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

УДК 519.242.33

На правах рукописи

Федоров Сергей Леонидович

МОДЕЛИРОВАНИЕ НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ И ПОСТРОЕНИЕ ОПЕРАТОРА ЭВОЛЮЦИИ ИХ ВЫБОРОЧНЫХ РАСПРЕДЕЛЕНИЙ НЕПАРАМЕТРИЧЕСКИМИ МЕТОДАМИ

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, 2017

ОГЛАВЛЕНИЕ

Общие сведения о работе....3
Введение....9
Глава I. Проблемы моделирования нестационарных временных рядов....14
1.1. Основные понятия в теории нестационарных временных рядов....14
1.2. Ограничения адаптивных методов прогнозирования временных рядов....22
1.3. Компьютерные программы для статистического анализа рядов....25
1.4. Кинетический подход к моделированию эволюции нестационарных функций распределения....31
1.5. Задача генерации нестационарного временного ряда....36
Глава II. Метод генерации ансамбля траекторий нестационарного временного ряда....40
2.1. Согласованный уровень стационарности и индекс нестационарности....40
2.2. Равномерное разбиение гистограммы и СУС в норме L1....47
2.3. Уравнение Фоккера - Планка для нестационарной ВПФР....58
2.4. Генерация выборки из нестационарной функции распределения....66
2.5. Статистический анализ функционалов, заданных на траектории случайного процесса....69
Глава III. Структура численного алгоритма моделирования нестационарных временных рядов....71
3.1. Алгоритм оптимального разбиения гистограммы....71
3.2. Алгоритм определения длины выборки для выявления нестационарности....72
3.3. Алгоритм решения уравнения Фоккера – Планка для ВПФР....73
3.4. Алгоритм генерации пучка нестационарных траекторий....76
3.5. Блок-схема программного комплекса....77
Глава IV. Результаты численных расчетов....85
4.1. Тестирование корректности модели прогнозирования ВПФР по уравнению Фоккера - Планка....85
4.2. Тестирование корректности модели генерации нестационарного временного ряда....87
4.3. Формирование паттернов и распознавание фрагментов траекторий....89
4.4. Пример статистического анализа функционала доходности торговой системы....93
4.5. Пример распознавания языка фрагмента текста....96
4.6. Анализ уровня стационарности сейсмограмм....98
Заключение....102
Список литературы....104

Общие сведения о работе

Актуальность темы.

Проблема моделирования нестационарных временных рядов, возникающих во многих областях человеческой деятельности, в настоящее время приобрела большое практическое значение в связи с развившимися возможностями вычислительной техники и резко возросшей детализацией описания самих процессов. Существует большое число примеров рядов данных, требующих моделирования с учетом нестационарных свойств, которые проявляют выборочные распределения наблюдаемых величин. Таковы биржевые ряды цен сделок на финансовые инструменты, кардиограммы и энцефалограммы в медицине, сейсмограммы, температурные кривые и показатели счетчиков радиоактивности, последовательности символов в текстах и цепочках геномов.

Анализ нестационарных случайных данных является частью проблемы так называемых Больших Данных, когда требуется разработать эффективный инструмент для сокращения описания, позволяющий тем не менее давать содержательные ответы на интересующие исследователя вопросы. Исторически существует важный пример эффективности применения кинетического подхода к анализу Больших Данных, в рамках которого оказалось возможным сведение большого числа уравнений механики к малому числу уравнений гидродинамики. Например, вместо точного решения уравнений движения для всех молекул газа в сосуде достаточно решить три уравнения относительно гидродинамических параметров, являющихся моментами локально-равновесной функции распределения этих молекул по скоростям, чтобы получить требуемые в большинстве практических задач ответы о давлении газа, его температуре и плотности. Однако к временным рядам, имеющим не только физическую, но и отчасти социальную природу, кинетический подход практически не применялся в силу отсутствия надежного динамического описания таких систем, а также и по причине вычислительной сложности возникающих статистических задач. Настоящая работа направлена на разработку и применение кинетического метода исследования Больших Данных, а также на создание программного продукта, достаточно универсального с точки зрения конкретной области его применения, для решения определенных задач стохастического управления, таких например, как оптимизация функционала штрафа, заданного на фрагменте траектории случайного процесса.

Традиционный подход к анализу нестационарных временных рядов состоит в том, что рассматриваются только такие ряды, которые с помощью различных линейных методов можно свести к стационарным. Соответствующие модели носят название

авторегрессионных интегрированных моделей скользящего среднего. Основы моделей такого типа были заложены в середине прошлого века Боксом и Дженкинсом [8]. Характерно, что эти модели оперируют не с функциями распределения, а непосредственно с элементами временного ряда. Ряды, не укладывающиеся в рамки регрессионного анализа, изучаются разными эвристическими методами, называемыми адаптивными, не имеющими четкого математического обоснования. В них предполагается, что ряды на некотором (правда, неизвестном) горизонте для выборки некоторой (правда, неизвестной) длины могут быть описаны той или иной стационарной моделью типа регрессии или авторегрессии, а потом (возможно, что прямо на следующем шаге) параметры такой модели должны быть пересчитаны с учетом новой информации или с учетом сравнения предсказанного значения с фактом. Недостатком этих подходов является то, что они применяются к единственной реализации случайного процесса, тогда как для эволюционирующих распределений методически более корректно изучать ансамбль возможных траекторий. Это требует использования кинетических уравнений – либо для генеральных совокупностей, либо для выборок. К преимуществам кинетического метода следует отнести также и то, что он не предполагает каких-то специальных свойств временных рядов, кроме естественного на практике требования равномерной ограниченности ряда по времени. Последнее нужно для того, чтобы при прогнозировании нестационарных распределений иметь возможность сравнения начального и конечного выборочного распределений на одной шкале значений случайной величины.

Кинетический подход к анализу нестационарных временных рядов развивается в настоящее время группой сотрудников в ИПМ им. М.В. Келдыша РАН под руководством д.ф.-м.н. Ю.Н. Орлова. Этот метод начал разрабатываться относительно недавно. Первая публикация [47] была сделана в 2007 г. Ю.Н. Орловым и К.П. Осмининым. В ней предлагалось использовать для прогноза нестационарной функции распределения уравнение Лиувилля с подходящей эмпирической скоростью переноса вероятности. Затем последовал ряд работ [10, 12, 48, 49, 50, 52, 53] по конструированию новых индикаторов нестационарности, ибо применение классических критериев к нестационарным процессам не вполне корректно. Подход с использованием уравнения Фоккера-Планка для описания эволюции выборочных функций распределения был предложен в [11, 13], однако соответствующий численный алгоритм реализован не был. Также не ставилась задача генерации виртуальных нестационарных траекторий, представляющая собой по сути реализацию нестационарного обобщения метода Монте-Карло. Эти вопросы и рассматриваются в представленной диссертационной работе.

Дадим теперь определения объекту, предмету и методу предпринимаемого научного исследования.

Объект исследования – нестационарные временные ряды.

Предмет исследования – кинетический подход к прогнозированию выборочной функции распределения нестационарного временного ряда.

Научная задача – разработка индикатора нестационарности временного ряда и создание численного алгоритма генерации ансамбля нестационарных траекторий, являющихся реализациями решения соответствующего кинетического уравнения.

Цель работы заключается в создании инструментария для тестирования функционалов, заданных на траектории нестационарного случайного процесса, и для изучения их статистических свойств.

Направления исследования. Для достижения поставленной цели в работе определена следующая последовательность исследований. Необходимо:

1. Разработать математическую модель индикатора нестационарности выборочных распределений временных рядов в разных нормах и реализовать ее в виде численного алгоритма.

2. Построить математическую модель эволюции выборочных функций распределения, такую, что уравнения эволюции моментов распределений заданных порядков, следующие из кинетического уравнения, совпадали бы с эмпирически наблюдаемыми их изменениями по элементам выборки.

3. Построить алгоритм численного решения кинетического уравнения относительно эмпирической функции распределения.

4. Предложить модель генерации нестационарной траектории, статистические свойства которой совпадают в пределах точности эксперимента с наблюдаемой выборочной функцией распределения временного ряда, и реализовать ее в виде алгоритма генерации ансамбля траекторий.

5. Разработать метод тестирования функционала, заданного на выборочной траектории нестационарного случайного процесса, с целью его возможной оптимизации и для анализа его статистических свойств.

Основные положения, выносимые на защиту, состоят в следующем.

1. Разработана математическая модель нестационарного временного ряда на основе численного решения эмпирического кинетического уравнения, которым описывается эволюция его выборочной функции распределения, и построена система индикаторов для идентификации уровня нестационарности в задачах статистического анализа нестационарных временных рядов.

2. Построен численный алгоритм генерации ансамбля траекторий нестационарного временного ряда и выборок из него в пределах заданного горизонта прогнозирования на основе решения уравнения Фоккера – Планка относительно выборочной плотности функции распределения, отвечающей данному временному ряду.

3. Разработан метод статистического анализа функционалов, заданных на траектории нестационарного случайного процесса, реализованный в виде программного комплекса с интерфейсом.

Научная новизна работы заключается в том, что впервые разработан и реализован в виде программы нестационарный аналог метода Монте-Карло применительно к анализу и прогнозированию временных рядов.

Личный вклад автора состоит в создании математических моделей временных рядов, разработке функционалов-индикаторов для описания нестационарности и создании программного комплекса, решающего вышеописанные задачи анализа временных рядов.

Апробация работы. Материалы диссертации докладывались на научных семинарах ИПМ им. М.В. Келдыша РАН, а также на конференциях:

1. Теоретические и прикладные аспекты современной науки. II научно-практическая международная конференция, август 2014, Белгород, Россия.

2. ICNAAM, September 19-25, 2016, Rhodes, Greece.

Публикации. По материалам диссертации опубликовано 10 работ. Из них 2 статьи в рецензируемых журналах, 2 статьи в трудах международных конференций и 6 препринтов ИПМ им. М.В. Келдыша РАН.

1. 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.

2. Клочкова Л.В., Орлов Ю.Н., Федоров С.Л. Моделирование ансамбля нестационарных траекторий с помощью уравнения Фоккера-Планка // Журнал Средневолжского математического общества, 2016. – Т.18. № 1.

3. Орлов Ю.Н., Федоров С.Л. Моделирование и статистический анализ функционалов, заданных на выборках из нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша. 2014. № 43. 26 с.

4. 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.

5. 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.

6. Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л. Метод базисных паттернов в анализе нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2016. № 7. 20 с.

7. Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А. Статистические закономерности европейских языков и анализ рукописи Войнича // Препринты ИПМ им. М.В. Келдыша. 2016. № 52. 36 с.

8. 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.

9. Федоров С.Л. Анализ функционалов, заданных на выборках из нестационарного временного ряда. // Труды II Международной научно-практической конференции Теоретические и прикладные аспекты современной науки, Белгород, август 2014. С. 9-16.

10. Yu. Orlov, S. Fedorov, A. Samouylov, Yu. Gaidamaka, D. Molchanov. Simulation of Devices Mobility to Estimate Wireless Channel Quality Metrics in 5G Network // Proc. ICNAAM, September 19-25, 2016, Rhodes, Greece.

Структура диссертации.

Диссертация «Моделирование нестационарных временных рядов и построение оператора эволюции их выборочных распределений непараметрическими методами» состоит из введения, четырех глав, заключения, приложения и списка литературы из 66 наименований, расположенных в алфавитном порядке. 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 сквозную нумерацию.

Во введении дается краткий обзор основных направлений исследований в области статистического анализа нестационарных временных рядов и формулируются проблемы, возникающие при разработке методов генерации нестационарных временных рядов, опирающихся на определенный закон эволюции выборочных функций распределения.

В первой главе представлен обзор основных методов аналитического и численного анализа и прогнозирования временных рядов. Рассмотрены подходы, развитые в теории стационарных и нестационарных временных рядов. Ставится задача о генерации пучка нестационарных траекторий случайного процесса, причем эволюция выборочной плотности функции распределения анализируется непараметрическими методами.

Во второй главе строится методика анализа нестационарных временных рядов, направленная на решение задач, которые возникают перед исследователем в практической работе при оценивании выборочной плотности. Эти задачи следующие: равномерное разбиение гистограммы; определение согласованного уровня стационарности выборочной плотности; нахождение длин выборок, на которых индекс нестационарности временного ряда максимален, а также длин, на которых распределение стационарно. Формулируется также метод статистической генерации нестационарной траектории.

В третьей главе приводятся алгоритмы решения поставленных задач, которые объединяются в единый программный комплекс. Для решения кинетического уравнения выписывается также соответствующая разностная схема. Дается общая блок-схема программного комплекса и описываются его возможности.

В четвертой главе приводятся результаты численного моделирования временных рядов, прототипы которых взяты из открытых данных ценовых рядов на финансовых рынках.

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

Введение

Настоящая диссертация посвящена развитию кинетического подхода к анализу и прогнозированию нестационарных временных рядов. Она продолжает направление исследований, проводимых в отделе кинетических уравнений и вычислительной физики ИПМ им. М.В. Келдыша РАН применительно к различным областям деятельности. Кинетический подход опирается на понятийный аппарат выборочных функций распределения, которые эволюционируют в соответствии с определенным модельным кинетическим уравнением. Построение такого модельного уравнения и его численное решение являются центральными задачами анализа временных рядов в рамках этого подхода. Основным отличием его от других методов является то, что в нем не делается попытки продолжить наблюдаемую в эксперименте траекторию некоторого случайного процесса, как это имеет место в моделях регрессионного типа, а предлагается исследовать ансамбль возможных траекторий, выборочные распределения которых ведут себя так же, как и наблюдаемые в эксперименте.

В эконометрических и экономико-математических моделях, применяемых при изучении и оптимизации процессов маркетинга и менеджмента, управления предприятием и регионом, точности и стабильности технологических процессов, в задачах надежности, обеспечения технологической и экологической безопасности, функционирования технических устройств и объектов, разработки организационных схем часто применяют понятия и результаты теории вероятностей и математической статистики, считая распределения стационарными. Чтобы применить эти распределения на практике на приемлемом уровне значимости, надо быть уверенным, что с заданной точностью выборочная функция распределения случайной величины будет близка к ее предполагаемому теоретическому распределению. Такая уверенность основана на том, что для стационарных в узком смысле случайных процессов выборочное распределение сходится по вероятности к теоретическому. Если есть основания считать, что процесс стационарен в широком смысле (т.е. существуют независящие от времени конечные моменты теоретического распределения нескольких первых порядков), то известно, что отклонения выборочных моментов от их теоретических значений распределены асимптотически нормально. Тем самым задача прогнозирования в стационарном случае может быть сведена к задаче аппроксимации средних величин.

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 (1933) для определения близости выборочной функции распределения случайной величины ξ к стационарному распределению, если оно есть. Именно, статистика $D_n = \sup_x |F_n(x) - F(x)|$ супремума модуля разности выборочной и точной интегральных функций распределения стационарной случайной величины ξ, принимающей значение x, по вероятности стремится к нулю с ростом объема выборки n так, что случайная величина $\sqrt{n}D_n$ имеет асимптотическое распределение в виде табулированной K-функции Колмогорова [19, 33, 35]. В дальнейшем на основе этого утверждения были получены различные широко применяемые асимптотические критерии [9, 28, 29, 57] о принадлежности двух выборочных распределений одной генеральной совокупности: критерий Колмогорова-Смирнова (1939), Вальда-Волфовица (1940), Вилкоксона (1945), Манна-Уитни (1947), Гнеденко-Королюка (1951) и другие критерии, применяемые к оценкам выборочных моментов (Стьюдента, Фишера, Крамера-Уэлча, «омега-квадрат» и др.). Большое число статистических критериев собрано в справочнике [33].

Другим фундаментальным утверждением является теорема Вольда (1938) о разложении, согласно которой любой стационарный случайный процесс представляется в виде суперпозиции детерминированного процесса и белого шума.

Еще одним методологически важным результатом является теорема Гофдинга (1948), утверждающая, что умноженные на $\sqrt{n}$ отклонения моментов эмпирического распределения, построенного по выборке объема n, от моментов генеральной совокупности для стационарной случайной величины распределены асимптотически нормально. Эта теорема позволяет определить скорость сходимости по вероятности выборочных моментов и вероятность отклонения их значений от теоретических, если

таковые известны. На основе этой теоремы определяются доверительные вероятности и доверительные интервалы для выборочных оценок параметров распределений.

Перечисленные утверждения математической статистики определяют основные принципы моделирования стационарных временных рядов. Обычно ряд представляется в виде суммы некоторой детерминированной составляющей и остатка, автокорреляционная функция которого с достаточной точностью близка к нулю, что свидетельствует о близости остатка к белому шуму. После этого ставится задача о нахождении наиболее близкой статистики, моделирующей поведение остатка.

Такой подход корректно обоснован только для стационарных рядов. Однако многие временные ряды, встречающиеся на практике, не являются стационарными. В этом 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.

Кроме того, в адаптивных методах исследования рядов, про которые априори не известно, являются ли они (ряды) стационарными или нет, не решен вопрос, по выборке какого объема следует проводить скользящее усреднение, чтобы получить наименьшую ошибку прогноза. Решение этой проблемы в существующих критериях оставляется на усмотрение пользователя в соответствии с его жизненным опытом.

Таким образом, 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.

Кроме того, на практике часто возникает задача тестирования некоторого функционала управления, заданного на траектории случайного процесса, для которого конкретная реализация траектории существенна. Типичным примером такого функционала является торговая система, применяемая на рынке финансовых

инструментов, на эффективность работы которой влияет, во-первых, правильное распознавание текущей ситуации, определяющей правила входа-выхода, и, во-вторых, численные значения таких правил должны быть оптимизированы не по одной траектории, а по ансамблю траекторий. В результате необходимо также иметь генератор случайных величин, распределение которых эволюционирует, причем закон эволюции не параметрический. Решению этих задач и посвящена настоящая диссертация.

Кроме того, существенным аспектом является и то, что на практике статистический анализ данных всегда связан с численным алгоритмом, реализующим ту или иную методику. Поэтому создание эффективных численных алгоритмов для целей математической статистики является практически важной задачей.

Применение статистических методов в практических исследованиях с помощью универсального или специализированного программного обеспечения рассматривалось во многих работах, посвященных оптимизации вычислительных алгоритмов для целей различных задач, решаемых средствами математической статистики: [15, 16, 17, 46, 62]. Необходимость включения в более или менее стандартные пакеты новых алгоритмов, позволяющих повысить точность статистических оценок при проверке вероятностных гипотез, обсуждалась в работах [1, 30, 64]. Перспективы развития программного обеспечения для решения задач математической статистики рассматривались в [2, 5, 56].

Отметим, что недостаточность существующих методов, как теоретических, так и численных, для прогнозирования временных рядов, встречающихся, в частности, на рынках ценных бумаг, обсуждается во многих публикациях. Например, как показано в работах [26, 38, 39, 45], встречающиеся на практике задачи параметрического оценивания не всегда могут быть решены с помощью асимптотики нормального распределения. Некоторые приемы изучения предельных распределений статистик содержатся в [3, 14, 40-43]. Методы корреляционного анализа данных подробно изложены в монографии [60], где рассмотрены также и ограничения их применимости на практике.

Подробный обзор статистических методов и моделей, применяемых на рынках ценных бумаг, содержится в монографии [44]. Во многих примерах, рассмотренных в [44], случайный фактор, определяющий стохастическое поведение цен, предполагается возможным описать в рамках винеровского или пуассоновского процессов. Как правило, остатки, т.е. разности между реальным и модельным поведением наблюдаемой величины, предполагаются в таких моделях распределенными нормально или с плотностью, позволяющей по эмпирическим выборкам определить параметры этих распределений.

Непараметрические критерии оценивания данных, использующие методы математического моделирования Монте-Карло [22, 23], приведены в [28, 33, 63].

Существенно, однако, что эти методы применимы только к стационарному распределению, и не могут корректно, т.е. с точностью, оцениваемой по стационарным критериям, использоваться для анализа нестационарных временных рядов. Например, в работе [34] представлено программное обеспечение для моделирования траекторий некоторого случайного процесса со скачками применительно к ценам акций предприятий авиационной отрасли РФ. Математическая постановка задачи в этой работе использует стохастические дифференциальные уравнения с зависящими от времени коэффициентами сноса и диффузии (что совершенно правильно) и эрланговский поток событий для описания скачков, но практическая реализация случайных траекторий основана на стационарном методе Монте-Карло при постоянных коэффициентах стохастического уравнения. Тем самым фактически решена задача генерации траекторий двумерного стационарного случайного процесса, отличающегося от практически наблюдаемого временного ряда движения цен на акции предприятий. Следовательно, генерация именно нестационарного временного ряда остается весьма актуальной.

