущего объема, или любые другие) не являются состоятельными оценками моментов распределения, так как сходимости по вероятности в общем случае нет.

Если в стационарном случае есть доказательная уверенность в асимптотической состоятельности оценок той или иной статистики, то в нестационарном случае отсутствует само понятие генеральной совокупности, что делает неприменимым весь развитый аппарат современной математической статистики, кроме тех случаев, когда априори известна функциональная принадлежность модели процесса. На практике же почти всегда не известно, к какому классу принадлежит распределение.

Следует подчеркнуть, что, как показывает анализ разладок в нестационарных потоках данных [45, 48], классические статистические критерии на практике имеют достаточно ограниченную область удовлетворительного применения. Следовательно, необходимо разработать инструментарий для адекватного анализа нестационарных распределений в скользящем окне наблюдения произвольной длины.

Применение статистических методов в практических исследованиях с помощью специально разработанного программного обеспечения рассматривалось во многих работах, посвященных оптимизации вычислительных алгоритмов (см. [1, 18, 24]). Необходимость включения в стандартные пакеты новых алгоритмов, позволяющих повысить точность статистических оценок при проверке вероятностных гипотез, отмечалась в работах [1, 14, 67].

Таким образом, обзор литературы в области прикладной математической статистики и моделирования процессов разладки показал, что существует проблема создания индикаторов, пригодных именно для нестационарных временных рядов. В настоящей работе строится индикатор – предиктор разладки достаточно общего вида, т.е. он может быть применен к процессам различной физической или экономической природы. Нетривиальным фактом является обнаруженная эмпирически стационарность распределения значений этого индикатора, в частности, для рядов электроэнцефалограмм, которые представляют практический пример применения разработанного метода.

Глава I. Согласованный уровень стационарности как индикатор разладки

1.1. Определение согласованного уровня значимости

Разладкой называется отклонение контролируемых параметров системы (изделия, программы, упорядоченной совокупности признаков и т.п.) от нормативных пределов. В статистическом смысле классической разладкой называется изменение одного стационарного распределения случайной величины на другое (тоже стационарное), произошедшее в случайный момент времени. Одной из задач математической статистики является обнаружение этого момента с наименьшей погрешностью на основе анализа наблюдаемых данных.

В настоящей диссертации рассматриваются практические аспекты обнаружения разладки в ее статистическом понимании – как изменения функции распределения случайной величины. Поскольку на практике анализируется не функция распределения в ее теоретико-вероятностном аспекте, а распределение выборки, т.е. выборочная функция распределения (далее ВФР), то задача о разладке сводится к определению того, принадлежат ли две наблюдаемые выборки одной и той же генеральной совокупности.

В стационарном случае задача о принадлежности двух выборок длины N одной генеральной совокупности решается непараметрической статистикой Колмогорова-Смирнова:

$$S_N = \sup_x |F_{1,N}(x) - F_{2,N}(x)|. \quad (1.1.1)$$

Для статистики (1.1.1) имеет место асимптотика

$$\lim_{N \rightarrow \infty} P \left\{ 0 < \sqrt{\frac{N}{2}} S_N < z \right\} = K(z), \quad (1.1.2)$$

где $K(z)$ есть табулированная функция Колмогорова [9, 21, 59].

На практике формула (1.1.2) применяется следующим образом. Если в процессе сравнения двух ВФР было найдено значение S_N в соответствии с (1.1.1) и вычислена

величина $z = \sqrt{\frac{N}{2}} S_N$, то величина $1 - K(z)$ приближенно считается равной вероятности

того, что $\sqrt{\frac{N}{2}} S_N \geq z$. Задав требуемый уровень значимости α , на котором принимается

статистическая гипотеза, считаем, что если $1 - K(z) < \alpha$, то осуществилось событие,

несовместимое с понятием случайности, и такие выборки следует считать различными. Если же $1 - K(z) \geq \alpha$, то на уровне значимости α считаем, что эти две выборки взяты из одной генеральной совокупности.

Более точно, задавая уровень значимости α для функции распределения статистики (1.1.1), требуется вычислить соответствующий α -квантиль, зависящий от длины выборки N . Обозначим его $\varepsilon_N(\alpha)$. Эта величина имеет следующее асимптотическое представление [5]:

$$\varepsilon_n(\alpha) = \sqrt{\frac{z}{2n}} - \frac{1}{6n} + o(1/n), \quad (1.1.3)$$

где z есть корень уравнения

$$1 - K\left(\sqrt{\frac{z}{2}}\right) = \alpha. \quad (1.1.4)$$

Однако имеется некоторая неопределенность такого статистического вывода. Она состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Но для этого надо знать, как часто две независимых выборки длины N отличаются в смысле ВФР в норме C больше, чем на некоторое число ε такое, что

$$1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) < \alpha. \quad (1.1.5)$$

Здесь уместно привести результаты следующего статистического эксперимента. Сгенерируем 1000 выборок длины 100 и 1000 из равномерного распределения на единичном отрезке. Затем вычислим попарные расстояния (1.1.1) между выборками для длин 100 и 1000 соответственно. Результаты показаны на Рис. 1.1 (а, б). Выяснилось, например, что все выборки длины 100 различаются между собой в норме C на величину, большую, чем 0,04 (разумеется, это касается только данных сгенерированных выборок). Этот результат означает, что если мы потребуем принятия решения на таком уровне значимости, для выполнения которого расстояния между выборками должны быть меньше 0,04 (этот уровень, впрочем, равен 10^{-6}), то в 100 % будет совершена ошибка отклонения верной гипотезы, что, очевидно, не будет правильным статистическим выводом с точки зрения практических нужд, хотя и совершенно верным теоретически. Следовательно, уровень значимости не может быть задан произвольно, если рассматриваются выборки определенной длины. Ведь по своему смыслу это есть число, равное доле принятых ошибочных решений в результате очень большого (в пределе бесконечного) количества статистических экспериментов. Таким экспериментом в данном случае является сравнение двух выборок. Возникает вопрос: на каком уровне z надо

провести «отсечку» в распределении расстояний между этими выборками, чтобы доля ошибочных отклонений верной гипотезы (о том, что выборки взяты из одного распределения), полученная в эксперименте, совпала бы с соответствующим квантилем распределения функции Колмогорова?

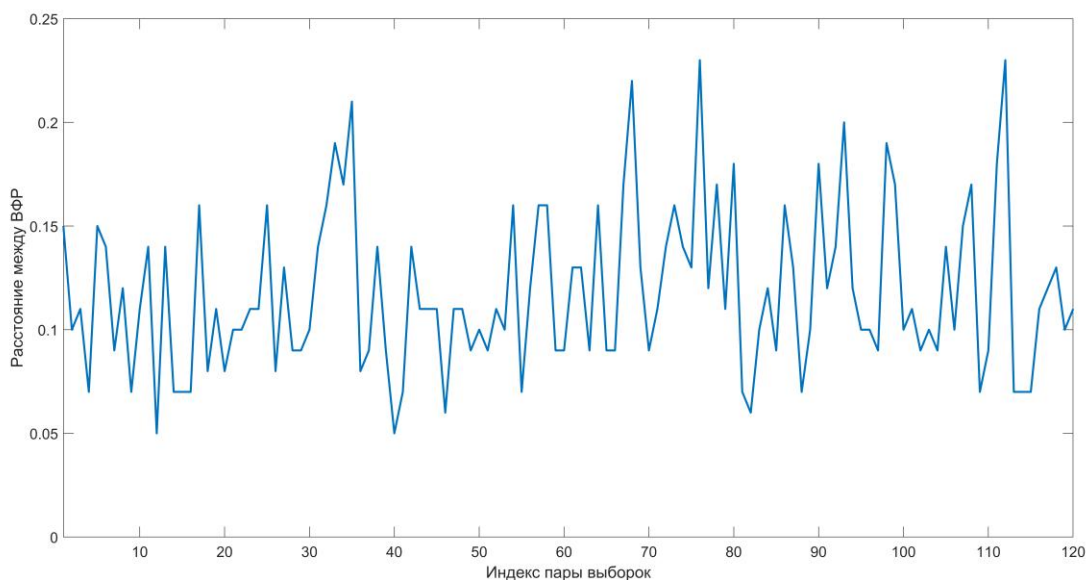


Рис. 1.1 (а) – Фрагмент временного ряда расстояний между независимыми ВФР из равномерного распределения, длина выборки 100

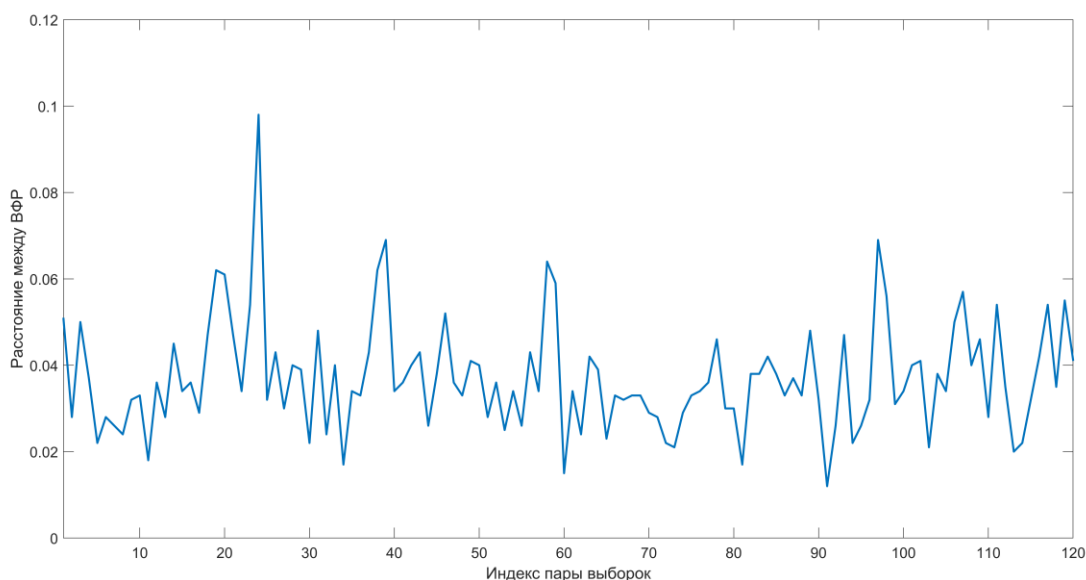


Рис. 1.1 (б) – Фрагмент временного ряда расстояний между независимыми ВФР из равномерного распределения, длина выборки 1000

Сравнивая между собой типичные расстояния между выборками для длин 100 и 1000 на Рис. 1.1 (а, б), видим, что они заметно различаются, поэтому подходить к этим двум статистикам с одним и тем же уровнем значимости (1.1.4) было бы неверно. Если

уровень значимости α в (1.1.4) выбран слишком малым, то может оказаться, что доля пар выборок, расстояние между которыми удовлетворяет этому условию, будет меньше, чем α , и тогда при отклонении верной гипотезы ошибка идентификации будет больше, чем α , тогда как критерий утверждает, что эта ошибка равна именно α . И наоборот, если уровень α выбран слишком большим, то на практике может оказаться, что число пар выборок, для которых условие (1.1.4) не выполнено, будет больше, чем $1 - \alpha$, т.е. больше уровня доверия, на котором предположительно выполнен эксперимент по сравнению с выбором.

На Рис. 1.2 приведена плотность выборочного распределения расстояний между сгенерированными выборками длины 100.

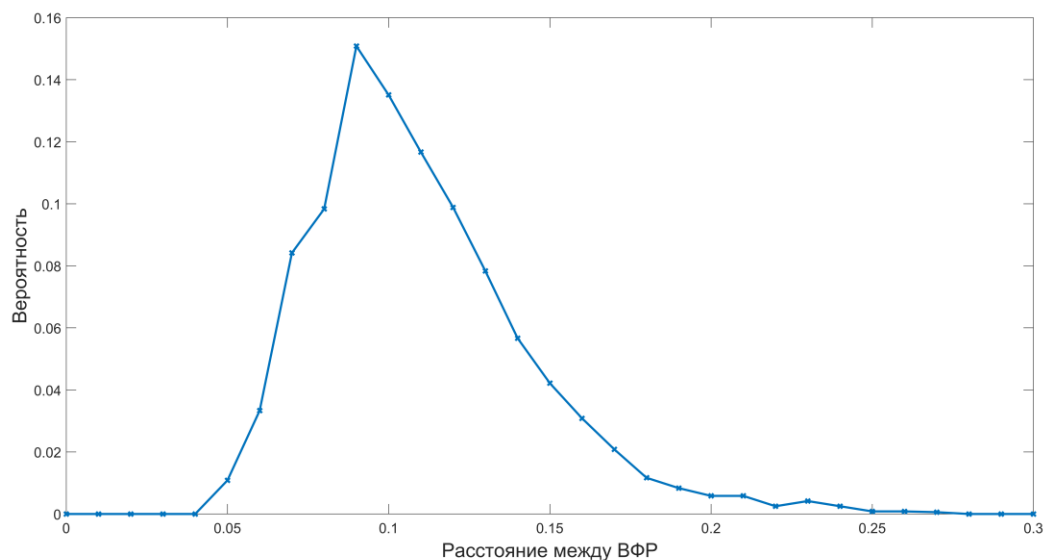


Рис. 1.2 – Выборочная плотность распределения расстояний между независимыми ВФР, длина выборки 100

В экспериментах по сравнению выборок случайной величиной является расстояние между парой выборок, а также и функция распределения этих расстояний, квантиль которой следует выбрать на практике в качестве нужного уровня значимости. Как известно (см., напр., [19, 35]), если случайная величина ξ (в данном примере это расстояние между выборками) имеет функцию распределения $F(x)$ (для расстояний между выборками асимптотически это есть функция Колмогорова), то случайная величина $\eta = F(\xi)$ (здесь это уровень значимости) имеет равномерное распределение на $[0; 1]$. Следовательно, согласованный с экспериментом уровень значимости α есть функция, линейно зависящая от расстояния ε между выборками. Поскольку же в норме

С это расстояние меняется от нуля до единицы, то следует положить $\alpha = \varepsilon$. В результате получаем, что критическое расстояние разделения выборок на уровне значимости, согласованном в вышеописанном смысле с экспериментом, определяется из уравнения

$$1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) = \varepsilon. \quad (1.1.6)$$

В результате приходим к выводу, что при анализе выборок определенной длины неправильно задавать априори желаемый уровень значимости, так как для заданной длины N выборки лишь при одном значении $\varepsilon = \varepsilon^*(N)$, определяемом из уравнения (1.1.6), вероятность превышения значения $\varepsilon^*(N)$ равна значимости используемого для этой цели критерия.

Решение уравнения (1.1.6) единственно, поскольку правая часть как функция ε монотонно возрастает от нуля до единицы, а левая монотонно убывает от единицы до нуля. Таблица численно определяемых решений этого уравнения приведена в [25]. Найденное решение было введено в [25] и названо согласованным отклонением для стационарных ВФР. Оно представляет собой стационарную точку уровня значимости критерия Колмогорова-Смирнова. Эта точка, как и функция Колмогорова, не зависит от вида распределения случайной величины. На Рис. 1.3 и в таблице 1.1 приведены результаты табуляции стационарной точки согласно [25].

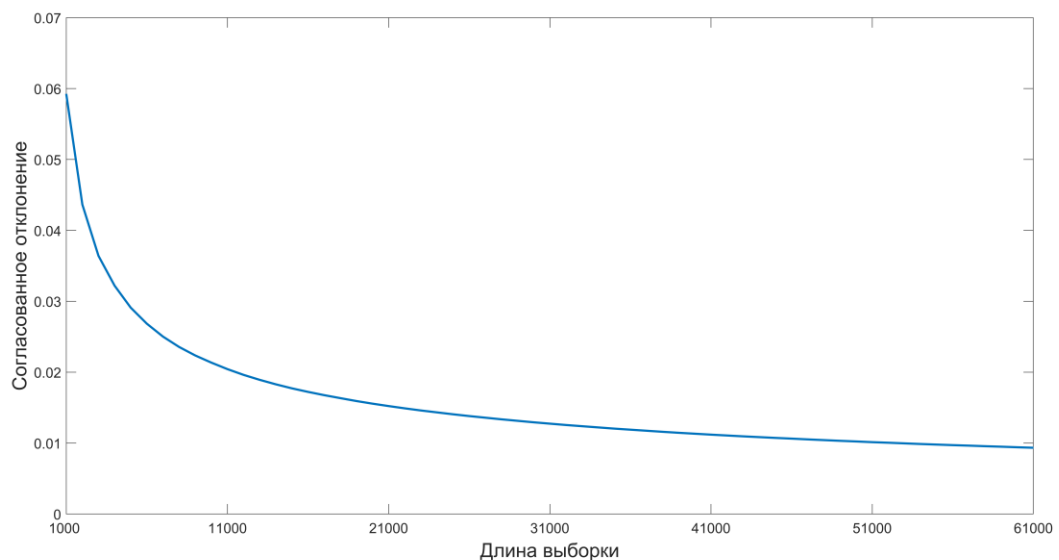


Рис. 1.3 – Зависимость стационарной точки уровня значимости от длины выборки [25]

Табл. 1.1. Значения стационарной точки для критерия Колмогорова-Смирнова [25]

Длина выборки N	СУС $\varepsilon^*(N)$
50	0,25
100	0,18
150	0,16
200	0,14
250	0,12
300	0,10
400	0,09
500	0,08
1000	0,05926
2000	0,04364
3000	0,03641
4000	0,03220
5000	0,02910
10000	0,02135
50000	0,01023
100000	0,00834

Например, если в вышеописанном эксперименте (Рис. 1.2) проинтегрировать (просуммировать) найденную эмпирически плотность распределения расстояний между выборочными распределениями для выборок длины 100 до такого значения расстояния $\varepsilon^*(100)$, при котором интеграл отличается от единицы на величину $\varepsilon^*(100)$, то численно обнаружится, что это значение приближенно (с точностью до третьего знака после запятой) равно 0,18.

Согласно [25], описанный здесь метод определения характерного расстояния между выборками конечной длины может быть использован как для стационарных, так и для нестационарных распределений. Если анализируется изменение распределения между выборками с течением времени, то для получения оценки следует построить статистику расстояний между так называемыми встык-выборками, т.е. между ВФР $F_N(x, t)$ и $F_N(x, t + N)$, сдвинутыми одна относительно другой на величину окна выборки:

$$\rho(N; t) = \|F_N(x, t) - F_N(x, t + N)\| \quad (1.1.7)$$

Далее строится функция распределения $G_N(\rho)$ расстояний (1.1.7), которая представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ . Определим теперь согласованный уровень стационарности (далее СУС) $\rho^*(N)$ так, что соответствующее расстояние равно значимости критерия, т.е. является решением уравнения

$$G_N(\rho) = 1 - \rho. \quad (1.1.8)$$

В стационарном случае уравнение (1.1.8) переходит в уравнение (1.1.6), поскольку тогда функция распределения $G_N(\rho)$ переходит в функцию из критерия Колмогорова-Смирнова.

В работе [25] был введен также индекс нестационарности временного ряда. Он определяется как отношение решения $\rho^*(N)$ уравнения (1.1.8) к стационарной точке $\varepsilon^*(N)$ уровня значимости стационарного критерия (1.1.6). Именно,

$$J(N) = \frac{\rho^*(N)}{\varepsilon^*(N)}. \quad (1.1.9)$$

Индекс нестационарности показывает, во сколько раз доля расстояний, больших СУС, превосходит аналогичный показатель для стационарных рядов. Если $J(N) \leq 1$, ряд считается стационарным, а если $J(N) > 1$, то ряд нестационарный. Индекс нестационарности позволяет проанализировать, на каких длинах выборки ряд ведет себя более или менее нестационарным образом, что важно при разработке других индикаторов, основанных на выборочных статистиках.

1.2. Примеры применения индикатора СУС

Использование одномерного СУС как индикатора уровня нестационарности временных рядов было применено в работах [3, 6, 12, 25, 26, 28, 29, 30, 32] для исследования различных практически важных вопросов. Кратко представим основные результаты этих исследований, демонстрирующие возможности индикатора СУС и объясняющие интерес к разработке более универсальных программных алгоритмов для его расчета.

Оценка реального уровня значимости при прогнозе землетрясений. В работе [3] были исследованы данные регионального каталога землетрясений по Японии. Рассматривалась статистика потока событий (момент времени, магнитуда). Традиционно считается, что функция распределения магнитуд имеет вид стационарного закона

(Гуттенберга-Рихтера), что позволяет прогнозировать магнитуду события на достаточно высоком (0,95) уровне доверия. Однако реальные данные за 10-летний промежуток времени показывают, что частота определенных событий оказывается заметно выше их теоретической оценки. Для анализа этой ситуации были построены графики СУС для временных рядов магнитуд и промежутков времени между событиями как функции длины выборки. Выяснилось, что оба ряда сильно нестационарны (Рис. 1.4).

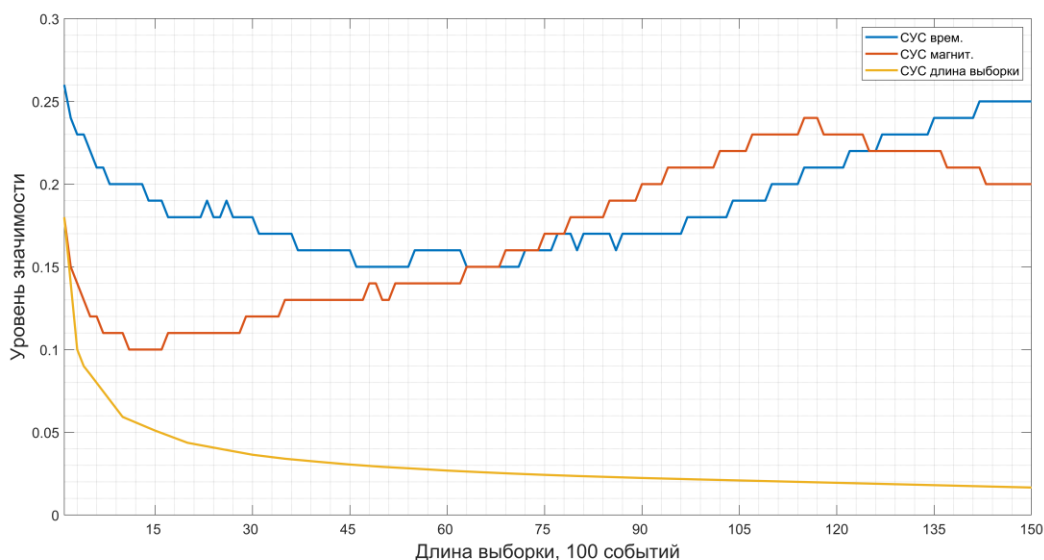


Рис. 1.4 – СУС для рядов магнитуд и промежутков времени между событиями в зависимости от длины выборки, [3]

Из Рис. 1.4 следует, что наилучшая точность (наименьший СУС) для ряда магнитуд и ряда промежутков времени между землетрясениями достигается на разных длинах. Для магнитуд наилучший уровень стационарности 0,10 достигается на выборках длин от 1100 до 1600 событий (это промежуток приблизительно от 3 до 5 лет). Для промежутков времени между событиями наилучшая точность 0,15 достигается на выборках длин от 4600 до 5400 (приблизительно от 12 до 15 лет). Видно также, что с увеличением длины выборки уровень значимости в оценке вероятностей не снижается, т.е. добавление новых событий не приводит к уточнению знаний о процессе в рамках статистических методов. Анализ, проведенный в [3], позволил также указать максимальный горизонт квазистационарного прогнозирования вероятности землетрясений. По магнитуде он составляет 500 событий (примерно 2 года), если оценивать вероятности по скользящей выборке 1100 событий, а по промежуткам времени – 800 событий (примерно 3 года) по скользящей выборке 4600 событий.

Оценка времени стабилизации случайного процесса. В работе [12] рассматривался функционал устойчивости беспроводного соединения (SIR, signal to

interference ratio) в группе периодически движущихся приемо-передающих устройств с несоизмеримыми частотами движения. Этот функционал является нелинейным по функции распределения положений этих устройств. В качестве реальных моделей движения рассматривались траектории поездов по веткам метро в Санкт-Петербурге, Москве и Лондоне. Результат вычислений представлен на Рис. 1.5. Наблюдается эффект снижения уровня нестационарности распределения SIR с увеличением линий в системе и с уменьшением среднего периода обращения поездов. Примерно через промежуток времени, равный трем средним периодам, распределение среднего по системе SIR становится стационарным.

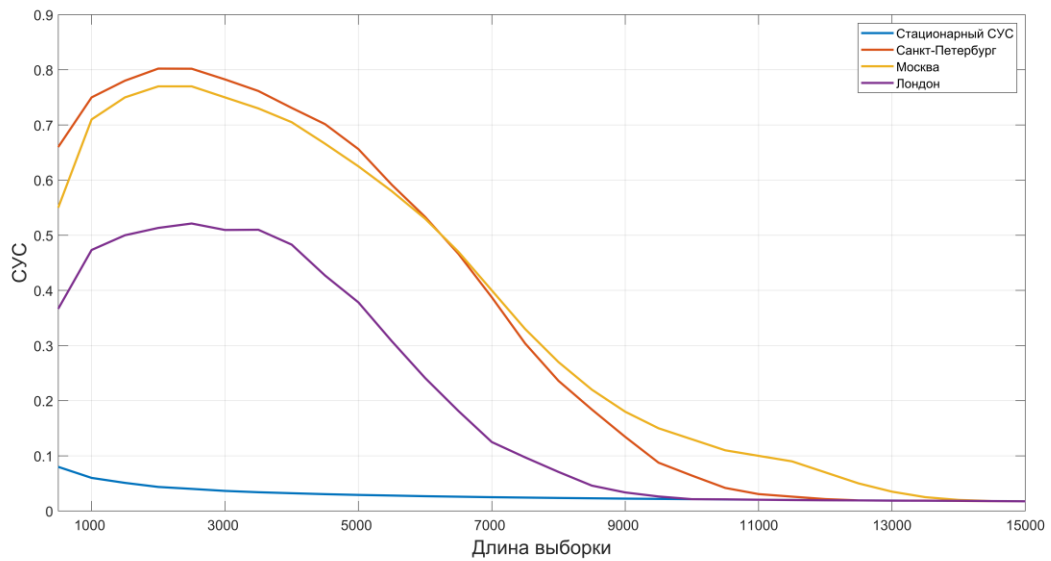


Рис. 1.5 – CUC для ряда SIR в зависимости от длины выборки

Поскольку скорость установления стационарного режима зависит не только от числа линий и станций на них, но и от геометрии области, то сравнивать между собой графики Рис. 1.5 на предмет выявления зависимости от числа абонентов не вполне корректно, хотя качественно они похожи один на другой. Более точным является сравнение скоростей установления стационарного режима внутри одной и той же системы метрополитена, но при разном количестве задействованных линий. В [12] была исследована зависимость индекса нестационарности (1.1.9) ряда средних значений SIR по ансамблю поездов от числа n линий и длины N выборки для метро Лондона. Выяснилось, что для малого числа линий (от 5 до 15) рассматриваемая зависимость приблизительно линейна, а с увеличением числа линий индекс нестационарности меняется в соответствии с полуэмпирической формулой (на уровне детерминации 0,9):

$$J(n, N) = 1 + \frac{A(N)}{\sqrt{n}}, \quad (1.1.10)$$

где $A(N)$ есть некоторый численный коэффициент, характерный для выборки определенной длины и геометрии области.

Таким образом, временной ряд SIR для системы поездов метро становится стационарным примерно через 3 часа после начала движения, причем индекс нестационарности стремится к своему стационарному значению для выборок любой длины по закону, близкому к центральной предельной теореме.

Определение минимальной длины выборки для статистического распознавания образов. В работе [26] ставилась задача распознавания функции распределения текста по буквам методом близости к определенному паттерну с целью идентификации автора. Для ее решения требовалось, чтобы, начиная с выборок достаточной длины, выборочные функции распределения были бы не случайны, а приближались бы к характерному авторскому паттерну. В [26] рассматривался аналог СУС, который был назван авторской длиной представительности и рассматривался для вероятностей трехбуквенных буквосочетаний (3-ПФР) в сравнении со стационарными значениями. Точность, с которой можно определять авторскую 3-ПФР в зависимости от длины текста, показана на Рис. 1.6.

Выяснилось, что на малых длинах ряды букв образуют нестационарный временной ряд (длины представительности лежат выше стационарной линии СУС), но после 10 тыс. знаков распределения становятся более чем стационарны. Это означает, что 3-ПФР текста стабилизируется в силу авторской специфики письма, что позволяет применить к текстам достаточной длины метод распознавания по принципу ближайшего эталона. На Рис. 1.6 приведены длины представительности для совокупности 60 текстов разных авторов.