Таким образом, обзор литературы в области математической статистики и статистического моделирования показывает, что существует проблема создания методики корректного анализа и моделирования нестационарных временных рядов. Такой подход, использующий для описания нестационарных временных рядов кинетические уравнения для выборочных функций распределения этих рядов, был предложен в 2008 г. в работах Ю.Н. Орлова и К.П. Осминина [48-50]. В этих работах были выведены эмпирические уравнения эволюции вероятностных распределений и сформулирован метод построения математической модели случайного процесса. При этом для ошибки прогноза эмпирических распределений на заданный горизонт была указана точная верхняя грань. Этот метод был доведен в [50] до стадии алгоритма в случае, когда модель эволюции выборочных распределений использует уравнение Лиувилля. Однако такая модель не вполне адекватно описывает эволюцию моментов выборочных распределений, если исходный ряд не является рядом с независимыми приращениями. В настоящей работе алгоритм статистического анализа и прогнозирования временных рядов строится для уравнения Фоккера-Планка относительно выборочных функций распределения, что представляется более адекватным практическим нуждам.

Глава I. Проблемы моделирования нестационарных временных рядов

1.1. Основные понятия в теории нестационарных временных рядов

В этом параграфе рассмотрены основные методы анализа временных рядов, часто применяемые на практике. Эти методы в силу своей общеупотребительности служат базисом для сравнения с ними вновь разрабатываемых статистических моделей. Поскольку в диссертации предлагается некоторая новая математическая модель прогнозирования временных рядов, то для методологического сравнения и оценки ее качества следует кратко описать существующие методы анализа и прогнозирования временных рядов. Представляемый обзор далеко не полон. Его цель – не перечислить все существующие модели, а обрисовать место результатов диссертации среди многообразия существующих направлений статистического анализа.

Основными статистическими методами исследования временных рядов являются: выделение временного тренда, регрессионный, автокорреляционный, адаптивный, построение периодограмм, выделение главных компонент. Ниже кратко описывается идеология этих методов, даются основные определения из математической статистики и приводятся базовые уравнения соответствующих моделей.

Напомним [6, 19, 35], что случайным процессом на некотором вероятностном пространстве называется параметрическое семейство случайных величин $x(t)$, принимающих значения из множества, называемого областью определения процесса. Если параметр t принимает дискретные значения, то процесс называется временным рядом.

Временной ряд называется стационарным в широком смысле, если его математическое ожидание не зависит от времени t, а корреляционная функция, являющаяся математическим ожиданием произведения отклонений значений ряда от среднего в различные моменты времени t_1 и t_2, зависит только от разности $t_1 - t_2$. Более общее определение стационарности в широком смысле [35] предполагает независимость от времени центральных моментов ряда вплоть до некоторого конечного порядка.

Случайный процесс $x(t)$ называется стационарным в узком смысле [35], если при любых t и τ случайная величина $x(t)$ распределена одинаково с величиной $x(t+\tau)$, т.е. стационарной является его функция распределения.

В диссертации используется определение стационарности в широком смысле, если речь идет о моментах ряда, и в узком смысле, если о его распределении.

Рассмотрение существующих подходов к анализу временных рядов начнем с метода временного сглаживания или выделения тренда. При исследовании временных рядов традиционно принято выделять несколько типов составляющих [28, 36]:

$$x(t) = x_{тренд}(t) + x_{цикл}(t) + \xi(t), \qquad (1.1.1)$$

где $x_{тренд}(t)$ – сравнительно плавно (медленно) меняющаяся компонента, определяемая долговременной тенденцией изменения ряда признаков, называемой трендом, $x_{цикл}(t)$ – циклическая или так называемая сезонная компонента, которая отражает повторяемость процессов на определенных промежутках времени, а $\xi(t)$ – случайная компонента, содержащая влияние прочих факторов, механизм которого (влияния) скрыт от наблюдателя. Первые две составляющих (тренд и цикл) в идеале должны быть описаны точно, т.к. это закономерные факторы, изучаемые в рамках детерминистских моделей. Однако следует заметить, что сами детерминистские модели представляют определенную идеализацию описываемых закономерностей, поэтому им также присуща некоторая неточность. В этом смысле представление (1.1.1) несколько условно, но оно бывает полезно на практике для интерпретации результатов статистического анализа данных.

Трендовая компонента временных рядов обычно не бывает известна точно, а, как и ряд в целом, является случайной величиной, но ее изменение из некоторых априорных суждений часто может быть качественно описано аналитически. Для описания тренда используются т.н. кривые роста, которые позволяют моделировать процессы трех основных качественных типов: без предела роста, с пределом роста без точки перегиба, а также с пределом роста и точкой перегиба.

Процессы развития без предела роста характерны в основном для абсолютных объемных показателей. Процессы с пределом роста характерны для относительных показателей, таких, как душевое потребление продуктов питания, внесение удобрений на единицу площади, затраты на единицу произведенной продукции и т.п. Процесс с пределом роста и точкой перегиба характерен, например, для описания изменения спроса на новые товары.

Для моделирования этих процессов используются полиномиальные или квазиполиномиальные (с экспоненциальными множителями и т.п.) зависимости, дробно-рациональные и линейно-логарифмические функции, кривые Гомперца и иные функциональные зависимости. В рамках многопараметрических моделей часто бывает возможно провести аппроксимацию данных с требуемой точностью, однако этот подход не всегда удовлетворителен при прогнозировании, поскольку подбираемые функции не

обязательно отражают реально обусловленную зависимость наблюдаемой величины от времени.

Таким образом, часто используемым методом моделирования нестационарных временных рядов является параметрическое оценивание. В этом случае подбираются параметры той или иной функциональной зависимости для трендовой составляющей, после исключения которой остается стационарный ряд. Оставшийся ряд может и не быть стационарным в смысле математического определения этого понятия, но на практике его удобно считать таковым с доверительной вероятностью, достаточной для исследователя. Для этой цели используются различные тесты на стационарность [33], которые, как правило, разработаны для применения к известным функциональным зависимостям (напр., нормального, экспоненциального или равномерного распределений).

Если нет оснований предполагать нетривиальную функциональную зависимость трендовой составляющей ряда, ее часто считают полиномиальной. В этом случае такой тренд может быть исключен путем перехода к первым, вторым и т.д. разностям в значениях ряда, т.е. вместо ряда $x(t)$ можно рассмотреть ряд $x(t)-x(t-1)$ или ряд из разностей более высокого порядка, называемый производным рядом. Такой метод достаточно эффективен, если функциональный тип тренда сохраняется во времени.

Целью сведения временного ряда к стационарному является появляющаяся тогда возможность использования теоремы Гливенко о сходимости эмпирической вероятности к распределению генеральной совокупности и критерия согласия Колмогорова о близости выборочной функции распределения и распределения генеральной совокупности [19, 35] для того, чтобы попытаться определить вид распределения, к которому относилась бы изучаемая выборка данных, после чего с известной доверительной вероятностью строить прогноз.

Именно, если n_i есть количество элементов выборки объема T, попавших в некоторый отрезок Δ_i из области значений случайной величины x, $\sum_i n_i = T$, и p_i есть априорная вероятность попадания результата наблюдения в данный отрезок, то, согласно теореме Гливенко, отношение n_i / T равномерно сходится по вероятности к p_i при $T \to \infty$, т.е.

$$P\left\{ \lim_{T\to\infty} \sup \left| \frac{n_i}{T} - p_i \right| = 0 \right\} = 1. \qquad (1.1.2)$$

Согласно критерию Колмогорова-Смирнова [19], если различные выборки с эмпирическим распределением $F_T(x)$ принадлежат одной и той же генеральной

совокупности с некоторым теоретическим распределением $F(x)$, то существует предел по вероятности

$$\lim_{T\to\infty}^{(P)} D_T = 0, \quad D_T = \sup_x \left|F_T(x) - F(x)\right|, \tag{1.1.3}$$

причем функция распределения величины $\sqrt{T}D_T$ стремится по вероятности к функции Колмогорова $K(z)$, так что выполнен критерий

$$\lim_{T\to\infty} P\left\{\sqrt{T}D_T < z\right\} = K(z), \quad K(z) = \begin{cases} 0, & z \le 0 \\ \sum_{-\infty}^{+\infty} (-1)^k \exp\left(-2k^2 z^2\right), & z > 0 \end{cases} \tag{1.1.4}$$

Желание иметь дело со стационарным рядом вызвано также возможностью обосновать прогнозные модели для такого ряда применением теоремы Вальда о разложении, согласно которой всякий стационарный процесс может быть единственным образом представлен в виде суммы двух некоррелированных между собой процессов: детерминированного (сингулярного процесса), прогноз которого на любое время вперед безошибочен, и чисто случайного (регулярного белого шума, т.е. стационарного процесса, фурье-разложение которого является константой). Поэтому, хотя реальные процессы, как правило, не являются стационарными, тем не менее, возникает желание в первом приближении считать их таковыми. Такой подход может дать удовлетворительный результат в задачах краткосрочного прогнозирования.

Ряды, которые после надлежащих приготовительных операций можно считать стационарными, далее изучаются методами регрессионного, корреляционного и гармонического анализов. Каждый из этих методов используется для создания некоторой прогнозной модели для изучаемых рядов. В зависимости от конкретной специфики ряда используются различные из перечисленных методов. Ниже кратко описаны их содержательные части.

Линейная регрессионная модель (ЛРМ) позволяет связать две величины Y и X линейной зависимостью вида $Y = aX + b$ по имеющимся N парам значений (x_k, y_k) методом наименьших квадратов (МНК):

$$\begin{gathered} Y - \bar{y} = a(X - \bar{x}), \\ a = \frac{\langle \Delta x \Delta y \rangle}{D^2(x)}, \quad \langle \Delta x \Delta y \rangle = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y}), \\ D^2(x) = \left\langle (\Delta x)^2 \right\rangle, \quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n. \end{gathered} \tag{1.1.5}$$

Зависимость (1.1.5) называется регрессионной. Ее можно записать в симметричном виде:

$$\frac{Y-\bar{y}}{D(y)} = Cor(x,y)\frac{X-\bar{x}}{D(x)}, \quad Cor(x,y) = \frac{\langle \Delta x \Delta y \rangle}{D(x)D(y)}. \tag{1.1.6}$$

Прогноз по модели (1.1.5) строится в предположении, что найденные параметры $\bar{x}, \bar{y}, a$ не зависят от времени. Такие модели называются гомоскедастичными. Если эти величины зависят от t, то модель называется гетероскедастичной. Определение такой зависимости проводится в рамках параметрических моделей, использующих для искомых величин определенные функциональные зависимости, либо посредством статистического анализа скользящих средних в виде авторегрессионных моделей или регрессий на время.

Разность между исходным рядом и его регрессионной аппроксимацией называется остатком. Если остатки коррелированы, то для уточнения исходной регрессионной модели применяется автокорреляционная модель для остатков ряда. Сам факт корреляции остатков определяется путем анализа автокорреляционной функции остатков либо с помощью критериев Стьюдента или Дарбина-Уотсона [33, 57].

Обобщение модели (1.1.5) на случай зависимости, возможно, нелинейной, от нескольких объясняющих переменных приводит к задаче выбора наиболее адекватной модели по числу переменных и виду регрессионных функций.

Регрессионные модели применяются в основном тогда, когда объясняющая переменная (в формулах (1.1.5) это величина X) не является случайной, а автокорреляция между значениями другой (объясняемой величины Y) мала. Для оценки автокорреляции используется выборочная автокорреляционная функция, определяемая по значениям ряда x_n как [35, 36]

$$F_N(n) = \frac{(N-n)\sum_{k=1}^{N-n} x_k x_{k+n} - \sum_{k=1}^{N-n} x_k \sum_{j=1}^{N-n} x_{j+n}}{\sqrt{(N-n)\sum_{k=1}^{N-n} (x_k)^2 - \left(\sum_{k=1}^{N-n} x_k\right)^2}\sqrt{(N-n)\sum_{k=1}^{N-n} (x_{k+n})^2 - \left(\sum_{k=1}^{N-n} x_{k+n}\right)^2}}. \tag{1.1.7}$$

Максимумы модуля автокорреляционной функции показывают наличие лагов, т.е. промежутков времени, на которых проявляется скрытая зависимость случайных величин. К примеру, в рядах с существенным влиянием циклической компоненты лаги выражены на графике выборочной автокорреляционной функции особенно сильно. Модели, использующие лаговую автокорреляцию, называются автокорреляционными (АМ) или авторегрессионными.

Первая АМ была построена Боксом и Дженкинсом [8]. Такая модель, называемая моделью порядка p, прогнозирует значение случайной величины $z(n)$ через p более ранних значений, используя имеющуюся выборку из N величин:

$$x_n = \sum_{k=1}^{p} U_k x_{n-k}\,, \quad x_n = z_n - \bar{z}\,, \quad \bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n\,, \tag{1.1.8}$$

Коэффициенты U_k находятся из определяющей системы уравнений Юла-Уокера, которая является прямым следствием (1.8):

$$\begin{cases} F_N(1) = U_1 + F_N(1)U_2 + \ldots + F_N(p-1)U_p \\ F_N(2) = F_N(1)U_1 + U_2 + \ldots + F_N(p-2)U_p \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ F_N(p) = F_N(p-1)U_1 + F_N(p-2)U_2 + \ldots + U_p \end{cases} \tag{1.1.9}$$

Как правило, используются АМ порядка не выше трех. Модель первого порядка имеет вид

$$x_n = F(1)x_{n-1}\,, \tag{1.1.10}$$

а модель второго порядка

$$x_n = \frac{F(1)\big(1-F(2)\big)x_{n-1} + \big(F(2)-F^2(1)\big)x_{n-2}}{1-F^2(1)}\,. \tag{1.1.11}$$

Для применения автокорреляционных моделей (АМ) желательно иметь временной ряд, автокорреляционная функция (1.1.7) которого имеет небольшое число максимумов и достаточно быстро спадает с ростом шага автокорреляции. Если имеется цикличность данных, которая меняется со временем, то полностью исключить ее различными методами сглаживания, как правило, не удается. В этом случае автокорреляционные модели применяются на этапе качественного анализа, точность которого должна быть улучшена с использованием других подходов.

Последовательное усложнение автокорреляционных моделей привело к появлению т.н. адаптивных моделей прогнозирования, которые базируются на объединении двух схем – скользящего среднего (СС) и авторегрессии. Модель СС состоит в том, что для определения свойств временного ряда с целью краткосрочного прогноза берется выборка последних данных за некоторый промежуток времени T. Например, среднее значение определяется по последней выборке заданного объема:

$$\bar{x}_T(t) = \frac{1}{T}\sum_{k=0}^{T-1} x(t-k)\,. \tag{1.1.12}$$

Затем применяются регрессионные или автокорреляционные модели вида (1.1.5) и (1.1.8), в которых средние величины определяются по формуле (1.1.12).

Модификацией (1.1.12) является взвешенная схема СС, когда оценкой текущего уровня является взвешенное среднее всех предшествующих уровней, причем веса при наблюдениях убывают по мере удаления от последнего уровня, т.е. информационная ценность наблюдений признается тем большей, чем ближе находятся они к концу интервала наблюдений.

Другим типом адаптивных прогнозных моделей является параметрическая модель сглаживания, когда значение в момент t+1 определяется не только статистиками самого ряда, но и прогнозом этого ряда за несколько предшествующих моментов времени, а также оценками текущих трендов (модели Брауна, Хольта и Уинтерса [36, 43, 44]). Как правило, адаптивные модели применяются для прогнозирования ряда на один шаг вперед. Например, двухпараметрическая модель Хольта использует сглаженные значения E ряда и значения его тренда H:

$$\begin{aligned} E_i &= a\left(E_{i-1} + H_{i-1}\right) + (1-a)x_i \\ H_i &= bH_{i-1} + (1-b)\left(E_{i-1} + E_{i-1}\right) \\ \hat{x}_{i+1} &= E_i + H_i \end{aligned} \qquad (1.1.13)$$

Здесь x_i - это текущее значение исходного ряда, $\hat{x}_{i+1}$ - прогнозное значение. Постоянные a и b в (1.1.13) называются константами сглаживания, относящимся к оценкам уровня и тренда соответственно. Выбор значений этих констант является достаточно субъективным. Возможно, например, их оценивание посредством МНК по серии тестовых аппроксимаций ряда путем отбора наилучших из них.

Одним из методов выделения т.н. главной части или главных компонентов временного ряда является метод сингулярного спектрального анализа (ССА) [20, 42], разработанного в теории нелинейных динамических систем [24, 37] и представляющего оригинальный подход к исследованию автокорреляционной зависимости. В этом подходе рассматривается выборка некоторого объема N и выбирается скользящее вдоль нее окно некоторого объема T. Получающиеся выборки $\{x_1, x_2, ..., x_T\}, \{x_2, x_3, ..., x_{T+1}\}$ и т.д. рассматриваются как векторы T-мерного линейного пространства. Расположив эти выборки в виде матрицы размером T*(N-T+1), можно найти ранг этой матрицы и определить базис (т.е. главные компоненты ряда) в пространстве скользящих окон. Затем можно попытаться связать значение x_{t+1} с базисными векторами типа $\mathbf{y}(t,T) = \{x_t, x_{t-1}, ..., x_{t-T+1}\}$ с помощью некоторой подходящей функции:

$$x_{t+1} = f\left(\mathbf{y}(t,T)\right). \qquad (1.1.14)$$

Динамическая система (1.1.14) порождает стационарный временной ряд, называемый главной частью исходного. Для этой системы можно определить траекторию, притягивающее множество и его размерность, функцию распределения точек траектории и другие характеристики.

Корреляционный анализ по моделям типа Бокса-Дженкинса (1.8) и их обобщениям позволяет выявить степень зависимости данных на выборках относительно небольшого объема. Если же требуется изучать выборки объемов свыше 10^6, как в биржевой торговле или статистической радиофизике, более адекватным является спектральное разложение стационарного случайного процесса, т.е. представление его в виде [35]

$$x(t) = \sum_k a_k \cos(\varphi_k + \omega_k t), \quad t \in Z, \tag{1.1.15}$$

где a_k - некоторые вещественные коэффициенты, а φ_k, ω_k - случайные равномерно распределенные величины из интервала $[-\pi;\pi]$. Если существует независящее от времени среднее значение $\bar{x}$ по генеральной совокупности, то ковариация $C(\tau)$ процесса (среднее значение произведения $(x(t+\tau)-\bar{x})(x(t)-\bar{x})$) представляется в виде

$$C(\tau) = \int_{-\pi}^{\pi} \cos \omega\tau dF(\omega), \quad F(\omega) = \frac{1}{2} \sum_{\omega_k \le \omega} a_k^2 . \tag{1.1.16}$$

Введенная в (1.1.16) функция $F(\omega)$ называется спектральной функцией процесса. Пусть $F(\omega)$ абсолютно непрерывна и $f(\omega) = F'(\omega)$ - спектральная плотность процесса. Тогда процесс называется белым шумом, если он имеет постоянную спектральную плотность: $f(\omega) = const$.

Спектральное представление корреляционной функции позволяет выявить скрытую периодичность значений ряда, но также может быть использовано и для оценки средних значений на основе данных выборки объема *n*. Например, асимптотика дисперсии оценки среднего значения имеет вид $nD(\bar{x}) \to 2\pi f(0)$. Однако оценка спектральной плотности на основе выборочной функции

$$f_n(\omega) = \frac{1}{\pi n}\left((x_1 - \bar{x})\cos\omega + (x_2 - \bar{x})\cos 2\omega + ... + (x_n - \bar{x})\cos n\omega\right)^2 \tag{1.1.17}$$

хотя и является асимптотически несмещенной, ее дисперсия не стремится к нулю при $n \to \infty$, т.е. она не является состоятельной. Можно получить состоятельные оценки для некоторых функционалов от $f(\omega)$, однако, их статистическая интерпретация не вполне очевидна. В то же время знание выборочной спектральной функции может дать дополнительную информацию о корреляционной функции, что полезно для

моделирования процесса. В этом и состоит основная ценность анализа периодограмм, т.е. фурье-образов функции ковариации.

Из проведенного рассмотрения следует, что ни одно из перечисленных основных направлений статистического анализа временных рядов не является универсальным: конкретный случайный процесс лучше всего моделируется методом одного из направлений. Нет утверждения о том, что некоторый метод при своем практическом применении дает наименьшую ошибку прогнозирования для любых временных рядов. Напротив, каждый из методов имеет определенные ограничения, препятствующие их эффективному применению к задачам прогнозирования. Проблемы, возникающие при анализе встречающихся на практике временных рядов, очерчены в следующем параграфе.

1.2. Ограничения адаптивных методов прогнозирования временных рядов

Методы, описанные в параграфе 1.1, корректно применимы только к стационарным рядам. Если ряд нестационарный, то теоремы об эффективности, состоятельности и асимптотической нормальности выборочных оценок и их дисперсий в общем случае не выполняются. Тем не менее, перечисленные методы применяются ко всем рядам, статистический анализ которых необходимо проводить для оптимизации той или иной практической деятельности. Новыми проблемами, возникающими при таком не вполне обоснованном применении, являются: задача минимизации ошибки прогнозирования для выбранного метода статистического анализа и задача выбора наиболее адекватной модели временного ряда. Последняя задача существует и при исследовании стационарных рядов, но в этом случае выбор модели может быть проведен по известным алгоритмам спецификации моделей [33, 57, 65], позволяющим отобрать оптимальное число параметров в рамках дисперсионного и корреляционного анализов.

При определении ошибки прогноза нестационарного временного ряда надо учесть два фактора: конечность выборки и различие распределений для разных выборок вследствие нестационарности процесса. Разные методы имеют неодинаковую чувствительность точности аппроксимации данных к действию указанных факторов.

Кроме того, горизонт прогноза нестационарного ряда ограничен, если при этом требуется сохранить заданную точность модели. В этой связи естественным является вопрос об оптимальном объеме выборки, на основе которой можно сделать прогноз с заданной точностью в интервале времени до указанного горизонта. Можно поставить и обратную задачу об определении допустимого горизонта прогноза, основанного на выборке фиксированного объема, а также задачу об определении точности прогноза на заданный промежуток времени по выборке заданного объема.

Таким образом, модели и методы прогнозирования стационарных рядов, такие, как регрессионные и корреляционные, нуждаются в адаптации при использовании их в нестационарном случае, поскольку тогда ошибка прогноза, получаемая этими методами, может не убывать с увеличением статистической базы. Рассмотрим в этой связи ограничения, присущие методам анализа и моделирования временных рядов, перечисленным в параграфе 1.1.

В моделях регрессионного анализа вида (1.1.5) средние величины (математическое ожидание, дисперсия, ковариация) постоянны. Уточнение этой модели в случае зависимости указанных величин от времени, т.е. от текущего значения *t*, может быть сделано посредством аналитического моделирования такой зависимости, либо переходом к первым, вторым и т.д. разностям в нестационарных временных рядах, выражающих зависимость средних величин от времени, либо тем же регрессионным анализом – но уже изучаемых величин на время. Окно усреднения становится при этом скользящим, как в формуле (1.1.10). Однако остается невыясненным, какой ширины должно быть это окно.

Те же проблемы возникают и при использовании АМ (1.1.8) или их обобщений. При этом возникают дополнительные трудности с анализом корреллограмм: например, необходимо отличать эффекты назначенной периодичности, связанной с суточным, недельным или иным циклом, и внутренне обусловленной зависимости между членами ряда. Увеличение промежутка усреднения в этом случае не приводит к успеху, поскольку зависимость, наблюдавшаяся в одной выборке, может исчезнуть в другой того же объема, но отнесенной к иному моменту времени. Усреднение корреллограммы по некоторому промежутку времени и переход к средней корреллограмме за период наблюдений увеличивает неточность прогноза на короткий промежуток времени, а на большом интервале такую задачу ставить вообще не очень осмысленно. Как и в случае с регрессионными моделями, наилучший период усреднения не известен.

Адаптивные модели типа (1.1.13), использующие весовые коэффициенты в обобщениях АМ, требуют весьма тонкой настройки сглаживающих функций в нестационарном случае, поскольку даже для стационарных процессов оптимальный выбор этих функций является отдельной достаточно сложной задачей. На практике адаптивные методы учета нестационарности тех или иных свойств данных применяются в скользящих окнах переменной длины. Эта длина и есть основной параметр оптимизации. Результатом оптимизации является фильтрация нестационарных компонент выборок из анализируемых временных рядов, получаемая локальным перебором. Тогда на тестовом массиве данных точность анализа может быть существенно повышена по сравнению с обычными корреляционными методами. Однако следует подчеркнуть, что на неоптимизированном

фрагменте данных точность прогноза, а не аппроксимации, резко снижается. Для стационарных процессов метод перебора выборок ограниченного объема, естественно, может быть достаточно эффективным. Если же процесс нестационарный, то в стороне остаются вопросы количественной оценки точности аппроксимации, которая меняется с течением времени, т.к. длина оптимальной выборки является локально переменной величиной с неизвестной статистикой.

Что касается гармонического анализа временных рядов, то он вообще имеет ценность только для стационарных в широком смысле процессов второго порядка, когда корреляционная функция зависит от разности моментов времен. Для нестационарных рядов большое число учитываемых членов ряда приводит к достаточно высокой погрешности в оценке статистических характеристик процесса в ближайшем будущем.

Метод сингулярного спектрального анализа представляется в этом контексте наиболее устойчивым к временному тренду, поскольку его задачей и является выделение соответствующих главных компонент ряда. Изменение с течением времени размерности пространства базисных векторов матрицы задержек, описанной в параграфе 1.1, представляется маловероятным событием: размерность является своеобразным индикатором данного процесса, обусловленного определенными физическими явлениями, и ее изменение будет свидетельствовать о том, что процесс изменился по своему качеству. Тем не менее, вопрос о размерности самой матрицы и количественной зависимости от этой размерности числа базисных векторов остается в этом методе открытым. Например, в работах [20, 43] говорится, что на первом этапе анализа, исходя из неких внутренних свойств системы (которые, заметим, еще не проанализированы), подбирается длина выборки и составляется матрица развертки, после чего применяется сингулярное разложение получающейся матрицы и производятся дальнейшие операции. Затем делается замечание, что качество выделяемых композиций разложения определяется вариативным параметром – длиной выборки, и предлагается использовать некий качественный графический критерий близости выделяемых компонент определенному идеальному сигналу для определения подходящей длины выборки. В указанных работах не дается количественной меры такой близости, а также не рассматривается вопрос о случайном совпадении результата применения графического критерия с желаемым результатом. Тем самым выбор длины скользящего окна исследователь должен делать, опираясь главным образом на свою интуицию.

Таким образом, задача об определении оптимального объема выборки является весьма важной для прогнозирования.

1.3. Компьютерные программы для статистического анализа рядов

Согласно материалам обзоров [1, 2, 5, 15, 66] по программному обеспечению для проведения статистического анализа данных на компьютере, существует более тысячи программ – статистических пакетов, выполняющих такие расчеты. Эти программы различаются, помимо удобства интерфейса, формата данных, более или менее подробного руководства и других непринципиальных для целей данной работы аспектов, полнотой доступных для пользователя методов анализа данных.

В зависимости от конкретной области приложений (физика элементарных частиц, геофизика, медицина, экономика и финансы, социология, психология и др.) соответствующие программы статистического анализа имеют определенные ограничения, чтобы не содержать в себе избыточных средств, которые не будут востребованы в данной области деятельности.

Кроме того, программное обеспечение в значительной степени ориентировано на определенного пользователя: это либо обучающие программы для студентов или работников без специального образования в математической статистике, в которых чаще всего заложены только элементарные операции и стандартные тесты, либо программы «среднего уровня», рассчитанные на пользователей, которые владеют специальной терминологией и знают преимущественные области применения того или иного метода, а также программы для специалистов. Специализированные программы достаточно универсальны, т.е. позволяют использовать практически все существующие в справочниках критерии и тесты, строить модели по выбранному алгоритму, прогнозировать. При этом во многих пакетах предусмотрена возможность программирования для решения той или иной конкретной статистической или теоретико-вероятностной задачи, поскольку сами задачи, которые не могут быть стандартизированы, естественно, не входят в пакет. Таким образом, имеющиеся программы подчинены трем главным целям: обучение, анализ, исследование.

Ниже перечислены наиболее востребованные статистические пакеты, в соответствии с классификацией, приведенной в [2]. Целью обзора, проводимого в этом параграфе, является не сравнительный анализ возможностей различного программного обеспечения в области математической статистики, а, во-первых, демонстрация того, что решение исследуемой в диссертации задачи не содержится ни в одном из модулей существующих программ, и, во-вторых, указание места разработанного соискателем алгоритма в ряду других программных продуктов. Разумеется, некоторые вопросы анализа нестационарных временных могут быть решены существующими средствами программного обеспечения, но при условии написания соответствующей процедуры как

составной части действующего алгоритма, если такая возможность предоставляется пакетом программ. Непосредственно же в статистических пакетах не дается информации, например, о распределении индекса нестационарности ряда или о доле стационарно объясняемых отклонений в статистиках тех или иных величин.

Универсальные пакеты имеют, как правило, название, совпадающее с фирмой-разработчиком: STATISTICA, SPSS, Minitab, Eviews, SAS, SYSTAT, Statgraphics (или, по другому, STSC), PSPP, SOFA, а также отечественный пакет Мезозавр (MESOSAUR) и др. Среди активно применяемых в последнее время универсальных пакетов, использующих язык C++, следует отметить такую программу как R, которая имеет открытый код и фактически представляет среду для проведения статистических исследований.

Полу-универсальные (т.е. имеющие определенный уклон в «специализацию») достаточно распространенные пакеты следующие: STADIA, ОЛИМП, РОСТАН, ODA, WinSTAT, Statit, UNISTAT, STATlab, Multivariance, JMP, SOLO. Алгоритмы и программы для многомерного статистического анализа содержатся в пакете ППСА.

Специализированные пакеты (классификация и снижение размерности, экспертные системы) следующие: КЛАСС-МАСТЕР, КВАЗАР, PALMODA, Stat-Media, STARC, BMDP/W, SigmaStat, Statistix, TURBO Spring-Stat-Win, СТАТЭКС, Statistical Navigator Pro, а также «Статистик-Консультант для Windows». Кроме того, существуют разнообразные нейросетевые пакеты [61, 63, 65].

Рассмотрим кратко структуру и возможности статистических программ на примере универсального программного пакета STATISTICA. Другие пакеты отличаются от него включением или, напротив, не включением более или менее стандартных тестов и критериев, которые содержатся в специализированных справочниках [33, 35].

В состав пакета STATISTICA входят следующие модули, согласно [2]: дисперсионного анализа, линейного и нелинейного регрессионного анализа, анализа временных рядов и прогнозирования, модули общих линейных и регрессионных моделей, а также модуль частных наименьших квадратов. Для целей настоящей работы интересен прежде всего модуль анализа временных рядов. Его структура описана ниже.

В модуле анализа временных рядов содержится блок дисперсионного анализа, предназначенный для применения в промышленных исследованиях, когда необходимо провести факторный анализ данных, которые представляют собой значения случайной величины, а не выбираются исследователем. В этом блоке можно анализировать производственные планы с любой комбинацией фиксированных и случайных эффектов, как стандартные факторные, так и иерархически вложенные планы. Реализованы методы оценки компонент дисперсии, исходя из принципов максимума правдоподобия и

ограниченного максимума правдоподобия. При оценке методом максимума правдоподобия применяются алгоритмы Ньютона-Рафсона и Фишера. Имеется набор функций для расчета взвешенных и невзвешенных маргинальных средних и их доверительных интервалов.

В модуле возможно также линейное и нелинейное оценивание параметров модели посредством подгонки. Пользователь может задать тип модели, вводя соответствующее уравнение в специальное окно редактора. Уравнения могут включать логические операторы, поэтому имеется возможность оценивать кусочно-разрывные модели регрессии и модели с индикаторами групп. В уравнениях могут быть использованы различные теоретические функции распределения (бета, биномиальное, Коши, хи-квадрат, экспоненциальное, экстремальных значений, Фишера, гамма, геометрическое, Лапласа, логистическое, нормальное, логнормальное, Парето, Пуассона, Рэлея, Стьюдента и Вейбулла). Пользователь может сам задать начальные значения, величину шага, критерий остановки итераций и т.д.

В набор результатов нелинейного оценивания входят: оценки параметров и их стандартные ошибки, матрица дисперсий и ковариаций для оценок параметров, предсказанные значения, остатки и соответствующие критерии согласия (лог-правдоподобие оцененной и «нулевой» моделей, критерий хи-квадрат для различий между средними, доля дисперсии, объясненная моделью, классификация наблюдений и отношение несогласия для выбранных моделей).

В модуле анализа временных рядов реализован широкий набор методов описания, построения моделей, декомпозиции и прогнозирования многомерных временных рядов, как во временной, так и в частотной области. Программа автоматически отмечает все этапы анализа временного ряда и сохраняет полную историю преобразований и полученные результаты (остатки ряда после применения той или иной модели, сезонную составляющую и т.д.), поэтому на любом этапе имеется возможность вернуться к более раннему шагу.

В модуле реализованы следующие преобразования: удаление тренда, удаление автокорреляций, сглаживание скользящими средними (невзвешенными или взвешенными, с весами, заданными пользователем или вычисленными по методам Даниеля, Тьюки, Хэмминга, Парзена и Бартлета), медианное сглаживание (когда среднее значение заменено медианой), простое экспоненциальное сглаживание, взятие разностей, суммирование, вычисление остатков, сдвиг, косинус-сглаживание, прямое и обратное преобразования Фурье. Можно выполнить анализ автокорреляций, частных автокорреляций и кросскорреляций.

Для задач прогнозирования реализована модель авторегрессии и проинтегрированного скользящего среднего. Перед построением модели ряд может быть подвергнут различным преобразованиям. В стандартный набор результатов входят оценки параметров, стандартные ошибки и корреляции. Предсказанные значения могут быть представлены в числовой и графической форме и добавлены к исходному ряду. Имеются многочисленные дополнительные функции для исследования остатков модели. Возможно применение регрессионных методов для переменных с запаздыванием.

Важной особенностью пакета является возможность проводить анализ прерванных временных рядов (т.н. рядов с интервенциями). Доступны следующие виды интервенций: однопараметрические скачкообразные, двупараметрические непрерывные, а также временные. Для всех прерванных моделей могут быть построены прогнозы по регрессионным моделям, которые (прогнозы) можно вывести на график вместе с исходным рядом или, если требуется, вычесть из исходного ряда.

В модуле реализованы все классические модели экспоненциального сглаживания. Задание модели может включать аддитивную или мультипликативную сезонную составляющую и/или линейный, экспоненциальный или демпфированый тренды; в частности, доступны популярные модели с линейным трендом Хольта-Винтерса. Пользователь может задавать начальное значение параметров сглаживания, начальное значение тренда и (если требуется) сезонные факторы. Для тренда и сезонной составляющей могут быть заданы независимые параметры сглаживания. Имеется возможность автоматического поиска лучшего набора этих параметров в смысле среднеквадратической, средней абсолютной или средней относительной ошибки.

В модуле имеется возможность задать произвольный сезонный лаг и выбрать либо аддитивную, либо мультипликативную сезонную модель. По выбранной модели вычисляются скользящие средние, отношения или разности, сезонные компоненты, ряд с сезонной поправкой, сглаженную тренд-циклическую и нерегулярную компоненты.

В модуле реализованы методы спектрального анализа одного ряда и кросс-спектрального анализа двух рядов. Результаты содержат коэффициенты при синусах и косинусах, периодограмму и оценку спектральной плотности.

Аналогичные возможности содержатся и в таких универсальных программных продуктах, как SAS или SPSS [2].

Таким образом, нетрудно заметить, что все представленные в перечисленных модулях алгоритмы рассчитаны скорее не на прогноз, а на аппроксимацию имеющихся данных. Прогноз же строится по алгоритмам, корректно применимым только к стационарным рядам, поскольку никакое скользящее окно не гарантирует сохранения

достигнутой сглаженности ряда в будущем, не говоря уже о подбираемых «на ощупь» модельных трендовых функциях.

Статистические пакеты, имеющие более узкую ориентацию, содержат ограниченное количество методов – тех, которые востребованы именно в данной области. Поскольку темой настоящей диссертации являются нестационарные временные ряды, то полезно рассмотреть те специализированные программы, которые применяются в соответствующих отраслях, где такие ряды возникают особенно часто – в экономической деятельности.

В настоящее время существует большое количество компьютерных эконометрических пакетов программ, позволяющих проводить регрессионный и корреляционный анализ временных рядов, встречающихся в практической деятельности экономиста или финансового аналитика. Как правило, эти пакеты рассчитаны на пользователя, владеющего элементарными знаниями из математической статистики, достаточными для того, чтобы построить количественную регрессионную модель взаимосвязи между сериями наблюдаемых данных или сделать оценку статистической гипотезы о принадлежности выборки некоторому функциональному классу. Однако такие программы не позволяют проводить более тонкие исследования, оптимизирующие точность аппроксимации данных или соответствующей прогнозной модели.

Одним из типичных активно используемых обучающих пакетов является «Econometric Views». Этот пакет позволяет находить по заданной выборке из стационарного ряда различные выборочные характеристики (первые четыре момента, коэффициенты асимметрии и эксцесса, медиану распределения и т.п.), а также проверить гипотезу о том, что выборка взята из той или иной генеральной совокупности (нормальной, лог-нормальной, пуассоновской и др.) Пакет позволяет строить регрессионные модели методом наименьших квадратов, взвешенным методом наименьших квадратов, моделями скользящего среднего и проинтегрированного скользящего среднего, тестировать различные гипотезы о характере связи между изучаемыми переменными. В пакете предусмотрена возможность генерировать методом Монте-Карло временные ряды с известным распределением (равномерным или нормальным).

Из более мощных специализированных программ следует отметить пакет ЭВРИСТА [66] для исследования временных рядов, в котором реализовано около 100 различных алгоритмов статистического анализа. Ниже кратко описаны возможности этой программы.

Программа позволяет провести сравнение близости двух выборок с помощью критериев Уилкоксона, Клотца, Колмогорова-Смирнова, хи-квадрат, Стьюдента, Фишера.

Возможна пред-обработка данных, которая включает в себя: преобразование Бокса-Кокса, взятие сезонных и несезонных разностей, вычисление автокорреляционной функции, вычисление аддитивной и мультипликативной сезонной компоненты, заполнение пропусков методом скользящего среднего.

Анализ тренда проводится методами простого или полиномиального скользящего среднего, а также по формулам Спенсера. Строятся доверительные интервалы для некоторых нелинейных моделей.

Возможно моделирование ряда с заданным законом распределения, а также тестирование выборок на соответствие заданному закону распределения по критериям хи-квадрат Пирсона и Колмогорова-Смирнова.

Прогнозирование временных рядов строится в программе на основе моделей Брауна, Хольта-Уинтерса, подбором тренда, авторегрессией-скользящего среднего, по сезонным и несезонным авторегрессионным моделям и др.

Спектральный анализ, проводимый системой ЭВРИСТА, включает в себя построение сглаженных оценок автокорреляционной функции и периодограммы временных рядов по частотной или временной шкале, фильтрацию данных нерекурсивным, рекурсивным тангенсным или рекурсивным синусным фильтрами.

Подводя итог, можно сделать вывод о том, что существующие программные пакеты отличаются один от другого, главным образом, количеством используемых в них классических критериев. Нестандартные критерии, такие как предлагаемый в работе интегральный критерий близости двух выборочных функций распределения, не применяются. Зависимость точности прогноза от объема выборки и горизонта прогнозирования не входит в перечень генерируемых результатов. Ни в одном из пакетов не предусмотрена возможность анализа ансамбля нестационарных траекторий, подчиненных определенному кинетическому уравнению.

В заключение этого параграфа следует подчеркнуть, что стационарные ряды возникают, как правило, в задачах естественно-физического содержания, когда результаты серии экспериментов распределены вокруг некоторого среднего значения, представляющего объективно существующее свойство физического объекта. Нестационарные же ряды сопутствуют обработке результатов случайных экспериментов в субъективной сфере человеческих отношений – в экономике, социологии, психологии, лингвистике. Из-за необходимости проводить их анализ исследователь вынужден использовать те методы, которые есть, т.е. методы анализа стационарных рядов. Если ряд

в некотором смысле «слабо нестационарен», то такой подход может быть успешен. В общем случае произвольного нестационарного ряда анализ, строго говоря, невозможен, поскольку нельзя выделить группу членов с некоторым сходным поведением. Однако можно придать понятиям «близость» и «сходное поведение» точный количественный смысл, позволяющий выделить из временного ряда те выборки, на которых ряд стационарен в смысле некоторого подходящего критерия. Такой критерий был введен в работах [48-50] и позволил перейти от нестационарных рядов к изучению их квазистационарных выборочных распределений.