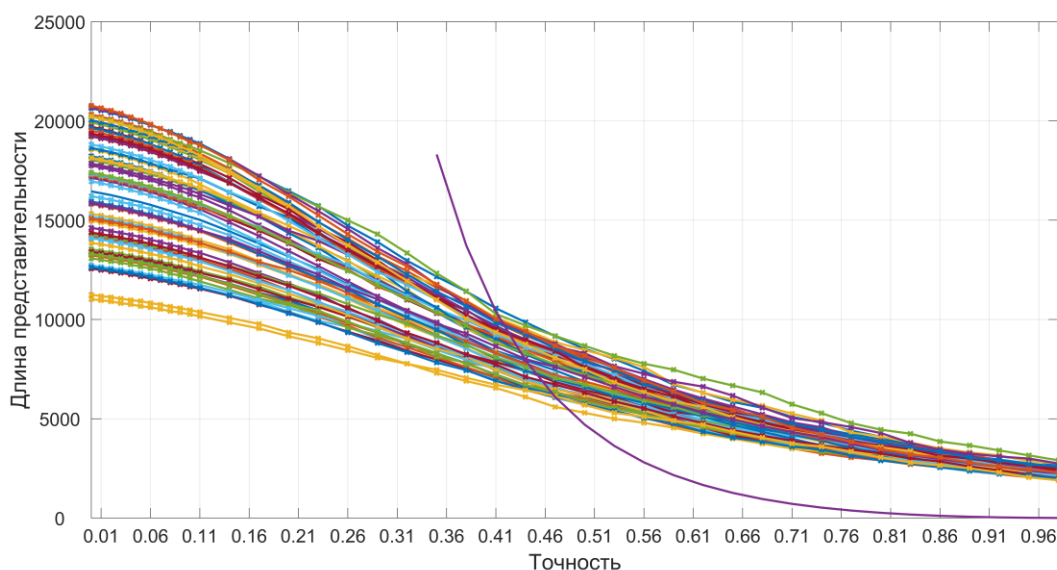


Рис. 1.6 – Длина 3-ПФР представительности текстов, [26]

Аналогичное поведение длины авторской представительности обнаруживается для большинства писателей, что позволяет сравнивать между собой произведения разной длины, ибо в статистическом смысле они оказываются эквивалентными: их распределения уже установились и дальнейшее увеличение длины текста не приведет к существенным отличиям.

Использование СУС как аналога нестационарной периодограммы. В работах [6, 25, 28, 29, 30, 32] проводился статистический анализ временных рядов, которые представляли собой биржевые цены закрытия (с заданным шагом по времени) на те или иные финансовые инструменты. Цель указанных работ состояла в том, чтобы выявить характерные временные периоды функционирования рынка, на которых ряды цен являются более (или наоборот, менее) стационарными, что важно для создания тестирующих алгоритмов торговых стратегий и для анализа результативности работы этих стратегий. Традиционный анализ с помощью разложения сигнала в ряд Фурье не давал желаемых результатов, поскольку ряды не были стационарными. Немонотонное же поведение СУС в зависимости от длины выборки (или, что то же, от времени наблюдения) позволило выделить характерные промежутки, на которых ряд цен ведет себя наиболее стационарно и, напротив, с резкими изменениями. На Рис. 1.7 приведен результат анализа приростов (первых разностей) индекса РТС с минутным отсчетом времени.

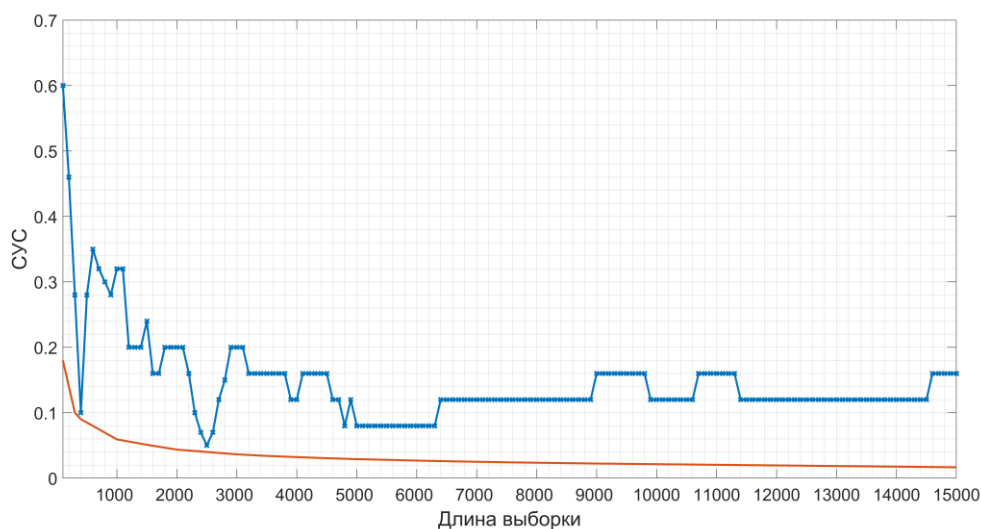


Рис. 1.7 – Зависимость СУС для минутных приростов индекса РТС, [28]

Из Рис. 1.7 видно, что с увеличением окна выборки до 400 мин ряд становится стационарным, после чего следует разладка, длящаяся примерно 4 суток. Затем ряд снова приближается к стационарному состоянию с последующей разладкой еще примерно на такое же время, и после очередного, но не так выраженного снижения (через 10 суток) уровень нестационарности опять возрастает, теперь уже окончательно.

Приведенные примеры показали, что СУС может представлять интерес как инструмент анализа в различных прикладных статистических задачах. Последний пример показал, что СУС можно попытаться использовать как индикатор или предиктор разладки.

Также выяснились ограничения вычислительных алгоритмов, применявшихся на первоначальных этапах вышеприведенных исследований. Индикатор СУС строился как функция длины выборки, но только один раз, т.е. он трактовался как характерное число для данного ряда. Вопрос изменчивости его распределения не рассматривался, а без этого трудно использовать СУС в качестве предиктора разладки, поскольку окно индикации, как правило, существенно меньше окна обучения. Вследствие этого надо иметь оценку вероятности события, состоящего в том, что СУС для данного фрагмента ряда будет выше, чем для всего ряда в целом.

Кроме того, поскольку добавление еще одной размерности (произвольного фрагмента ряда для вычисления СУС) существенно увеличивает время счета, так что актуальной задачей является разработка приближенных методов оценки СУС.

1.3. Компьютерные программы для моделирования индикаторов разладки

Большая часть программного обеспечения в области статистического анализа не является узко специализированным. Такие программы как SPSS и Statistica и подобные им продукты (Minitab, Eviews, Statgraphics, SAS и др. – см. обзоры программного обеспечения в [1, 7, 8, 14, 24]) позволяют проводить анализ временных рядов в очень широком спектре задач. Вычисление разладки в них осуществляется как некоторый побочный результат, поскольку в основе расчета лежит проверка того или иного статистического критерия. Важно подчеркнуть, что критерии, табулированные в статистических пакетах, относятся к стационарным случайным процессам. Нестационарность учитывается только в контексте авторегрессионных моделей.

Непосредственно задачу обнаружения разладки решают сравнительно немногие программные комплексы, созданные для применения в конкретных производственных областях или для решения конкретных технических задач. Как правило, они представляют собой определенную модификацию стандартной программы из статистического пакета, т.е. по-прежнему предназначены для анализа стационарного ряда.

Например, в [20] рассматривается задача идентификации разладки с несколькими скачками, для чего применяется модифицированный алгоритм кумулятивных сумм в скользящем окне. Длина окна выбирается в соответствии с корреляционным интервалом изучаемой функции, т.е. базовый процесс предполагается стационарным.

Большой класс представляют задачи обнаружения разладки в виде сбоев в режиме функционирования программного обеспечения, которые решаются посредством анализа данных, собираемых при эксплуатации систем массового обслуживания, таких как количество обработанных запросов или средняя длина очереди. Применительно к различным практическим приложениям соответствующие методы описаны в [2, 42, 43, 45, 47, 48, 60, 61, 62, 64, 67, 70, 72, 73, 75, 76]. Применяемые подходы сводятся к выявлению момента резкого изменения либо непосредственно наблюдаемых характеристик рассматриваемой системы, либо наблюдаемых статистических данных о других характеристиках этой системы [22, 36, 63]. Задачи такого типа первоначально рассматривались А.Н. Колмогоровым, А.Н. Ширяевым, М. Гиршиком, С. Робертсом и др. (см. [37, 38, 39, 50, 68]). Все перечисленные задачи относятся к стационарному типу.

Непараметрические критерии для распознавания разладки в нестационарных временных рядах в настоящее время не имеют программной реализации.

1.4. Задача разработки программного комплекса для построения СУС

Как показал обзор современных программ для идентификации разладки, преимущественно используются параметрические методы оценивания, т.е. определяются доверительные интервалы, которым принадлежит оцениваемый параметр распределения изучаемой случайной величины, функциональный вид распределения которой считается известным.

Из непараметрических методов оценивания близости распределений как таковых, а не функционалов от них (близость средних, дисперсий и т.п.), программно реализованы лишь классические стационарные методы, использующие критерии Колмогорова-Смирнова (норма C), Джини (норма L_1), Крамера – фон Мизеса (норма L_2) и разные взвешенные нормы, применяемые в критериях Андерсона-Дарлинга, Ватсона, Фроцини (см. [19, 21]).

Для анализа нестационарных и не коинтегрированных рядов имеется только один программный комплекс [31], в котором реализован расчет СУС в зависимости от длины выборки для анализа пошаговой эволюции выборок внутри заданного одномерного фрагмента данных. Это означает, что для заданной длины встык-выборки алгоритм выдает ровно одно число. После вариации длины выборки результатом работы алгоритма является график СУС в виде $\rho^*(N)$, где N меняется в заданных пределах в соответствии с длиной L_{tot} полного набора анализируемых данных.

Для задачи распознавания разладки с указанием вероятности ошибки первого рода (ложной тревоги) необходимо дополнить вышеуказанный численный алгоритм расчета СУС возможностью проводить анализ по различным фрагментам исходного ряда, т.е. в качестве выходных данных он должен выдавать функцию распределения случайной величины СУС на основе статистики значений последовательности СУС $\rho_n^*(N, L)$, построенной для встык-выборок длины N в окнах длины L . Таким образом, Задача 1 при разработке программного комплекса формулируется как увеличение размерности блока статистического анализа СУС.

Кроме того, надо предусмотреть возможность многоканального анализа, для чего требуется провести распараллеливание алгоритма. Это важно в задачах анализа высокочастотных биометрических данных, которые записываются одновременно по нескольким отведениям (вплоть до нескольких тысяч различных рядов). Следовательно, возникает Задача 2 об использовании высокопроизводительных вычислений и проведении оценки их эффективности на основе вычислительных экспериментов для конкретных прикладных задач. Эта задача естественным образом сопряжена с развитием так называемых систем с интенсивным программным обеспечением. Согласно стандарту ISO/IEC/IEEE 42010:2011(E) это системы, функциональность которых определяется в основном их программными средствами [52]. Согласно Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года такими системами являются аппаратно-программные комплексы с большим удельным весом программной части [23].

Задача 3 состоит в разработке метода сокращенных вычислений для построения статистики СУС, поскольку в условиях больших данных алгоритмы, основанные на полном переборе, неэффективны, а часто и не реализуемы в силу естественно ограниченных возможностей суперкомпьютеров.

Выводы по Главе I.

В Главе I проведен обзор результатов по моделированию и распознаванию разладки, введено понятие согласованного уровня стационарности как кандидата на роль такого индикатора для существенно нестационарных процессов и поставлена задача о разработке программного продукта, обладающего возможностью проводить необходимый многомерный параметрический анализ с распределением статистики СУС.

Результаты Главы I опубликованы в работах [12, 33, 34, 54, 69].

Глава II. Эквивалентность по Чернову в задачах математической статистики

2.1. Эквивалентные по Чернову операторные функции и полугруппы

В этой главе метод так называемых конечнократных аппроксимаций для построения решающих операторов задачи Коши для дифференциальных уравнений в частных производных, известный как формализм Фейнмана [49], будет применен к анализу уровня нестационарности выборочных функций распределения. Этот метод устанавливает связь между статистикой и динамикой, что позволяет построить математически корректную процедуру аппроксимации случайных процессов с помощью динамических систем.

Формулами Фейнмана называются представления однопараметрической группы Шредингера $e^{it\hat{H}}$, где $\hat{H}$ есть генератор группы (квантовый оператор Гамильтона), с помощью пределов интегралов по декартовым степеням пространства, связанного с классической гамильтоновой системой, гамильтониан которой H есть символ оператора $\hat{H}$ в терминологии псевдодифференциальных операторов.

Ключевым математическим утверждением, обосновывающим применение формализма Фейнмана, является теорема Чернова [46], устанавливающая сходимость определенного итерационного процесса к полугруппе, которая представляет решение задачи Коши. Эта теорема состоит в следующем.

Теорема (P. Chernoff, [46]).

Пусть X – банахово пространство, $B(X)$ – банахово пространство линейных ограниченных операторов в X и пусть функция $F: [0; +\infty) \rightarrow B(X)$ удовлетворяет условиям $F(0)=I$, $F(t)$ непрерывна в сильной операторной топологии и имеет место оценка $\|F(t)\|_{B(X)} \leq \exp(\alpha t), t \geq 0$ при некотором $\alpha \geq 0$. Тогда, если оператор $F(t)$ замыкаем и его замыкание является генератором сильно непрерывной полугруппы операторов $G_t, t > 0$, то для любого $u \in X$ и любого $T > 0$ справедлив предельный переход

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \left\| G_t u - \left(F\left(\frac{t}{n}\right) \right)^n u \right\|_X = 0. \quad (2.1.1)$$

Развитие метода, основанного на применении теоремы Чернова, в последнее время связано с получением явных аналитических выражений [27] в виде сходящегося итерационного процесса (2.1.1) для решения задач Коши для уравнений теплопроводности или Шредингера.

Теорема Чернова, обосновывающая процедуру построения генераторов полугрупп, позволяет придать новый смысл получающимся предельным выражениям. Например, при построении равновесного оператора плотности квантовой системы надо определить ядро оператора $\hat{\rho}(\beta) = e^{-\beta\hat{H}}$. Если известен классический символ $H(q, p)$ оператора $\hat{H}$, то равновесная плотность функции распределения фазовых переменных с точностью до нормировки определяется выражением $f(\beta; q, p) = e^{-\beta H(q, p)}$, которое, однако, не является символом оператора $\hat{\rho}(\beta)$. Тем не менее, если ввести оператор $\hat{D}_n = \hat{f}(\beta/n)$, т.е. оператор, для которого функция $f(\beta/n; q, p) = e^{-\beta H(q, p)/n}$ является символом, то в соответствии с теоремой Чернова итерационный процесс $(\hat{D}_n)^n$ сходится к некоторой полугруппе. В работе [71] было показано, что этот предельный оператор совпадает с оператором плотности, т.е.

$$\hat{\rho}(\beta) = \lim_{n \rightarrow \infty} (\hat{D}_n)^n. \quad (2.1.2)$$

Следовательно, чтобы получить статистический оператор, надо проквантовать классическую плотность функции распределения как динамическую величину и затем бесконечное число раз свернуть полученный оператор с самим собой, уменьшая при этом полугрупповой параметр β в соответствующее число раз. Тем самым квантовая статистика получается как предел итераций квантовой динамики. Поскольку же ядро оператора $\hat{D}_n$ определено явно и представляет собой сверточное произведение, его можно рассматривать как усреднение известных функций, заданных в классическом фазовом пространстве, по известной мере. В пределе при $n \rightarrow \infty$ это позволяет трактовать квантовый статистический оператор как результат статистического усреднения бесконечного числа независимых случайных сдвигов в классическом пространстве, т.е. как результат усреднения случайных полугрупп. Указанная трактовка теоремы Чернова была рассмотрена в [27], что позволило придать математический смысл усреднению полугрупп. В общем случае линейная комбинация полугрупп не обязана быть полугруппой. Но если в результате итераций, аналогичных описанному выше, можно получить сходимость к полугруппе, то такая полугруппа объявляется эквивалентной по Чернову данной линейной комбинации полугрупп.

Определение эквивалентности по Чернову [27]. Операторнозначная функция $\hat{f}(t)$, действующая из некоторой правой полуокрестности нуля на числовой оси в банахово пространство $B(X)$ линейных ограниченных операторов, действующих в банаховом пространстве X , называется эквивалентной по Чернову полугруппе $\hat{g}(t)$, если $\forall T > 0$ и $\forall u \in X$ выполняется условие:

$$\lim_{n \rightarrow \infty} \sup_{t \in [0; T]} \left\| \left(\hat{f}(t/n) \right)^n - \hat{g}(t) \right\|_X u = 0. \quad (2.1.3)$$

На основе определения эквивалентности по Чернову в работе [27] были построены примеры эквивалентных по Чернову операторов, отвечающих статистической смеси генераторов полугрупп, что позволило дать следующее определение обобщенного среднего значения случайной полугруппы.

Определение среднего значения случайной полугруппы [27]. Сильно непрерывная однопараметрическая полугруппа U ограниченных линейных преобразований банахова пространства X является обобщенным средним значением случайной полугруппы ξ , если полугруппа U эквивалентна по Чернову математическому ожиданию $M\xi$.

Также в работе [27] была доказана теорема об эквивалентности по Чернову для средних значений случайных полугрупп, в которых сформулированы достаточные условия того, чтобы эти средние значения, не будучи сами полугруппами, порождали полугруппу, эквивалентную им по Чернову.

Теорема об эквивалентности средних значений полугрупп [27]. Пусть $\{\hat{H}_n\}$ – последовательность генераторов сильно непрерывных полугрупп в банаховом пространстве X . Пусть также $\{p_n\}$ – последовательность неотрицательных чисел, сумма

ряда из которых равна единице: $\sum_{n=1}^{\infty} p_n = 1$. Пусть также существует линейное

подпространство $D \subset X$, являющееся существенной областью определения каждого из генераторов $\hat{H}_n$ и такое, что для любого $g \in D$ ряд $\sum_{n=1}^{\infty} p_n \|\hat{H}_n g\|_X$ сходится. Тогда, если

оператор $\hat{H}$ определен на D формулой $\hat{H}g = \sum_{n=1}^{\infty} p_n \hat{H}_n g$ и его замыкание является

генератором сильно непрерывной полугруппы $\hat{U}(t) = e^{t\hat{H}}$, $t \geq 0$, то среднее значение $M\hat{U}$

случайной полугруппы $\hat{U}_n(t) = e^{t\hat{H}_n}$, определяемое формулой $M\hat{U} = \sum_{n=1}^{\infty} p_n \hat{U}_n$,

эквивалентно по Чернову полугруппе $\hat{U}(t) = e^{t\hat{H}}$.

Описанные конструкции позволяют сформулировать единый подход к решению многих эволюционных задач. Он состоит в том, что вместо поиска точного решения ищется подходящая операторнозначная функция, не обладающая, вообще говоря, полугрупповым свойством, после чего строится эквивалентная ей по Чернову полугруппа, которая и является решением соответствующей задачи. Аналоги этого метода применительно к задачам математической статистики, относящимся к теме настоящего исследования, формулируются в нижеследующих разделах.

2.2. Применение эквивалентности по Чернову к уравнению Лиувилля для выборочных распределений

При построении прогнозной модели временного ряда часто используются регрессионные аппроксимации [25, 35]. Дадим определения соответствующих понятий.

Пусть в результате наблюдений за случайными величинами ξ и η были получены ряды значений $x_1, \dots, x_N$ и $y_1, \dots, y_N$ соответственно. Пусть также $\rho(y, y')$ – функция расстояния в пространстве Y , которому принадлежат значения случайной величины η . Пространство значений случайной величины ξ обозначим через X .

Регрессией случайной величины ξ на случайную величину η называется такое отображение $\varphi: X \rightarrow Y$, для которого математическое ожидание расстояния $M\rho(y, \varphi(x))$ минимально. Пусть X и Y – линейные пространства с евклидовой нормой. Тогда для нахождения регрессии используется метод наименьших квадратов.

Чтобы отличать значения случайной величины η , полученные в результате аппроксимации $\varphi(x)$, от фактических, будем снабжать их тильдой ($\tilde{y}$), а прогнозные значения – шляпкой ($\hat{y}$).

Введем обозначения для выборочных средних величин:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_k, \quad \bar{y} = \frac{1}{N} \sum_{n=1}^N y_k, \quad \sigma_x^2 = \frac{1}{N} \sum_{n=1}^N (x_k - \bar{x})^2, \quad \text{cov}_{xy} = \frac{1}{N} \sum_{n=1}^N (x_k - \bar{x})(y_k - \bar{y}), \quad (2.2.1)$$

В этих обозначениях выборочная прямая регрессии записывается в виде:

$$\tilde{y}_n - \bar{y} = \frac{\text{cov}_{xy}}{\sigma_x^2} \cdot (x_n - \bar{x}). \quad (2.2.2)$$

В частности, если рассматривается выборка из N значений $x(n)$, $n=1,2,\dots,N$ случайной величины ξ , то регрессия этой величины на «время», то есть на дискретный аргумент n , имеет вид

$$\tilde{x}(n) - \bar{x} = \frac{12 \text{cov}_{xn}}{N^2 - 1} \cdot (n - (N+1)/2). \quad (2.2.3)$$

Если рассматривать модель (2.2.3) как прогнозную при $n > N$, т.е. записывать прогнозную оценку в виде

$$\hat{x}(n) - \bar{x} = \frac{12 \text{cov}_{xn}}{N^2 - 1} \cdot (n - (N+1)/2),$$

то становится ясно, что приближенная модель временного ряда имеет вид линейной динамической системы с постоянной скоростью $\frac{dx}{dt} \equiv \hat{x}(n+1) - \hat{x}(n) = v = \frac{12 \text{cov}_{xn}}{N^2 - 1}$.

Также и во многих других практических случаях модели временных рядов являются динамическими системами. Следовательно, одна из задач статистического анализа состоит в том, чтобы по имеющимся наблюдаемым данным построить подходящую динамическую систему, моделирующую временной ряд.

Непрерывным пределом соответствующей дискретной динамической системы является гладкий динамический поток. С каждым таким потоком может быть ассоциировано уравнение Лиувилля, описывающее эволюцию плотности фазовых точек, представляющих данную динамическую систему. Интересно поставить обратную задачу: по имеющемуся уравнению Лиувилля, моделирующему эволюцию выборочных распределений случайной величины, построить динамическую систему, которая ему соответствует, и дискретный аналог которой мог бы тогда служить моделью временного ряда. Построим обоснование такого подхода.

Пусть имеется некоторая динамическая система, т.е. однопараметрическая группа (или полугруппа) G_t преобразований

$$x(t) = G_t x(0), \quad (2.2.4)$$

действующая на некотором множестве точек $x \in \Gamma$, называемом фазовым пространством системы. Согласно [4] уравнением Лиувилля для этой системы называется уравнение эволюции плотности $\rho_0(x(0))$ распределения начальных условий в силу уравнения движения системы, определяемого фазовой скоростью

$$v(x) = \frac{dx}{dt} = \frac{d}{dt} G_t x|_{t=0}. \quad (2.2.5)$$

Если взять любую динамическую величину, т.е. некоторую функцию $B(x(t))$, аргумент которой зависит от параметра t в соответствии с уравнением $dx/dt = v(x)$, то ее значение в момент времени t записывается в виде

$$B_t \equiv B(x(t)) = B(G_t x(0)) = G_t(B(x(0))). \quad (2.2.6)$$

Среднее значение динамической величины $B(x(t))$ по начальному распределению получается усреднением выражения (2.2.5) по мере с плотностью $\rho_0(x)$:

$$\langle B_t \rangle_0 = \int_{\Gamma} B(G_t(x)) \rho_0(x) dx. \quad (2.2.7)$$

Введем обозначение

$$\rho_t(y) = \int_{\Gamma} \rho_0(x) \delta(G_t(x) - y) dx. \quad (2.2.8)$$

Так как $B(G_t(x)) = \int_{\Gamma} B(y) \delta(G_t(x) - y) dy$, то с учетом (2.2.8) выражение (2.2.7) для

среднего значения динамической величины приводится к виду

$$\begin{aligned} \langle B_t \rangle_0 &= \int_{\Gamma} B(G_t(x)) \rho_0(x) dx = \int_{\Gamma} \rho_0(x) dx \int_{\Gamma} B(y) \delta(G_t(x) - y) dy = \\ &= \int_{\Gamma} B(y) dy \int_{\Gamma} \delta(G_t(x) - y) \rho_0(x) dx = \int_{\Gamma} B(y) \rho_t(y) dy \equiv \langle B \rangle_t. \end{aligned} \quad (2.2.9)$$

Формула (2.2.9) показывает, что действие эволюционного потока G_t можно перенести с динамической величины на плотность функции распределения и получить, что среднее значение динамической величины в момент времени t по начальному распределению равно среднему значению той же динамической величины, но как функции просто точки в фазовом пространстве, по распределению в момент времени t .

Уравнение эволюции для $\rho_t(y)$ получается дифференцированием по времени выражения (2.2.8) с учетом определения (2.2.5) динамического потока:

$$\frac{\partial \rho_t(y)}{\partial t} = \int_{\Gamma} \rho_0(x) \frac{\partial}{\partial t} \delta(G_t(x) - y) dx = - \int_{\Gamma} \rho_0(x) v(x) \frac{\partial}{\partial y} \delta(G_t(x) - y) dx = - \frac{\partial}{\partial y} (\rho v)_t(y). \quad (2.2.10)$$

Пусть фазовое пространство динамической системы (2.2.4) определяется положениями x точки и их независимыми приращениями $x^{(k)}$ вплоть до n -го порядка. Тогда поток (2.2.5) запишется в виде:

$$x^{(k)} = \frac{d^k x}{dt^k}, \quad k = 1, 2, \dots, n-1; \quad \frac{dx^{(n)}}{dt} = x^{(n+1)} = v_{n+1}(x, \dot{x}, \dots, x^{(n)}). \quad (2.2.11)$$

Обозначим соответствующую плотность функции распределения (ПФР), зависящую от $n+1$ фазовых переменных и времени, через $f_{n+1}(x, \dot{x}, \dots, x^{(n)}; t)$. Для нее уравнение Лиувилля (2.2.10) примет вид:

$$\begin{aligned} \frac{\partial f_{n+1}(x, \dot{x}, \dots, x^{(n)}; t)}{\partial t} = & - \sum_{k=0}^{n-1} x^{(k+1)} \frac{\partial}{\partial x^{(k)}} f_{n+1}(x, \dot{x}, \dots, x^{(n)}; t) - \\ & - \frac{\partial}{\partial x^{(n)}} v_{n+1}(x, \dot{x}, \dots, x^{(n)}) f_{n+1}(x, \dot{x}, \dots, x^{(n)}; t). \end{aligned} \quad (2.2.12)$$

Введем частичные ПФР согласно определению:

$$f_s(x, \dot{x}, \dots, x^{(s-1)}; t) = \int f_{n+1}(x, \dot{x}, \dots, x^{(n)}; t) dx^{(s)} dx^{(s+1)} \dots dx^{(n)}. \quad (2.2.13)$$

Здесь интегралы берутся по соответствующим пределам изменения фазовых переменных. При нулевых граничных условиях для ПФР f_{n+1} и для скорости v_{n+1} эволюция частичных ПФР (2.2.13) имеет вид:

$$\begin{aligned} \frac{\partial f_s(x, \dot{x}, \dots, x^{(s-1)}; t)}{\partial t} = & - \sum_{k=0}^{s-2} x^{(k+1)} \frac{\partial}{\partial x^{(k)}} f_s(x, \dot{x}, \dots, x^{(s-1)}; t) - \\ & - \frac{\partial}{\partial x^{(s-1)}} \int x^{(s)} f_{s+1}(x, \dot{x}, \dots, x^{(s)}; t) dx^{(s)}. \end{aligned} \quad (2.2.14)$$

Система уравнений (2.2.14) представляет собой цепочку Боголюбова [4] для частичных распределений. Введем среднее смещение u_s порядка s :

$$u_s(x, \dot{x}, \dots, x^{(s-1)}; t) = \frac{\int x^{(s)} f_{s+1}(x, \dot{x}, \dots, x^{(s)}; t) dx^{(s)}}{f_s(x, \dot{x}, \dots, x^{(s-1)}; t)}. \quad (2.2.15)$$

Тогда первое уравнение цепочки (2.2.14) (т.е. уравнение для $s = 1$) примет вид:

$$\begin{aligned} \frac{\partial f_1(x, t)}{\partial t} = & - \frac{\partial}{\partial x} (u(x, t) f_1(x, t)); \\ f_1(x, t) = & \int f_2(x, \dot{x}; t) d\dot{x}; \quad u(x, t) f_1(x, t) = \int \dot{x} f_2(x, \dot{x}; t) d\dot{x}. \end{aligned} \quad (2.2.16)$$

Для второго уравнения цепочки имеем

$$\begin{aligned} \frac{\partial f_2(x, \dot{x}; t)}{\partial t} = & - \dot{x} \frac{\partial f_2(x, \dot{x}; t)}{\partial x} - \frac{\partial}{\partial \dot{x}} (w(x, \dot{x}; t) f_2(x, \dot{x}; t)); \\ f_2(x, \dot{x}; t) = & \int f_3(x, \dot{x}, \ddot{x}; t) d\ddot{x}; \quad w(x, \dot{x}; t) f_2(x, \dot{x}; t) = \int \ddot{x} f_3(x, \dot{x}, \ddot{x}; t) d\ddot{x}. \end{aligned} \quad (2.2.17)$$

Здесь в терминах (2.2.15) введены обозначения $u = u_1$; $w = u_2$. Аналогично строятся уравнения для остальных плотностей совместных распределений фазовых переменных.