1.4. Кинетический подход к моделированию эволюции нестационарных функций распределения

Исследуем вопрос о том, какими уравнениями имеет смысл описывать эволюцию плотности выборочного распределения. С одной стороны, поскольку ВПФР строится по конечной выборке данных, то уравнение ее эволюции по своему существу должно быть дискретным в соответствии с разбиением гистограммы на классовые интервалы. С другой стороны, удобнее использовать уравнения эволюции в дифференциальной форме, поскольку тогда более четко можно проследить аналогию с динамической системой, которой бывает удобно моделировать временной ряд, а дискретная запись будет представлять собой численную схему расчетов в каждом конкретном случае.

Поскольку имеется закон сохранения вероятности со временем, то, используя аналогию со статистической механикой, можно предположить, что модель эволюции ВПФР должна основываться на некотором аналоге уравнения Лиувилля. Тем самым выборочной функции распределения будет локально по времени сопоставлена некоторая динамическая система. Этот подход был предложен в [48] и развит в [10, 11]. Отметим, что кинетическое уравнение типа Лиувилля с источником было введено в [51] для моделирования демографических процессов. В работе [11] было отмечено, что если уравнение Лиувилля дает не достаточно точное приближение, то следует использовать уравнения большего порядка по производным относительно изменений случайной величины, например, уравнение Фоккера-Планка.

Однако, имея в виду применение теории к нестационарным процессам, следует указать на существенное отличие того уравнения неразрывности, которое может быть построено в фазовом пространстве координат (значений ряда $x(t)$), скоростей (приращений этого ряда $v(t) \equiv \dot{x} = x(t+1) - x(t)$), и, возможно, ускорений высших порядков [21], от уравнения Лиувилля классической статистической механики.

В случае динамических систем, заданных аналитически, во многих случаях удается в явном виде определить динамически-инвариантную меру. В случае же численной реализации некоторой системы (возможно, что и не динамической) в виде временного ряда априорно выбираемая мера не обязана быть инвариантной. Напомним, что вероятностная трактовка уравнения Лиувилля как уравнения, сохраняющего нормировку, есть просто переформулированная теорема Лиувилля классической механики о сохранении фазового объема для консервативных динамических систем [7].

Для статистической выборки фазовый объем, определяемый разностными производными, доступными по этой выборке, может и не сохраняться точно. Чтобы устранить это неудобное обстоятельство, вводится требование равномерной ограниченности временного ряда. Тем не менее, выборочная скорость изменения значений ряда даже и в этом случае может не обладать требуемым свойством, поскольку численное определение дивергенции фазовой скорости проводится по конечному, а не бесконечному набору данных. В то же время ВПФР любой размерности нормирована на единицу по построению, поэтому оператор, переводящий ВПФР из одного момента времени в другой в том же самом фазовом пространстве, должен иметь вид закона сохранения нормировки. Но изучаемый временной ряд не обязательно порожден динамической системой, для которой и выводится уравнение Лиувилля. Следовательно, необходимо согласовать «выборочную фазовую скорость», появляющуюся в математической статистике, со скоростью в статистической механике, для чего следует провести подробный анализ того, как возникает уравнение эволюции для выборочной плотности функции распределения.

Предположим, что имеется достаточно большое количество данных $\{x(t)\}$, анализ которых позволяет определить оптимальный объем T выборки для построения ВПФР $f_T(x,t)$ и описания ее эволюции, скажем, на один шаг вперед. Рассматривая набор таких ВПФР $f_T(x,t),\ f_T(x,t-1),\ldots$ в предшествующие моменты времени, можно определить соответствующие эмпирические изменения этих плотностей за один шаг:

$$\frac{\partial f_T(x,t-k)}{\partial t} = f_T(x,t-k+1) - f_T(x,t-k), \quad k = 1, 2, \ldots . \tag{1.4.1}$$

Выражение, стоящее в правой части (1.4.1), определяет частную производную по времени в терминах наблюдения за временным рядом в скользящем окне с единичным шагом по времени.

Допустим, что знаний о таких изменениях первого порядка по времени оказалось достаточно для того, чтобы сложилось представление об эволюции ВПФР за один шаг по времени с точностью ε, для которой и был найден оптимальный объем выборки T.

Определим тогда подходящую «скорость» изменения ВПФР $u_T(x,t)$ так, чтобы изменение (1.4.1) можно было записать в виде уравнения

$$\frac{\partial f_T(x,t)}{\partial t} + \frac{\partial}{\partial x} u_T(x,t) f_T(x,t) = 0 . \qquad (1.4.2)$$

Скорость $u_T(x,t)$, удовлетворяющую уравнению (1.4.2) совместно с (1.4.1), была введена в [48] и названа эмпирической лиувиллевой скоростью (ЭЛС). Далее считаем объем выборки *T* фиксированным и соответствующий индекс для краткости опускаем.

Таким образом, правая часть (1.4.1) представляет собой изменение по времени выборочной функции распределения в ячейке, предшествующей той, в которой вычисляется значение эмпирической скорости. Видно, что скорость изменения ВПФР в данный момент времени определяется значениями распределения в этот и последующий моменты. Это означает, что в текущий момент времени t ЭЛС известна для предыдущего момента $t-1$, что следует учитывать при составлении эволюционной модели. Именно, чтобы определить ВПФР в момент времени $t+1$, надо построить ЭЛС в момент времени *t*.

Выясним, какой статистический смысл можно придать введенной эмпирической скорости изменения ВПФР. Для этого рассмотрим уравнение эволюции первого выборочного момента, определяемого как

$$m(t) = \langle x \rangle_t = \int x f(x,t) dx . \qquad (1.4.3)$$

Из (1.4.2) следует, что эволюция первого момента записывается в виде

$$\frac{dm(t)}{dt} = \int x \frac{\partial f(x,t)}{\partial t} dx = -\int x \frac{\partial}{\partial x} u(x,t) f(x,t) dx .$$

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

$$\frac{dm(t)}{dt} = \int u(x,t) f(x,t) dx = \langle u \rangle_t = U(t) , \qquad (1.4.4)$$

где $U(t)$ есть среднее значение эмпирической скорости $u(x,t)$ по распределению, которое задано в момент времени *t*.

С другой стороны, то же самое изменение первого выборочного момента, вычисляемое непосредственно по элементам выборки временного ряда, есть

$$\frac{dm(t)}{dt} = m(t+1) - m(t) = \frac{1}{T}\sum_{k=1}^{T} x(t-T+k+1) - \frac{1}{T}\sum_{k=1}^{T} x(t-T+k) = \frac{1}{T}\sum_{k=1}^{T} \dot{x}(t-T+k) , \qquad (1.4.5)$$

т.е. $U(t)$, как показывает последнее равенство, есть среднее значение выборочной скорости, определяемой по приращениям значений исходного временного ряда.

Выражения (1.4.4) и (1.4.5) представляют среднее значение одной и той же скорости. Но, в отличие от (1.4.4), среднее в (1.4.5), будучи средним значением скоростей $\dot{x}$, должно вычисляться не через распределение $f(x,t)$ значений ряда, а через совместное распределение ряда и его приращений. Введем тогда двумерную ВПФР $F(x,\dot{x},t)$ совместного распределения случайных величин x и $\dot{x}$ исходного ряда и ряда его производной. Производная $\dot{x} = x(t+1) - x(t)$ трактуется как разность между значениями ряда в соседние моменты времени. Эта разность и есть аналог микроскопической скорости системы из T «частиц» (значений ряда). Для построения двумерной ВПФР $F(x,\dot{x},t)$ в момент времени t необходимо иметь значение случайной величины в момент $t+1$, чтобы можно было построить микроскопическую скорость в момент времени t.

Одномерная ВПФР $f(x,t)$ определяется затем по $F(x,\dot{x},t)$ формулой

$$f(x,t) = \int F(x,\dot{x},t)d\dot{x}. \qquad (1.4.6)$$

Заметим, что среднее значение эмпирической скорости из правой части (1.4.5) равно

$$U(t) = \int \dot{x}F(x,\dot{x},t)dxd\dot{x},$$

и эта же величина равна выражению $\int u(x,t)f(x,t)dx$ в (1.4.4).

Следовательно, средняя локальная скорость $u(x,t)$ определяется как

$$u(x,t)f(x,t) = \int \dot{x}F(x,\dot{x},t)d\dot{x}. \qquad (1.4.7)$$

Чтобы более корректно подойти к вопросу замыкания уравнения Лиувилля (1.4.2), будем его решать совместно с уравнением (1.4.7). Полученная система все еще не является замкнутой в смысле кинетического подхода, поскольку возникшая в (1.4.7) двумерная ВПФР $F(x,\dot{x},t)$ пока еще не определена как функция времени. Для нее также следует записать уравнение Лиувилля, в котором x и $\dot{x}$ формально считаются независимыми переменными.

По аналогии с производной случайной координаты по времени в уравнении Лиувилля (1.4.2), производную случайной скорости по времени также следует рассматривать в некотором усредненном смысле, для чего надо расширить пространство фазовых переменных. Тогда уравнение Лиувилля для $F(x,\dot{x},t)$ должно иметь следующий вид, вытекающий из тех же представлений о сохранении нормировки:

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x}(F\dot{x}) + \frac{\partial}{\partial \dot{x}}\left(F\left\langle\frac{\delta\dot{x}}{\delta t}\right\rangle\right) = 0. \qquad (1.4.8)$$

Смысл этого уравнения определяется тем, как трактуется величина $\langle\delta\dot{x}/\delta t\rangle$. Если рассматривается динамическая система, для которой известна зависимость ускорения

$\ddot{x} = w(x, \dot{x}, t)$ от координат, скоростей и времени, то следует считать, что $\langle \delta\ddot{x} / \delta t \rangle$ и есть это ускорение. В таком случае из (1.4.8) вытекает

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x}(F\dot{x}) + \frac{\partial}{\partial \dot{x}}(Fw) = 0 . \qquad (1.4.9)$$

Если же считать, что $\delta\ddot{x} / \delta t$ является случайной величиной, то в уравнении Лиувилля аналог динамического ускорения если и можно ввести, то в некотором усредненном смысле, т.е. в виде $\langle \delta\ddot{x} / \delta t \rangle$. Разумеется, это не «вывод» уравнения эволюции, а всего лишь эвристические соображения относительно того, каким могло бы быть кинетическое уравнение выборочной плотности функции распределения для нестационарного временного ряда.

Тем самым для определения величины $\langle \delta\ddot{x} / \delta t \rangle$ требуется дальнейшее расширение фазового пространства. Введем среднее ускорение как функцию координаты, скорости и времени:

$$w(x, \dot{x}, t)F(x, \dot{x}, t) = \int \ddot{x} F_3(x, \dot{x}, \ddot{x}, t) d\ddot{x} . \qquad (1.4.10)$$

Видно, что в этом случае возникает зацепление многочастичных распределений. Замыкание цепочки можно провести, задав априори выражение для интеграла от неизвестной величины $\int \ddot{x} F_3(x, \dot{x}, \ddot{x}, t) d\ddot{x}$ в правой части уравнения (1.4.10). Характерно, что такое замыкание совершенно не повлияет на вид уравнения для одномерной ВПФР в силу дивергентного характера слагаемых.

Если же нет оснований считать интеграл $\int \ddot{x} F_3(x, \dot{x}, \ddot{x}, t) d\ddot{x}$ известной величиной, то цепочка уравнений Лиувилля выписывается для ВПФР следующего порядка, и т.д. Однако, чем выше порядок ВПФР, тем труднее найти разумное основание для того или иного способа обрыва цепочки кинетических уравнений. Следовательно, желательно построить систему уравнений, замкнутую относительно одномерной ВПФР. Чтобы замкнуть уравнение Лиувилля (1.4.2), которое используется для прогнозирования ВПФР $f(x,t)$, следует независимо определить величину $u(x,t)$ как значение средней локальной скорости.

Таким образом, кинетический подход к изучению временного ряда состоит в том, что строится подходящее кинетическое уравнение для ВПФР определенной размерности и предлагается тот или иной метод его замыкания. Настоящая работа развивает этот подход, используя в качестве модельного уравнения не уравнение Лиувилля, а уравнение Фоккера-Планка, что включает в рассмотрение диффузионные процессы.

1.5. Задача генерации нестационарного временного ряда

Во многих задачах прикладного статистического анализа существует необходимость тестирования тех или иных индикаторов локального поведения временного ряда с целью оценки вероятности их правильного срабатывания. Как правило, индикаторы представляют собой функционалы от фрагментов траектории случайного процесса. Примеры индикаторов: отношение числа положительных приростов значений временного ряда к числу отрицательных за определенный период времени; угловой коэффициент прямой регрессии для выборки заданной длины; расстояние между выборочными плотностями функции распределения ряда в той или иной норме, дисперсия накопленного размаха за определенный период и т.п.

Чтобы оценить эмпирическую условную вероятность того, что определенный интервал значений индикатора отвечает ожидаемому исследователем поведению ряда в настоящем или будущем, нужно иметь много реализаций изучаемого процесса, тогда как в наличии имеется лишь одна фактически наблюденная траектория. На практике берется фрагмент траектории, который представляется достаточно большим, и на нем собирается требуемая статистика по индикатору: число ошибочных срабатываний, число ошибочных несрабатываний и число правильных срабатываний. Однако, если временной ряд нестационарный, то, например, оптимизация длины выборки для получения минимальной ошибки индикатора по некоторому фрагменту прошлой траектории не имеет особого смысла, поскольку оптимум искался для конкретного фрагмента траектории случайного процесса. При другой последовательности тех же самых значений временного ряда возможен иной результат оптимизации, если индикатор является не функционалом от функции распределения, а определяется последовательностью значений случайной величины на выборочной траектории. Возникает задача тестирования индикатора на устойчивость относительно различных реализаций случайного процесса, имеющего близкие (с точки зрения исследователя – квазистационарные) выборочные распределения. Для этого требуется сгенерировать пучок возможных траекторий временного ряда, исходящий из заданного текущего состояния, и проверить на нем устойчивость срабатывания индикатора.

Если процесс стационарный, то набор его траекторий может быть получен, исходя из следующих соображений.

Рассмотрим равномерно распределенную на конечном отрезке $[a; b]$ случайную величину ξ. Она имеет ПФР

$$f_U(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a,b]\\ 0, & x\notin[a,b]\end{cases} \tag{1.5.1}$$

Распределение (1.5.1) бывает удобно задать в виде, когда явно выделен центр μ промежутка длины $\omega=|b-a|$:

$$f_U(x;\mu,\omega)=\begin{cases}\dfrac{1}{\omega}, & \mu-\dfrac{\omega}{2}\le x\le\mu+\dfrac{\omega}{2}\\ 0, & x\notin[\mu-\omega/2,\ \mu+\omega/2]\end{cases} \tag{1.5.2}$$

Среднее значение этого распределения равно μ, а дисперсия равна $\omega^2/12$. Стандартизованным равномерным распределением является распределение на отрезке $[0;1]$, ПФР которого в терминах (1.5.2) есть $f_U(x;1/2,1)$. Распределение (1.5.2) переводится в $f_U(y;1/2,1)$ посредством линейного преобразования $y=\dfrac{1}{2}+\dfrac{x-\mu}{\omega}$.

Равномерное распределение используется для того, чтобы определить, является ли выборка $\{x_1,x_2,...,x_n\}$ выборкой из распределения с некоторой предполагаемой функцией распределения $F(x)$. В основе различных статистических критериев согласия с гипотезой о виде распределения $F(x)$ лежит следующее утверждение [19]. Пусть ξ есть случайная величина с непрерывной ФР $F(x)$. Тогда случайная величина $\eta=F(\xi)$ имеет ПФР $f_U(y;1/2,1)$. Действительно, поскольку по определению функции распределения для случайной величины, которая сама является функцией распределения, имеет место система равенств

$$P\big(F(x)\le y\big)=\begin{cases}1, & y>1\\ y, & 0<y\le 1\\ 0, & y\le 0\end{cases}$$

то плотность этой функции распределения является $f_U(y;1/2,1)$.

На основе этого утверждения строятся алгоритмы генерации временного ряда $\{x_k\}$ с заданной непрерывной плотностью функции распределения $f(x)$, для которой $F(x)$ есть ее функция распределения. Для этого генерируется произвольная последовательность чисел $\{y_k\}$, равномерно распределенных на $[0;1]$, после чего по формуле

$$y_k=F(x_k),\quad x_k=F^{-1}(y_k) \tag{1.5.3}$$

находятся элементы ряда $\{x_k\}$. Обращение функции распределения в (1.5.3) возможно в силу ее строгой монотонности.

На практике условие непрерывности ПФР формально нарушается, потому что наблюденные значения получены с конечной точностью измерения, но это не является принципиальным ограничением. Процесс считается непрерывным с той точностью, с которой проводятся измерения. Поэтому, несмотря на то, что гистограмма, дающая оценку плотности вероятности, формально отвечает некоторой дискретной случайной величине, функция распределения может быть сделана непрерывной, если внутри ячейки классового интервала гистограммы постулировать некоторое – например, равномерное – распределение.

Используя описанный метод, можно провести его обобщение для построения численной модели нестационарного временного ряда в соответствии с определенными требованиями, накладываемыми на его ВПФР. Сложность анализа и, следовательно, моделирования нестационарного временного ряда с заданными статистическими свойствами его распределения состоит в том, что на практике в каждый момент времени наблюдается только одно значение из предполагаемой генеральной совокупности текущего распределения вероятностей. Поэтому с увеличением периода наблюдения выборочное распределение, построенное по увеличивающемуся количеству данных, не сходится к генеральной совокупности, из которой эти данные получены. Следовательно, анализу доступны только выборочные распределения, и задача анализа состоит не в установлении вида генеральной совокупности, что невозможно, а в нахождении эмпирических правил изменения ВПФР на определенном промежутке времени. Эти правила и представляют собой те свойства, которые могут быть приписаны выборочному распределению и протестированы на различных реализациях случайного процесса с близкими распределениями.

Чтобы имитировать процесс, близкий к реальным наблюдениям, например, за динамикой цен на бирже на какой-либо финансовый инструмент (цена акции, индекс набора ценных бумаг, курс валюты и т.п.), предлагается следующая схема действий. На первом этапе по имеющимся историческим данным строятся выборочные распределения приростов $x(t)$ цен на этот инструмент за тот промежуток времени T, который представляет интерес. Например, строятся распределения приростов цен закрытия в определенном временном интервале для какого-нибудь инструмента за два соседних месяца по скользящей выборке длиной в месяц. Тем самым в каждый момент времени t определена ВПФР $f_T(x,t)$, допустим, 5-минутных приростов цен, и соответствующая ей ВФР $F_T(x,t)$ согласно. Здесь длина T выражена, естественно, в единицах 5-минутных интервалов. Затем генерируется стационарный равномерно распределенный на [0;1] ряд

чисел $\{y_k\}$ длиной T. Пусть t_0 есть начальный момент времени, в который известна функция распределения за первый месяц наблюдения. Тогда в последующие моменты времени с 5-минутным шагом одна из возможных траекторий случайного процесса, для которого ВПФР меняется от $f_T(x,t_0)$ до $f_T(x,t_0+T)$, строится по формуле обращения соответствующей локальной по времени функции распределения, движущейся в скользящем окне длины T:

$$y_k = F_T(x_k, t_0 + k). \qquad (1.5.4)$$

Подчеркнем, что, согласно (1.6.4), в каждый момент времени t из распределения $F_T(x,t)$ генерируется только одно значение ряда. Сама же $F_T(x,t)$ выступает в этот момент времени как генеральная совокупность. Тем самым имитируется процесс наблюдения за динамикой нестационарного временного ряда.

Задавая различные равномерно распределенные ряды $\{y_k\}_j$, где $j = 1,..., S$ есть номер генерации, можно получить пучок траекторий, ассоциированных с ВПФР текущих выборок. Эта конструкция и будет реализована в данной работе. Соответствующие алгоритмы описываются далее в главах II и III.

Глава II. Метод генерации ансамбля траекторий нестационарного временного ряда

2.1. Согласованный уровень стационарности и индекс нестационарности

Сравним между собой две ВФР в условиях, когда генеральное распределение не известно. В стационарном случае задача о принадлежности двух выборок одной генеральной совокупности решается непараметрической статистикой Колмогорова-Смирнова:

$$S_N = \sup_x \left| F_{1,N}(x) - F_{2,N}(x) \right|. \qquad (2.1.1)$$

Для статистики (2.2.1) имеет место асимптотика

$$\lim_{N\to\infty} P\left\{ 0 < \sqrt{\frac{N}{2}}\, S_N < z \right\} = K(z), \qquad (2.1.2)$$

где $K(z)$ есть табулированная функция Колмогорова [2, 3].

На практике формула (2.1.2) применяется следующим образом. Если в процессе сравнения двух ВФР было найдено значение S_N в соответствии с (2.1.1) 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 уровень значимости α, считаем, что если $1 - K(z) < \alpha$, то осуществилось маловероятное событие, несовместимое с понятием случайности, и эти 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. Некоторая неопределенность вывода состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Суть проблемы в том, что необходимо выяснить, как часто две независимых выборки длины N отличаются в смысле ВФР в норме C больше, чем на некоторое число ε такое, что

$$1 - K\left(\sqrt{\frac{N}{2}} \varepsilon \right) < \alpha. \qquad (2.1.3)$$

Если число α в (2.1.3) выбрано слишком малым, то может так случиться, что доля пар выборок, расстояние между которыми удовлетворяет этому условию, будет меньше, чем α, и тогда по факту ложного отбрасывания «своей» выборки ошибка идентификации

будет больше, чем α, тогда как критерий утверждает, что эта ошибка равна именно α. Обратно, если α выбрано слишком большим, то на практике может оказаться, что число пар выборок, для которых условие (2.1.3) не выполнено, больше, чем $1-\alpha$, т.е. больше уровня доверия, на котором предположительно выполнен эксперимент по сравнению выборок. Для иллюстрации описываемой проведем статистический эксперимент по генерации 10 тыс. независимых выборок длиной 1000 из нормального распределения. Для удобства сравнения вся сгенерированная совокупность чисел была нормирована на отрезок $[0;1]$. Далее эта совокупность была разбита на пары выборок, после чего для каждой пары строилось расстояние (2.1.1). Фрагмент ряда расстояний между выборками приведен на рис. 1. Видно, что доля очень близких выборок, как и очень далеких, весьма мала. Если назначить уровень значимости слишком малым, то есть риск постоянно совершать ошибку «пропуска цели», отклоняя «свою» выборку.

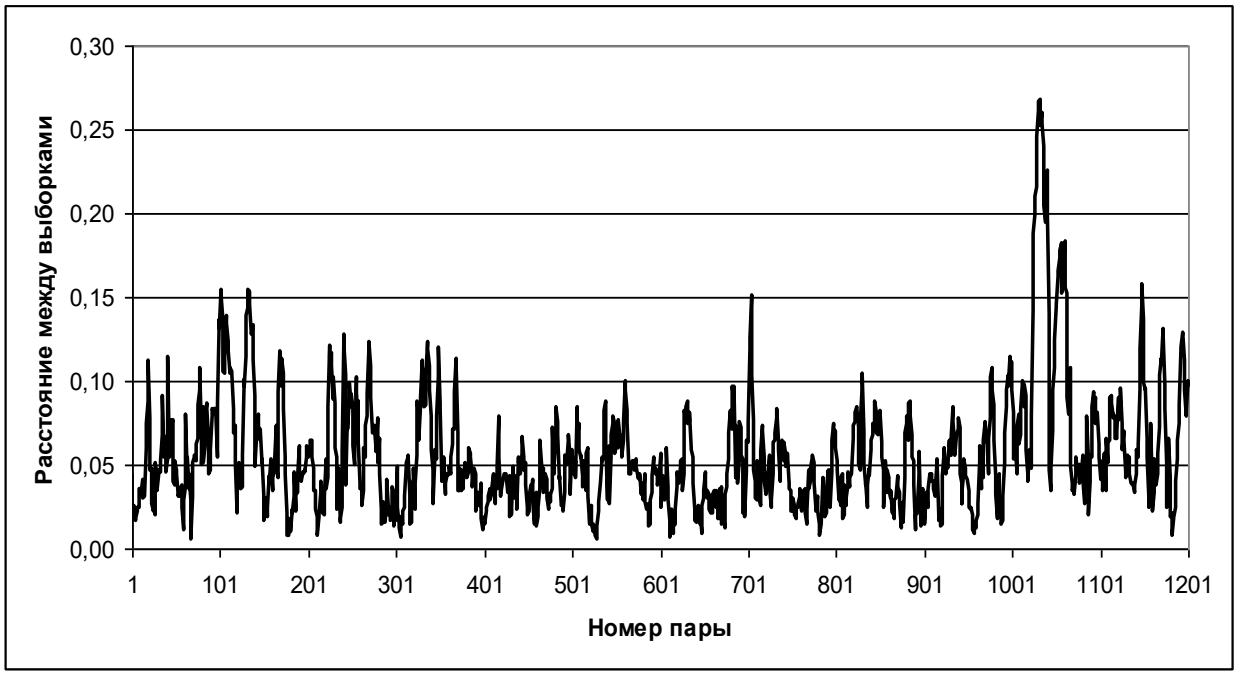

Рис. 1. Фрагмент временного ряда расстояний между независимыми ВФР из нормальной совокупности длиной 1000 данных

На рис. 2 приведена плотность распределения расстояний между выборками, временной ряд которых показан на рис. 1, по результатам 5000 вычисленных расстояний между парами.

Рис. 2. Эмпирическое распределение расстояний между независимыми выборочными распределениями из нормальной совокупности длиной 1000 данных

Возникает вопрос: на каком уровне на рис. 2 следует «отцепить хвост» распределения расстояний так, чтобы интеграл от отцепленной части был бы равен эмпирически наблюденной вероятности превышения этого уровня? Именно, ищется некоторый согласованный с экспериментом уровень значимости, равный эмпирической доле ошибочных отклонений верной гипотезы, которые (отклонения) принимаются тогда, когда расстояние между выборками больше некоторого критического уровня $\varepsilon^*(N)$.

Для ответа на этот вопрос заметим, что 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. [19, 33]), если СВ ξ (в данном примере это расстояние между выборками) имеет ФР $F(x)$ (для расстояний между выборками асимптотически это есть функция Колмогорова), то СВ $\eta = F(\xi)$ (здесь это уровень значимости) имеет равномерное распределение на $[0;1]$. Следовательно, согласованный с экспериментом уровень значимости α как квантиль равномерно распределенной СВ есть функция, линейно зависящая от расстояния ε между выборками. Поскольку же в норме C 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

$$1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) = \varepsilon\,. \qquad (2.1.4)$$

Таким образом, при анализе выборок определенной длины даже из стационарного временного ряда было бы неправильно задавать априори желаемый уровень значимости, так как для заданной длины N выборки лишь при одном значении $\varepsilon = \varepsilon^*(N)$, определяемом из уравнения (2.1.4), вероятность превышения значения $\varepsilon^*(N)$ равна значимости используемого для этой цели критерия.

Решение уравнения (2.1.4) единственно, поскольку правая часть как функция ε монотонно возрастает от нуля до единицы, а левая монотонно убывает от единицы до нуля. Численно определяемые решения этого уравнения приведены на рис. 3 и в табл. 1. Найденное решение будем называть согласованным уровнем стационарности (СУС) или согласованным отклонением между ВФР.

Введенное понятие СУС отвечает современным потребностям статистического анализа данных. Если в «докомпьютерную» эпоху статистик анализировал лишь одну выборку, для чего и разрабатывались критерии значимости, то в настоящее время часто приходится проводить анализ в скользящем окне, так что в априорных предположениях о стационарности ряда просто нет нужды, они могут быть проверены непосредственно.

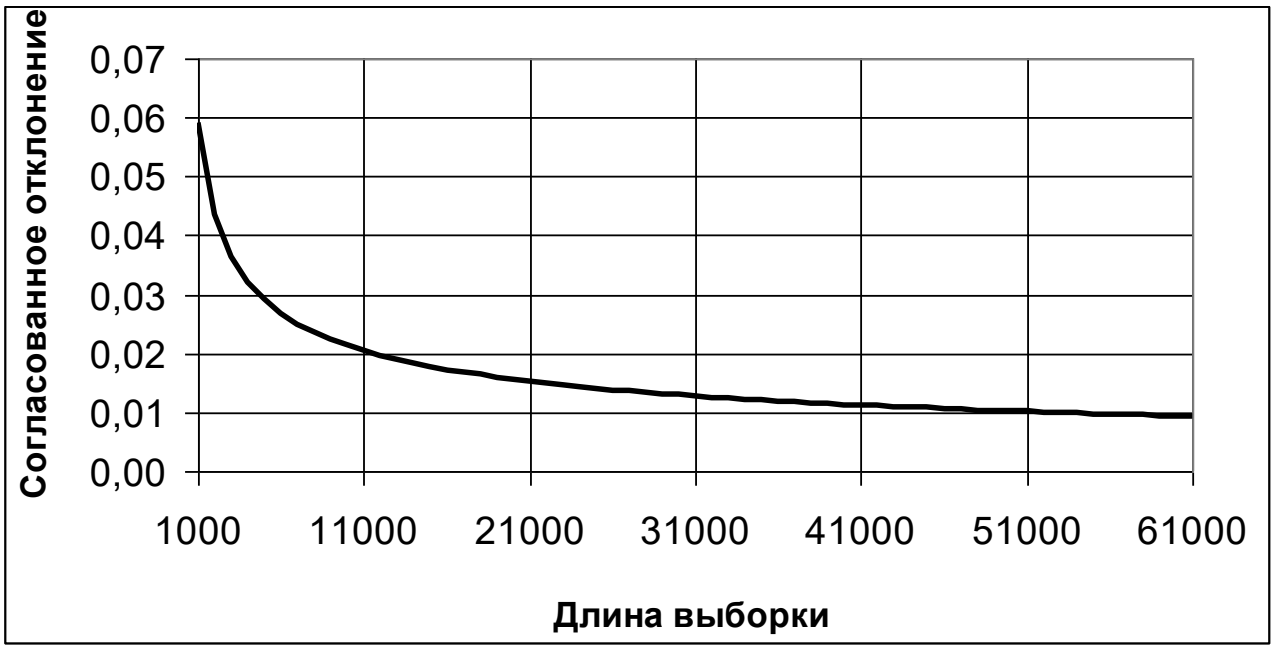

Рис. 3. Зависимость СУС от длины выборки

Табл. 1. Значения СУС для критерия Колмогорова-Смирнова

Длина выборки N	СУС $\varepsilon^*(N)$
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

Аналогично можно рассмотреть расстояния между выборочными распределениями и в других нормах. Пусть, в частности, $g_N^p(\rho)$ есть плотность функции распределения расстояний ρ между двумя независимыми выборками длины N в норме L_p. Супремум этого расстояния равен $2^{1/p}$. Определим для каждой нормы СУС $\rho^*(N)$ как решение уравнения

$$\int_0^{\rho^*(N)} g_N^p(\rho)d\rho = 1 - \frac{\rho^*(N)}{2^{1/p}}. \qquad (2.1.5)$$

Это расстояние, нормированное на свое максимальное отклонение, определяет уровень значимости, на котором в принципе можно различить между собой выборки из одного и того же распределения. Для краткости в (2.1.5) опущено указание на то, что рассматриваемый СУС $\rho^*(N)$ определен для специфицированной нормы. Отметим здесь, что для норм в терминах плотностей распределений расстояние между выборками существенно зависит от способа разбиения области значений СВ на классовые интервалы. Определение СУС для таких норм будет дано в следующем параграфе.

Описанный здесь метод определения характерного уровня статистического шума конечной выборки универсален, так как может быть использован как для стационарных, так и для нестационарных распределений. В любом случае сначала анализируется статистика 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:

$$\rho(N;t) = \left\| F_N(x,t) - F_N(x,t+N) \right\| \tag{2.1.6}$$

Далее строится функция распределения $G(\rho;N)$ расстояний (2.1.6), которая представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ. Определим теперь согласованный уровень стационарности $\rho^*(N)$ так, что соответствующее расстояние равно значимости критерия, т.е. является решением уравнения

$$G(\rho;N) = 1-\rho\,. \tag{2.1.7}$$

В стационарном случае уравнение (2.1.7) переходит в уравнение (2.1.4), поскольку тогда функция распределения $G(\rho;N)$ переходит в функцию Колмогорова.

Итак, стационарный СУС $\varepsilon^*(N)$ известен как решение уравнения (2.1.4). Пусть также вычислен и СУС $\rho^*(N)$ из (2.1.7) для изучаемого ряда. Индексом нестационарности временного ряда будем называть отношение

$$J(N) = \frac{\rho^*(N)}{\varepsilon^*(N)}. \tag{2.1.8}$$

Этот индекс показывает, во сколько раз доля расстояний, больших СУС, превосходит аналогичный показатель для стационарных рядов. Если $J(N) \le 1$, 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. Отметим, что на практике пороговое значение, равное единице, отделяющее стационарное поведение ряда от нестационарного, заключено в коридоре ширины $\varepsilon^*(N)$, поскольку по построению сам этот индекс определен с указанной точностью.

На рис. 4 показан пример построения СУС для стационарных временных рядов, имеющих различные плотности распределения: равномерное и распределение арксинуса,

порожденное известной логистической динамической системой $x_{n+1} = 4x_n(1-x_n)$. Как и должно быть, стационарный СУС в норме С не зависит от вида распределения.

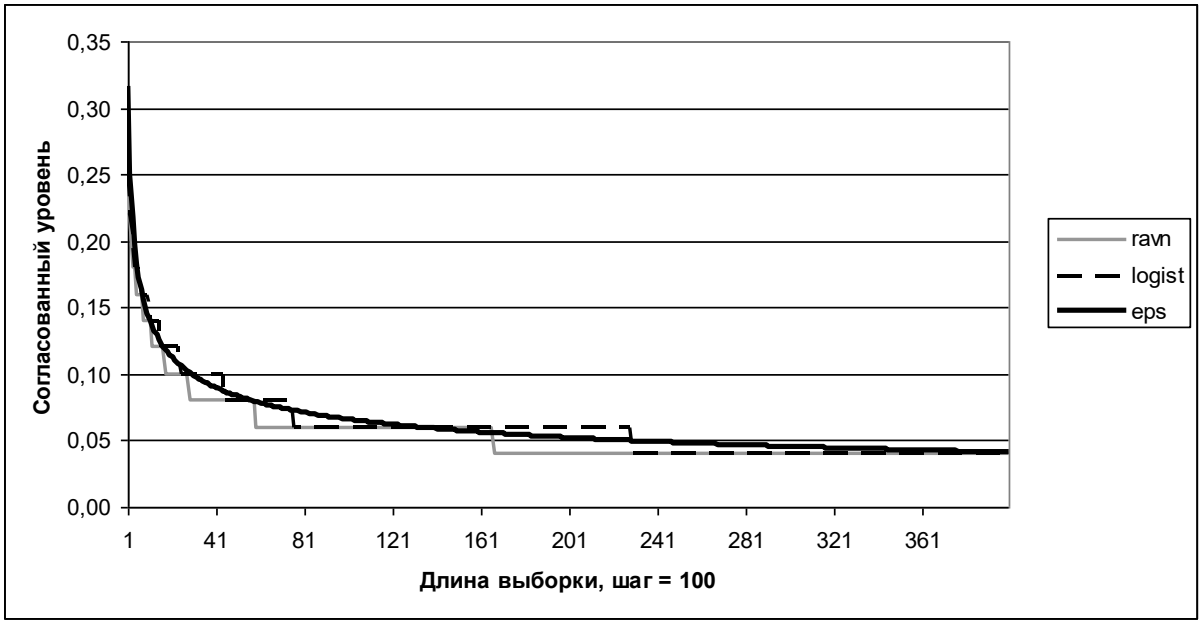

Рис. 4. Согласованный уровень стационарности в норме С для стационарных ВПФР

На рис. 5 и рис. 6 приведены примеры построения СУС для нестационарного временного ряда – биржевого индекса РТС по минутным ценам закрытия. Хорошо видны квазистационарная область на малых длинах выборок, а также рост нестационарности с увеличением промежутка наблюдения.

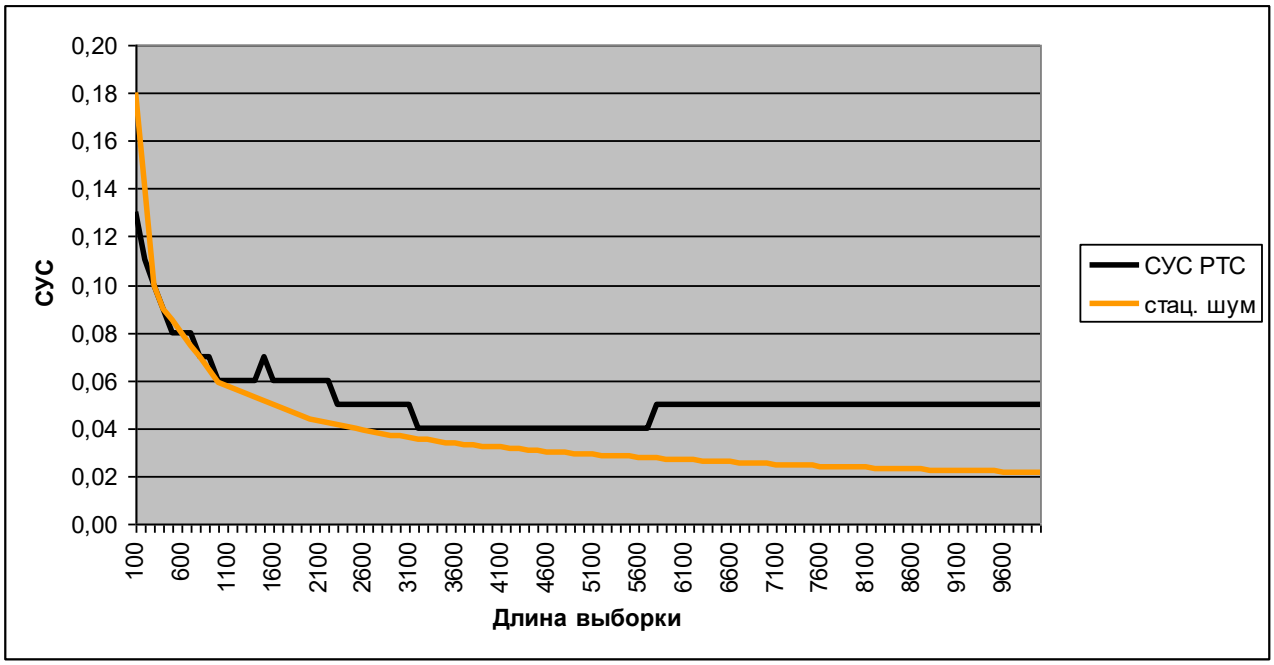

Рис. 5. СУС минутных цен закрытия индекса РТС
в сравнении со стационарным уровнем шума

Разделив согласно (2.1.8) СУС (черный график рис. 5) на уровень шума (оранжевый график рис. 5), получаем график индекса нестационарности рассматриваемого временного ряда как функцию длины выборки (рис. 6).