Предположим теперь, что наблюдаемый в дискретные моменты времени $t = 1, 2, 3, \dots$ временной ряд значений $x(t)$ случайной величины ξ порожден некоторой

неизвестной динамической системой с фазовым потоком G_t . Задача состоит в том, чтобы аппроксимировать этот поток по конечному набору наблюдаемых значений.

Рассмотрим выборку длины N значений $x(k)$ с единичным шагом по времени на промежутке от $k = t - N + 1$ до $k = t : \{x(t - N + 1), \dots, x(t)\}$. Соответствующую выборочную плотность функции распределения (ВПФР), представляемую, например, в виде гистограммы, обозначаем $f_N(x, t)$. Считаем, что значения $x(k)$ известны во все более ранние моменты времени для $k \leq t$. Будем также предполагать для удобства сравнения с рассмотренным выше случаем динамических систем, что гистограмма ВПФР является выборочной оценкой некоторой непрерывно дифференцируемой ПФР для подходящей динамической системы. Проанализируем, в какой мере статистическая схема, развитая для механических систем, может быть перенесена на описание эволюции ВПФР случайных величин.

Поскольку рассматриваются именно выборочные распределения, то фазовое пространство такой системы ограничено размахом выборки. Следовательно, если фазовое пространство ограничено, как оно и есть для фактически наблюдаемых значений, и рассматривается случайный процесс с независимыми приращениями, то уравнение Лиувилля не может быть применено для описания эволюции ВПФР с нулевыми граничными условиями, так как тогда решение должно иметь вид сдвига заданного распределения по оси координат:

$$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = 0, \quad f|_{t=0} = f_0(x) \Rightarrow f = f_0(x - vt). \quad (2.2.18)$$

Тогда, либо мы рассматриваем эволюцию на относительно коротких временах, т.е. на горизонте $\tau \ll N$, чтобы можно было пренебречь концевыми эффектами, либо скорость зависит от x , являясь в этом смысле средней по двумерному распределению. Аналогом уравнения (2.2.16) в этом случае является уравнение:

$$\begin{aligned} \frac{\partial f_N(x, t)}{\partial t} &= - \frac{\partial}{\partial x} (u_N(x, t) f_N(x, t)); \\ f_N(x, t) &= \int \Phi_N(x, \dot{x}; t) d\dot{x}, \\ u_N(x, t) f_N(x, t) &= \int \dot{x} \Phi_N(x, \dot{x}; t) d\dot{x}. \end{aligned} \quad (2.2.19)$$

Следовательно, если по предположению эволюция ВПФР $f_N(x, t)$ основывается на уравнении Лиувилля, то этому уравнению можно сопоставить динамическую систему, средняя выборочная скорость которой есть $u_N(x, t)$.

Определим статистический смысл скорости $u_N(x, t)$ в уравнении Лиувилля (2.2.19). Для этого, как и выше, рассмотрим уравнение эволюции средних значений

аналогов динамических величин, например, первого выборочного момента, определяемого как

$$m_N(t) = \langle x \rangle_t = \int x f_N(x, t) dx. \quad (2.2.20)$$

После дифференцирования величины (2.2.20) по времени и подстановки вместо $\partial f_N / \partial t$ соответствующего выражения из (2.2.19) получаем

$$\frac{dm_N(t)}{dt} = \int x \frac{\partial f_N(x, t)}{\partial t} dx = - \int x \frac{\partial}{\partial x} u_N(x, t) f_N(x, t) dx. \quad (2.2.21)$$

Затем после интегрирования по частям с учетом того, что в граничных ячейках плотность распределения равна нулю, получаем

$$\frac{dm_N(t)}{dt} = \int u_N(x, t) f_N(x, t) dx = \langle u \rangle_t = U_N(t), \quad (2.2.22)$$

где $U_N(t)$ есть среднее значение выборочной скорости $u_N(x, t)$ по выборочному же распределению, которое задано в момент времени t .

С другой стороны, то же самое изменение первого выборочного момента может быть вычислено непосредственно по значениям временного ряда:

$$\begin{aligned} \frac{dm_N(t)}{dt} &= m_N(t+1) - m_N(t) = \frac{1}{N} \sum_{k=1}^N x(t-N+k+1) - \frac{1}{N} \sum_{k=1}^N x(t-N+k) = \\ &= \frac{1}{N} \sum_{k=1}^N \dot{x}(t-N+k), \quad \dot{x}(t) = x(t+1) - x(t). \end{aligned} \quad (2.2.23)$$

В свою очередь, среднее значение приращений выражается через двумерную ВПФР $\Phi_N(x, \dot{x}, t)$ совместного распределения случайных величин x и $\dot{x}$ исходного ряда и ряда его производной, полученного взятием разностей $x(t+1) - x(t)$ в соседние моменты времени:

$$U_N(t) = \int \dot{x} \Phi_N(x, \dot{x}, t) dx d\dot{x}. \quad (2.2.24)$$

Но, как можно видеть, средняя скорость из (2.2.24) есть просто результат интегрирования выражения $u_N(x, t) f_N(x, t)$ из (2.2.19), и, таким образом, выражения (2.2.19) и (2.2.22) согласованы. Тогда разность $x(t+1) - x(t)$ есть аналог микроскопической скорости системы из N «частиц» (значений ряда).

Таким образом, $U_N(t)$ есть среднее значение выборочной скорости, определяемой по приращениям значений исходного временного ряда. Поскольку же формула (2.2.19) совпадает с представлением статистической механики (2.2.16), ей в этом случае можно придать корректный механический смысл.

Далее объем выборки считаем фиксированным и индекс N в обозначениях выборочных функций для краткости опускаем.

Как уже говорилось выше, на практике выборочные плотности представляются в виде гистограмм, которые фактически являются сеточными функциями при численном решении уравнения Лиувилля. Соответствующая дискретная схема выглядит следующим образом. Считаем, что область изменения случайной величины разбита на $n+2$ классовых интервалов по каждой переменной, шаг по времени совпадает с отсчетом нумерации значений случайной величины. Тогда само уравнение эволюции с граничными условиями принимает вид:

$$\begin{aligned} f(i, t+1) &= f(i, t) + u(i, t) f(i, t) - u(i+1, t) f(i+1, t); \\ f(0, t) &= 0; f(n+1, t) = 0; u(0, t) = 0. \end{aligned} \quad (2.2.25)$$

Граничные условия в (2.2.25) удовлетворяют представлению входящих в это уравнение величин через суммирование по двумерному распределению согласно (2.2.19):

$$f_N(i, t) = \sum_{j=1}^n \Phi_N(i, j; t); \quad \sum_{i=1}^n f_N(i, t) = 1; \quad u_N(i, t) f_N(i, t) = \sum_{j=1}^n j \Phi_N(i, j; t). \quad (2.2.26)$$

Однако уравнение эволюции для двумерного распределения $\Phi_N(i, j; t)$ нам не известно, а если для этой цели исходить из цепочки Боголюбова, то цепочка формально бесконечна и ее замыкание требует дополнительных гипотез. Поэтому на практике удобнее моделировать эволюцию ВПФР $f_N(i, t)$, не расширяя фазовое пространство без необходимости, а сравнивая между собой одномерные распределения вероятностей в различные моменты времени.

В частности, интерес представляет эволюция ВПФР на горизонте, равном окну выборки, для чего надо сравнить между собой в норме L_1 распределения $f_N(i, t-N)$ и $f_N(i, t)$, отвечающие так называемым встык-выборкам. Предположим, что пара таких распределений нам известна. Существует ли в таком случае дискретная модель перехода одного из них в другое в виде эволюции по Лиувиллю в соответствии со схемой (2.2.25)? Макроэволюцией в этом примере будем считать один временной гипершаг длины N , соответствующую скорость изменения ВПФР в i -ом классовом интервале обозначим через $v(i)$. Целью моделирования является представление такого перехода в виде N последовательно выполняемых единичных шагов. Соответствующая микроэволюция, определяемая изменением ВПФР на единичном шаге по времени, на каждом таком шаге происходит со своей скоростью, обозначаемой $u(i, k)$, $t-N \leq k \leq t-1$. Это и есть та самая скорость, которая фигурирует в уравнении Лиувилля (2.2.19). Возникает вопрос: можно ли считать, что

$$v(i) = \sum_{k=t-N}^t u(i, k), \quad (2.2.27)$$

то есть что скорость макроэволюции пропорциональна средней скорости микроэволюции? В случае утвердительного ответа на этот вопрос можно было бы предложить в качестве одной из простейших моделей микроэволюции равномерное изменение ВПФР с постоянной по времени скоростью $u(i, k) = v(i) / N$.

Рассмотрим сначала вопрос о существовании скорости макроэволюции.

Теорема 2.1. Пусть для фиксированного момента времени t начальная ВПФР $f(i, t - N), 1 \leq i \leq n$ во всех внутренних ячейках строго положительна, а нулевые значения принимает только в граничных ячейках с номерами $i = 0; n + 1$. Тогда существует единственная скорость $v(i)$, удовлетворяющая дискретной аппроксимации уравнения Лиувилля (2.2.25), отвечающего переходу от заданной начальной ВПФР $f(i, t - N)$ к заданной конечной ВПФР $f(i, t)$.

Доказательство. Перепишем уравнение (2.2.25) для гипершага длины N в виде

$$v(i+1)f(i+1, t - N) = f(i, t - N) + v(i)f(i, t - N) - f(i, t)$$

и будем рекуррентно разрешать его относительно скоростей $v(i)$, считая известной скорость в нулевой граничной ячейке: $v(0) = 0$. Тогда получаем для скорости в первой ячейке значение

$$v(1) = \frac{f(0, t - N) \cdot (1 + v(0)) - f(0, t)}{f(1, t - N)} = \frac{f(0, t - N) - f(0, t)}{f(1, t - N)} = \frac{0 - 0}{f(1, t - N)} = 0.$$

Аналогично для скорости во второй ячейке получаем

$$v(2) = \frac{f(1, t - N) \cdot (1 + v(1)) - f(1, t)}{f(2, t - N)} = \frac{f(1, t - N) - f(1, t)}{f(2, t - N)}.$$

Таким образом, зная $v(i - 1)$, находим $v(i)$:

$$v(i) = \frac{F(i - 1, t - N) - F(i - 1, t)}{f(i, t - N)}, 1 \leq i \leq n; \quad F(i, t) = \sum_{j=1}^i f(j, t). \quad (2.2.28)$$

Поскольку же $F(n, t - N) = F(n, t) = 1$, то из (2.2.28) следует, что скорость в правой граничной ячейке равна нулю, как и должно быть по условию. Теорема 2.1 доказана.

Следствие. Если начальная ВПФР $f(i, t - N)$ равна нулю в некотором внутреннем классическом интервале $1 < i < n$, то переход в заданное конечное состояние с ВПФР $f(i, t)$ по уравнению Лиувилля (2.2.25) возможен только при наличии закона сохранения для выборочных функций распределения в предыдущей ячейке: $F(i - 1, t) = F(i - 1, t - N)$.

Доказательство. Пусть в одной внутренней ячейке начальная ВПФР обращается в ноль:

$$f(i-1, t-N) > 0; \quad f(i, t-N) = 0; \quad f(i+1, t-N) > 0.$$

Через один гипершаг по времени имеем для плотности в левой соседней ячейке:

$$f(i-1, t) = f(i-1, t-N) + v(i-1)f(i-1, t-N) - v(i)f(i, t-N).$$

Но $f(i, t-N) = 0$, поэтому остается

$$f(i-1, t) = f(i-1, t-N) \cdot (1 + v(i-1)),$$

откуда следует, что скорость в этой ячейке должна быть равна

$$v(i-1) = \frac{f(i-1, t) - f(i-1, t-N)}{f(i-1, t-N)}.$$

Поскольку $f(i, t) = F(i, t) - F(i-1, t)$, то

$$v(i-1) = \frac{F(i-1, t) - F(i-2, t) - F(i-1, t-N) + F(i-2, t-N)}{f(i-1, t-N)}.$$

Но по общим формулам (2.2.28) должно быть

$$v(i-1) = \frac{F(i-2, t-N) - F(i-2, t)}{f(i-1, t-N)}.$$

Следовательно, последние две формулы совместны, если только выполнено условие $F(i-1, t) = F(i-1, t-N)$. В противном случае скорость в ячейке $i-1$ не определена и уравнение Лиувилля не применимо. Следствие доказано.

Таким образом, локальное обращение в ноль ВПФР согласуется с эволюцией по уравнению Лиувилля только в том случае, если фактически оказалось, что временной ряд образован двумя независимыми случайными величинами с непересекающимися носителями. Тогда для каждой такой случайной величины можно перенормировать ВПФР и свести задачу к предыдущей, то есть к утверждению теоремы 2.1.

Рассмотрим теперь вопрос о существовании аналогичной скорости микроэволюции, которая за N шагов переведет начальное распределение в конечное. Пусть для простоты у нас всего два шага по времени, так что известны скорость $v(i)$ и две промежуточные скорости $u(i)$ и $w(i)$. Тогда в соответствии с (2.2.25) имеем

$$\begin{aligned} f(i, t-N/2) &= f(i, t-N)(1+u(i)) - u(i+1)f(i+1, t-N); \\ f(i, t) &= f(i, t-N/2)(1+w(i)) - w(i+1)f(i+1, t-N/2). \end{aligned}$$

Подставляя выражение для $f(i, t-N/2)$ из первого этих уравнений во второе, получаем, что переход из начального состояния в конечное осуществляется по модели

$$\begin{aligned} f(i, t) &= f(i, t-N)(1+w(i))(1+u(i)) - f(i+1, t-N)(1+w(i))u(i+1) - \\ &- f(i+1, t-N)(1+u(i))w(i+1) + f(i+2, t-N)w(i+1)u(i+2). \end{aligned}$$

С другой стороны, найденная ранее макроэволюция определяется формулой

$$f(i, t) = f(i, t-N)(1+v(i)) - f(i+1, t-N)v(i+1).$$

Легко проверить, что обе схемы совпадают в линейном приближении по скоростям, но, вообще говоря, эти схемы различны. Если в первой схеме ввести скорость

$$v(i) = u(i) + w(i),$$

то получится уравнение

$$f(i, t) = f(i, t - N)(1 + v(i)) - f(i + 1, t - N)v(i + 1) + \\ + f(i, t - N)u(i)w(i) - f(i + 1, t - N)(u(i + 1)w(i) + u(i)w(i + 1)) + w(i + 1)u(i + 2)f(i + 2, t - N).$$

Первая строка написанного равенства отвечает уравнению Лиувилля (2.2.25), а вторая строка дает квадратичную по скоростям добавку и выражает тот факт, что в общем случае сумма двух динамических потоков вида (2.2.4) не есть динамический поток. В данном примере эта добавка представляет правую разностную аппроксимацию второй производной

$$\frac{\partial}{\partial x} \left(w u \frac{\partial f}{\partial x} \right).$$

Следовательно, гипотеза (2.2.27) не реализуется. Это означает, что если макроэволюция определяется по уравнению Лиувилля со скоростью $v(x)$, то на микроуровне нельзя использовать аналогичную модель со скоростью $u(x) = v(x)/N$, если численный расчет проводить в предположении сохранения полугруппового свойства для потоков. Это свойство предполагает, что найденное на временном шаге t значение ВПФР может служить начальным условием для того же самого уравнения эволюции, но решаемого на следующем шаге $t + 1$. Как показывает рассмотренный пример, в данной разностной схеме полугрупповое свойство динамики не выполняется и возникает диффузионный член. При большом количестве шагов по времени это приводит к расходимости решения: вероятности в отдельных ячейках могут стать отрицательными, а их сумма может значительно отличаться от единицы.

Таким образом, скорость макроэволюции в правосторонней разностной схеме первого порядка аппроксимации уравнения Лиувилля в общем случае не может быть представлена как сумма скоростей микроэволюции. Поэтому для ВПФ временного ряда, который порожден не динамической системой, представление о скорости изменения выборочной плотности как о среднем значении скоростей одношаговой эволюции не является достаточно обоснованным. Тем не менее, существует математически корректно определенная величина, которую можно рассматривать как предельное состояние некоторой специальной линейной комбинации динамических потоков вида (2.2.4). Соответствующее обоснование дается с помощью теоремы Чернова и теоремы об эквивалентности средней полугруппы.

Из определения эквивалентности по Чернову следует, что если $F(t)$ есть оператор полугруппы, то эквивалентный ему по Чернову оператор совпадает с ним. Это, в частности, относится к оператору потока G_t из (2.2.4).

Пусть теперь $G_{1t}, G_{2t}, \dots, G_{kt}$ - некоторые динамические потоки вида (2.2.4). Рассмотрим их линейную комбинацию с неотрицательными весами, в сумме дающими единицу:

$$F(t) = \sum_{l=1}^k p_l G_{lt}, \quad p_l \geq 0, \quad \sum_{l=1}^k p_l = 1. \quad (2.2.29)$$

Естественно, оператор $F(t)$ в (2.2.29) в общем случае не является полугруппой. Однако полугруппой является оператор, эквивалентный ему по Чернову. Именно этот оператор и является предельным состоянием конечнократных аппроксимаций подходящего динамического потока, предположительно стоящего за уравнением Лиувилля для нестационарных ВПФР. Докажем это утверждение.

Теорема 2.2. Пусть $G_{lt}, l=1, 2, \dots, k$ - гладкие динамические потоки вида (2.2.4), $v_l(x)$ - соответствующие скорости из (2.2.5), и $\{p_l\}_{l=1}^k$ - неотрицательные действительные числа с единичной суммой. Тогда математическое ожидание MF потока (2.2.29) эквивалентно по Чернову потоку $\tilde{G}_t$, которому отвечает средняя скорость

$$\tilde{v}(x) = \sum_{l=1}^k p_l v_l(x).$$

Доказательство. Рассмотрим действие среднего потока на фазовую точку $x(0)$. Согласно (2.2.4) имеем $G_{lt}x(0) = x(0) + v_l t + o(t)$. Из (2.2.28) получаем

$$F(t)x(0) = x(0) \sum_{l=1}^k p_l + t \sum_{l=1}^k p_l v_l + o(t) = x(0) + \tilde{v}t + o(t). \quad (2.2.30)$$

Поскольку оператор $F(t)$ удовлетворяет условиям теоремы Чернова, то для него предел (2.1.1) существует. Обозначим его $\tilde{G}_t$. Из (2.2.30) следует, что $\tilde{G}_0 = I$ и

$$\frac{d}{dt} \tilde{G}_t x|_{t=0} = \tilde{v}(x). \text{ Теорема 2.2 доказана.}$$

В сущности, теорема 2.2 является прямым следствием из доказанной в [27] для операторных функций теоремы об эквивалентном по Чернову усреднении. Однако ее методологическая роль в обосновании корректности построения эвристических динамических моделей для аппроксимаций временных рядов в задачах математической

статистики ранее не была выявлена. Практическое применение этой конструкции демонстрируется следующим утверждением.

Теорема 2.3. Пусть начальная выборочная плотность функции распределения строго положительна во всех внутренних классовых интервалах. Тогда средняя по ансамблю скорость микроразвития по уравнению Лиувилля равна скорости макроэволюции, деленной на число шагов, в окне, равном длине выборки.

Доказательство. Пусть в каждый момент t известна ВПФР в виде положительной гистограммы $f_N(j, t)$, разбитой по n классовым интервалам, по выборке длины N . Вычислим вероятность реализации определенной траектории временного ряда на горизонте, равном длине выборки. В момент времени $t+1$ наблюдается значение $x(t+1)$, которое с вероятностью $f_N(j, t+1)$ попало в j -ый кластер. Если значение $x(t-N+1)$ тоже попало в j -ый кластер, то $f_N(j, t+1)$ будет такой же, как и $f_N(j, t)$. Вероятность этого события (т.е. того, что на одном шаге по времени ВПФР не изменится), равна

$$p_N(t+1) = \sum_{j=1}^n f_N(j, t+1) f_N(j, t-N+1).$$

Если $x(t+1)$ и $x(t-N+1)$ попали в разные кластеры (соответственно j и $k \neq j$), то в кластере j ВПФР увеличится, а в кластере k уменьшится на $1/N$. Вероятность такого события равна

$$q_N(t+1) = \sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq j}}^n f_N(j, t+1) f_N(k, t-N+1) = 1 - p_N(t+1).$$

Тогда в момент времени $t+1$ ВПФР, построенная в скользящем окне длины N , выражается через ВПФР на предыдущем шаге по формуле:

$$f_N(i, t+1) = f_N(i, t)(1 - \delta_{ij})(1 - \delta_{ik}) + \left(f_N(i, t) + \frac{1}{N}\right)\delta_{ij} + \left(f_N(i, t) - \frac{1}{N}\right)\delta_{ik}.$$

Соответствующая скорость $u_N(i, t)$ в i -ой ячейке определяется из (2.2.24):

$$f_N(i, t+1) - f_N(i, t) = u_N(i, t) f_N(i, t) - u_N(i+1, t) f_N(i+1, t) = \frac{1}{N}(\delta_{ij} - \delta_{ik}).$$

Тогда уравнения для $u_N(i, t)$ рекуррентно разрешимы по формулам (2.2.28):

$$\begin{aligned}
u_N(0, t) &= 0; \\
u_N(1, t) &= \frac{u_N(0, t) f_N(0, t) - (\delta_{0j} - \delta_{0k})/N}{f_N(1, t)} = 0; \\
u_N(2, t) &= \frac{\delta_{1k} - \delta_{1j}}{N f_N(2, t)}; \dots; \quad u_N(i, t) = \frac{\sum_{m=1}^{i-1} (\delta_{mk} - \delta_{mj})}{N f_N(i, t)}.
\end{aligned}$$

Средняя скорость микроэволюции ВПФР по нестационарному выборочному распределению вероятностей, которое реализовалось на заданном промежутке времени, равна:

$$\begin{aligned}
\langle u_N(i, t) \rangle &= \frac{\sum_{m=1}^{i-1} \sum_{j,k=1}^n f_N(j, t+1) f_N(k, t-N+1) (\delta_{mk} - \delta_{mj})}{N f_N(i, t)} = \\
&= \frac{\sum_{m=1}^{i-1} (f_N(m, t-N+1) - f_N(m, t+1))}{N f_N(i, t)} = \frac{F_N(i-1, t-N+1) - F_N(i-1, t+1)}{N f_N(i, t)} = \frac{v(i, t)}{N}.
\end{aligned}$$

Теорема 2.3 доказана.

В силу доказанной выше теоремы 2.2 эта средняя скорость имеет смысл для конечнократной аппроксимации по Чернову подходящего динамического потока (2.2.5).

Используем теперь концепцию средней скорости микроэволюции для модификации численной схемы решения уравнения Лиувилля, чтобы во всяком случае аппроксимация этой эволюции была бы устойчивой на горизонте, равном окну выборки.

Пусть скорость макроэволюции $v(x)$ найдена. Тогда будем аппроксимировать не одношаговую микроэволюцию, а эволюцию ВПФР от момента времени $t - N$ до момента $t - N + k$. Введем соответствующую скорость для гипершага длины k в модели линейной аппроксимации по времени:

$$u(x, k) = v(x) \frac{k}{N}. \quad (2.2.31)$$

Из (2.2.25) тогда следует, что

$$\begin{aligned}
f(i, t - N + k) &= f(i, t - N) (1 + u(i, k)) - u(i + 1, k) f(i + 1, t - N) = \\
&= f(i, t - N) + \frac{k}{N} (v(i) f(i, t - N) - v(i + 1) f(i + 1, t - N)).
\end{aligned} \quad (2.2.32)$$

Поскольку $1 \leq k \leq N$ и при $k = N$ из уравнения (2.2.32) получается наблюдаемая эволюция от $f(i, t - N)$ к положительной функции $f(i, t)$, то и на промежуточных этапах получается положительная ВПФР, причем условия нормировки выполняются автоматически. Получим уравнение для микроэволюции. Соответствующая

микроскорость $w(i, k)$ должна описывать переход ВПФР от момента k к моменту $k+1$ согласно (2.2.25), т.е.

$$f(i, t - N + k + 1) = f(i, t - N + k)(1 + w(i, k)) - w(i + 1, k)f(i + 1, t - N + k). \quad (2.2.33)$$

С другой стороны, в любой момент времени k ВПФР $f(i, t - N + k)$ определяется по скорости $kv(i)/N$ через одну и ту же функцию $f(i, t - N)$. Это дает возможность определить скорость $w(i, k)$. Из (2.2.31), (2.2.33) получаем:

$$\begin{aligned} w(0, k) &= 0; \\ w(1, k) &= -\frac{1}{N} \frac{f(0, t) - f(0, t - N)}{f(1, t - N) + \frac{k}{N}(f(1, t) - f(1, t - N))} = 0; \\ w(i, k) &= -\frac{1}{N} \frac{F(i - 1, t) - F(i - 1, t - N)}{f(i, t - N) + \frac{k}{N}(f(i, t) - f(i, t - N))}, \quad 1 < i \leq n. \end{aligned} \quad (2.2.34)$$

Знаменатель в выражении (2.2.34) для $w(i, k)$ всегда положителен (если только ВПФР не обращается в какой-либо ячейке в ноль, что исключено по теореме 2.1), и тогда корректно определена микроскорость в соответствующем уравнении Лиувилля (2.2.33). Построенная разностная схема позволяет получить аппроксимацию решения уравнения Лиувилля на заданном временном горизонте в задаче с закрепленными концами, то есть при заданных начальном и конечном распределениях. Это решение дается нелинейной аппроксимацией плотности функции распределения по времени в виде:

$$\begin{aligned} f(i, t - N + k + 1) &= f(i, t - N + k) + \\ &+ \frac{f(i, t - N + k)}{N} \frac{F(i - 1, t - N) - F(i - 1, t)}{\left(1 - \frac{k}{N}\right)f(i, t - N) + \frac{k}{N}f(i, t)} - \\ &- \frac{f(i + 1, t - N + k)}{N} \frac{F(i, t - N) - F(i, t)}{\left(1 - \frac{k}{N}\right)f(i + 1, t - N) + \frac{k}{N}f(i + 1, t)} \end{aligned} \quad (2.2.35)$$

Скорость $w(i, k)$ можно трактовать как дискретную динамическую систему, непрерывный аналог которой дается уравнением:

$$\frac{dx}{dt} = w(x, t) = \frac{F(x) - G(x)}{(N - t)F'(x) + tG'(x)}, \quad (2.2.36)$$

где $F(x)$ и $G(x)$ представляют начальную и конечную функции распределения соответственно, N есть горизонт аппроксимации.

Можно заметить, что модель (2.2.36) представляет собой линейное дифференциальное уравнение первого порядка относительно функции $t(x)$. Его решение имеет следующий вид:

$$t(x) = N \frac{F(x) - F(x(0))}{F(x) - G(x)}, \quad (2.2.37)$$

где $x(0)$ есть начальное значение координаты в момент времени $t = 0$.

Соответствующая дискретная динамическая система первого порядка имеет вид:

$$x(t+1) - x(t) = w(x(t), t). \quad (2.2.38)$$

Из такой интерпретации следует, что N величин $x(k)$ образуют фрагмент временного ряда с ВПФР, отвечающей определенной траектории динамической системы, реализующейся с вероятностью $f(x(0))$. Ансамбль траекторий, усредненный по начальному состоянию с ВФР $F(x(0))$, дает ВФР $G(x)$ конечного состояния. Однако такой подход к моделированию временного ряда как динамической системы представляет интерес только в том случае, если скорость в правой части (2.2.38) отлична от нуля. Из (2.2.37) следует, что для этого требуется, чтобы начальное и конечное распределения не перекрывались. В противном случае точка пересечения кривых ВФР будет отвечать переходу динамической системы к тривиальной форме постоянной. Следовательно, класс таких временных рядов сравнительно беден, но не пуст. В нем, в частности, содержатся примеры монотонных по времени трендов.

2.3. Эквивалентный по Чернову согласованный уровень значимости

С уравнением Лиувилля, рассмотренным выше, естественно связан СУС, построенный по совокупности расстояний между эволюционирующими встык-выборками. Если считать, что величина СУС представляет собой характеристику данного случайного процесса, то изменение его можно было бы трактовать как разладку. Практический вопрос, возникающий в таком случае, состоит в длине окна, где считается этот индикатор.

Пусть по длине L_{tot} фактически известного фрагмента временного ряда построен СУС как решение уравнения (1.1.8)

$$G_N(\rho) = 1 - \rho \quad (2.3.1)$$

для встык-выборок длины N . Следует учесть, что $L_{tot} \gg 2N$, так что на длине L_{tot} укладывается достаточно много длин N , чтобы можно было набрать статистику для построения распределения $G_N(\rho)$. На этой длине L_{tot} в доле случаев $\rho^*(N)$ расстояние между встык-выборками больше, чем $\rho^*(N)$. Идея идентификации разладки состоит в том, что если теперь в определенном скользящем окне длины L доля расстояний между встык-выборками длины N оказалась больше, чем вычисленная ранее величина $\rho^*(N)$, то

в этом окне фиксируется разладка. Для такого вывода требуется, чтобы СУС нестационарного ряда был бы случайной величиной, имеющей стационарное распределение.

Исследуя локальное значение СУС – то есть значение, полученное по промежутку времени L , существенно меньшему, чем первоначальный объем данных L_{tot} , естественно возникает вопрос о флуктуациях этого локального СУС относительно базового значения, отвечающего множеству в целом. Важно понимать, что СУС всего множества $\rho_{tot}^*(N)$ не является средним значением последовательности СУС $\rho_n^*(N, L)$, построенной для встык-выборок длины N в окнах длины L . Рассмотрим функцию

$$\Psi_N(\rho) = 1 - G_N(\rho). \quad (2.3.2)$$

Предположим, что ВФР $G_N(\rho)$ аппроксимирует дифференцируемую функцию распределения $G(\rho)$ соответствующей генеральной совокупности. Поскольку $G_N(\rho)$ монотонно возрастает от 0 до 1, то функция $\Psi_N(\rho)$ обладает свойствами:

$$\Psi_N(0) = 1, \Psi'_N(0) = -a_N \leq 0.$$

В таком случае существует предел

$$\lim_{n \rightarrow \infty} \left(\Psi_N \left(\frac{\rho}{n} \right) \right)^n = e^{-\rho a_N} \equiv \Phi_N(\rho). \quad (2.3.3)$$

Следуя определениям, данным выше для операторных функций, будем называть предельную функцию $\Phi_N(\rho)$ эквивалентной по Чернову функции $\Psi_N(\rho)$ (то есть уровню значимости распределения $G_N(\rho)$). Будем обозначать эту эквивалентность как

$$\overset{Ch}{\Phi_N(\rho) \propto \Psi_N(\rho)}. \text{ Очевидно, } |\Phi_N(\rho) - \Psi_N(\rho)| = o(\rho).$$

Согласно теореме об эквивалентности по Чернову средней полугруппы, если имеется (конечный или бесконечный) набор функций $\Psi_N^k(\rho)$ вида (2.3.2), каждая из которых эквивалентна в смысле (2.3.3) функции $\Phi_N^k(\rho)$ с коэффициентом $-a_N^k$ в показателе экспоненты, и если задан соответствующий набор неотрицательных коэффициентов p_k таких, что $\sum_k p_k = 1$, то средняя функция $\bar{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho)$ эквивалентна в смысле (2.3.3) функции $\bar{\Phi}_N(\rho) = e^{-\rho \bar{a}_N}$, где

$$\bar{a}_N = \sum_k p_k a_N^k. \quad (2.3.4)$$

Рассмотрим далее последовательность непересекающихся промежутков длины L . Для каждого k -го промежутка построим эмпирическое распределение $G_N^k(\rho, L)$ расстояний между встык-выборками длины N . Пусть таких промежутков n , так что $nL = L_{tot}$ есть полная длина изучаемого фрагмента ряда. Тогда распределение расстояний, построенное по всему объему данных, есть среднее распределение, полученное усреднением по отдельным выборкам:

$$G_N(\rho, L_{tot}) = \frac{1}{n} \sum_{k=1}^n G_N^k(\rho, L),$$

$$\bar{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho) = 1 - G_N(\rho, L_{tot}), \quad p_k = \frac{1}{n}. \quad (2.3.5)$$

Пусть $\rho_k^*(N)$ есть стационарная точка для функции $\Psi_N^k(\rho)$, а $\tilde{\rho}_k(N)$ - для эквивалентной ей функции $\Phi_N^k(\rho)$. В силу того, что $a_N^k \geq 0$, эта стационарная точка единственная. Обозначим через $\tilde{\rho}(N)$ стационарную точку эквивалентной средней функции $\bar{\Phi}_N(\rho) = e^{-\rho \bar{a}_N}$. С точностью до $o(\rho)$ имеем из (2.3.3), (2.3.4):

$$\tilde{\rho}_k(N) = \frac{1}{1 + a_N^k}, \quad \tilde{\rho}(N) = \frac{1}{1 + \bar{a}_N} = \frac{1}{\frac{1}{n} \sum_{k=1}^n \frac{1}{\tilde{\rho}_k(N)}}. \quad (2.3.6)$$

Поскольку эквивалентные функции отличаются от оригиналов на величину $o(\rho)$, то на такую же величину отличаются и соответствующие стационарные точки.

Таким образом, доказана следующая теорема о стационарных точках распределений с непрерывными плотностями.

Теорема 2.4. Пусть распределения случайных величин имеют непрерывные плотности и пусть на множестве этих случайных величин задана некоторая неотрицательная мера. Тогда стационарная точка функции, эквивалентной по Чернову среднему уровню значимости данных распределений, с точностью до бесконечно малой второго порядка совпадает с обратной величиной к среднему значению обратных величин стационарных точек функций, эквивалентных по Чернову уровням значимости данных распределений:

$$\frac{1}{\tilde{\rho}(N)} = \sum_{k=1}^n \frac{p_k}{\tilde{\rho}_k(N)} \equiv \frac{1}{\bar{\tilde{\rho}}(N)}. \quad (2.3.7)$$

Если не ограничиваться линейным приближением в представлении эквивалентной функции (2.3.6), а записать в общем случае решение уравнения $e^{-a\rho} = \rho$ в виде $\tilde{\rho} = R(a)$,

где R есть соответствующий разрешающий оператор для единственной стационарной точки, то вместо (2.3.7) имеем

$$R^{-1}(\tilde{\rho}) = \sum_{k=1}^n p_k R^{-1}(\tilde{\rho}_k). \quad (2.3.8)$$

В силу сказанного ранее соотношения (2.3.7), (2.3.8) с точностью $o(\rho)$ могут быть распространены и на стационарные точки распределений $\Psi_N^k(\rho)$. Формула (2.3.7) принципиальна, поскольку она позволяет значительно уменьшить количество вычислительных процедур, если требуется изучить поведение СУС на множествах, представляющих собой объединение некоторых других множеств, для которых, собственно, и считается СУС. То есть для каждой длины N встык-выборки СУС вычисляется всего один раз в наименьших доступных для этой цели непересекающихся окнах длины L , покрывающих все исследуемое множество данных L_{tot} .

Такой подход существенно сокращает время расчета в задачах анализа Больших Данных. Он позволяет оценивать важную статистическую характеристику нестационарного временного ряда по аппроксимационной процедуре, сходящейся к точному значению с гарантированной точностью.

Выводы по Главе II.

Основные результаты Главы II состоят в теоремах 2.1-2.4, обосновывающих применение понятия эквивалентных по Чернову средних полугрупп для создания моделей нестационарных временных рядов и для сокращения объема вычислительных процедур при расчете статистик, связанных с определением многомерного согласованного уровня стационарности.

Построена устойчивая разностная схема аппроксимации решения уравнения Лиувилля посредством использования среднего динамического потока и доказано, что этот поток эквивалентен по Чернову полугруппе, реализующей локальное по времени состояние нестационарного временного ряда.

Построен эквивалентный по Чернову согласованный уровень стационарности как стационарная точка уровня значимости распределения расстояний между функциями распределения.

Результаты Главы II опубликованы в работах [15, 16, 17, 55, 56, 57, 58].

Глава III. Численный алгоритм моделирования разладки в нестационарных временных рядах и программный комплекс

3.1. Алгоритм построения СУС для анализа нестационарности

В Главе III описывается численный алгоритм для моделирования разладки в нестационарных временных рядах в терминах выборочных распределений и приводится блок-схема реализующего программного комплекса.

Поскольку данные для анализа временных рядов могут быть получены в различных форматах и с использованием различного количества электродов, то для унификации работы с такими данными используется определенный формат (При обработке данных, поступающих в формате edf, использовался модуль библиотеки EDFbrowser, который служит для конвертации исходных файлов в кодировку ASCII (github.com/Teuniz/EDFbrowser))

В начале работы алгоритма либо задаются, либо считываются из исходного файла параметры, характеризующие входные данные, и дополнительные настройки для алгоритма. К таким параметрам, в частности, относятся: общее число используемых в расчетах выборок; количество данных по каждой компоненте; шаг для встык-выборок внутри заданного расчетного окна; длина одной встык-выборки; размер интервала в котором считается одно значение СУС.

На Рис. 3.1 показан фрагмент исходных данных, поступающих на вход в блок статистического анализа. Самый левый столбец представляет из себя временной ряд, значения которого соответствуют определенному моменту времени в который были сделаны измерения. Остальные столбцы являются отдельными выборками, на которых проводятся численные расчеты.

Time	1	2	3	4	5	6	7	8
13363	43.00296	76.29788	17.73898	-17.0617	18.0736	183.0424	107.0831	58.06096
13363	49.02616	67.43039	33.96816	-15.054	4.019467	196.9292	119.6315	69.94006
13363.01	48.85885	60.0687	43.00296	-16.5598	-5.01533	202.1159	135.1914	78.30561
13363.01	40.32598	46.01456	33.96816	-59.2241	-6.85576	182.8751	129.1682	76.79981
13363.02	42.16641	56.55516	34.63741	-59.7261	2.513667	183.8789	124.1489	75.29401
13363.02	48.35692	78.64023	40.82792	-9.53274	24.59874	195.4234	115.9506	92.35975
13363.02	57.72634	100.3907	51.36852	1.677111	24.26412	202.9524	111.4332	94.53479
13363.03	73.45359	109.7601	63.41492	32.29505	20.41596	208.9756	115.2814	90.01739
13363.03	83.49226	108.087	72.1151	85.16537	39.32212	218.0104	130.1721	97.54639
13363.04	67.26308	86.16924	41.49716	83.82688	34.80472	205.1275	122.4758	84.66344

13363.04	47.68767	57.39172	1.007867	73.28628	9.206112	196.26	85.33268	58.89752
13363.04	31.96043	32.62967	-21.7464	65.25534	-9.36542	190.906	77.63637	43.00296
13363.05	31.12387	36.81245	-26.2638	34.30278	-12.0424	199.4389	99.72144	41.16254
13363.05	54.38012	61.07257	-11.3732	4.186778	-0.33062	220.8547	112.2698	45.34532
13363.05	72.95166	55.71861	-2.84029	6.361823	22.08907	218.5123	119.4642	33.96816
13363.06	87.1731	58.73021	10.87922	17.40436	58.89752	216.8392	139.8761	39.15481
13363.06	87.1731	60.40332	17.23705	17.73898	71.78048	196.4273	142.8877	31.79312
13363.07	88.84621	75.79595	27.27572	31.29118	81.1499	187.2252	155.2687	9.206112
13363.07	108.087	104.2388	60.0687	62.24374	103.0677	212.1545	179.1942	21.0852

Рис. 3.1 – Пример форматного представления фрагмента рядов данных для алгоритма

Численный алгоритм реализован на языке C++ в среде разработки MSVS 2017.

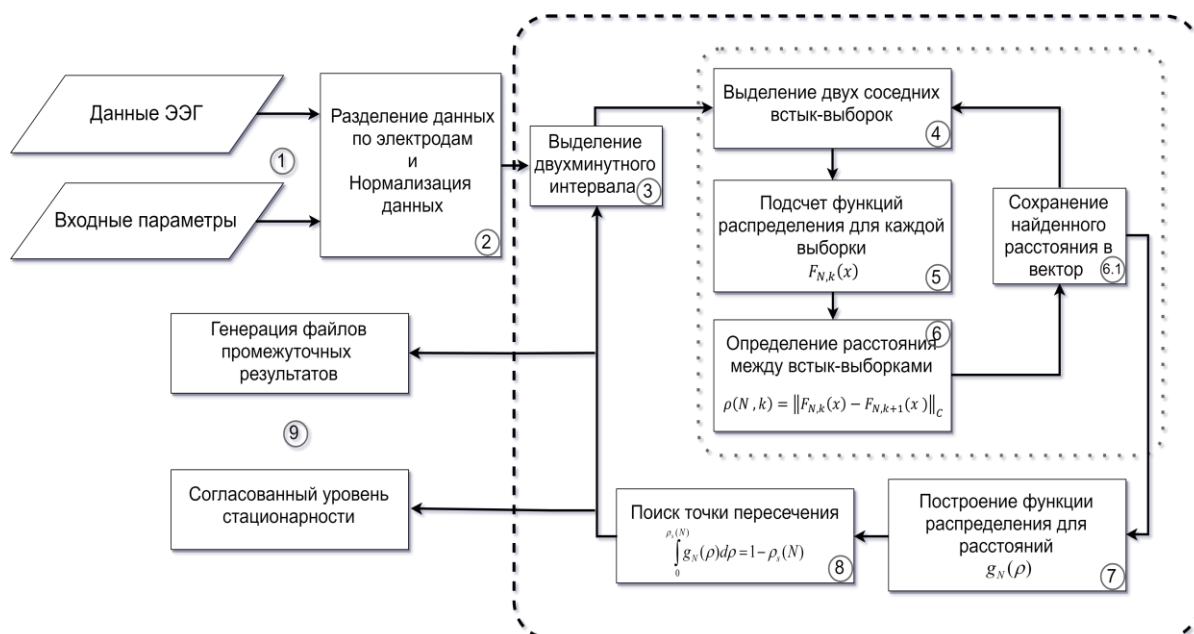


Рис. 3.2 – Блок-схема вычислительного алгоритма
для согласованного уровня стационарности

Численный алгоритм, реализующий построение СУС для нестационарного временного ряда в зависимости от длины встык-выборки $2N$, длины фрагмента ряда L и размерности d случайной величины, устроен следующим образом.

Шаг 1. Читается файл исходных данных. Определяется длина общей выборки и количество столбцов данных.

Шаг 2. Для того, чтобы полученные результаты можно было сравнивать между собой происходит нормализация данных на отрезке $[-1, 1]$.

Шаг 3. Читается массив упорядоченных данных $x_i(n)$, $i = 1, 2, \dots, d$; $n = 1, 2, \dots, L_{tot}$.

Определяется точность ε различения элементов ряда. Этой точностью будет определяться количество классовых интервалов

$$K_{\text{int}} = [1/\varepsilon] + 1, \quad (3.1.1)$$

т.е. мелкость расчетной сетки при вычислении ВФР. Точность либо задается пользователем из практических соображений, либо, если она не задана, определяется следующим образом. Для каждой компоненты ряда рассматривается последовательность модулей первых разностей $v_i(n) = |x_i(n+1) - x_i(n)|$. Дальнейшее зависит от вида распределения этих величин. Если оно унимодально, в качестве точности берется отношение моды к размаху ряда. Если же распределение монотонно или не имеет выраженной моды, берется его среднее значение и делится на размах ряда.

Задаются параметры: длина выборки N , шаг смещения по времени τ , длина фрагмента ряда L для вычисления статистики по набору встык-выборок, длина окна l для идентификации разладки. Каждый из этих параметров является независимым, так что для каждой i -ой компоненты временного ряда вычисляемая статистика образует 4-мерный массив. Задаются пределы изменения параметров: $N \in [N_{\text{ini}}, N_{\text{fin}}]$ и аналогично для остальных параметров τ, L, l . По умолчанию, если нет специальных указаний, шаг смещения выбирается равным длине выборки: $\tau = N$.

Шаг 4. В векторе данных для каждой компоненты выделяются выборки, соответствующие заданному интервалу, после чего происходит инициализация основного вычислительного модуля алгоритма. Нумеруются элементы ряда, отвечающие конечным элементам X_{ik} выборок длины N , которые сдвигаются с заданным шагом h :

$$X_{i1} = x_i(N); X_{i2} = x_i(N + \tau); \dots; X_{in} = x_i(N + (n-1)\tau).$$

Встык-выборкой с номером k при выбранных значениях параметров N и τ считается множество элементов, заключенных между $x_i(1 + (k-1)\tau)$ и $x_i(2N + (k-1)\tau)$.

Шаг 5. Для каждой компоненты ряда независимо вычисляется ВФР. Область изменения случайной величины разбивается на $K_{\text{int}} = [1/\varepsilon] + 1$ ячеек, после чего вычисляется индикатор принадлежности элемента ряда одной из классовых ячеек для данной выборки. В итоге ВФР $F(k)$, где $k = 1, 2, \dots, K_{\text{int}}$ есть номер классического интервала, для множества элементов $x_1, \dots, x_N$ вычисляется как ступенчатая функция со скачками в точках, равных концам классовых интервалов (в отличие от формального определения ВФР как ступенчатой функции со скачками в точках наблюдаемых значений случайной величины). Каждая новая пара встык-выборок выделяется со сдвигом, определенным на шаге 3, относительно начала предыдущей пары.

Шаг 6. Вычисляется вектор расстояний (т.е. совокупность расстояний, вычисленных в скользящем окне с заданным шагом по времени) между эмпирическими функциями распределения двух встык-выборок, посчитанными на предыдущем шаге, в норме C . Каждое расстояние определяется по формуле

$$\rho = \max_k |F_1(k) - F_2(k)|. \quad (3.1.2)$$

Шаг 7. Численно находится СУС $\rho^*(N)$ для распределения расстояний между встык-выборками по функции распределения $G_N(\rho) = \int_0^\rho g_N(r) dr$ этих расстояний, которая вычисляется с сеточным шагом 0,01. Именно, находится такой номер $k(N)$ ячейки равномерного разбиения отрезка $[0; 1]$, при котором совместно выполняются неравенства

$$k(N): \sum_{k=1}^{k(N)} g_N(k) \geq 1 - 0,01k(N), \quad \sum_{k=1}^{k(N)+1} g_N(k) < 1 - 0,01(k(N) + 1). \quad (3.1.3)$$

Шаг 8. Вычисляется эквивалентная по Чернову функция для уровня значимости $\Psi_N(\rho) = 1 - G_N(\rho)$ и с заданной точностью строится оператор R , отыскивающий ее стационарную точку.

Далее происходит возврат к шагу 3, после чего расчёты повторяются для новой пары встык-выборок другой длины.

Шаг 9. Финалом работы алгоритма являются результирующие значения СУС.

3.2. Блок-схема программного комплекса

Программный комплекс, построенного в диссертации вычислительного алгоритма применительно к конкретным задачам анализа временных рядов, можно представить в виде блок-схемы, изображенной на Рис. 3.3.

Программный комплекс имеет возможность работы в двух режимах: во-первых, подбор оптимальной ширины окна для дальнейшего анализа нестационарного случайного процесса; во-вторых, реализация расчета СУС по экспериментальным данным, с задаваемыми параметрами разбиения выборок данных. Схемы работы каждого из перечисленных выше режимов приведены на Рис 3.4 и 3.5.



Рис 3.3 – Общая схема работы программного комплекса NSSAT

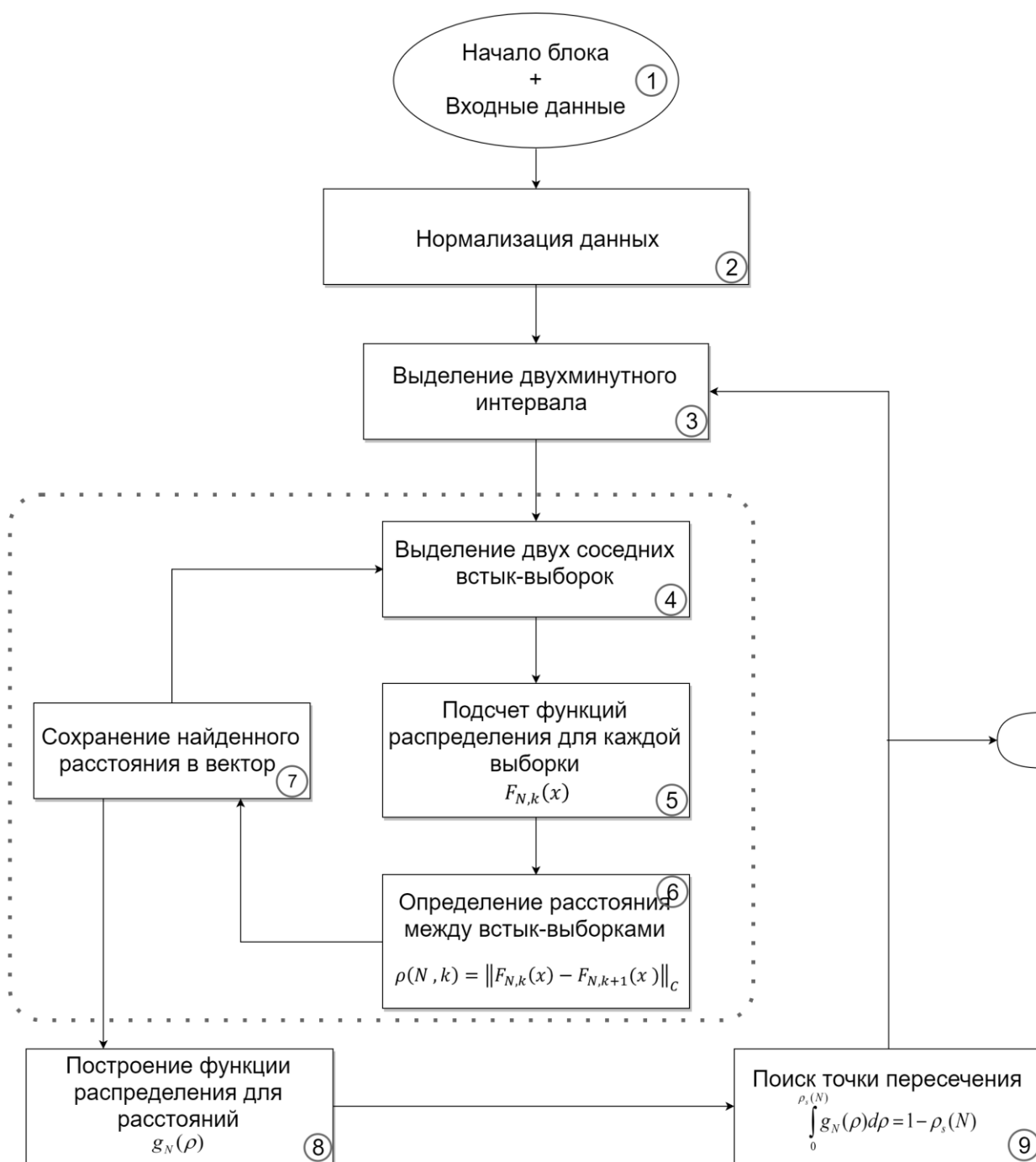


Рис 3.4 – Схема работы расчетного модуля комплекса NSSAT в режиме расчета СУС с задаваемыми параметрами разбиения выборок данных.

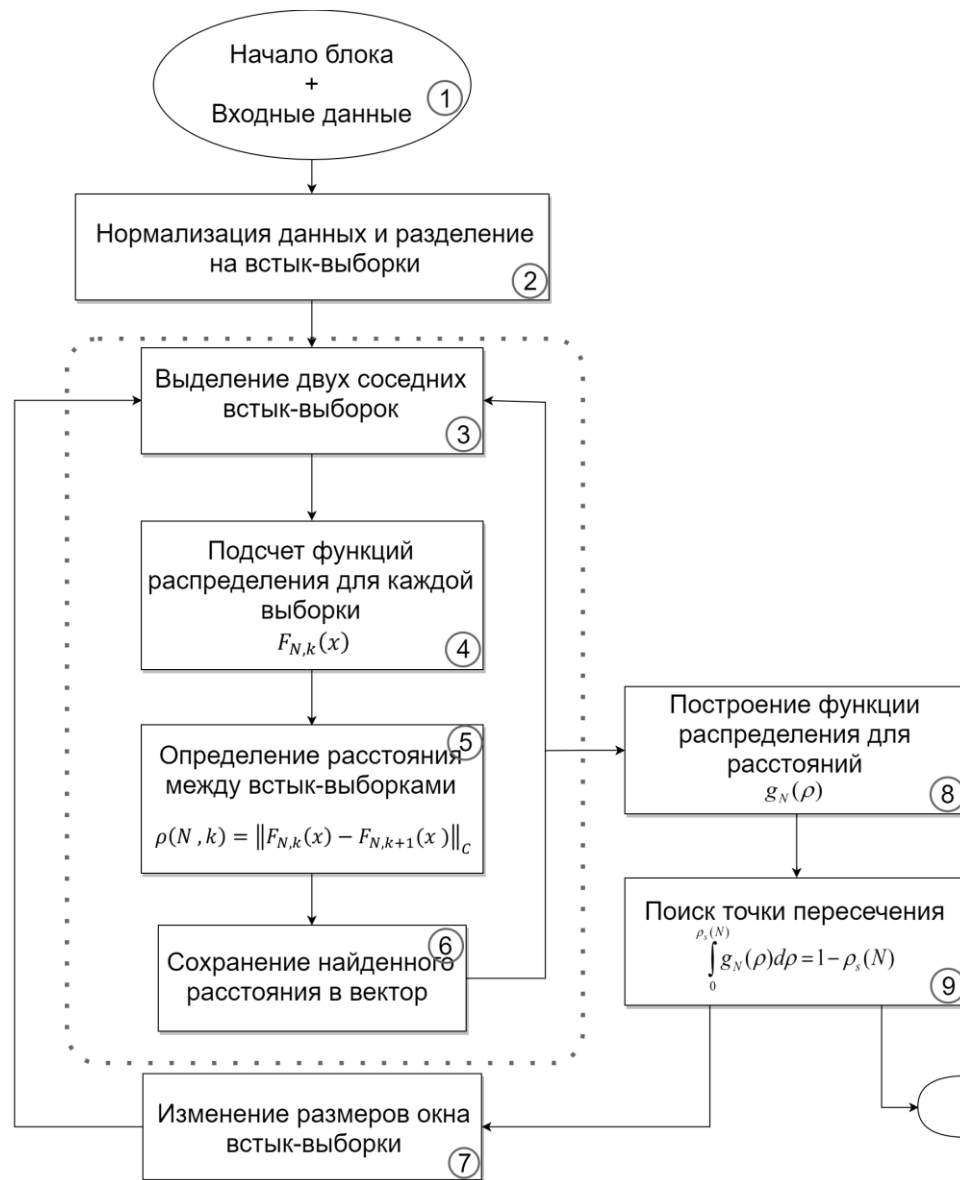


Рис 3.5 – Схема работы расчетного модуля комплекса NSSAT в режиме подбора оптимальной ширины окна для дальнейшего анализа нестационарного случайного процесса

3.3. Распараллеливание вычислений

При обработке большого количества векторов данных применялось распараллеливание вычислений. Была проведена программная оптимизация времени вычисления описанного выше алгоритма с использованием интерфейса OpenMP. Расчеты производились на 2, 4, 6 и 8 параллельных потоках для двух вариантов входных данных, которые представляли собой 64 временных ряда длиной 690000 и 2340000 данных (стандартные размеры файлов данных, используемых в исследовании). Каждому потоку выделялся определенный набор временных рядов, которые в дальнейшем на нем

обрабатывались, и результаты вычислений записывались в отдельный файл. Собранные статистика представлена на Рис. 3.6 (а, б).