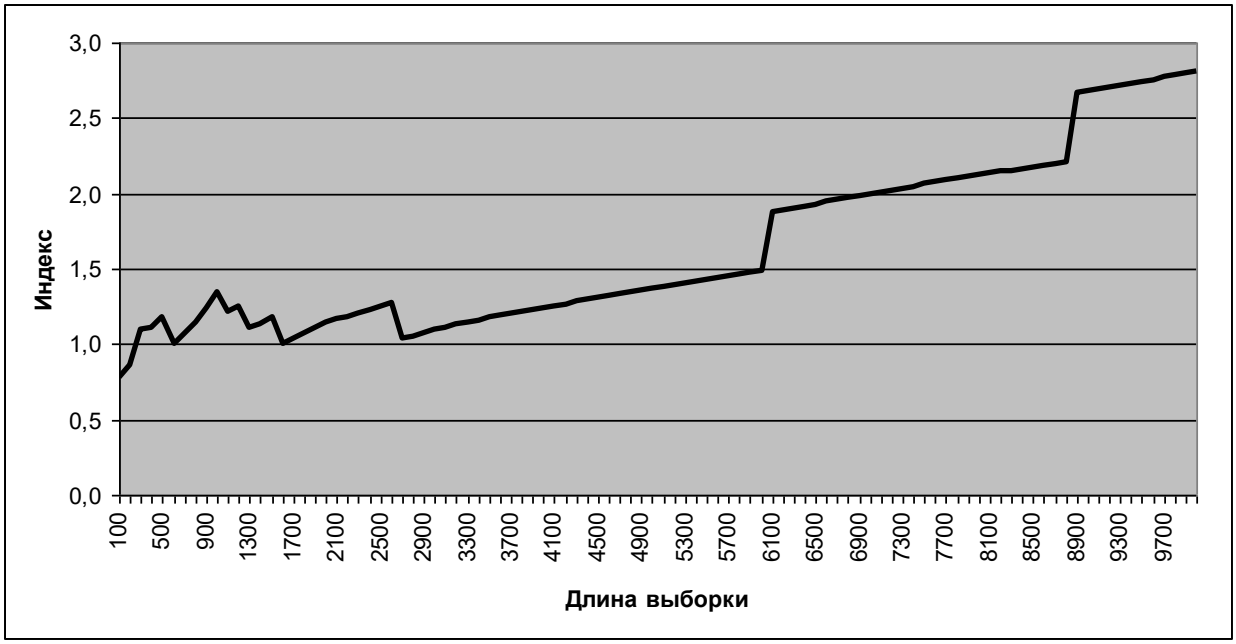

Рис. 6. Индекс нестационарности ряда минутных цен закрытия индекса РТС

Таким образом, построенный индикатор является эффективным и корректным инструментом для анализа нестационарности временных рядов.

2.2. Равномерное разбиение гистограммы и СУС в норме L1

При проведении статистических экспериментов, совершаемых в дискретные моменты времени, удобно исходить из представления о некоторой непрерывной случайной величине, значения которой могут быть наблюдены в любой момент времени. Это, например, температура воздуха, артериальное давление, биржевая цена акции, показания кардиограмм или энцефалограмм и т.п. Тогда представляется естественным оценить предполагаемую непрерывную функцию распределения изучаемой величины через построение выборочной плотности распределения, агрегируя наблюдаемые данные в некоторые классовые интервалы. Поэтому, даже будучи непрерывной, на практике СВ ξ принимает дискретное множество значений или принадлежит конечному множеству промежутков значений $\Delta x_1, \Delta x_2, ..., \Delta x_n$, внутри которых, однако, тоже может быть некоторое свое распределение, которым мы пока интересоваться не будем.

Количество n классовых интервалов, на которые разбит промежуток $[0;1]$, определяется двумя факторами: во-первых, практически достаточной точностью

измерения наблюдаемых величин и, во-вторых, точностью, с которой желательно иметь представление о распределении. Говоря об аппроксимации гистограммой, т.е. набором кусочно-постоянных функций, некоторой непрерывной ПФР, будем считать, что внутри каждого классового интервала распределение равномерно.

Эмпирическим или выборочным распределением СВ ξ, построенным по набору из N данных, называется совокупность частот $p_i(N) = k_i / N$, где k_i есть количество данных, попавших в i-ый классовый интервал. Соответствующая гистограмма образует выборочную плотность функции распределения (ВПФР).

Если СВ принципиально дискретна, как в экспериментах с бросанием монеты, кубика и т.п. объектов, то выборочная функция распределения (ВФР) определяется как ступенчатая кусочно-постоянная функция со скачками в точках $x = x_k$, являющихся наблюденными значениями СВ. Нам будет удобно определять ВФР как вероятность того, что значение СВ не превосходит заданной величины, т.е. $F_\xi(x) = P(\xi \le x)$. Тогда ВФР, построенная по N наблюдениям, определяется формулой

$$F(x; N) = \frac{1}{N} \sum_{k=1}^{N} [x_k \le x], \qquad (2.2.1)$$

где

$$[z \le x] = \begin{cases} 1, & z \le x \\ 0, & z > x \end{cases}.$$

Если ВПФР получается в результате кластеризации наблюдаемых значений по n классовым интервалам $\Delta_i = [a_{i-1}; a_i)$, $a_0 = 0$, $a_n = 1$, внутри которых распределение имеет вид гистограммы, т.е. равномерно, то используются следующие определения ВПФР $f(x; N)$ и ВФР $F(x; N)$:

$$\begin{aligned} f(x; N) &= p_i(N), \quad x \in \Delta_i; \\ F(x; N) &= F(a_{i-1}; N) + p_i(N) \frac{x - a_{i-1}}{a_i - a_{i-1}}, \quad x \in \Delta_i. \end{aligned} \qquad (2.2.2)$$

При этом последний, т.е. n-ый, классовый интервал является отрезком $\Delta_n = [a_{n-1}; 1]$.

Далее будем рассматривать равномерное разбиение гистограммы, когда ширина классового интервала равна $1/n$.

Рассмотрим последовательность из N значений случайной величины ξ, которые попали в классовые интервалы $\Delta_i = [a_{i-1}; a_i)$, $a_i = (i-1)/n$. Элементы этой последовательности обозначим x_j, $j = 1, 2, ..., N$. Возникает вопрос: какова должна быть

мелкость разбиения, чтобы с нужной точностью получить оценку ПФР, составив также и адекватное представление о виде распределения по виду ВПФР?

Очевидно, наилучшая точность в оценке вероятностей достигается при разбиении на один классовый интервал: тогда внутри него ВПФР совпадает с ВФР и равна единице. Однако практически этот результат бесполезен, ибо не дает представления о виде распределения. Если же взять число промежутков разбиения слишком большим, например, $n > N$, то существуют такие промежутки, в которых оказывается невозможным трактовать эмпирические частоты как оценки соответствующих вероятностей, поскольку тогда не может быть реализовано условие теоремы Гливенко (см. [19]) о сходимости эмпирической вероятности к теоретической при $N \to \infty$. Следовательно, для каждой длины выборки N существует оптимальное равномерное разбиение гистограммы, при котором и детализация распределения достаточна, и точность оценки ПФР оказывается приемлемой. Чтобы формализовать условие оптимальности, обратимся к классическому решению статистической задачи об оценивании доверительного интервала для статистики среднего значения на основе наблюдаемых данных.

Как известно [19, 33, 35], когда генеральная дисперсия σ^2 теоретического распределения не известна, а оценивается только по выборочной дисперсии $s_2(N)$, для получения доверительного интервала генерального среднего значения μ следует рассматривать *t*-статистику Стьюдента, которая применительно к этой задаче имеет вид:

$$t = \sqrt{N-1}\,\frac{\left|m_1(N)-\mu\right|}{\sqrt{s_2(N)}}. \qquad (2.2.3)$$

Здесь

$$m_1(N) = \frac{1}{N}\sum_{k=1}^{N} x_k \equiv \langle x \rangle_N, \quad s_2(N) = \left\langle \left(x - m_1(N)\right)^2 \right\rangle_N. \qquad (2.2.4)$$

Смысл статистики (2.2.3) в том, что если задать некоторый уровень значимости ε, то с вероятностью ε выражение $\left|m_1(N)-\mu\right|$ не превосходит величины

$$t_{1-\varepsilon/2}(N-1)\sqrt{s_2(N)} \,/\, \sqrt{N-1}\,, \qquad (2.2.5)$$

где $t_\alpha(N-1)$ есть α-квантиль распределения Стьюдента с $N-1$ степенью свободы. При больших N можно считать $N-1 \approx N$ и число степеней свободы в квантиле распределения Стьюдента взять для простоты бесконечным (тогда это распределение совпадает с нормальным).

Согласно центральной предельной теореме (ЦПТ, [19, 35]) отклонение выборочного среднего значения $m_1(N)$, определяемого в (2.2.4) по выборке длины N, от генерального

среднего μ распределено асимптотически нормально с нулевым средним и стремящейся к нулю дисперсией σ^2 / N, где σ^2 есть дисперсия исходной СВ по гипотетической генеральной совокупности. Рассмотрим в качестве такой СВ саму эмпирическую частоту $p_i(N)$ попадания в i-ый классовый интервал. Формально эмпирическая частота есть среднее выборочное значение случайной величины, называемой индикатором I_{ik} принадлежности результата k-го наблюдения в i-ому классовому интервалу. Индикатор I_{ik} определяется формулой

$$I_{ik} = \begin{cases} 1, & x_k \in \Delta_i; \\ 0, & x_k \notin \Delta_i. \end{cases} \qquad (2.2.6)$$

Как следует из формулы (2.2.6), эмпирическую частоту можно записать в виде

$$p_i(N) = \langle I \rangle = \frac{1}{N} \sum_{k=1}^{N} I_{ik} .$$

Выборочная дисперсия эмпирической частоты $p_i(N)$ равна

$$s_2(N;i) = \left\langle \left(I - \langle I \rangle\right)^2 \right\rangle = \left\langle I^2 \right\rangle - \langle I \rangle^2 = \langle I \rangle - \langle I \rangle^2 = p_i(N) \cdot \left(1 - p_i(N)\right). \qquad (2.2.7)$$

Таким образом, из (2.2.5) следует, что оценка генеральной частоты f_i^* заключена в ε-доверительном интервале

$$\left| p_i(N) - f_i^* \right| \le t_{1-\varepsilon/2}(N-1) \sqrt{s_2(N;i)} / \sqrt{N-1} . \qquad (2.2.8)$$

Зададим уровень значимости ε равным уровню неопределенности в позиционировании доверительного интервала $\left| p_i(N) - f_i^* \right|$. Поскольку исходным требованием является близость оценки к генеральной частоте, то естественно потребовать выполнения условия

$$\sum_{j=1}^{n} \left| p_i(N) - f_i^* \right| \le \varepsilon . \qquad (2.2.9)$$

Очевидно, условию (2.2.9) можно удовлетворить, потребовав выполнения более жесткого условия – близости каждой из частот в отдельности:

$$\left| p_i(N) - f_i^* \right| \le \varepsilon f_i^* . \qquad (2.2.10)$$

Таким образом, если для каждой эмпирической частоты выполнить условие $t_{1-\varepsilon/2} \dfrac{\sqrt{s_2(N;i)}}{\sqrt{N}} \le \varepsilon f_i^*$, то одновременно с выполнением условия (2.2.9) будет достигнут и требуемый уровень значимости для статистики Стьюдента в (2.2.8). Однако если

некоторые вероятности в результате выбранного разбиения на классовые интервалы сами оказались малы, много меньше ε, то нет необходимости требовать, чтобы и они были оценены с той же точностью. Поэтому уместно для каждой вероятности выбрать свою точность аппроксимации ε_i и считать, что требуемый в целом уровень значимости определяется средневзвешенной по разбиению точностью, так что

$$t_{1-\varepsilon/2} = \frac{1}{\Sigma_N(n)} \sum_{i=1}^{n} \sqrt{s_2(N;i)}\, t_{1-\varepsilon_i/2}, \qquad (2.2.11)$$

где сумма, определяющая влияние мелкости разбиения гистограммы на точность оценки эмпирических вероятностей, равна

$$\Sigma_N(n) = \sum_{i=1}^{n} \sqrt{s_2(N;i)} = \sum_{i=1}^{n} \sqrt{p_i(N) \cdot \left(1 - p_i(N)\right)}. \qquad (2.2.12)$$

Таким образом, величина $\sqrt{s_2(N;i)} / \Sigma_N(n)$ представляет весовой коэффициент для квантилей распределения Стьюдента в данной задаче. Сумма (2.2.12) выражает качество приближения плотности гистограммой, поскольку, чем меньше сумма, тем выше точность оценки ВПФР, т.е. тем меньше число ε. С увеличением числа интервалов сумма (2.2.12) не убывает, что означает снижение точности оценки ВПФР. В стационарном случае эта сумма представляет собой некий эффективный функционал учета особенностей графика функции плотности при аппроксимации плотности генерального распределения кусочно-постоянной гистограммой.

Теперь из (2.2.11) получаем, что

$$\sum_{i=1}^{n} t_{1-\varepsilon_j/2} \frac{\sqrt{s_2(N;i)}}{\sqrt{N}} = t_{1-\varepsilon/2} \frac{\Sigma_N(n)}{\sqrt{N}} \le \varepsilon \sum_{i=1}^{n} f_i^* = \varepsilon,$$

откуда на уровне значимости ε следует оценка

$$\frac{t_{1-\varepsilon/2}}{\varepsilon} \le \frac{\sqrt{N}}{\Sigma_N(n)}. \qquad (2.2.13)$$

При заданной точности ε и способе разбиения гистограммы формула (2.2.13) в случае знака равенства дает оценку на минимальную длину выборки, при которой эта точность достигается в среднем. Поскольку функция $t_{1-\varepsilon}$ табулирована (см., напр., [35]), то функция $t_{1-\varepsilon} / \varepsilon$ известна. Она монотонно убывает с ростом ε, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей по заданному разбиению гистограммы. Некоторые значения функции $t_{1-\varepsilon} / \varepsilon$ приведены в табл. 2.

Табл. 2. Зависимость функции $t_{1-\varepsilon}/\varepsilon$ от ε

ε	$t_{1-\varepsilon}$	$t_{1-\varepsilon}/\varepsilon$	ε	$t_{1-\varepsilon}$	$t_{1-\varepsilon}/\varepsilon$
0,0005	3,291	6582,0	**0,025**	1,960	78,4
0,0025	2,807	1122,8	**0,05**	1,645	32,9
0,0050	2,576	515,2	**0,10**	1,282	12,8
0,0075	2,432	324,3	**0,15**	1,036	6,9
0,010	2,326	232,6	**0,20**	0,842	4,2
0,015	2,170	144,7	**0,30**	0,524	1,7
0,020	2,054	102,7	**0,40**	0,253	0,6

Обозначим для краткости

$$\varphi(\varepsilon)=\frac{t_{1-\varepsilon}}{\varepsilon},\quad \psi=\varphi^{-1},\quad z\equiv z(N,n)=\frac{\sqrt{N}}{\Sigma_N(n)}. \tag{2.2.14}$$

Тогда точность оценки ВПФР определяется формулой

$$\varepsilon=2\psi(2z). \tag{2.2.15}$$

Отметим, что при $\varepsilon>0{,}01$ имеет место аппроксимация [33] квантиля нормального распределения, с которым при больших N совпадает квантиль распределения Стьюдента:

$$t_{1-\varepsilon/2}=\sqrt{-\frac{\pi}{2}\ln\left(1-(1-\varepsilon)^2\right)}, \tag{2.2.16}$$

что позволяет проводить вычисления без использования интерполяции. Вид функции $t_{1-\varepsilon}/\varepsilon$ показан на рис. 7.

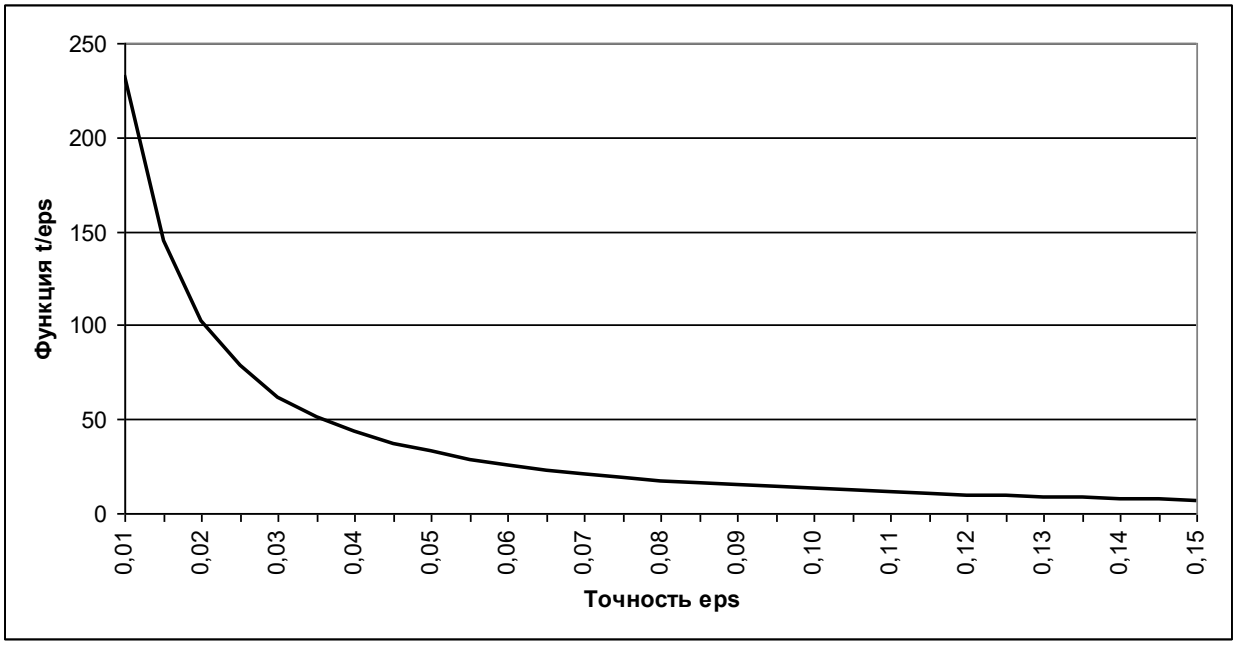

Рис. 7. Вид функции $t_{1-\varepsilon}/\varepsilon$

Заметим теперь, что если вероятности оценены с точностью ε, то, как легко показать, с той же точностью оценивается и среднее значение СВ. Поэтому нахождение средних значений по выборке фактических величин с большей точностью будет излишним в смысле превышения точности несмотря на то, что собственно измерения наблюдений могут быть сделаны весьма точно, но это точность измерительного инструмента, а не статистики. С другой стороны, если выбрано разбиение на n классовых интервалов, то среднее значение по выборке отличается от среднего значения по гистограмме на величину порядка $1/n$.

Исходя из этих качественных соображений, можно сделать вывод, что на практике не имеет смысла выбирать мелкость разбиения гистограммы, превышающую точность оценки вероятностей по этой самой гистограмме. Так как число интервалов желательно выбрать наибольшим из возможных, то оптимальным равномерным разбиением гистограммы будем считать ближайшее натуральное число n (классовых интервалов) к решению уравнения, согласующего уровень значимости (2.2.15) и мелкость разбиения:

$$\frac{1}{n} = 2\psi\left(\frac{2\sqrt{N}}{\Sigma_N(n)}\right). \qquad (2.2.17)$$

Уравнение (2.2.17) может не иметь решений в натуральных числах относительно n при заданной длине выборки N, но оно всегда существует и единственно относительно точности $\varepsilon = 1/n$, так как с уменьшением ε число разбиений $n = 1 + [1/\varepsilon]$ не убывает, как и аргумент функции ψ в правой части (2.2.17), причем сама функция ψ при этом монотонно возрастает, а области значений обеих этих непрерывных функций ε и $\psi(\varepsilon)$ совпадают.

Итак, построено разбиение гистограммы, согласованное с видом плотности функции распределения, определяемой этой гистограммой. При этом нормированная длина классового интервала совпадает с точностью оценки вероятности в гистограммной норме L1. После того, как определен такой объект, как гистограмма, можно приступать к следующей задаче: сравнению двух гистограмм на предмет выяснения того, из одной и той же генеральной совокупности были взяты выборки для построения оценок ПФР.

Пусть $f_i(N)$ есть значение гистограммы в i-ом классовом интервале, причем число таких интервалов n выбрано в соответствии с вышеописанным согласованным равномерным разбиением. Расстояние между двумя гистограммами будем вычислять в норме L1:

$$\rho(N) = \left\| f^{(1)}(x;N) - f^{(2)}(x;N) \right\| = \sum_{i=1}^{n} \left| f_i^{(1)}(N) - f_i^{(2)}(N) \right| \qquad (2.2.18)$$

В соответствии с идеологией п. 2.1 строим функцию распределения $G(\rho;N)$ расстояний (2.2.18), которая представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ.

Заметим теперь, что если каждая гистограмма имеет уровень статистической неопределенности, равный $\varepsilon=1/n$, то их разность в смысле (2.2.18) имеет неопределенность порядка 2ε. Следовательно, на уровне значимости ε для согласованного расстояния между гистограммами должно выполняться условие $G(2\varepsilon;N)=1-\varepsilon$. Таким образом, в стационарном случае СУС совпадает с точностью гистограммы, т.е. равен $\varepsilon=1/n$.