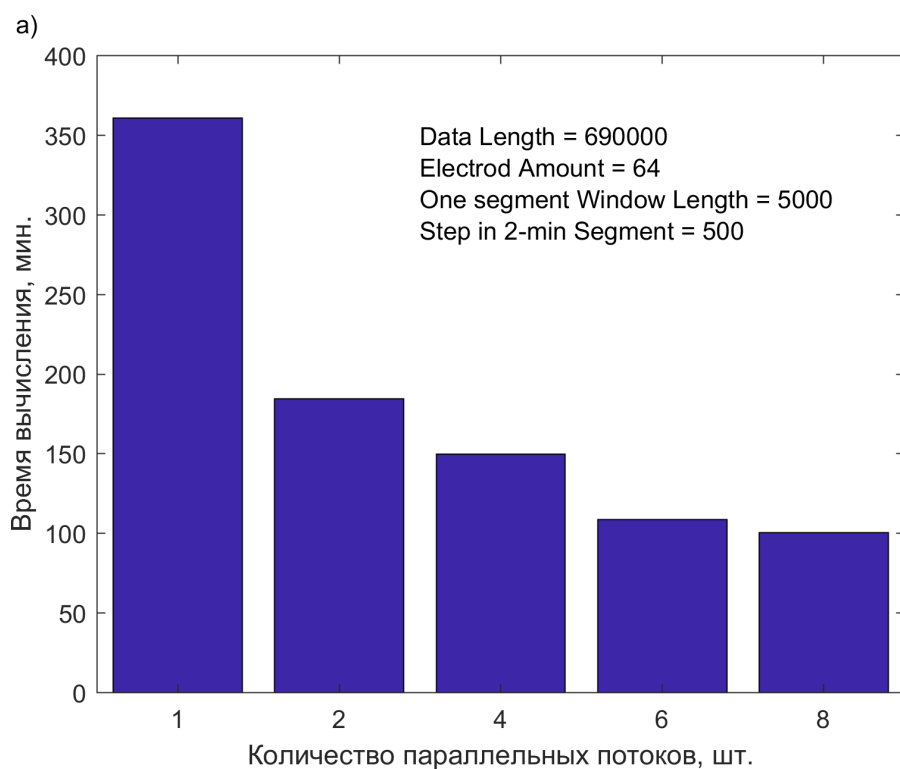


Рис. 3.6 (а) – Время работы алгоритма при разном количестве параллельных потоков для входного файла размером 64x690000 данных

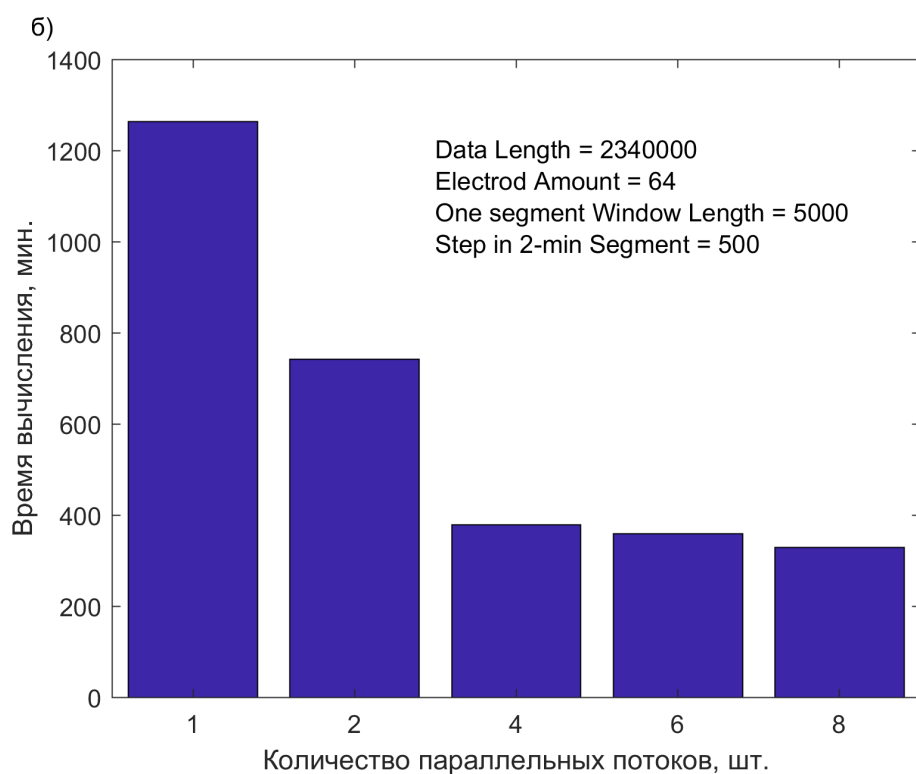


Рис. 3.6 (б) – Время работы алгоритма при разном количестве параллельных потоков для входного файла размером 64x2340000 данных

Тестирование проводилось на системе со следующими характеристиками: Процессор Intel® Core™ i7-2670QM 6 МБ кэш-памяти, 2.20 ГГц, ОЗУ 16 ГБ.

Из Рис. 3.6 (а, б) видно, что с увеличением числа данных качественно картина не изменилась, а время счета оказалось пропорциональным количеству данных (то есть во втором случае в 4 раза больше). Видно также, что на 8 потоках удалось достигнуть ускорения алгоритма примерно в 3,6 раз по сравнению с однопоточной последовательной обработкой.

На Рис. 3.7 представлена зависимость ускорения вычислительного процесса от количества потоков.

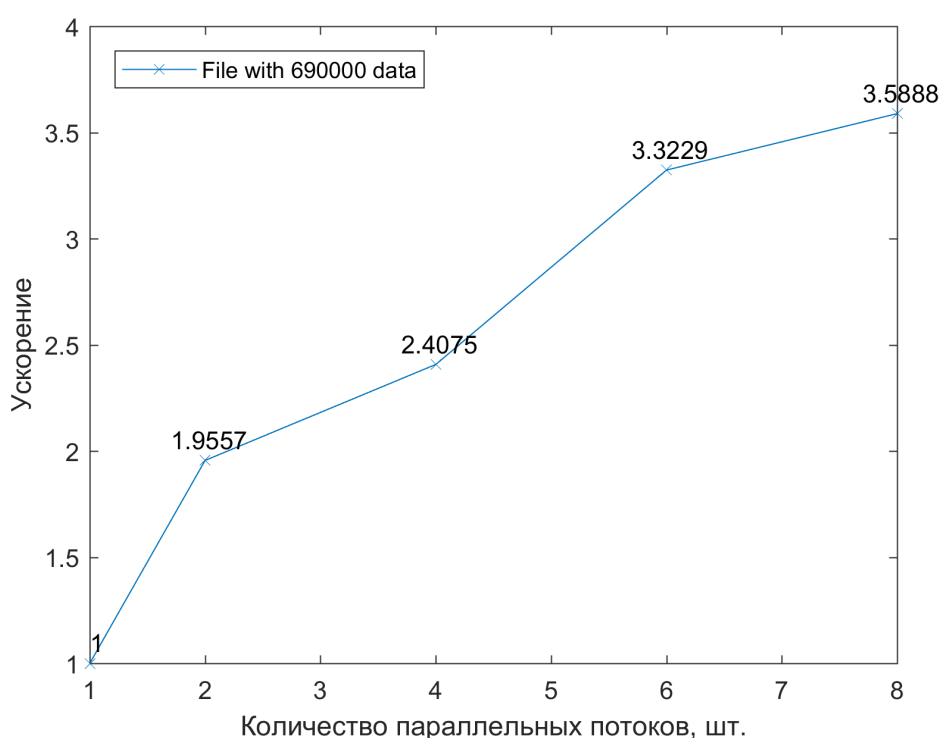


Рис. 3.7 – Ускорение работы алгоритма при разном количестве параллельных потоков

3.4. Программирование пользовательского интерфейса

Поскольку программный код, решающий перечисленный комплекс задач, весьма объемный, а также для целей практического использования разработан интерфейс, позволяющий управлять решением блоков задач. Скрин-шот рабочего окна программы представлен на Рис. 3.8. Интерфейс программного комплекса был написан на языке C++/CLI.

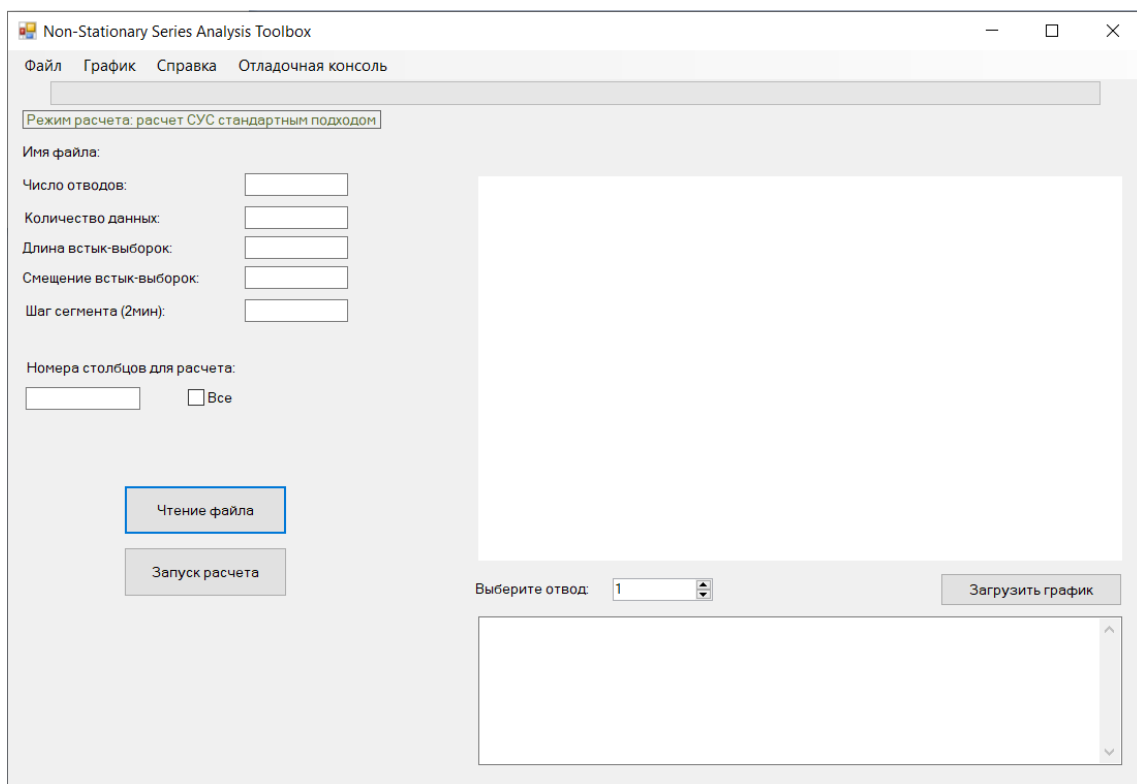


Рис. 3.8 – Скрин-шот рабочего окна программы

В верхней части окна программы располагается стандартный элемент интерфейса программ, предназначенных для работы в операционных системах семейства Windows, строка меню.

В пункте меню «Файл» находятся элементы управления, относящиеся к работе с текущей сессией расчетов (Рис. 3.9).

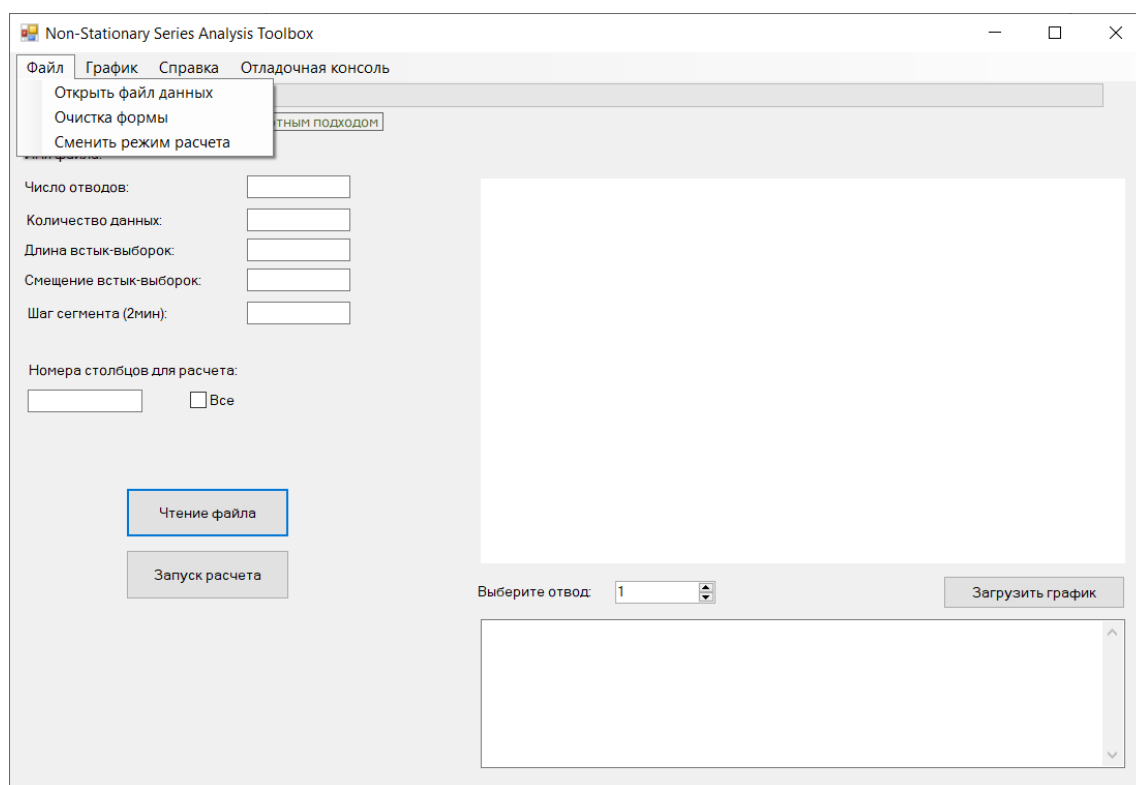


Рис. 3.9 – Скрин-шот рабочего окна программы

В пункте меню «График» находятся элементы управления, относящиеся к работе с графической областью, где отображаются изображения, построенные по полученным результирующим данным (Рис. 3.10).

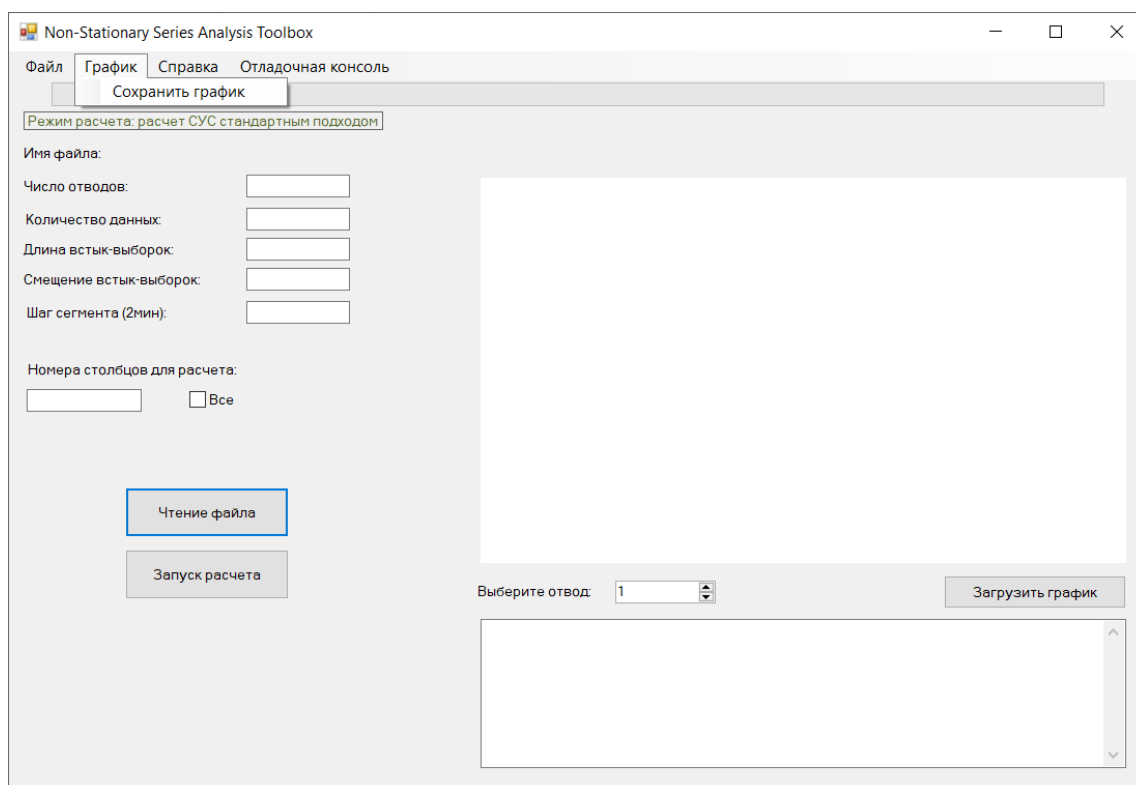


Рис. 3.10 – Скрин-шот рабочего окна программы

В пункте меню «Справка» позволяет получить детальную информацию о каждом элементе управления программного комплекса (Рис. 3.11). Также, в пункте меню «Справка» можно найти информацию о том, в каких форматах должны быть входные данные.

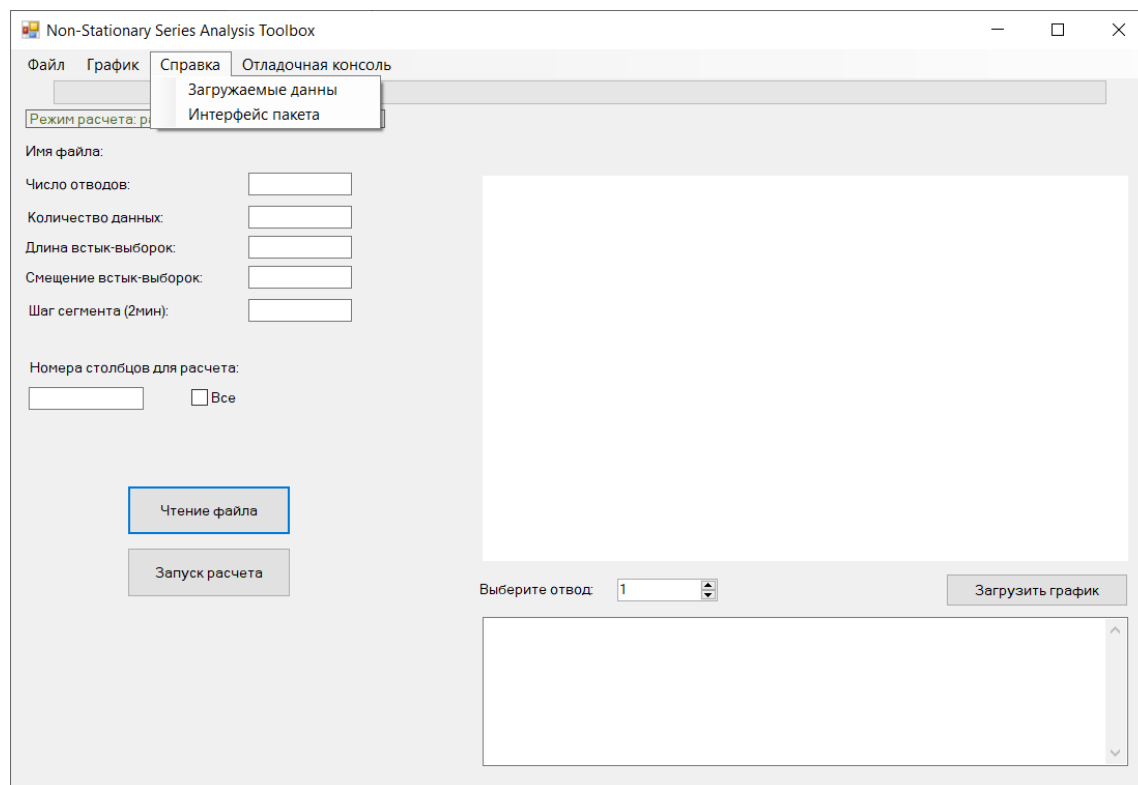


Рис. 3.11 – Скрин-шот рабочего окна программы

В пункте меню «Отладочная консоль» позволяет вызвать командную консоль программного комплекса для детального анализа отладочной информации по происходящим расчетам в программном комплексе (Рис. 3.12).

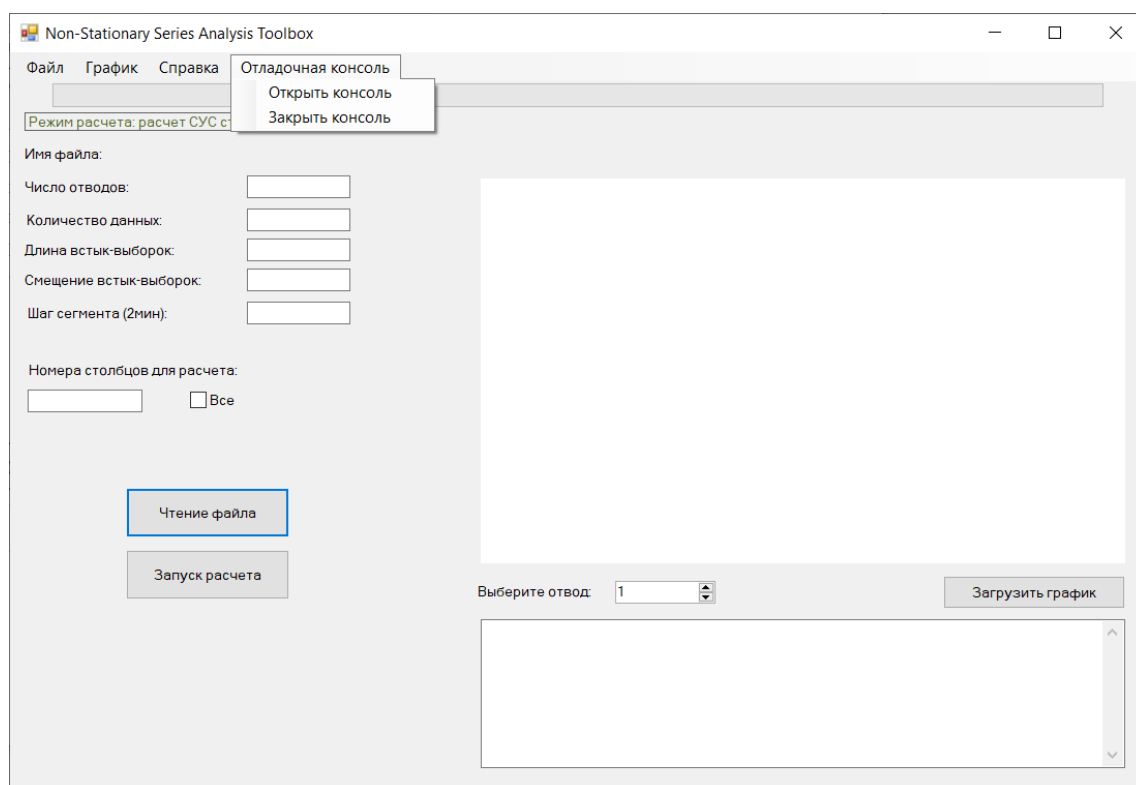


Рис. 3.12 – Скрин-шот рабочего окна программы

При нажатии пункта меню «Файл» >>> «Открыть файл данных», появляется окно, где можно выбрать файл исходных данных определенного формата. После выбора файла происходит подсчет количества временных рядов в файле и расчет их длины, соответственно заполняются значения полей «Число отводов» и «Количество данных» в области задания инициализирующих параметров. Также происходит оценка предполагаемого времени, которое понадобится для расчета, и выводится пользователю, а в информационной области появляется полный путь до открытого файла (Рис. 3.13).

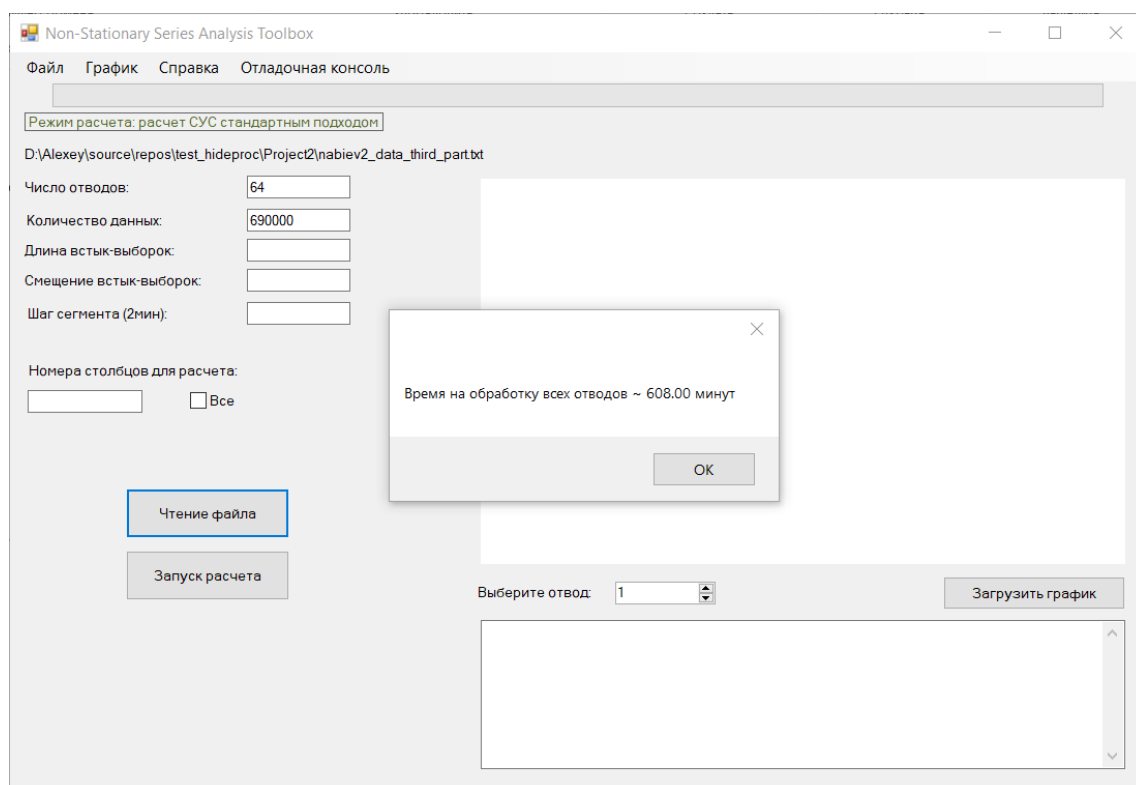


Рис. 3.13 – Скрин-шот рабочего окна программы после выбора файла данных для расчетов

При нажатии пункта меню «Файл» >>> «Очистить форму» происходит очистка памяти приложения, сброс всех настроек и очистка графической области.

При нажатии пункта меню «Файл» >>> «Сменить режим расчета», меняются поля в области задания инициализирующих параметров и алгоритм расчета. Текущий режим отображается в информационной области.

При нажатии пункта меню «График» >>> «Сохранить график», можно сохранить построенный график из результирующих данных.

В информационной области расположен индикатор выполнения, который отображает процесс загрузки файла данных (Рис. 3.14).