В отличие от СУС в норме С, теперь даже для стационарных ВПФР оптимальное разбиение гистограммы для оценки плотности зависит от вида распределения. Чем шире распределение, т.е. чем больше его дисперсия, тем меньшим числом интервалов можно обойтись при аппроксимации его плотности. И наоборот, чем уже пик унимодального распределения, тем больше интервалов требуется для адекватного воспроизведения этой плотности. Эта ситуация иллюстрируется на рис. 8 примерами разбиений в зависимости от длины выборки для двух принципиально разных стационарных распределений.

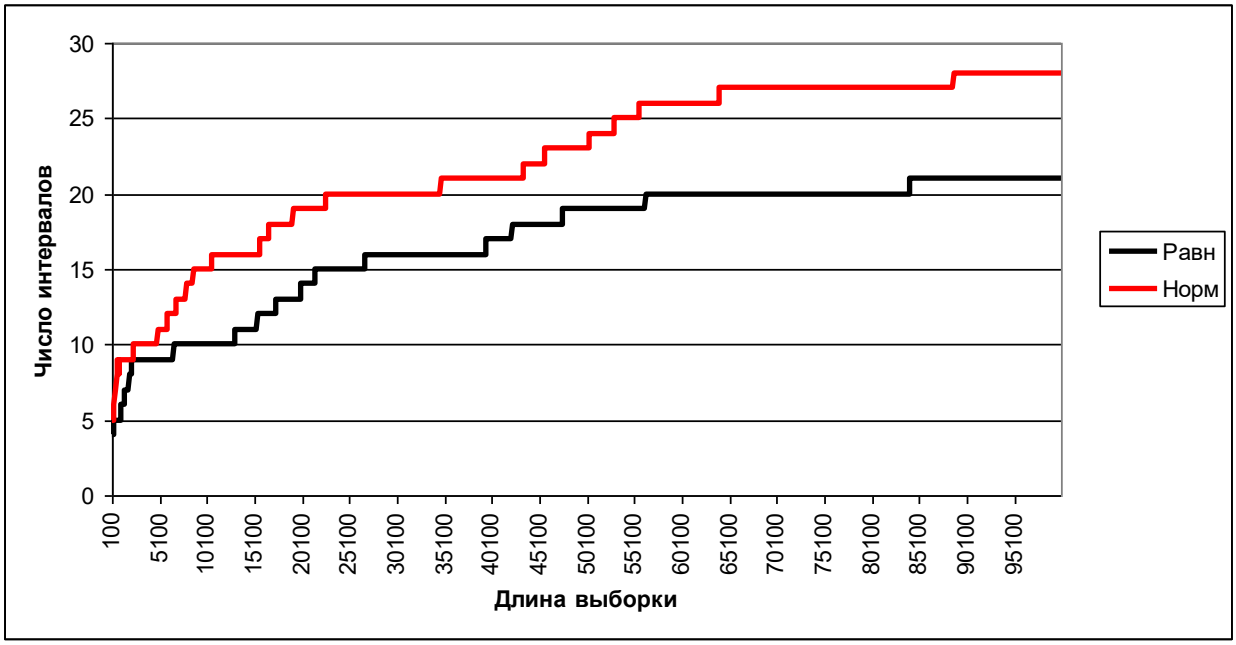

Рис. 8. Оптимальное число интервалов в зависимости от длины выборки для стационарных распределений

В нестационарном случае СУС $\rho^*(N)$ в норме L1 определяется как решение обобщения стационарного уравнения $G(2\varepsilon; N) = 1 - \varepsilon$ относительно расстояния $\rho = 2\varepsilon$:

$$G(\rho; N) = 1 - \rho / 2 . \qquad (2.2.19)$$

Здесь $G(\rho; N)$ есть эмпирическая функция расстояний (2.2.18) между независимыми выборками длины N.

На рис. 9 приведено оптимальное разбиение гистограммы для оценивания ВПФР нестационарного временного ряда на примере ряда цен сделок на нефтяной фьючерс.

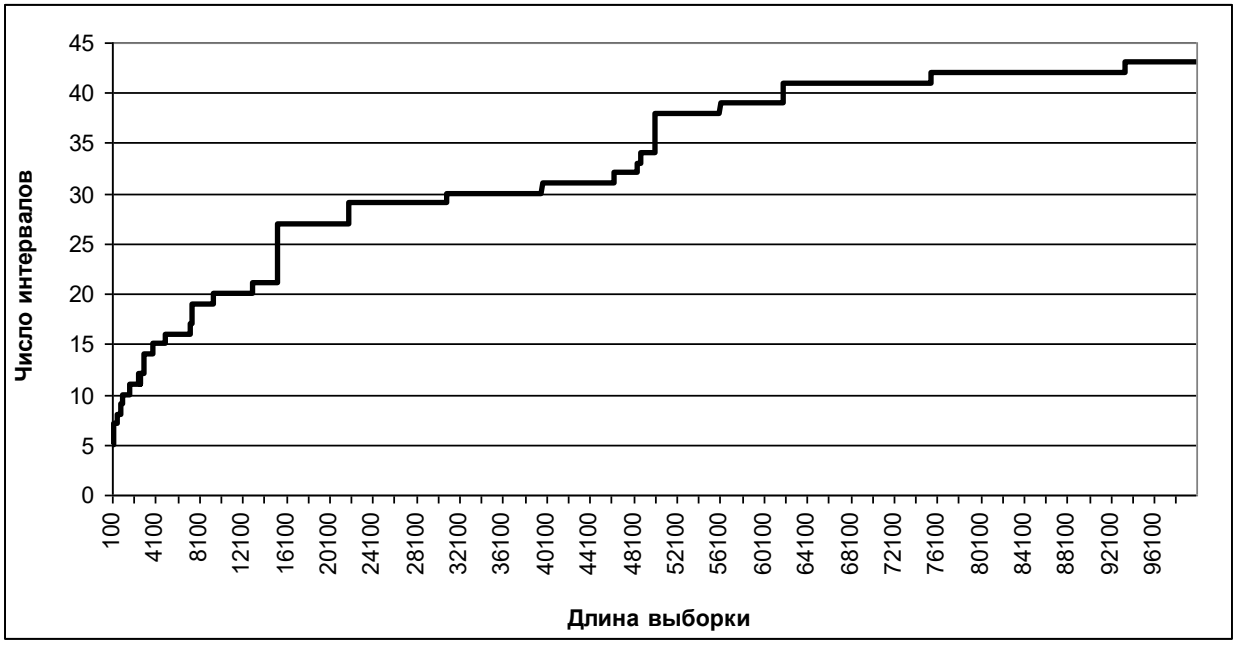

Рис. 9. Разбиение гистограммы для ряда цен сделок на нефтяной фьючерс

В соответствии с этим разбиением далее находятся расстояния между независимыми выборками из рассматриваемого временного ряда. На рис. 10 показан пример эмпирической плотности функции распределения расстояний между выборками в 10 тыс. точек для ряда приростов цен сделок на нефтяной фьючерс по данным за год. Интеграл от этой функции по ρ при фиксированном N дает функцию распределения $G(\rho; N)$.

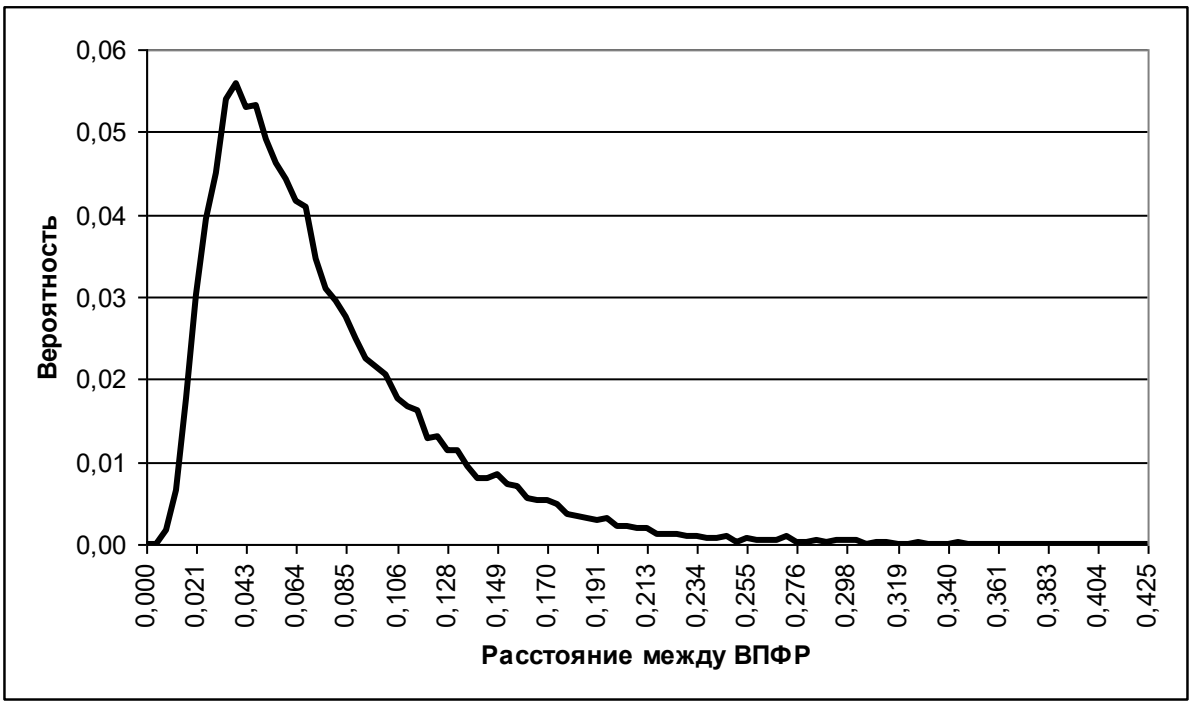

Рис. 10. Распределение расстояний между встык-выборками в 10 тыс. точек ряда цен сделок на нефтяной фьючерс (CL)

Решая для этого примера уравнение (2.2.19), получаем зависимость СУС от длины выборки. Эта зависимость приведена на рис. 11 вместе с точностью расстояния между гистограммами, т.е. вместе с графиком $2/n(N)$, где график $n(N)$ показан на рис. 9.

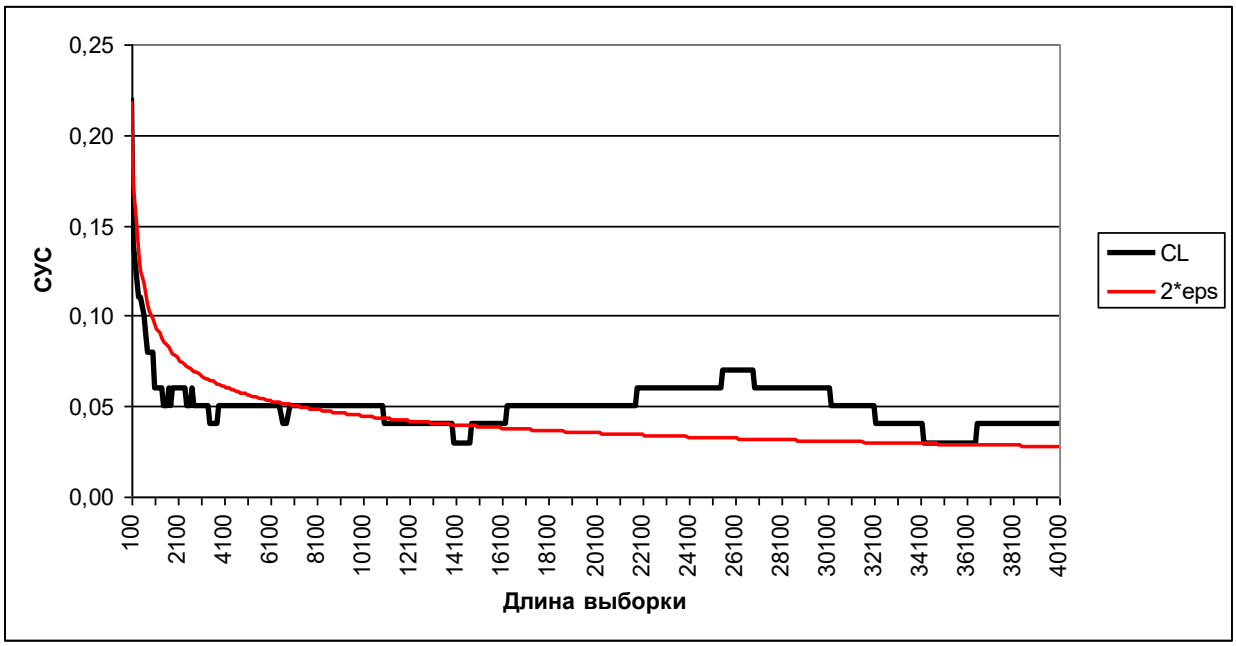

Рис. 11. СУС ряда цен на фьючерс CL в сравнении с точностью разности гистограмм

Зависимость СУС от длины выборки несет информацию относительно нестационарных свойств временного ряда. Для многих нестационарных рядов характерно наличие статистически значимых локальных минимумов СУС в зависимости от длины выборок, тогда как для стационарного ряда наблюдается монотонно убывающая зависимость. Аргументы этих локальных минимумов можно трактовать как типовые для данного ряда промежутки времени, на которых происходит смена режима работы наблюдаемой системы. Аналогично, аргументы локальных максимумов отвечают наибольшей разладке между распределениями.

Далее как и в (2.1.8), вводим индекс нестационарности в виде отношения доли выборок, превосходящих СУС, к доле выборок, превосходящих уровень статистической неопределенности разности гистограмм:

$$J(N) = \frac{\rho^*(N)}{2\varepsilon^*(N)} = n(N) \cdot \rho^*(N)/2\,. \qquad (2.2.20)$$

Пример для ряда CL, получающийся по формуле (2.2.20) из рис. 11, показан на рис. 12.

Рис. 12. Индекс нестационарности ряда CL

Из рис. 12 четко виден максимум индекса нестационарности, чему отвечают длины выборок, на которых наиболее заметна разладка между распределениями. Также хорошо идентифицируются устойчивые длины выборок, на которых ряд имеет стационарное поведение.

2.3. Уравнение Фоккера - Планка для нестационарной ВПФР

После того, как построен метод корректного определения уровня нестационарности временного ряда по расстояниям между его выборочными распределениями, можно поставить вопрос об эволюции этих распределений, рассматривая сначала встык-выборки, а затем и выборки, смещенные на произвольный промежуток. Целью эволюционного анализа выборочных распределений является разработка метода генерации ансамбля траекторий нестационарного случайного процесса по наблюдениям за одной, единственно доступной из эксперимента, траекторией. В такой постановке это есть обратная задача статистической механики: по кинетическому уравнению восстановить одну из возможных фазовых траекторий исходной динамической системы. Прямая задача – из уравнений динамики для систем многих частиц получить уравнение, описывающее эволюцию плотности их начального распределения в фазовом пространстве. Если решить такое уравнение, то в каждый момент времени будет известна функция распределения частиц по координатам и скоростям (или в более общем случае по их независимым кинематическим переменным). Тогда в каждый момент времени можно сгенерировать выборку из этого распределения, которая и будет представлять одну из возможных реализаций исходной механической системы.

В работах [10, 11, 13] аналогичный подход был предложен для анализа нестационарных временных рядов. Именно, если окажется возможным описать изменение выборочной функции распределения с помощью кинетического уравнения, то откроется перспектива генерации траекторий нестационарных случайных процессов, что имеет большую практическую важность и фактически представляет собой нестационарный вариант метода Монте-Карло. Нетривиальность задачи состоит в том, что решение обратной задачи статистической механики для нестационарных процессов оказывается не однопараметрическим (с зависимостью от времени), а двупараметрическим, когда добавляется еще и зависимость от длины выборки. Именно эта зависимость и была проанализирована в предыдущих разделах диссертации 2.1 и 2.2.

Чтобы создать кинетическую модель нестационарного временного ряда, надо, следуя работам [11, 13], вывести уравнение эволюции для его функции распределения. Однако в отличие от задач статистической механики, где такое уравнение получается на основе уравнений движения механической системы, здесь такая система отсутствует. Можно говорить только об эволюции выборочных функций распределения. В работе [48] в качестве простейшего эволюционного уравнения для ВПФР было предложено уравнение Лиувилля. Однако дальнейший анализ показал, что оно адекватно описывает только процессы с независимыми приращениями, что на практике может и не выполняться. В

результате в [11] было получено уравнение типа Фоккера-Планка (т.е. диффузии со сносом) для эмпирических распределений и показано, как коэффициенты этого уравнения могут быть определены по наблюденным историческим данным. Но вычислительной схемы для его решения применительно к выборочным распределениям, а также алгоритма генерации по этому решению ансамбля соответствующих траекторий предложено не было. В настоящей работе восполняется этот пробел.

Как известно [6], уравнение Фоккера-Планка представляет собой обрыв разложения обратного преобразования Фурье для условной вероятности перехода вдоль фазовой траектории на втором порядке, что дает уравнение относительно плотности (уже не условной) вероятности:

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(u(x,t) f(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\lambda^2(x,t) f(x,t)\right). \qquad (2.3.1)$$

По определению, процесс $x(t)$ называется диффузионным, если коэффициент диффузии $\lambda^2(x,t)$ неотрицательный. Если $\lambda^2(x,t)$ строго больше нуля, непрерывный марковский процесс называется регулярным.

Рассмотрим теперь, в какой мере эмпирическую ВПФР, построенную по конкретному временному ряду, можно описывать уравнением (2.3.1). Считаем здесь длину выборки N фиксированной и для краткости не указываем зависимость от нее ВПФР. Суть проблемы в том, что выбранная модель эволюции ВПФР должна в виде следствий корректно описывать эволюцию эмпирических моментов. Например, легко проверить, что уравнение Лиувилля не описывает в общем случае эволюцию выборочной дисперсии.

Действительно, если исходить из уравнения $\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(u(x,t) f(x,t)\right)$, то эволюция первого момента $m(t) = \langle x \rangle_t = \int x f(x,t) dx$ определяется уравнением

$$\frac{dm(t)}{dt} = \int x \frac{\partial f(x,t)}{\partial t} dx = -\int x \frac{\partial}{\partial x} u(x,t) f(x,t) dx .$$

которое после интегрирования по частям с учетом того, что в граничных ячейках плотность распределения равна нулю, приводится к виду

$$\frac{dm(t)}{dt} = \int u(x,t) f(x,t) dx = \langle u \rangle_t = U(t) , \qquad (2.3.2)$$

где $U(t)$ есть среднее значение эмпирической скорости $u(x,t)$ по распределению, которое задано в момент времени *t*. Уравнение (2.3.2) имеет ясный статистический смысл, если записать его применительно к элементам временного ряда: изменение первого

выборочного момента, вычисляемое непосредственно по элементам выборки временного ряда, есть

$$\frac{dm(t)}{dt} = m(t+1) - m(t) = \frac{1}{N}\sum_{k=1}^{N} x(t-N+k+1) - \frac{1}{N}\sum_{k=1}^{N} x(t-N+k) = \frac{1}{N}\sum_{k=1}^{N} \dot{x}(t-N+k),$$

что является ничем иным, как средним значением $U(t)$ выборочной скорости, определяемой по приращениям значений исходного временного ряда. В вышенаписанной цепочке равенств точкой обозначена разностная производная значений временного ряда: $\dot{x}(t) = x(t+1) - x(t)$. Выясним теперь, как эволюционируют в силу уравнения Лиувилля высшие моменты ВПФР, определяемые как $m_r(t) = \int x^r f(x,t)dx$. Для них тем же путем, что и для первого момента, получаем

$$\frac{dm_r}{dt} = \int x^r \frac{\partial f}{\partial t} dx = -\int x^r \frac{\partial uf}{\partial x} dx = r\int x^{r-1} uf dx = r\,\mathrm{cov}\left(x^{r-1}, u\right) + r m_{r-1} U . \qquad (2.3.3)$$

Оказывается, что при $r \geq 2$ уравнение (2.3.3) не является непрерывным аналогом соответствующей разностной производной выборочных моментов. Выборочные моменты, вычисленные по элементам ряда, будем снабжать верхним индексом e, чтобы отличать их от моментов ВПФР. Тогда имеем

$$m_r^e(t) = \frac{1}{N}\sum_{k=1}^{N} x_{t-T+k}^r , \quad \dot{m}_r^e = m_r^e(t+1) - m_r^e(t) \neq r\,\mathrm{cov}\left(x, \dot{x}\right) + r m_{r-1}^e(t) U(t) . \qquad (2.3.4)$$

Например, при $r = 2$ эволюция второго выборочного момента непосредственно по элементам ряда дается формулой

$$\dot{m}_2^e(t) = \frac{1}{N}\sum_{k=1}^{N}\left(x_{t-N+k+1}^2 - x_{t-N+k}^2\right) = \frac{x_{t+1}^2 - x_{t-N+1}^2}{N} = U(t)\left(x_{t+1} + x_{t-N+1}\right).$$

В то же время эволюция второго момента ВПФР согласно (2.3.3) имеет вид

$$\dot{m}_2 = 2\,\mathrm{cov}(x, \dot{x}) + 2mU = 2\left\langle x\dot{x}\right\rangle =$$
$$= \frac{2}{N}\sum_{k=1}^{N} x_{t-N+k}\dot{x}_{t-N+k} = \frac{2}{N}\sum_{k=1}^{N} x_{t-N+k}\left(x_{t-N+k+1} - x_{t-N+k}\right),$$

которое в общем случае не равно выражению для $\dot{m}_2^e$, написанному строкой выше.