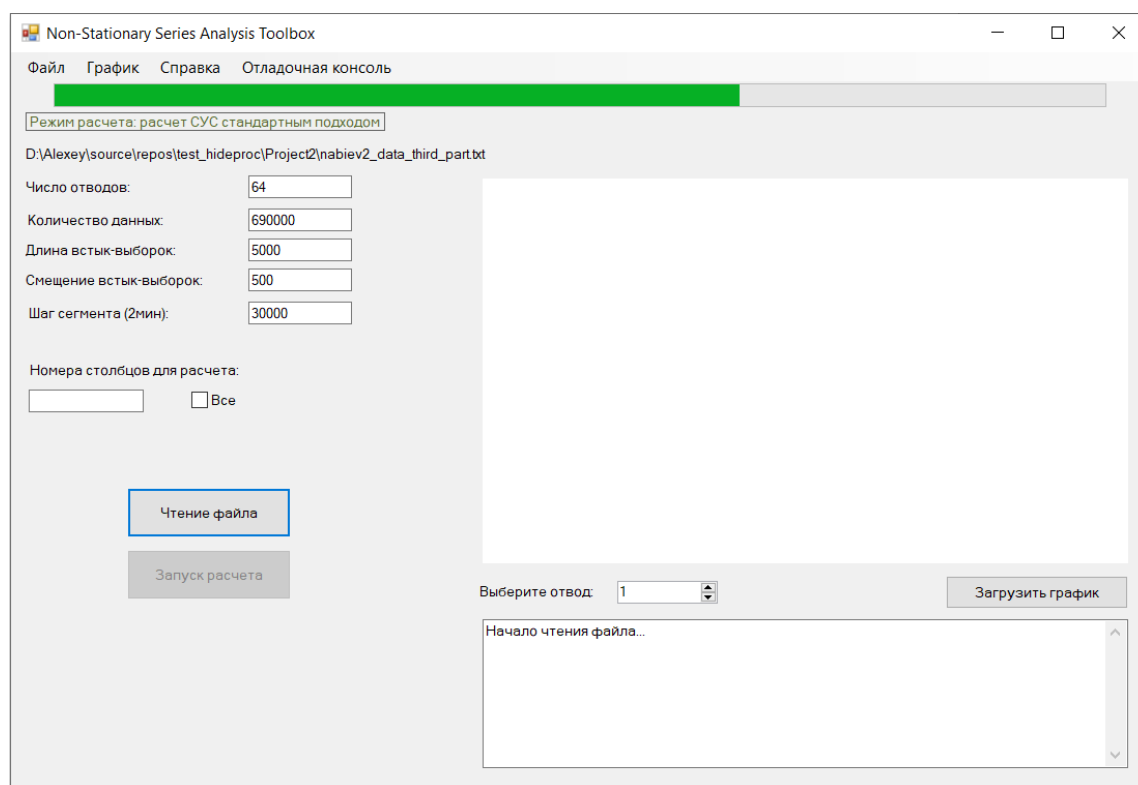


Рис. 3.14 – Скрин-шот рабочего окна программы во время чтения файла данных в память программного комплекса

При работе программы в режиме расчета СУС с задаваемыми параметрами разбиения выборок данных, в область задания инициализирующих параметров доступны для заполнения следующие пункты: число отводов, количество данных, длина встык-выборок, смещение встык-выборок, шаг сегмента (2 мин). Число отводов – количество столбцов данных в исходном файле данных. Количество данных – количество точек в столбце данных; все столбцы данных должны быть одинаковой длины в одном файле. Длина встык-выборки – ширина окна, в котором считается функция распределения. Смещение встык-выборок – выборки могут наслаиваться друг на друга для повышения качества вычисления СУС. Шаг сегмента (2 мин) – размер выборки в котором вычисляется 1 точка СУС.

При работе программы в режиме подбора оптимальной ширины окна для дальнейшего анализа нестационарного случайного процесса, в область задания инициализирующих параметров доступны для заполнения следующие пункты: число отводов, количество данных, шаг увеличения окна, максимальная величина шага. Шаг увеличения окна – шаг, с которым будет увеличиваться ширина окна встык-выборки; изначально ширина окна встык-выборки равна 1000 точек. Максимальная величина шага – задает максимальное значения ширины окна встык-выборки.

В область задания инициализирующих параметров можно также задавать специфичные номера столбцов файла с исходными данными для расчета или можно указать, что требуется произвести расчет для всех столбцов. Заданные номера разделяются пробелом. Если в окне не указывать никаких значений, то будет происходить расчет для всех столбцов данных.

Кнопка «Чтение файла» используется для того, чтобы загрузить файл с исходными данными в память программного комплекса для дальнейших расчетов.

Кнопка «Запуск расчета» позволяет после заполнения всех необходимых полей в области задания инициализирующих параметров запустить расчет. Если какие-то поля были не заполнены, то программа сообщит пользователю об этом (Рис. 3.15).

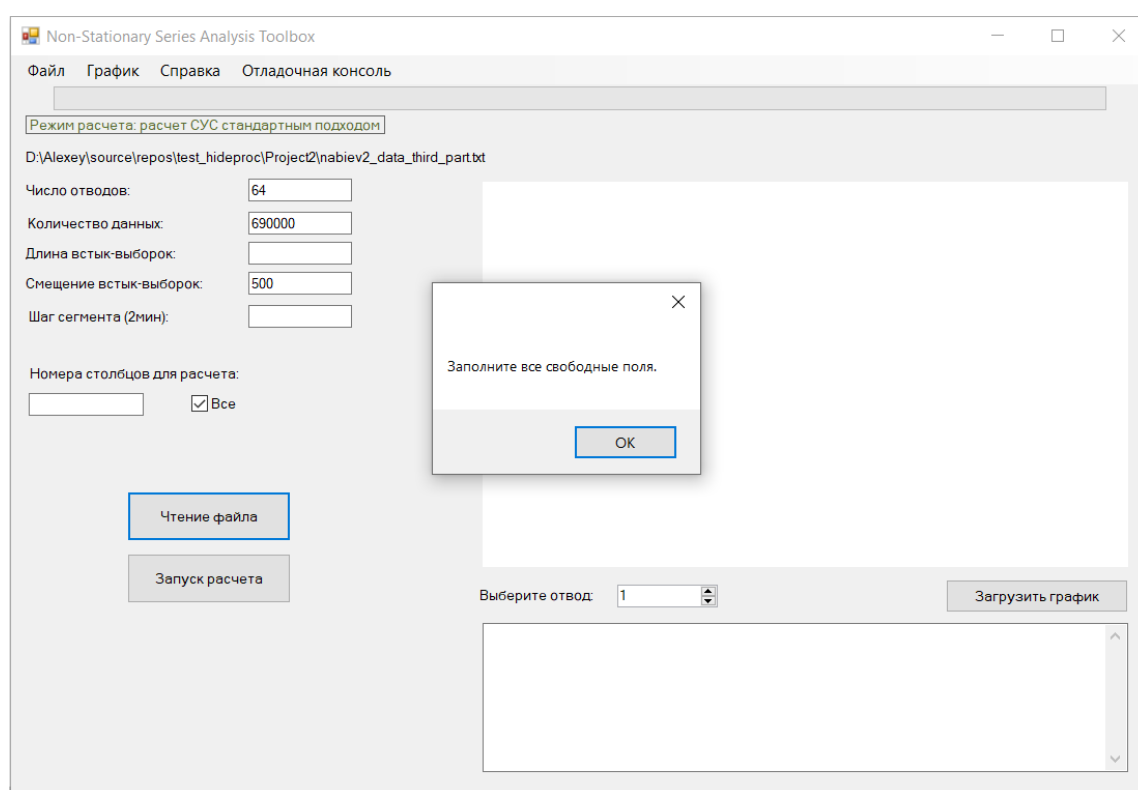


Рис. 3.15 – Скрин-шот рабочего окна программы при начале расчетов с незаполненными полями в области задания инициализирующих параметров

Программный комплекс может также проводить визуализацию полученных результатов. После проведенных расчетов в папке сессии будут находиться результирующие файлы по каждому отводу, для того чтобы посмотреть графическое представление данных необходимо, при помощи стрелочек, выбрать номер отвода и нажать кнопку «Загрузить график». Для режима расчета СУС с задаваемыми параметрами разбиения выборок данных, график будет выглядеть следующим образом – Рис. 3.16. В данном режиме необходимо, чтобы поле «Число отводов» было заполнено.

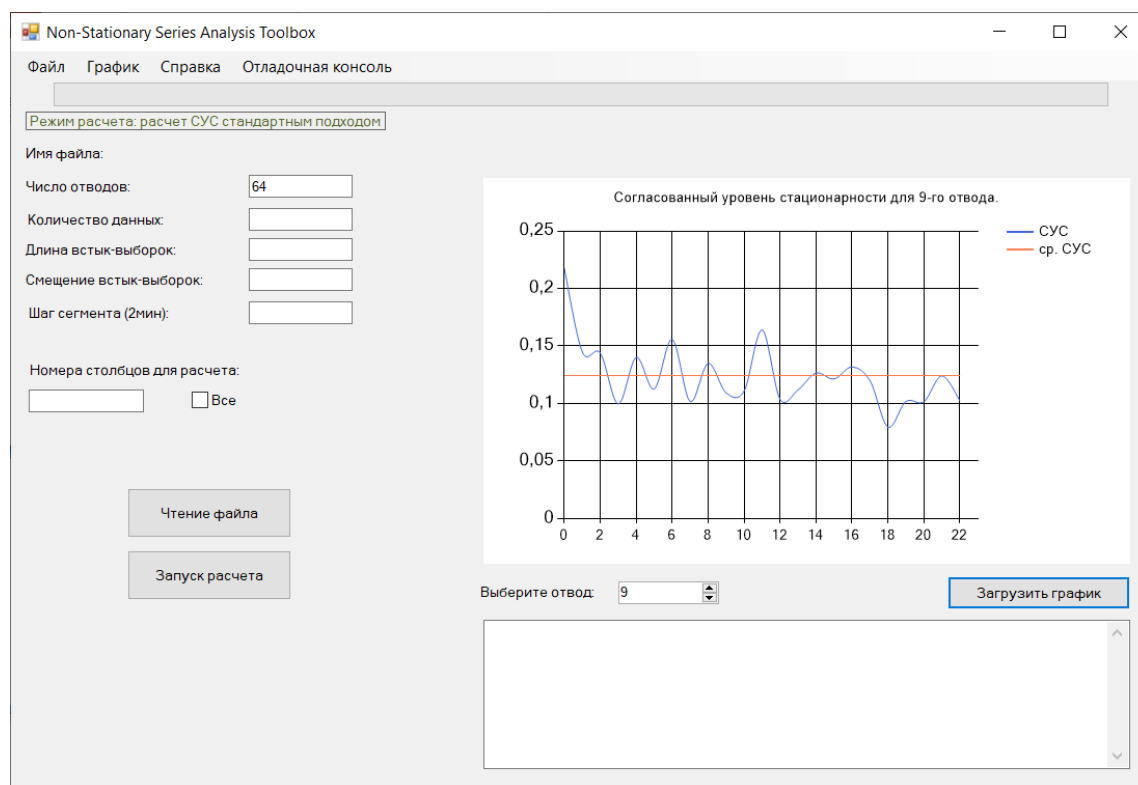


Рис. 3.16 – Скрин-шот рабочего окна программы в режиме расчета СУС с задаваемыми параметрами разбиения выборок данных

При работе в режиме подбора оптимальной ширины окна для дальнейшего анализа нестационарного случайного процесса, рабочее окно программы будет иметь следующий вид – Рис. 3.17. В данном режиме необходимо, чтобы поля «Число отводов» и «Шаг увеличения окна» были заполнены.

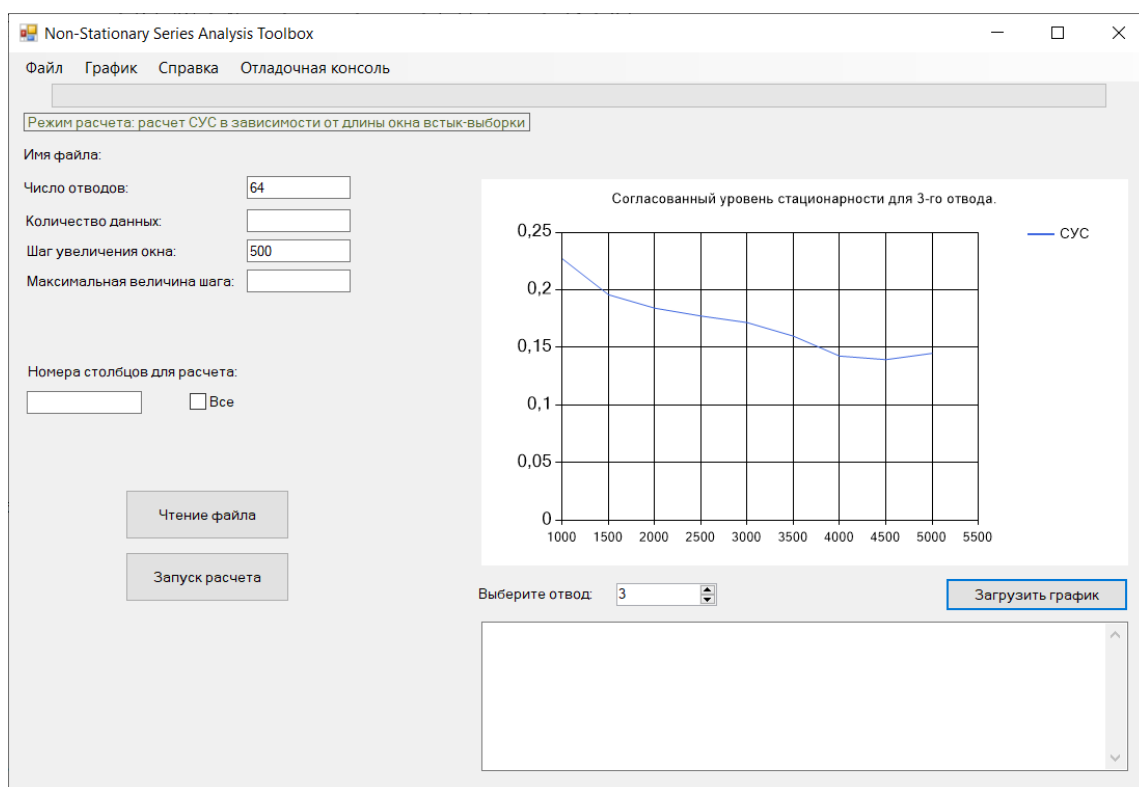


Рис. 3.17 – Скрин-шот рабочего окна программы в режиме подбора оптимальной ширины окна

В области информирования пользователя выводится информация о текущем состоянии программного комплекса, как это показано на Рис. 3.13.

Вычислительный алгоритм и пользовательский интерфейс написаны на языке C++/CLI в среде разработки MS Visual Studio 2017. Для работы программного комплекса необходима операционная система Windows 7 или более новые версии с поддержкой ASP.NET Core Runtime 2.1.509; 10 Mb свободного места на диске.

Выводы по Главе III.

Разработан программный комплекс, решающий многомерную параметрическую задачу вычисления СУС, а также реализован интерфейс пользователя, позволяющий работать с программой в режимах исследования и статистического анализа нестационарных временных рядов.

Результаты Главы III опубликованы в работах [14, 17, 52].

Глава IV. СУС как предиктор приступа эпилепсии при анализе ЭЭГ

4.1. Гипотеза разладки распределения СУС как предиктора приступа эпилепсии

Построение методики предсказания приступа эпилепсии является одной из актуальных задач практической нейрохирургии. Эта задача в настоящее время не имеет ясного решения в виде установления причинно-следственных связей, поэтому применяются различные статистические методы, основанные на анализе данных, получаемых с отведений (электродов) при снятии электроэнцефалограмм (далее ЭЭГ).

Предсказание приступа эпилепсии по данным ЭЭГ является довольно сложной задачей, затрагивающей различные области науки: нейрофизиологию, математику, физику, статистику и машинное обучение.

Эпилепсия сама по себе является уникальным заболеванием, в котором процентное количество времени, когда проявляются симптомы, очень мало, но может играть жизненно важную роль для пациента. Например, в самом продолжительном наблюдении, выполненном в NeuroVistaCorporation (SeattleWA, USA), времени, когда пациент находится в состоянии приступа, соответствует менее 0,05 % от общего времени наблюдения [47]. Большую часть времени (~ 99,95 %), когда больные не находятся в состоянии приступа, они живут относительно нормальной, повседневной жизнью. Однако эпилептический приступ может начаться в любой момент их жизни: в транспорте, во время работы/учебы, занятиями физическими упражнениями и т.д., что может подвергнуть жизнь пациента опасности.

Существуют различные способы предотвращения эпилептических приступов.

Одним из них является прием антиэпилептических препаратов, которые предназначены для смягчения приступов эпилепсии [65], воздействуя на химию, управляющую клеточными процессами, лежащими в основе эпилепсии. Однако они оказывают большое влияние на организм, которое не всегда может быть предсказуемо.

В тех случаях, когда препараты оказываются малоэффективными, применяется хирургическое вмешательство с последующей резекцией области возникновения эпилепсии, чтобы поврежденные нейроны больше не участвовали в сетевой активности. Несмотря на свою радикальность, данный метод имеет очень древнее происхождение [72]

Если же вышеприведенные способы не могут помочь справиться с приступами эпилепсии, тогда в мировой практике применяются электрические стимуляторы для

прерывания или контроля судорог, вызванных приступом. Электрическая стимуляция производится регулярно через определенные промежутки времени. Последние же исследования в этой области нацелены не на контролирование приступа посредством постоянного стимулирования, а на прерывание в момент его зарождения. Проблема заключается в том, что для этой цели необходимо весьма точно определять или предсказывать момент наступления приступа [74].

Биоэлектрическая активность мозга во время приступа эпилепсии резко усиливается, поэтому использование записи ЭЭГ для определения «разладки» между нормальным состоянием и приступом выглядит вполне естественным. Однако поскольку целью является не констатация разладки, а ее предсказание, задача существенно усложняется. Выясняется, что непосредственная связь между характеристиками сигнала ЭЭГ в нормальном состоянии и будущим приступом не является статистически достоверной. Это связано с тем, что использование традиционных статистических процедур (таких как метод контрольных карт, метод кумулятивных сумм, метод максимального правдоподобия) для обнаружения разладок в реальной системе с интенсивным потоком событий, каковой и является запись ЭЭГ, неэффективно. Причина этого состоит в том, что, хотя наблюдаемые сигналы имеют определенную цикличность, на достаточно большом промежутке времени система проявляет свойства нестационарной хаотичности, тогда как стандартные методы обнаружения разладки требуют стационарности процесса, характеризующего систему в некотором стабильном состоянии.

Для записи ЭЭГ применяется система электродов, которые размещаются на скальпе пациента. Они подключаются к усилителю с высоким входным сопротивлением, после чего регистрируется разность потенциалов между двумя точками, одна из которых фиксирована (уровень нулевого потенциала), а другая определяется одним из электродов на скальпе. Обычно на скальпе размещается 19 электродов, но в отдельных случаях используется меньшее число электродов или добавляются дополнительные [11, 40, 70].

Биоэлектрическая активность фиксируется в достаточно широкой частотной полосе. В то же время низкие и высокие частоты подвергаются весьма сильным воздействиям из-за изменения проводимости кожного покрова и при произвольной мышечной активности. Для снятия клинической ЭЭГ обычно используется диапазон частот от 0,5 до 70 Гц. Компьютерные энцефалографы работают на высокой частоте в 256 Гц, 512 Гц или 1024 Гц.

Согласно обзору [40], наиболее часто применяемым в настоящее время методом анализа ЭЭГ является расчёт спектра мощности [66, 73]. Также используются корреляционный и регрессионный методы, но многомерность сигнала ЭЭГ и большой

объем обрабатываемых данных не позволяют достоверно определить уровень связи сигналов по отдельным отведениям в зависимости от состояния пациента. Эта связь может быть обусловлена различными физическими механизмами, достоверно идентифицировать влияние которых в настоящее время не удастся. Традиционным методом исследования спектра сигнала является когерентный анализ, основанный на преобразовании Фурье сигналов по отдельным отведениям.

Открытым является также вопрос о характеристиках нестационарности сигналов ЭЭГ по различным отведениям. Согласно [41], изменение амплитуды и частотных характеристик по отведениям могут представлять интерес как клинически значимые параметры и индикаторы изменения состояния пациента. Основной математической трудностью тут представляется построение достоверных статистических оценок по малым временным отрезкам.

Таким образом, для идентификации разрядки как предвестника приступа эпилепсии в нестационарном потоке данных ЭЭГ необходимо найти специфический «закон сохранения» (если он, конечно, существует и обладает индикативным свойством) в виде функционала, т.е. некоторой выборочной статистики, построенной на траектории изучаемого случайного процесса и такой, что ее выборочное распределение, в свою очередь, стационарно.

Для нестационарных процессов основным объектом изучения являются выборочные распределения расстояний между выборочными распределениями в некоторых нормах. В настоящей работе изучается распределение расстояний в норме непрерывных функций.

Предположим, что нестационарность свойственна самой природе наблюдаемой системы. Возможно, что она вызвана случайным вложением нескольких стационарных случайных процессов, отвечающих за определенные скрытые состояния этой системы. Хотя эти состояния на относительно коротких промежутках наблюдения не могут быть выявлены статистически, гипотеза об их соответствии скрытым параметрам может быть проверена. Если смена состояния системы соответствует изменению состава случайных процессов, то, предположительно, уровень нестационарности изменится. Индикацией разрядки в таком случае может служить СУС, вычисляемый в скользящем окне определенной длины.

Если допустить, что состояние пациента перед приступом эпилепсии меняется не мгновенно, то индикатор изменения распределения СУС мог бы тогда использоваться как предиктор.

Поскольку разладку требуется распознать как можно быстрее, окно наблюдения должно быть достаточно малым, поэтому сравнивать между собой два распределения СУС, построенных для встык-выборок некоторой длины N по фрагменту длины L не эффективно. Каждый фрагмент длины L дает только одно значение СУС, так что для построения распределения потребуется длина много большая, чем L , тогда как индикатор разладки должен сработать на промежутке $l \ll L$. В то же время промежуток индикации разладки l должен быть достаточно большим по сравнению с длиной выборки N , чтобы по нему можно было бы сделать оценку текущего значения СУС.

Сама оценка СУС делается очень легко в соответствии с определением этого индикатора. Например, пусть выяснилось, что текущему нормальному состоянию пациента отвечает СУС ρ_0 . Выбираем некоторую длину l , на которой укладывается K выборок длины N , т.е. $l = KN$. Выбираем также шаг τ , с которым в окне длины l будем сдвигать встык-выборки. Тогда на длине l можно будет вычислить $\frac{l-2N}{\tau} + 1$ расстояний между соседними выборками. Далее считаем количество превышений этими расстояниями уровня ρ_0 . Как только оно превысит критическое значение

$$n_{cr} = \left(\frac{l-2N}{\tau} + 1 \right) \rho_0, \quad (4.1.1)$$

фиксируется разладка. Основной проблемой здесь является выбор параметров l и τ , при которых индикатор срабатывает наиболее достоверно.

4.2. Статистический анализ рядов данных по отведениям ЭЭГ

Исследуем возможность предсказания приступа эпилепсии на основе анализа записей ЭЭГ с помощью индикатора СУС. Проверяемое предположение заключается в том, что уровни стационарности пациента в спокойном состоянии и на некотором подбираемом эмпирически промежутке l перед приступом различаются.

Типичная зависимость СУС и индекса нестационарности от длины выборки для различных отведений для пациента в спокойном состоянии показана на Рис. 4.1 и 4.2. Из этих рисунков следует, что для каждого электрода характерен свой уровень нестационарности, причем этот уровень существенно превышает табулированные стационарные значения (см. табл. 1.1) и не убывает с увеличением длины выборки. Этот факт находит отражение в росте индекса нестационарности (Рис. 4.2) с увеличением длины выборки, который значительно превосходит единицу.

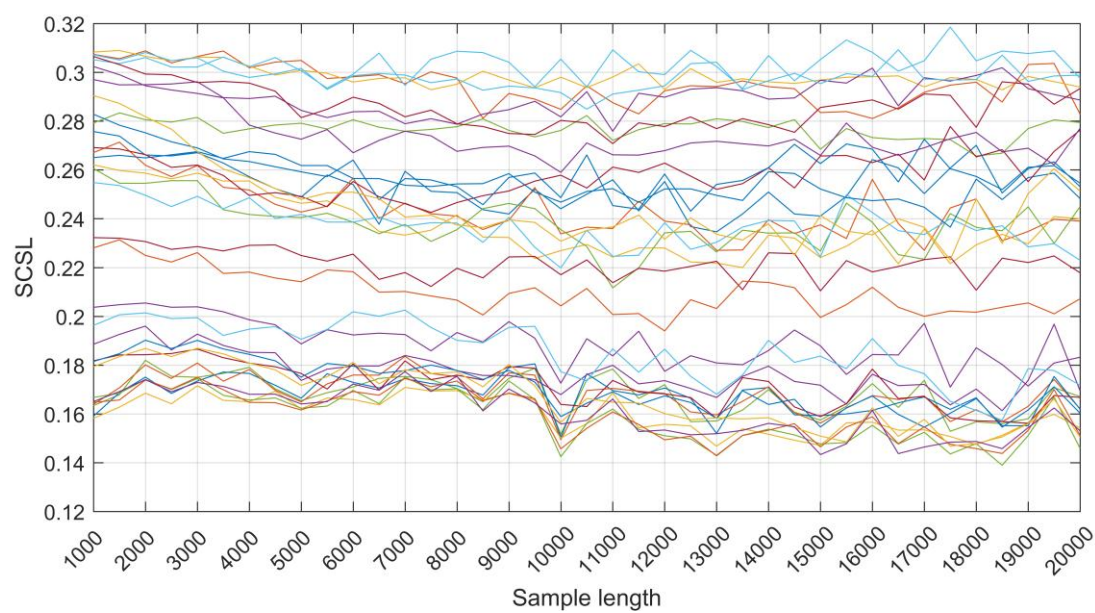


Рис. 4.1 – Зависимость СУС от длины выборки для различных отведений

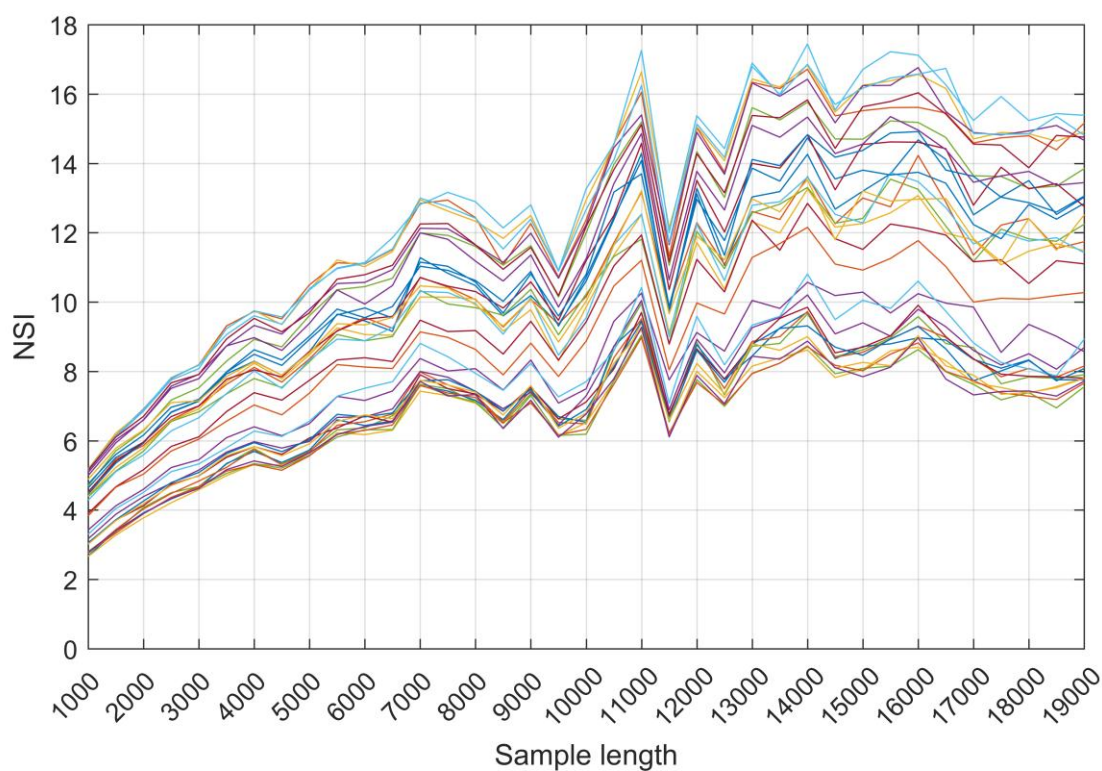


Рис. 4.2 – Зависимость индекса нестационарности от длины выборки
для различных отведений

Определим теперь по имеющимся эмпирическим данным оптимальную длину встык-выборки, для которой наиболее эффективно вычислять расстояния между ними с целью идентификации разладки.

С одной стороны, из практических соображений длина выборки, по которой определяется разладка, должна быть по возможности минимальной, поскольку эта длина равна задержке, с которой принимается решение. С другой стороны, на выборках слишком малых длин может наступить хаотическое поведение индикатора, так что точность его предсказаний ухудшится.

Нас будут интересовать следующие критерии устойчивости индикации разладки.

Во-первых, количество отведений, по которым увеличивается уровень нестационарности перед приступом, должно быть по возможности наибольшим, причем этот максимум должен быть устойчивым относительно вариации длины выборки.

Во-вторых, индекс нестационарности желательно иметь наибольшим, чтобы минимизировать ошибку ложного срабатывания.

В-третьих, по тем отведениям, которые фиксируют разладку, превышение стационарного СУС должно иметь наибольший коэффициент устойчивости, каковым является отношение среднего индекса нестационарности в момент разладки к его стандартному отклонению.

На Рис. 4.3 представлены результаты анализа данных для конкретных пациентов с целью оптимизации указанных параметров.

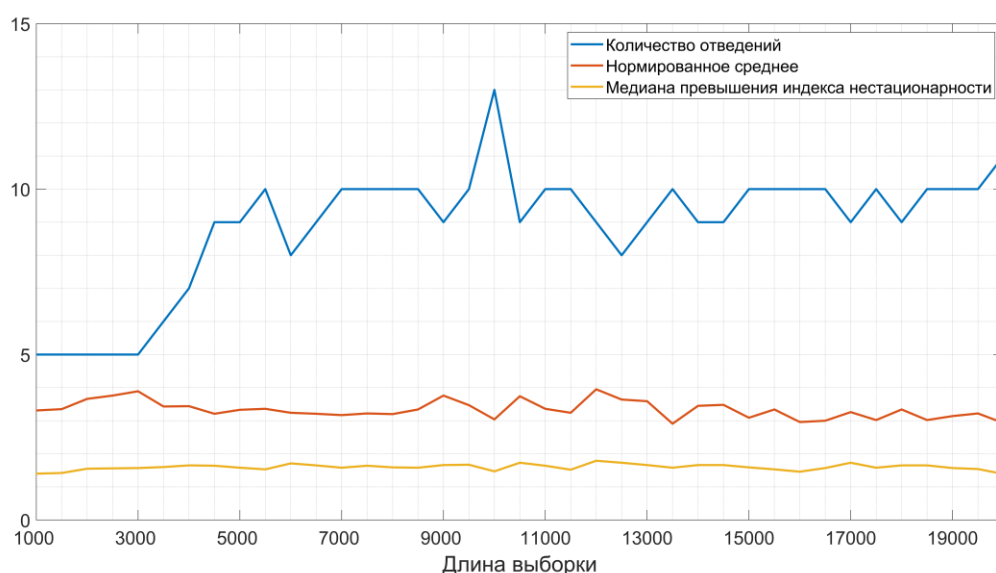


Рис. 4.3 – Значения параметров в зависимости от длины выборки

Из Рис. 4.3 следует, что и нормированное среднее значение индекса нестационарности, и медиана его превышения по отведениям, зафиксировавшим разладку,

приблизительно постоянны и не имеют выраженного максимума в зависимости от длины выборки. Единственным значимо меняющимся параметром является число индикативных отведений. Если длина выборки меньше 5000 (20 секунд наблюдений), то число отведений, по которым можно зафиксировать разладку, сравнительно невелико. Если же длина выборки больше 5000, то указанное количество отведений не имеет четкого тренда к росту, а остается приблизительно постоянным. Следовательно, минимально достаточная длина выборки для анализа разладки равна 5000.

Таким образом, в качестве рабочей гипотезы выбирается анализ индекса нестационарности по определенным отведениям для выборок длины 5000. Число отведений, по которым наблюдается существенное превышение СУС, ожидается в этом случае на уровне 8-10.

Далее, из Рис. 4.1 следует, что типичный СУС равен примерно 0,22. Чтобы не было ложного срабатывания индикатора в силу малого количества данных, длина окна индикации разладки должна быть не меньше, чем необходимая для достижения стационарного согласованного уровня значимости, равного этому значению. Согласно данным табл. 1.1 это означает, что в окне наблюдения l должно быть хотя бы 80 значений расстояний между встык-выборками. Из практических соображений это окно должно находиться в промежутке 2-5 минут. Для 2-х-минутного промежутка указанные требования можно соблюсти, если встык-выборки длины 5000 сдвигать с шагом 500. При частоте 250 Гц в этом промежутке укладывается $L = 30$ тыс. данных ЭЭГ по каждому отведению

Используя эти оценки, будем считать предиктором приступа эпилепсии превышение уровня СУС в окне 2 мин для встык-выборок длины 5000, сдвигаемых с указанным выше шагом 500. Предиктор применяется к тем эмпирически определенным отведениям для конкретного пациента, по которым СУС в последние две минуты перед приступом заметно (более, чем на величину самого СУС) превосходил СУС спокойного состояния.

Чтобы выяснить, действительно ли СУС является индикатором состояния пациента, или же последовательность $\rho_k^*(N)$ просто представляет собой реализацию некоторого случайного процесса, не связанного с этим состоянием напрямую, а вызванного случайным изменением окружающей среды, на которое и реагирует мозг, необходимо проверить на стационарность выборку $\rho_k^*(N)$ на k -ом промежутке, где состояние пациента стабильно.

Ранее в Главе II были введены распределения $G_N^k(\rho, L)$ расстояний между встык-выборками длины N , построенные по промежуткам длины L , для которых $\rho_k^*(N) \equiv \rho_k^*(N, L)$ есть СУС k -го промежутка длины L . Если объем данных $nL = L_{tot}$ достаточен для того, чтобы построить распределения величин СУС $\rho_k^*(N, L)$, то можно сравнить эти распределения для одного и для разных состояний пациента и при фиксированной длине встык-выборки N исследовать зависимость этих распределений от длины фрагмента L .

Исследования были проведены для 10 разных пациентов, и результаты качественно совпали. Приведем для примера результаты обработки данных для одного из пациентов. Общий объем данных, отвечающих визуально не меняющемуся состоянию бодрствования пациента, составил 2,5 часа, т.е. примерно 2,3 млн данных или $n = 78$ независимых двухминутных отрезков. Для каждого из этих двухминутных промежутков и для каждого отведения вычислялся СУС $\rho_k^*(5000, 30000)$, $k = 1, \dots, n = 78$. Типичный пример ряда указанных величин СУС для одного из электродов приведен ниже на Рис. 4.4

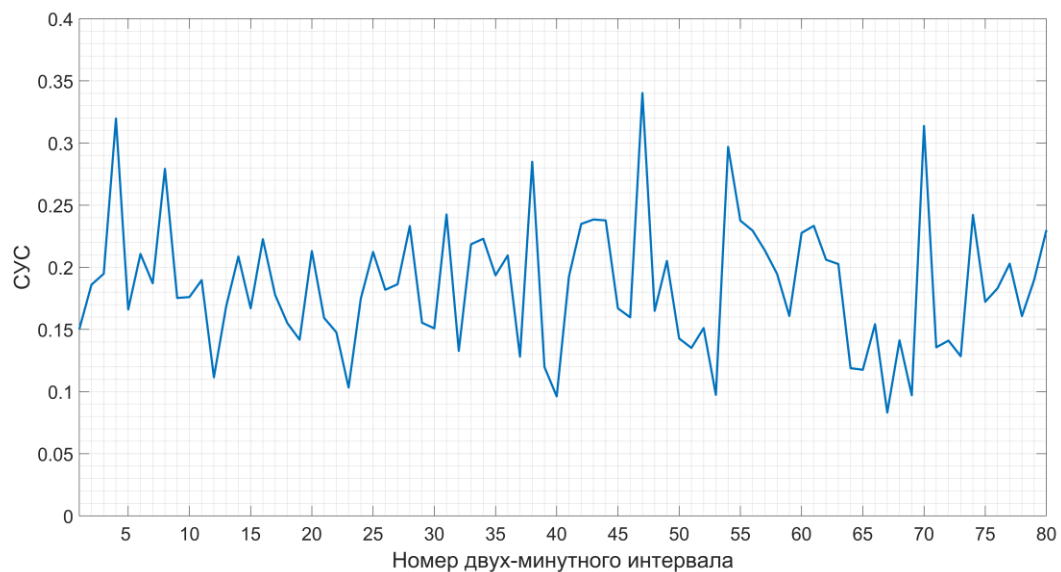


Рис. 4.4 – Пример ряда СУС $\rho_k^*(5000, 30000)$

Эмпирические распределения значений СУС по первой и второй половинам данных, а также по всей выборке в целом приведены на Рис 4.5.

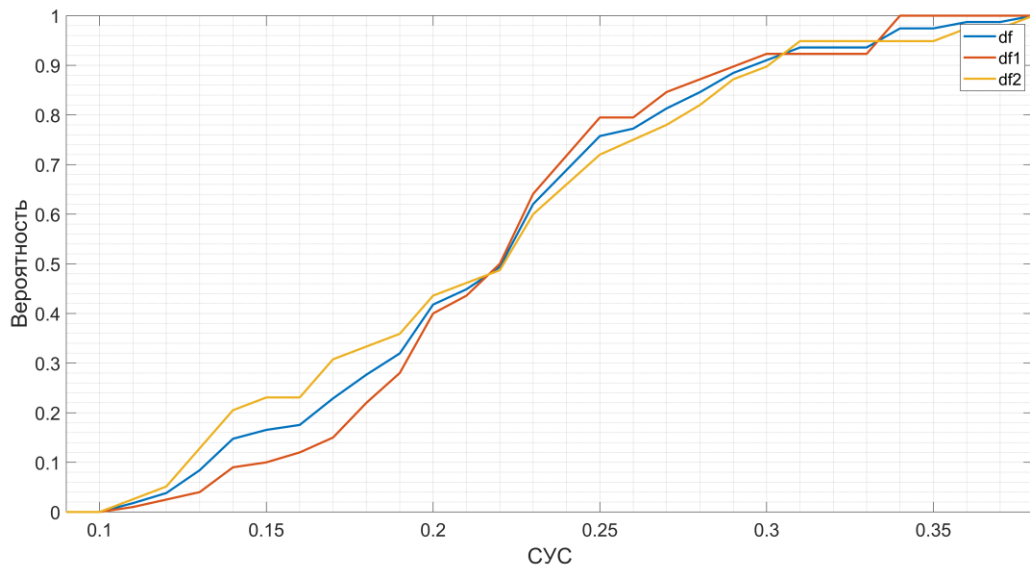


Рис. 4.5 – Эмпирические распределения СУС $\rho_k^*(5000, 30000)$

Для стационарных выборок длиной 40 данных теоретический согласованный уровень значимости составляет $\varepsilon_0(40) = 0,25$. В приведенном примере (как и для всех 64 отведений рассматриваемых данных ЭЭГ) СУС меняется от 0,245 до 0,255. Это означает, что ряд СУС стационарный, хотя сам исходный ряд показаний ЭЭГ, для которых этот СУС посчитан, является сильно нестационарным. Индекс нестационарности для встык-выборок ряда ЭЭГ длиной 5000 на промежутке L_{tot} варьируется по отведениям от 5 до 10 (см. Рис. 4.2).

Поскольку же ряд значений СУС стационарный, имеет смысл исследовать его статистические характеристики в зависимости от длины окна L , в котором этот СУС вычисляется. Это необходимо для того, чтобы понимать, с какой вероятностью такой локальный СУС превосходит СУС, посчитанный по всему множеству. Длина встык-выборок при этом фиксирована и равна $N = 5000$.

Как было указано выше, лишь некоторые отведения (в количестве порядка 10) могут служить индикаторами-предвестниками приступа эпилепсии. Это те отведения, для которых относительное превышение индекса нестационарности в последние две минуты перед приступом по сравнению со спокойным состоянием составляет величину не менее, чем СУС. Поэтому имеет смысл сравнить статистические свойства распределений СУС для индикативных и остальных отведений.

Введем среднее значение и стандартное отклонение для ряда СУС как функции длины окна L :

$$avr(n) = \frac{1}{n} \sum_{k=1}^n \rho_k^*(N, L), \quad sd(n) = \sqrt{\frac{1}{n} \sum_{k=1}^n (\rho_k^*(N, L) - avr(n))^2}, \quad n = \frac{L_{tot}}{L}. \quad (4.2.1)$$

На Рис. 4.6 представлена зависимость нормированного среднего значения СУС от количества промежутков n (т.е. в зависимости от обратной длины окна) для индикативного электрода и дана ее степенная аппроксимация.

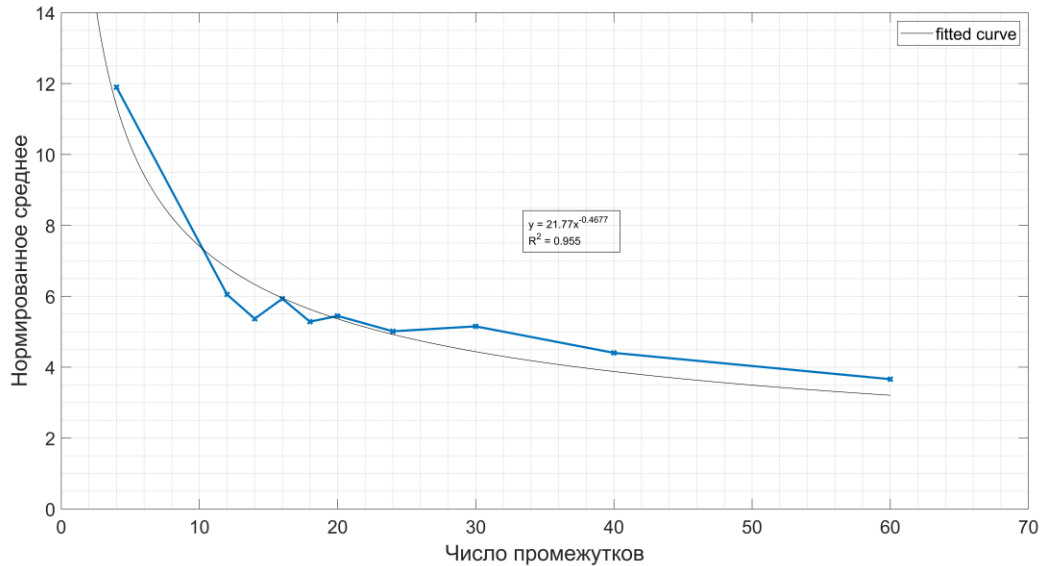


Рис. 4.6 – Зависимость нормированного среднего СУС от длины окна, индикативный электрод

Отметим, что, хотя длина выборки ряда СУС невелика (менее 100), все же степенная аппроксимация близка к зависимости $\frac{avr(n)}{sd(n)} \sqrt{n} \approx const$. Это согласуется со сделанным ранее выводом о стационарности ряда СУС, поскольку есть снование считать, что отклонение локального СУС от среднего значения распределено асимптотически нормально в соответствии с центральной предельной теоремой. Плотность функции распределения СУС, обозначенной как df на Рис. 4.5, представлена на Рис. 4.7. Она имеет приближенно симметричный треугольный вид.

Согласно данным Рис. 4.7, вероятность превышения уровня $\rho_{tot}^*(5000) \approx 0,22$ в окне длины L на величину, равную $(\rho_{tot}^*(5000))^2 \approx 0,05$, есть $1 - df(0,27) \approx 0,23$. Эта величина равна вероятности ошибки ложной тревоги.

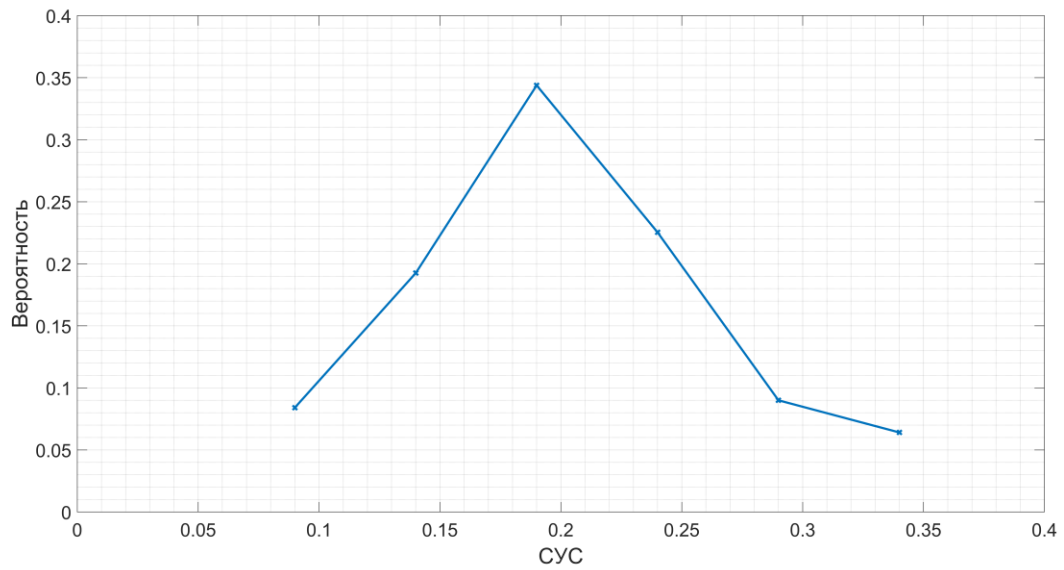


Рис. 4.7. Стационарная плотность распределения СУС для окна 30000 данных (2 мин.), индикативный электрод

Полезно сравнить уровень СУС на всем промежутке (СУС-tot в легенде Рис. 4.8) с эквивалентной по Чернову стационарной точкой функции значимости. Для всего множества данных стационарная точка уровня значимости $\Psi_{5000}(\rho)$ равна $\rho_{tot}^*(5000) \approx 0,22$. Значение же стационарной точки эквивалентной ей по Чернову функции $\Phi_{5000}(\rho)$ составило величину $\tilde{\rho}_{tot}(5000) \approx 0,20$.

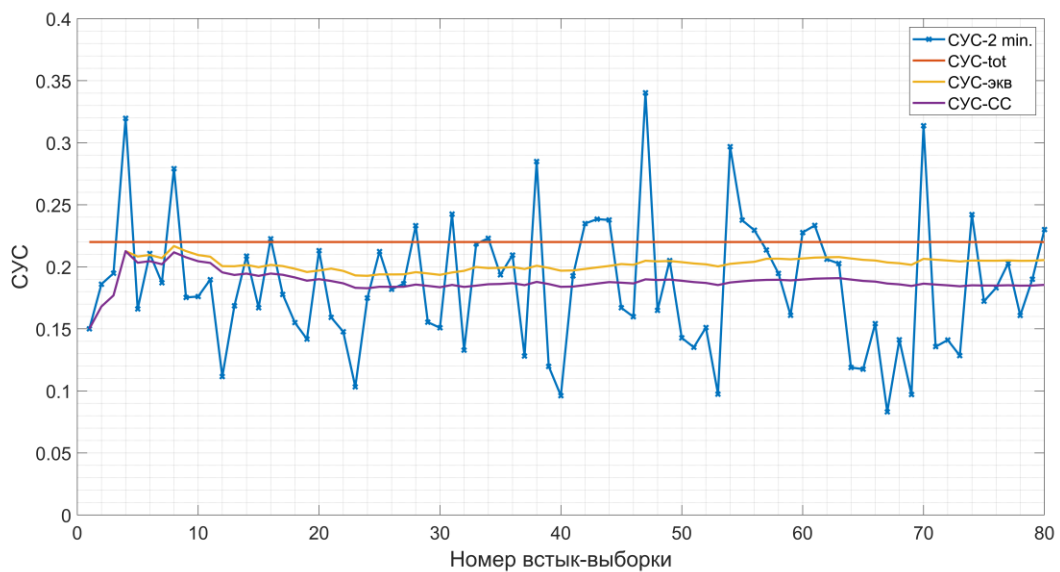


Рис. 4.8 – Сравнение стационарных точек уровня значимости для данных, представленных на Рис. 4.4

Для сравнения на Рис. 4.8 показана также кривая скользящего среднего уровня СУС (обозначение СУС-СС). Видно, что она заметно уступает в точности модели, использующей аппроксимацию по Чернову. Следовательно, предложенное в Главе II приближение более точное, чем кажущееся естественным выборочное среднее.

Рассмотрим теперь, как меняется максимальное значение СУС, наблюдаемое на полном объеме данных L_{tot} для встык-выборок длины 5000, в зависимости от длины окна L . На Рис. 4.9 показана соответствующая эмпирическая зависимость и ее аппроксимация.

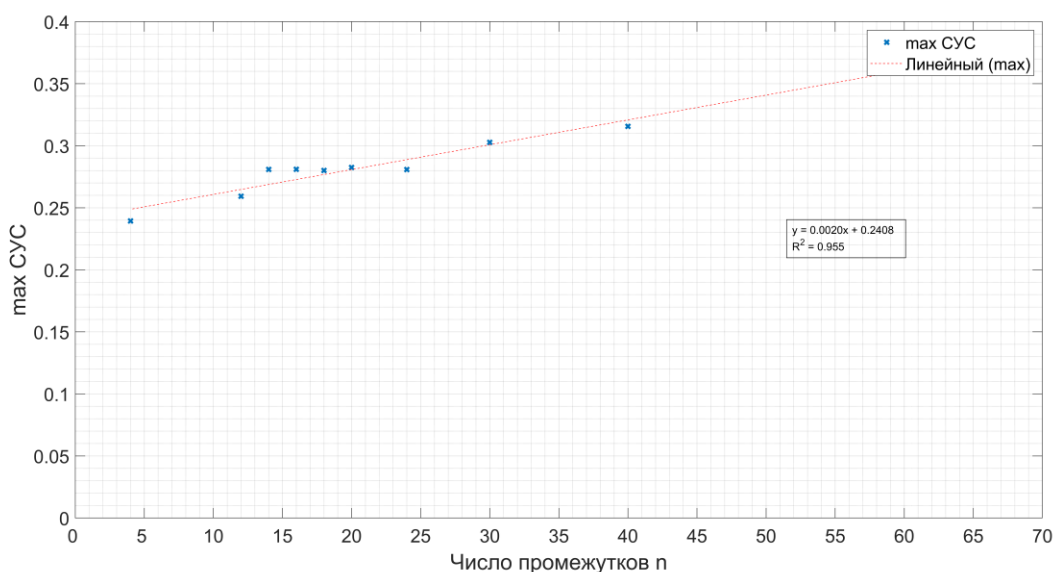


Рис. 4.9. Зависимость максимального значения СУС для встык-выборок длины 5000 от длины окна, индикативный электрод

Из Рис. 4.9 видно, что указанная зависимость носит линейный характер (обратная пропорциональность длине L).

Рассмотрим теперь аналогичные зависимости для обычных электродов, не служащих индикаторами разрядки. На рис. 4.10 приведено нормированное среднее (т.е. отношение среднего значения к среднеквадратичному отклонению) для таких электродов, а на Рис. 4.11 соответствующая зависимость максимального значения СУС от длины окна, в котором он вычислялся.

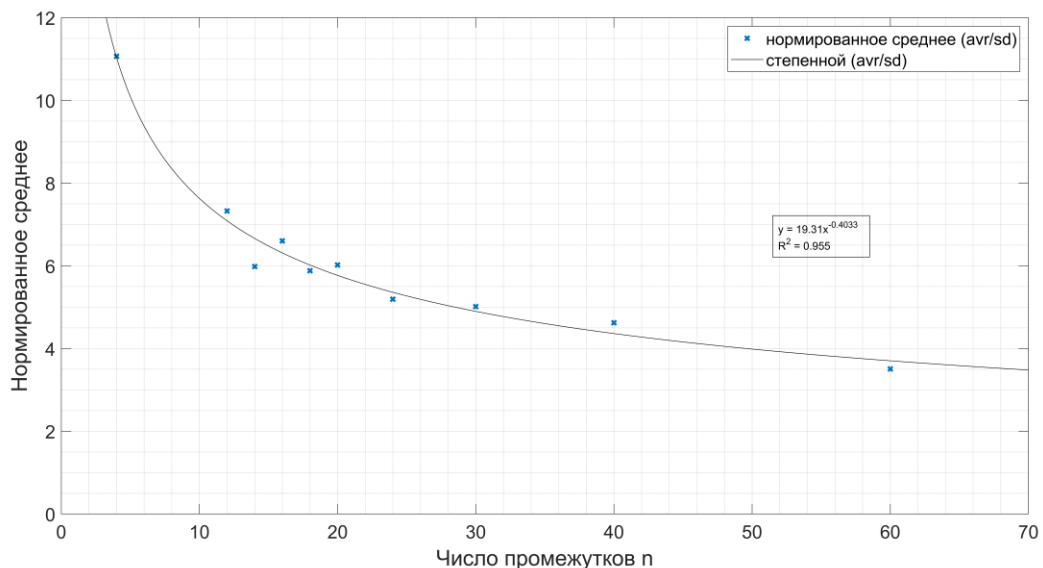


Рис. 4.10 – Зависимость нормированного среднего СУС от длины окна, «обычный» электрод

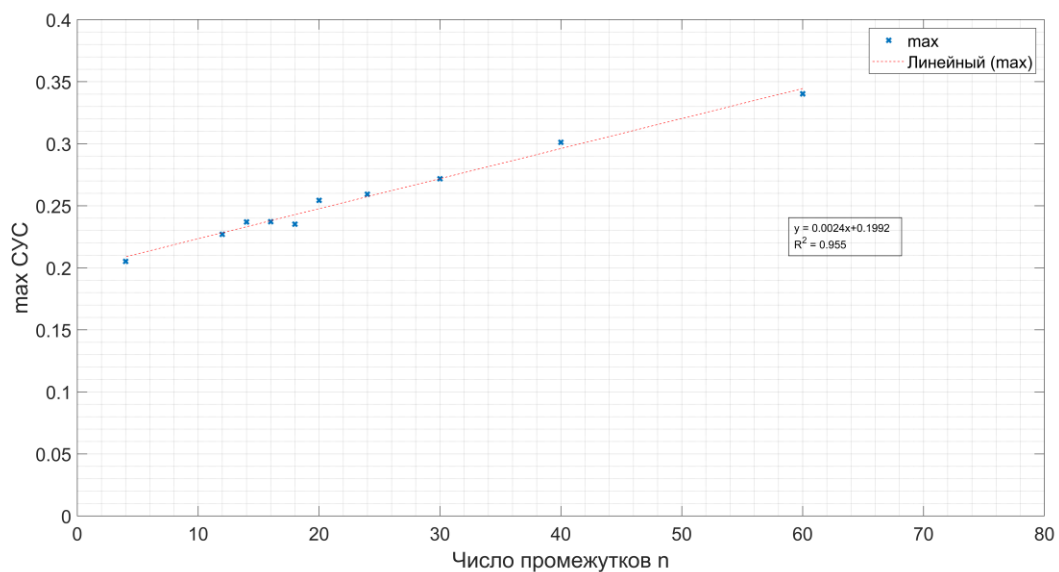


Рис. 4.11 – Зависимость максимального значения СУС для встык-выборки длины 5000 от длины окна, «обычный» электрод

Таким образом, выяснилось, что по каждому типу электродов – индикативному и не индикативному – распределение согласованного уровня стационарности является стационарным и имеет примерно одинаковый вид. Иными словами, тип нестационарного поведения временного ряда не меняется в течение того промежутка времени, пока состояние пациента визуально считается стабильным. При этом детерминация аппроксимаций рассмотренных зависимостей для «обычных» электродов заметно выше,

чем для индикативных, то есть эти «обычные» электроды порождают более стационарный ряд данных. Важно подчеркнуть, что при этом сам ряд данных ЭЭГ для визуально стабильного состояния пациента является нестационарным, поэтому индикатором смены состояния пациента следует считать не отклонение показаний ЭЭГ (ибо стационарные критерии оценки уровня значимости оказываются некорректными), а изменение уровня нестационарности выборочных распределений показаний ЭЭГ, если этот уровень действительно различен для разных состояний. В следующем разделе приводятся данные по расчетам СУС для разных, но условно стабильных состояний.

4.3. Стационарные паттерны функций распределения СУС

Рассмотрим уровни нестационарности для состояний, которые предположительно можно считать различными: это, во-первых, состояния бодрствования и сна без приступов эпилепсии, и, во-вторых, состояния, предшествующие двум приступам эпилепсии. Выяснилось, что распределения СУС этих состояний различаются значительно сильнее, чем это можно ожидать для выборок из одной генеральной совокупности. При этом существенно, что для каждого состояния показания ЭЭГ имеют стационарно распределенный ряд значений СУС. Поскольку плотности распределения визуальнее более информативны, чем сами ВФР, мы приведем на Рис. 4.12 именно эти функции, хотя расстояния между распределениями различных состояний указаны в норме S для ВФР.

Расстояния между этими состояниями (ВФР в норме S) следующие: $\delta_{12} = 0,36$; $\delta_{23} = 0,33$; $\delta_{31} = 0,45$. Количество данных, по которым построены эти распределения: pdf1 – 78 значений, pdf2 – 90 значений, pdf3 – 55 значений. Типичными же значениями расстояний между стационарными распределениями таких длин являются величины порядка 0,25. Следовательно, можно сделать вывод, что величина СУС действительно различна для разных чередующихся состояний пациента, так что она может служить индикатором разладки. Разладкой в данном случае следует считать смену одного нестационарного режима показаний ЭЭГ другим нестационарным же, но с отличающимся от предыдущего состоянием уровнем СУС.

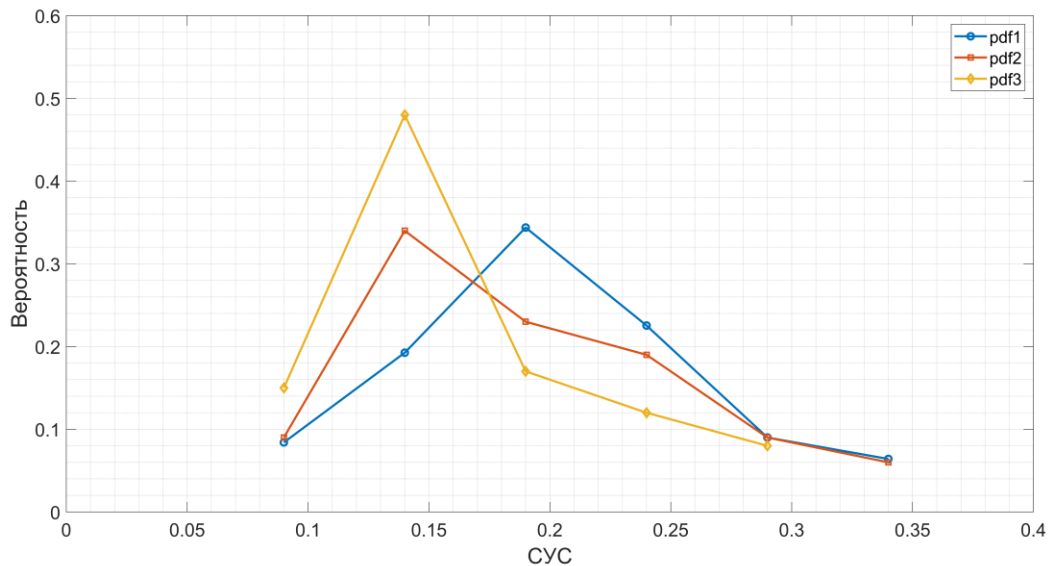


Рис. 4.12 – Плотности распределений СУС для встык-выборок длины 5000 в окне длиной 2 минуты для трех состояний:
1- бодрствование до засыпания, 2- сон,
3- бодрствование между приступами эпилепсии

В состоянии сна показания ЭЭГ также нестационарны, только уровень СУС несколько ниже, чем в состоянии бодрствования. Состояние же между приступами эпилепсии по среднему уровню напоминает состояние сна, т.е. релаксации, которое прерывается аномальными всплесками СУС.

4.4. Верификация индикатора разладки

При непрерывном наблюдении за состоянием пациента важно учитывать время после приступа, необходимое для релаксации СУС к условно спокойному уровню. То, что этот уровень условный, связано с тем, что уровень нестационарности ЭЭГ в разных состояниях, вообще говоря, разный, и в этом смысле отсутствует некоторый характерный уровень СУС, к которому возвращается организм. Уровни СУС до приступа и после приступа могут быть совершенно различными. Это означает, что статистика для новой разладки после каждого приступа собирается заново. Для оценки времени условной релаксации анализировались данные по 64 отведениям в течение достаточно длительного времени, охватывающего несколько приступов одного и того же пациента.

На Рис. 4.13 приведен график изменения СУС 20-секундных встык-выборок в двухминутных промежутках. Перед моментом наступления приступа эпилепсии наблюдается увеличение частоты превышения глобального уровня СУС локальными значениями СУС. Тем самым индикатором является не амплитуда графика СУС, а частота

превышения заданного индикативного уровня, определяемого по спокойному участку траектории.

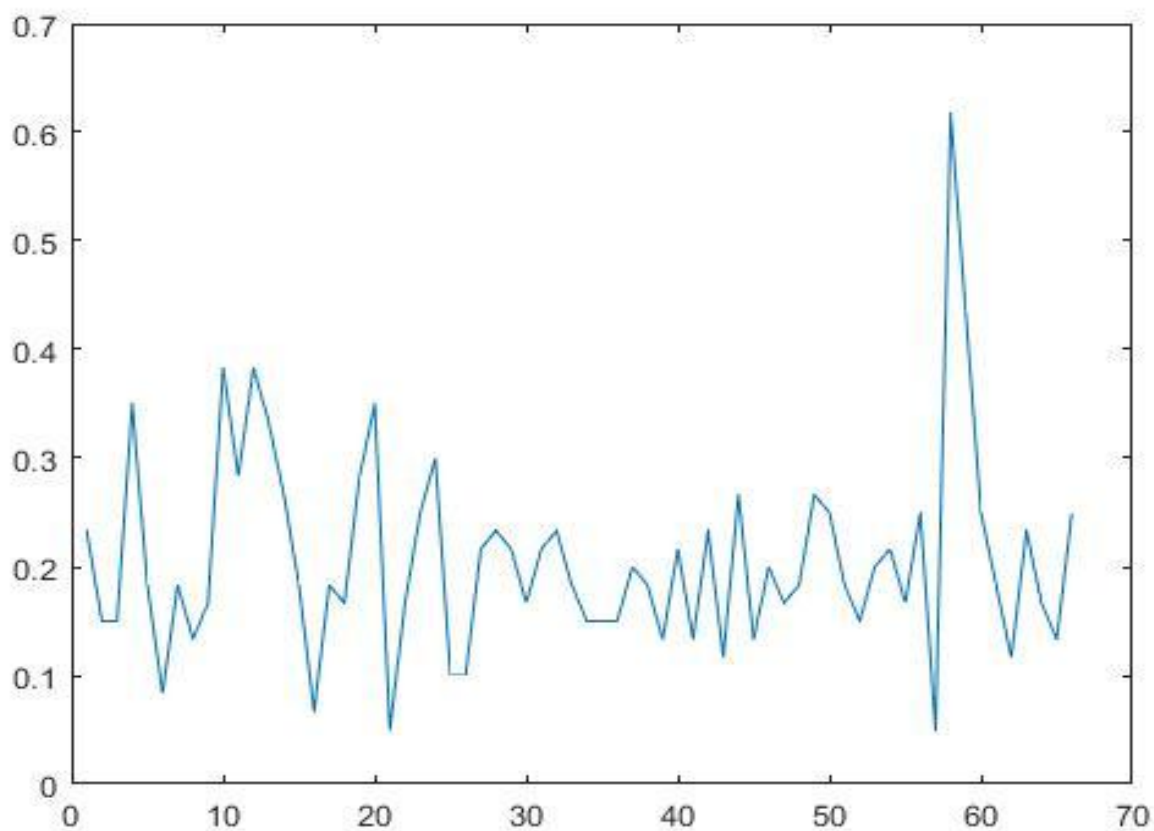


Рис. 4.13 – Типичный график СУС для отведения – индикатора приступа

Выяснилось, что, во-первых, приступ всегда проходит на фоне индикации разрядки, каковой считается превышение в двухминутном окне сканирования уровня СУС промежутка в целом; во-вторых, после приступа есть некоторый период, в течение которого СУС снижается до следующего характерного значения, отмеченного горизонтальной линией. Периоды сразу перед приступом и сразу после него должны быть исключены из статистики по определению характерного СУС условно-спокойного состояния.

Индикативные отведения ведут себя в указанном смысле достаточно синхронно. Как уже говорилось выше (см. Рис. 4.3), для каждого пациента обнаруживается устойчивое количество из восьми-десяти отведений, по которым фиксируется приступ эпилепсии как момент превышения СУС условно-спокойного состояния. На Рис. 4.14 приведены соответствующие графики изменения СУС, оканчивающие приступом эпилепсии, для конкретного пациента.

Индекс нестационарности двухминутных интервалов для выборок длины 5000 по данной группе отведений весьма высок: он изменяется от 3 до 15, имея среднее значение на уровне 7. Это означает, что на практике применение классических статистических критериев на уровне значимости, допустим, 0,05 приводит к ошибке в 7 раз большей, порядка 0,35.

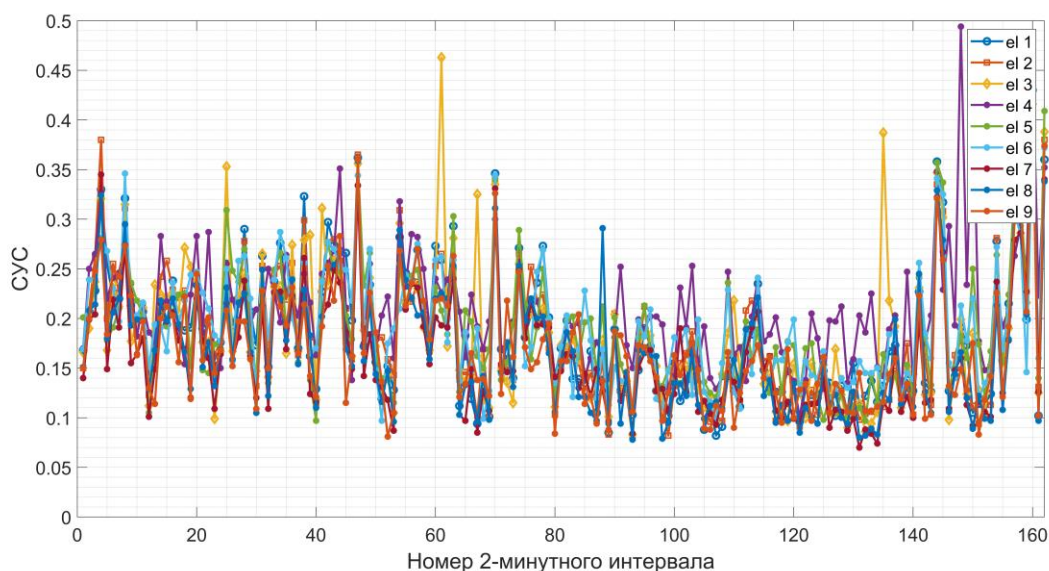


Рис. 4.14 – Графики СУС для девяти отведений – индикаторов приступа

Описанная методика была протестирована на ЭЭГ десяти пациентов. В общей сложности были проанализированы 20 достаточно длинных фрагментов, оканчивающихся приступом. Критический уровень СУС для разных пациентов оказывался различным, причем после приступа он существенно менялся и для одного пациента. Однако выборки длины 5000 каждый раз оказывались оптимальны для обнаружения разрядки.

Таким образом, на основе проведенного анализа можно предложить эмпирическое правило набора статистики для расчета СУС как индикатора разрядки применительно к прогнозированию приступа эпилепсии. Данные начинают учитываться для расчета СУС примерно через 8-10 минут после приступа и собираются в течение последующих 30 минут. Вычисляется СУС встык-выборок оптимальной длины (в нашем примере это 5000) на всем промежутке сбора данных, после чего в скользящем двухминутном окне определяется доля событий, когда расстояние между встык-выборками превосходит СУС. Как только эта доля становится больше самого СУС, фиксируется очередная разрядка.

Выводы по Главе IV.

В работе построен индикатор разрядки, который может служить эффективным предиктором наступления приступа эпилепсии. Для его построения использовалась нестационарная статистка распределения расстояний между выборочными функциями распределения ЭЭГ-сигнала. Было выяснено, что ряды данных по каждому отведению нестационарны. На наборах данных ЭЭГ реальных пациентов была подтверждена гипотеза о том, что существуют участки головного мозга, нестационарная активность которых перед приступом возрастает. Этот факт позволил применить модель согласованного уровня стационарности для индикации разрядки нестационарного случайного процесса. Были определены оптимальная длина встык-выборки для вычисления СУС и длина окна для фиксации разрядки, и показано, что выборочная функция распределения СУС стационарна. Показана также эффективность аппроксимации эквивалентной стационарной точки уровня значимости по Чернову.

Результаты Главы IV опубликованы в работах [15, 53, 55].

Заключение

В настоящей диссертации разработан новый метод построения непараметрического индикатора разладки в нестационарных многомерных временных рядах и построен программный комплекс, реализующий соответствующие численные алгоритмы.

Разработанный в диссертации метод анализа с помощью аппроксимаций Фейнмана-Чернова позволяет значительно сократить объем вычислений при тестировании индикаторов-предикторов изменения состояния временного ряда и организовать расчет разладки в реальном времени.

Новизна предложенного подхода к трактовке понятия разладки в нестационарном ряде по сравнению с классическими стационарными аналогами этой задачи состоит в том, что индикатором является не отклонение текущих выборочных распределений одно от другого или от некоего паттерна, а отклонение уровня нестационарности ряда, который является статистикой следующего уровня. Примеры применения такого подхода показали его оправданность. В частности, с помощью построенного индикатора оказалось возможным повысить точность предсказания приступа эпилепсии на основе данных ЭЭГ по сравнению с традиционными методами анализа спектра мощности сигнала.

Таким образом, численный код, генерирующий по фрагменту траектории нестационарного ряда величину предиктора разладки и определяющий одновременно оптимальную длину сканирования ряда, представляет практическую важность.

Основные результаты диссертационной работы состоят в следующем.

1. Разработана математическая модель индикатора разладки нестационарного временного ряда на основе статистики согласованного уровня стационарности.
2. Разработан метод анализа зависимости индикатора разладки от двух параметров – длины фрагмента ряда и длины окна встык-выборки. Разработан также метод приближенного вычисления значений индикатора разладки на основе использования формул Фейнмана-Чернова конечнократной аппроксимации полугрупп.
3. Построен численный алгоритм для определения многомерного индикатора разладки в реальном времени и программный комплекс с пользовательским интерфейсом.
4. Проведены численные эксперименты по индикации разладки применительно к рядам электроэнцефалограмм для пациентов с приступами эпилепсии и показано, что построенный индикатор имеет стационарное распределение, а его применение характеризуется ошибкой первого рода, равной значению индикатора разладки, и нулевой эмпирической ошибкой второго рода.

Дальнейшей задачей в области построения и анализа индикаторов разладки в теоретической области является исследование возможности уменьшения порога ошибки первого рода, который в данной работе равен СУС, за счет анализа самого ряда СУС в окне индикации и построения модели соответствующего временного ряда.

В практической области следует рассмотреть возможность увеличения размерности вычислительной задачи за счет анализа расстояний между выборками, сдвинутыми на расстояния, превосходящие окно выборки. Этот аспект также может снизить частоту ложных срабатываний индикатора разладки, не увеличивая при этом вероятность ошибки второго рода.

Таким образом, результаты диссертации и их теоретическая и прикладная значимость показывают перспективность дальнейшего развития кинетического направления в области анализа и моделирования нестационарных временных рядов.

Список литературы

1. Айвазян С.А. Методы эконометрики: учеб. – М. Магистр. ИНФРА-М, 2014.
2. Бассвиль М. и др. Обнаружение изменения свойств сигналов и динамических систем (пер. с англ.) – М.: Мир, 1989. – 280 с.
3. Блянкинштейн Н.И., Орлов Ю.Н., Родкин М.В., Федоров С.Л. Об оценке уровня стационарности каталогов землетрясений // Препринты ИПМ им. М.В. Келдыша. 2017. № 40. 18 с.
4. Боголюбов Н.Н. Проблемы динамической теории в статистической физике. М.-Л.: ГИТТЛ, 1946.
5. Большев Л.Н. Асимптотически пирсоновские преобразования // Теория вероятностей и ее применения, 1963. Т.8. С. 129-155.
6. Босов А.Д., Орлов Ю.Н., Федоров С.Л. О распределении рядов абсолютных приростов цен на финансовых рынках // Препринты ИПМ им. М.В. Келдыша. 2014. № 96. 15 с.
7. Вуколов Э.А. Основы статистического анализа. – М.: ФОРУМ ИНФРА-М, 2004. – 464с.
8. Гафарова Е.А. Применение прикладных программ при обучении эконометрическим дисциплинам // Современные проблемы науки и образования. – 2014. – № 6.
9. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.
10. Жиглявский А. А., Красковский А. Е. Обнаружение разладки случайных процессов в задачах радиотехники. – Л.: Изд-во ЛГУ, 1988. – 224 с.
11. Иванов Л.Б. Прикладная компьютерная электроэнцефалография. 2-е изд. – М.: MBN, 2004. – 352 с.
12. Ивченко А.Ю., Кислицын А.А., Орлов Ю.Н. Распределение SIR для траекторий детерминированного движения на примере метрополитена // Препринты ИПМ им. М.В. Келдыша. 2018. № 155. 21 с.
13. Казанцева Н.Н. Статистический контроль и статистические методы управления качеством. – Томск: Изд-во ТПУ, 2004. – 116 с.
14. Кислицын А.А. Программный комплекс для анализа статистики согласованного уровня стационарности временных рядов // Препринты ИПМ им. М.В. Келдыша. 2020. № 26. 22 с.
15. Кислицын А.А., Козлова А.Б., Корсакова М.Б., Машеров Е.Л., Орлов Ю.Н. Стационарная точка уровня значимости для нестационарных функций распределения // Препринты ИПМ им. М.В. Келдыша. 2018. № 113. 20 с.

16. Кислицын А.А., Козлова А.Б., Корсакова М.Б., Орлов Ю.Н. Индикатор разладки для нестационарных случайных процессов // Доклады РАН, сер. математическая, 2019. Т. 484. № 4. С. 393-396.
17. Кислицын А.А., Орлов Ю.Н. Моделирование эволюции выборочных распределений случайных величин с помощью уравнения Лиувилля // Математическое моделирование, 2020. Т. 32. №1. С. 111-128.
18. Кислицын А.А., Орлов Ю.Н. Свидетельство о регистрации программы для ЭВМ «Программный комплекс NSSAT (Non-Stationary Series Analysis Toolbox) для определения и визуализации характеристик нестационарности временных рядов». Правообладатель: ИПМ им. М.В. Келдыша РАН. Свидетельство о государственной регистрации № 2019660374 от 05.08.2019.
19. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.
20. Козинев И.А., Мальцев Г.Н. Модифицированный алгоритм обнаружения разладки случайного процесса и его применение при обработке многоспектральных данных // Обработка информации и управление, 2012. № 3. С. 9-17.
21. Корольюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
22. Никифоров И. В. Последовательное обнаружение изменения свойств временных рядов. – М.: Наука, 1983. – 200 с.
23. Об утверждении Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года. – 2013. URL: <http://government.ru/docs/8024/> (дата обр. 06.05.2016). Проверен: 07.05.2016.
24. Орлов А.И. Эконометрика: учебник для вузов. – Ростов н/Д : Феникс, 2009. – 276 с.
25. Орлов Ю.Н. Кинетические методы исследования нестационарных временных рядов. – М.: МФТИ, 2014. – 276 с.
26. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 312 с.
27. Орлов Ю.Н., Сакбаев В.Ж., Смолянов О.Г. Формулы Фейнмана как метод усреднения случайных гамильтонианов // Труды МИРАН, 2014. Т. 285. С. 232-243.
28. Орлов Ю.Н., Федоров С.Л., Давидько В.А. К вопросу классификации нестационарных временных рядов: состав индекса РТС // Препринты ИПМ им. М.В. Келдыша. 2014. № 54. 18с.
29. Орлов Ю.Н., Федоров С.Л. Генерация нестационарных траекторий временного ряда на основе уравнения Фоккера-Планка // Труды МФТИ, 2016. Т. 8. № 2. С. 126-133.

30. Орлов Ю.Н., Федоров С.Л. Моделирование распределений функционалов на ансамбле траекторий нестационарного случайного процесса // Препринты ИПМ им. М.В. Келдыша. 2016. № 101. 14 с.
31. Орлов Ю.Н., Федоров С.Л. Свидетельство № 2017619117 от 15.08.2017 о государственной регистрации программы для ЭВМ «Программный комплекс NonStatBox для статистического анализа и моделирования нестационарных временных рядов». Правообладатели: Орлов Ю.Н., Федоров С.Л.
32. Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2011. № 53. 20 с.
33. Топтыгина А.П., Азиатцева В.В., Савкин И.А., Кислицын А.А., Семикина Е.Л., Гребенников Д.С., Алешкин А.В., Сулимов А.В., Сулимов В.Б., Бочаров Г.А. Прогнозирование специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита // Иммунология. 2015. 36(1): 22 – 30.
34. Топтыгина А.П., Азиатцева В.В., Кислицын А.А., Бочаров Г.А. Способ прогнозирования первичных и вторичных вакцинальных неудач при вакцинации против вирусов кори, краснухи и эпидемического паротита у детей с помощью вакцины Приорикс и способ персонифицированного подхода к коррекции вакцинальных неудач. Патент № 2599506, 10.10.2016.
35. Уилкс С. Математическая статистика. – М.: Наука, 1967. – 632 с.
36. Ширяев А. Н. Статистический последовательный анализ. Оптимальные правила остановки. – М.: Наука, 1976. – 272 с.
37. Ширяев А. Н. Минимаксная оптимальность метода кумулятивных сумм в случае непрерывного времени // Успехи математических наук, 1996. Т. 51, № 4. С. 173-174.
38. Ширяев А. Н. Об оптимальных методах в задачах скорейшего обнаружения // Теория вероятностей и ее применения, 1963. Т. 8, № 1. С. 26-51.
39. Ширяев А. Задача скорейшего обнаружения нарушения стационарного режима // ДАН СССР, 1961. Т. 138. С. 1039-1042.
40. Щекутьев Г.А. Нейрофизиологические исследования в клинике. – М.: Антидор, 2001. – 236 с.
41. Adeli H., Ghosh-Dastidar S. Automated EEG-Based Diagnosis of Neurological Disorders: Inventing the Future of Neurology – Boca Raton: CRC Press, 2010. – 423p.
42. Basseville M., Nikiforov I. V. Detection of abrupt changes: theory and application. – N.J. : Prentice Hall Englewood Cliffs, 1993.

43. Ben-Gal I., Morag G., Shmilovici A. Context-Based Statistical Process Control // *Technometrics*, 2003. Vol. 45, no. 4. P. 293–311.
44. Ben-Gal I., Singer G. Statistical process control via context modeling of finite-state processes: an application to production monitoring // *IIE Transactions*, 2004. Vol. 36, no. 5. P. 401 – 415.
45. Casas P., Vaton S., Fillatre L., Nikiforov I. Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements // *Computer Networks*, 2010. Vol. 54, no. 11. P. 1750 – 1766.
46. Chernoff P. Note on product formulas for operator semigroups // *J. Funct. Anal.*, 84, 1968. P. 238-242.
47. Cook M.J., O'Brien T.J., Berkovic S.F., et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study // *Lancet Neurol.*, 2013. Vol. 12. P. 563 – 571.
48. Erramilli A., Narayan O., Willinger W. Experimental queueing analysis with long-range dependent packet traffic // *IEEE/ACM Transactions on Networking (TON)*, 1996. Vol. 4, no. 2. P. 209 – 223.
49. Feynman R.P. Space-time approach to nonrelativistic quantum mechanics // *Rev. Mod. Phys.*, 1948. Vol. 20, P. 367 – 387.
50. Girshick M. A., Rubin H. A Bayes approach to a quality control model // *The Annals of Mathematical Statistics*. – 1952. P. 114 – 125.
51. Hassani H. Singular spectrum analysis: methodology and comparison // *Journal of Data Science*, 2007. Vol. 5, no. 2. P. 239 – 257.
52. ISO/IEC/IEEE Systems and software engineering – Architecture description // *ISO/IEC/IEEE 42010:2011(E)* (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000). – 2011. – Jan. – P. 1–46.
53. Kislitsyn A.A., Kozlova A.B., Masherov E.L., Orlov Yu.N. Numerical Algorithm for Self-consistent Stationary Level for Multidimensional Non-stationary Time-series // *Keldysh Institute Preprints*. 2017. № 124. 14 p.
54. Kislitsyn A., Savinkov R., Novkovic M., Onder L., Bocharov G. Computational Approach to 3D Modeling of the Lymph Node Geometry // *Communication*. 2015. 3(2): 222 – 234.
55. Kislitsyn A.A., Orlov Yu.N. On the Distribution of the Stationary Point of Significance Level for Empirical Distribution Function // *IEEE Transactions*, 2019.
DOI:10.1109/ICUMT.2018.8631234
56. Kislitsyn A.A., Orlov Yu.N. Nonstationary stochastic motion modeling by dynamical systems // *ECMS Conference Proc.*, 2019. P. 466-472.

DOI: 10.7148/2019-0466.

57. Kislitsyn A.A, Orlov Yu.N. Chernoff approximation for non-stationary random walk modeling // Lobachevsky Journal of Mathematics, 2019. Vol. 40. No 12. P. 2095-2102.

58. Kislitsyn A.A, Orlov Yu.N. Dynamical System Model with the use of Liouville Equation for Empirical Distribution Function Densities // Discontinuity, Nonlinearity and Complexity, 2020. Vol. 9. No 4. P. 529-540.

DOI: 10.5890/DNC.2020.12.006

59. Kolmogoroff A.N. Sulla determinazione empirica di una legge di distribuzione // Giornale dell' Istituto Italiano degli Attuari. 1933. – Vol. 4. – № 1. – P. 83-91.

60. Lakhina A., Crovella M., Diot C. Characterization of network-wide anomalies in traffic flows // Proceedings of the 4th ACM SIGCOMM conference on Internet measurement - IMC '04. – 2004. Vol. 6. P. 201.

61. Lakhina A., Crovella M., Diot C. Detecting distributed attacks using networkwide flow traffic // Proceedings of FloCon 2005 Analysis Workshop, 2005.

62. Lakhina A., Crovella M., Diot C. Diagnosing network-wide traffic anomalies // ACM SIGCOMM Computer Communication Review, 2004. Vol. 34, no. 4. P. 219.

63. Lorden G. Procedures for Reacting to a Change in Distribution. 1971.

64. Moustakides G. V. Optimal Stopping Times for Detecting Changes in Distributions // The Annals of Statistics, 1986. Vol. 14, no. 4. P. 1379 – 1387.

65. Perucca E. An introduction to antiepileptic drugs // Epilepsia, 2005. Vol. 46, Suppl. 4. P. 31-37.

66. Peters T.M., Williams J. (eds.) The Fourier Transform in Biomedical Engineering. – Basel: Birkhäuser, 1998. – 199 p.

67. Pham D.-S., Venkatesh S., Lazarescu M., Budhaditya S. Anomaly detection in large-scale data stream networks // Data Mining and Knowledge Discovery. 2014. Vol. 28. P. 145–189.

68. Roberts S. A comparison of some control chart procedures // Technometrics, 1966. Vol. 8, no. 3. P. 411 – 430.

69. Savinkov R., Kislitsyn A., Watson D., Loon R., Novkovic M., Onder L., Bocharov G. Data-driven modelling of the FRC network for studying the fluid flow in the conduit system // Engineering Applications of Artificial Intelligence. 2016, V. 62. DOI: 10.1016/j.engappai.2016.10.007

70. Schomer D.H., Lopes da Silva F.H. Nidermeyer's Electroencephalography. – Philadelphia, Lippincott Williams & Wilkins, 2011. – 1296 p.

71. Smolyanov O.G., Tokarev A.G., Truman A. Hamiltonian Feynman path integrals via the Chernoff formula // J. Math. Phys., 43, 2002. P. 5161-5171.

72. Temkin O. Falling sickness: History of epilepsy from the Greeks to the beginnings of modern neurology / 2nd ed. John Hopkins U.P., 1994.
73. Theis F.J., Meyer-Base A. Biomedical Signal Analysis: Contemporary Methods and Applications – Cambridge: MIT Press, 2010. – 428p.
74. Varsavsky A., Mareels I., Cook M. Epileptic seizures and the EEG: Measurement, models, detection and prediction. – CRC Press, 2011. 370 p.
75. Vautard R., Ghil M. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series // *Physica D: Nonlinear Phenomena*, 1989. Vol. 35, no. 3. P. 395–424.
76. Vautard R., Yiou P., Ghil M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals // *Physica D: Nonlinear Phenomena*, 1992. Vol. 58, no. 1. P. 95–126.
77. Yigitbasi N., Gallet M., Kondo D., Iosup A., Epema D. Analysis and modeling of time-correlated failures in large-scale distributed systems // *Proceedings IEEE/ACM International Workshop on Grid Computing*. 2010. P. 65–72.