Точно так же и эволюция центральных выборочных моментов ряда, определяемых по ВПФР как $\mu_r(t) = \int \left(x - m(t)\right)^r f(x,t)dx$, $r \geq 2$, отличается от эволюции моментов, вычисленных по элементам ряда:

$$\mu_r^e(t) = \frac{1}{N}\sum_{k=t-N+1}^{t}\left(x(k) - m(t)\right)^r , \quad m(t) = \frac{1}{N}\sum_{k=t-N+1}^{t} x(k) ,$$

но

$$\dot{\mu}_r = -r\mu_{r-1}\frac{dm}{dt} + \int (x-m)^r \frac{\partial f}{\partial t}dx = -r\mu_{r-1}\frac{dm}{dt} - \int (x-m)^r \frac{\partial uf}{\partial x}dx =$$

$$= -r\mu_{r-1}\frac{dm}{dt} + r\mu_{k-1}U + r\int (x-m)^{r-1}(u-U)\,fdx = r\left\langle (\Delta x)^{r-1}\Delta u \right\rangle \neq \dot{\mu}_r^e(t).$$

Следовательно, уравнение Лиувилля применимо как модель эволюции к весьма узкому классу временных рядов. В работе [11] было показано, что ели мы хотим, чтобы уравнения эволюции выборочных моментов, равные эмпирическим, получались бы и из кинетического уравнения для ВПФР, то порядок производной по x в кинетическом уравнении должен совпадать с порядком момента. В настоящей работе будет рассмотрен первый содержательно нетривиальный случай, когда порядок «правильно» эволюционирующего момента равен двум. Ему отвечает уравнение относительно ВПФР типа Фоккера-Планка. Ниже выводятся соотношения на параметры этого уравнения (коэффициенты сноса и диффузии), которые могут быть оценены по элементам временного ряда.

Итак, пусть $f(x,t)$ – одномерная ВПФР, построенная по выборке некоторого объема, который далее в этом параграфе считается постоянным и потому не указывается в аргументах ВПФР. Обозначим также $F(x,v,t)$ двумерную ВПФР значений ряда и его приращений, и введем еще трехмерную ВПФР $\Phi(x,v,w,t)$ значений ряда, его первых и, соответственно, вторых разностей. Тогда в приближении Фоккера-Планка функция $F(x,v,t)$ удовлетворяет двумерному уравнению диффузионного типа. Если матрица диффузионных коэффициентов не зависит от x,v, а зависит только от времени t, то уравнение Фоккера-Планка для $F(x,v,t)$ имеет вид

$$\begin{aligned} &\frac{\partial F}{\partial t} + \frac{\partial}{\partial x}(vF) + \frac{\partial}{\partial v}(WF) - \frac{\lambda(t)}{2}\frac{\partial^2 F}{\partial x^2} - \chi(t)\frac{\partial^2 F}{\partial x \partial v} - \frac{\mu(t)}{2}\frac{\partial^2 F}{\partial v^2} = 0, \\ &W(x,v,t) = \frac{1}{F(x,v,t)}\int w\Phi(x,v,w,t)dw. \end{aligned} \qquad (2.3.5)$$

Эта модель дополняет уравнение Лиувилля эффектом случайного нестационарного блуждания. Используя связь

$$f(x,t) = \int F(x,v,t)dv$$

и учитывая обращение в ноль ВПФР всех порядков на границе области интегрирования, получаем из (2.3.5) интегрированием по скорости одномерное уравнение Фоккера-Планка:

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x}(uf) - \frac{\lambda(t)}{2}\frac{\partial^2 f}{\partial x^2} = 0, \quad u(x,t) = \frac{1}{f(x,t)}\int vF(x,v,t)dv. \qquad (2.3.6)$$

Подчеркнем, что в уравнении (2.3.5) коэффициент λ должен определяться по элементам ряда в соответствии с уравнениями эволюции моментов в силу этого самого уравнения. Важно учитывать, что, хотя одномерное уравнение (2.3.6) не содержит в явном виде некоторых коэффициентов, которые входят в двумерное уравнение (2.3.5), эти коэффициенты могут влиять на эволюцию моментов распределения $F(x,v,t)$ по скоростям.

Если мы обрываем цепочку совместных распределений на бинарной ВПФР $F(x,v,t)$, то среднее ускорение $W(x,v,t)$ в (2.3.5) следует задать независимо от трехмерной ВПФР. Это можно сделать в духе наивного прогноза, считая ускорение известным с предыдущего шага по времени (производная по времени трактуется в разностном смысле): $W(x,v,t) = W(x,v,t-1)$. Это же замечание относится и к коэффициентам λ, χ, μ. Если окажется, что такое приближение слишком грубо, то и обрывать на двумерной ВПФР нельзя, а требуется использовать аналогичное уравнение для $\Phi(x,v,w,t)$, и тогда моменты четырехмерного распределения считать известными, и т.д.

Подчеркнем, что нашей целью является сведение задачи к одномерному уравнению (2.3.6), а двумерное уравнение (2.3.5) будет использоваться для того, чтобы замкнуть моментную систему. Рассмотрим уравнение эволюции первого выборочного момента, определяемого как

$$m(t) = \langle x \rangle_t = \int x f(x,t) dx,$$

но эволюция будет следовать не уравнению Лиувилля, как в первоначальной простой постановке, а уравнению Фоккера-Планка (2.3.6). Из (2.3.6) получаем

$$\frac{dm(t)}{dt} = \int x \frac{\partial f(x,t)}{\partial t} dx = -\int x \frac{\partial}{\partial x} u(x,t) f(x,t) dx + \frac{\lambda(t)}{2} \int x \frac{\partial^2 f(x,t)}{\partial x^2} dx .$$

После интегрирования по частям с учетом того, что в граничных ячейках плотность распределения равна нулю, получаем результат, аналогичный следствию из уравнения Лиувилля:

$$\frac{dm(t)}{dt} = \int u(x,t) f(x,t) dx = \langle u \rangle_t = U(t), \qquad (2.3.7)$$

где $U(t)$, как и выше, есть среднее значение эмпирической скорости $u(x,t)$ по одномерному распределению в момент времени t. Рассмотрим теперь эволюцию самой эмпирической скорости. С одной стороны, из (2.3.6) следует, что

$$\frac{\partial}{\partial t}\left(f(x,t) u(x,t) \right) = u \frac{\partial f}{\partial t} + f \frac{\partial u}{\partial t} = u \left(-\frac{\partial (uf)}{\partial x} + \frac{\lambda}{2} \frac{\partial^2 f}{\partial x^2} \right) + f \frac{\partial u}{\partial t}. \qquad (2.3.8)$$

С другой стороны, используя (2.3.5), получаем, что та же самая правая часть (2.3.8) равна

$$\frac{\partial(uf)}{\partial t} = \int v \frac{\partial F(x,v,t)}{\partial t}\, dv = \int v \left(-v \frac{\partial F}{\partial x} - \frac{\partial (WF)}{\partial v} + \frac{\lambda}{2} \frac{\partial^2 F}{\partial x^2} + \chi \frac{\partial^2 F}{\partial x \partial v} + \frac{\mu}{2} \frac{\partial^2 F}{\partial v^2} \right) dv .$$

После интегрирования ее по частям получаем

$$\frac{\partial(uf)}{\partial t} = -\frac{\partial}{\partial x} \int v^2 F(x,v,t) dv + \int W(x,v,t) F(x,v,t) dv + \frac{\lambda}{2} \frac{\partial^2 (uf)}{\partial x^2} - \chi \frac{\partial f}{\partial x} . \qquad (2.3.9)$$

Введем обозначения для входящих в (2.3.9) интегралов:

$$e(x,t) f(x,t) = \int v^2 F(x,v,t) dv , \quad a(x,t) f(x,t) = \int W(x,v,t) F(x,v,t) dv . \qquad (2.3.10)$$

Если мы имеем дело с динамической системой, то статистико-механический смысл интегралов в (2.3.10) следующий: $e(x,t)$ есть удвоенная плотность средней кинетической энергии, а $a(x,t)$ представляет среднее локальное (по координате и времени) ускорение от внешнего поля. В результате из (2.3.8) и (2.3.9) получаем уравнение эволюции локальной скорости $u(x,t)$:

$$f \frac{\partial u}{\partial t} = u \left(\frac{\partial (uf)}{\partial x} - \frac{\lambda}{2} \frac{\partial^2 f}{\partial x^2} \right) - \frac{\partial (ef)}{\partial x} + af + \frac{\lambda}{2} \frac{\partial^2 (uf)}{\partial x^2} - \chi \frac{\partial f}{\partial x} . \qquad (2.3.11)$$

Уравнение (2.3.11) отличается от выводимых из уравнения Лиувилля уравнений эволюции моментов ВПФР. Из него следует, что изменение со временем средней макроскопической скорости $U(t)$ равно среднему ускорению, обусловленному внешними причинами, что согласуется с традиционными механическими представлениями:

$$\frac{dU(t)}{dt} = A(t) \equiv \int a(x,t) f(x,t) dx = \iint W(x,v,t) F(x,v,t) dx dv .$$

Заметим теперь, что в уравнение (2.3.11) входит величина $e(x,t)$, эволюция которой также должна быть определена. Используя тот же метод, что и при выводе уравнения (2.3.11), получаем из (2.3.6) и (2.3.10) уравнение

$$\frac{\partial (ef)}{\partial t} = \int v^2 \frac{\partial F(x,v,t)}{\partial t}\, dv = -\frac{\partial (e_3 f)}{\partial x} - 2\chi \frac{\partial (uf)}{\partial x} + \frac{\lambda}{2} \frac{\partial^2 (ef)}{\partial x^2} + \mu + \beta f . \qquad (2.12)$$

Здесь введены две новые величины:

$$\begin{aligned} e_3(x,t) f(x,t) &= \int v^3 F(x,v,t) dv , \\ \beta(x,t) f(x,t) &= \int \left(2vW + v^2 \frac{\partial W}{\partial v} \right) F(x,v,t) dv . \end{aligned} \qquad (2.3.13)$$

Таким образом, эволюция момента второго порядка выражается, кроме ранее введенных величин, через градиент момента третьего порядка и новый потоковый член.

Если обозначить через $e_k(x,t)$ плотность момента k-го порядка двумерной ВПФР по скоростям

$$e_k(x,t)\,f(x,t)=\int v^k F(x,v,t)dv\,, \tag{2.3.14}$$

то уравнение эволюции этой величины имеет следующий вид:

$$\begin{aligned}&\frac{\partial(e_k f)}{\partial t}=\int v^k\,\frac{\partial F(x,v,t)}{\partial t}\,dv=\\&=\int v^k\left(-v\frac{\partial F}{\partial x}-\frac{\partial(WF)}{\partial v}+\frac{\lambda}{2}\frac{\partial^2 F}{\partial x^2}+\chi\frac{\partial^2 F}{\partial x\partial v}+\frac{\mu}{2}\frac{\partial^2 F}{\partial v^2}\right)dv=\\&=-\frac{\partial(e_{k+1}f)}{\partial x}-k\chi\frac{\partial(e_{k-1}f)}{\partial x}+\frac{\lambda}{2}\frac{\partial^2(e_k f)}{\partial x^2}+k(k-1)\frac{\mu}{2}e_{k-2}f+\beta_k f\,,\end{aligned} \tag{2.3.15}$$

где

$$\beta_k(x,t)\,f(x,t)=\int\left(kW+v\frac{\partial W}{\partial v}\right)v^{k-1}F(x,v,t)dv\,.$$

Во многих случаях можно приближенно считать, что локальные ускорения W малы, либо что локальное ускорение является однородной формой скорости порядка s. Тогда без учета W эволюция момента k-го порядка выражается через уже известные величины и момент $(k+1)$-го порядка:

$$\begin{aligned}&\frac{\partial(e_k f)}{\partial t}=-\frac{\partial(e_{k+1}f)}{\partial x}+\frac{\lambda}{2}\frac{\partial^2(e_k f)}{\partial x^2}-k\chi\frac{\partial(e_{k-1}f)}{\partial x}+\\&+(k+s)ae_{k-1}f+k(k-1)\frac{\mu}{2}e_{k-2}f\,.\end{aligned} \tag{2.3.16}$$

Если на каком-либо порядке оборвать моментную систему, т.е. задать $e_{k+1}(x,t)$ независимо, то уравнение для момента k-го порядка будет представлять собой уравнение диффузии со сносом и источником.

Выясним теперь статистический смысл остальных коэффициентов, входящих в двумерное уравнение Фоккера-Планка (2.3.5). Рассмотрим сначала, как эволюционирует второй центральный выборочный момент одномерного распределения

$$\sigma^2(t)=\int\big(x-m(t)\big)^2\,f(x,t)dx\,.$$

В соответствии с (2.3.7) и (2.3.8) имеем

$$\frac{d\sigma^2}{dt}=-\int\big(x-m(t)\big)^2\frac{\partial(uf)}{\partial x}\,dx+\frac{\lambda}{2}\int\big(x-m(t)\big)^2\frac{\partial^2 f}{\partial x^2}\,dx=\lambda+2\int uxfdx-2mU\,.$$

Поскольку ковариация координаты и скорости равна

$$R(t) = \mathrm{cov}_{x,v}(t) = \int xvF(x,v,t)dxdv - \int xF(x,v,t)dxdv \cdot \int vF(x,v,t)dxdv =$$
$$= \int xu(x,t)f(x,t)dx - m(t)U(t) \equiv \mathrm{cov}_{x,u}(t),$$

то эволюция выборочной дисперсии в силу уравнения Фоккера-Планка имеет вид

$$\frac{d\sigma^2}{dt} = \lambda + 2R. \tag{2.3.17}$$

Тем самым выявлен статистический смысл параметра $\lambda(t)$ в уравнении (2.3.5). Это есть разность между производной текущей выборочной дисперсии ряда и удвоенной выборочной ковариацией исходного ряда и ряда его первых разностей. Сам же коэффициент ковариации $R(t)$ эволюционирует следующим образом:

$$\begin{aligned}
&\frac{dR}{dt} = -\frac{d(mU)}{dt} + \int x\frac{\partial(uf)}{\partial t}dx = \\
&= -U^2 - mA + \int x\left(-\frac{\partial(ef)}{\partial x} + af + \frac{\lambda}{2}\frac{\partial^2(uf)}{\partial x^2} - \chi\frac{\partial f}{\partial x}\right)dx = \\
&= -U^2 + \mathrm{cov}_{x,a} + E + \chi.
\end{aligned} \tag{2.3.18}$$

Здесь введено обозначение

$$E = E(t) = \int e(x,t)f(x,t)dx.$$

Следовательно, коэффициент $\chi(t)$ в уравнении (2.3.5) представляет собой разность между производной по времени ковариации координаты-скорости и ковариацией координаты-ускорения, минус разность между средней энергией $E(t)$ и энергией среднего движения $U^2(t)$.

По аналогии с (2.3.17), эволюция выборочной дисперсии скорости имеет вид

$$\frac{d\sigma_v^2}{dt} = \mu + 2\,\mathrm{cov}_{v,a}. \tag{2.3.19}$$

Таким образом, если задать левые части выражений (2.3.17-2.3.19), исходя из информации, известной к текущему моменту времени, то в рамках приближения двумерного уравнения Фоккера-Планка (2.3.5) его коэффициенты полностью определены.

Аналогично можно рассмотреть эволюцию и других моментов одномерного выборочного распределения, но получающиеся уравнения для высших моментов уже будут отличаться от эмпирических, строящихся непосредственно по выборкам. Например, коэффициент асимметрии γ одномерного распределения вычисляется по формуле

$$\gamma(t) = \frac{1}{\sigma^3(t)}\int \big(x - m(t)\big)^3 f(x,t)dx.$$

Эволюция этого коэффициента в силу уравнения Фоккера-Планка (2.3.6) дается формулой

$$\frac{d\gamma}{dt} = -3\gamma\frac{\lambda+2R}{2\sigma^2} + \frac{3}{\sigma^3}\int (x-m(t))^2 u(x,t) f(x,t)dx = \\ = -3\gamma\frac{\lambda+2R}{2\sigma^2} + 3\frac{U}{\sigma} + \frac{3}{\sigma^3}\left\langle (x-\bar{x})^2 (u-U)\right\rangle.$$

В работе [11] показано, что уравнение Фоккера-Планка (2.3.6), в котором коэффициенты сноса и диффузии определяются по формулам (2.3.6), (2.3.17), корректно, т.е. коэффициент диффузии положителен. Это позволяет использовать эмпирическое уравнение эволюции ВПФР для моделирования ансамбля траекторий нестационарного случайного процесса, эволюция выборочного распределения которого описывается данным кинетическим уравнением.

2.4. Генерация выборки из нестационарной функции распределения

Если известна ФР временного ряда $x(t)$ в каждый момент времени на некотором временном интервале, то можно сгенерировать набор траекторий, формирующих на рассматриваемом промежутке времени выборочное распределение, близкое к той ВФР, которая отвечает фактической траектории временного ряда. Уровень близости определяется статистической неопределенностью, с которой реализуется выборка из нестационарного временного ряда. Как было показано в п. 2.1, эта неопределенность отличается от классической, отвечающей стационарному случаю и определяемой по критерию Колмогорова-Смирнова, а является решением трансцендентного уравнения (2.1.7), которое было названо СУС.

Пусть ФР $F(x,t),\ t\in[t_0+1, t_0+N]$, известна на указанном промежутке времени либо как фактическая ВФР, построенная в скользящем окне некоторой заданной длины N, например, в соответствии с формулой (2.2.1), либо как решение уравнения Фоккера-Планка по методу, описанному в п. 2.3, и вычисленная через найденную плотность по формуле (2.2.2).

Затем генерируется стационарный равномерно распределенный на [0;1] ряд чисел $\{y_k\}$ длиной N. Генерируя теперь из известной ФР $F(x,t)$ по одному значению x_k в момент $t_k = t_0 + k$, получаем одну из возможных траекторий, имеющих на указанном промежутке распределение, близкое в плане СУС к распределению $F(x,t)$ на выборке заданной длины. Генерируя набор равномерно распределенных выборок, получаем в

результате набор траекторий, которые можно рассматривать как ансамбль решений кинетического уравнения.

Пусть t_0 есть начальный момент времени, в который ВФР $F_N(x,t_0)$ известна. Тогда в последующие моменты времени одна из возможных траекторий случайного процесса, для которого ВПФР меняется известным образом от $f_N(x,t_0)$ до $f_N(x,t_0+N)$, строится по формуле обращения соответствующей локальной по времени функции распределения, движущейся в скользящем окне длины N:

$$y_k = F_N(x_k, t_0 + k). \tag{2.4.1}$$

Подчеркнем, что, согласно (2.4.1), в каждый момент времени t из распределения $F_N(x,t)$ генерируется только одно значение ряда. Сама же $F_N(x,t)$ выступает в этот момент времени как генеральная совокупность. Тем самым имитируется процесс наблюдения за динамикой нестационарного временного ряда.

Пусть имеется набор равномерно распределенных рядов $\{y_k\}_j$, $j = 1,..., s$ длиной N. По формуле (2.4.1) из него можно получить пучок из s траекторий, ассоциированных с двумя ВПФР: $f_N(x,t)$ и $f_N(x,t+N)$, согласно наблюдаемой эволюции этих распределений. Каждая j-ая траектория из набора траекторий построенного пучка порождает на отрезке $[t_0+1; t_0+T]$ ВПФР $\tilde{f}_N(\{y\}_j; x, t_0+N)$, отличную, вообще говоря, от наблюденной $f_N(x,t_0+N)$. Однако по построению все эти выборочные траектории являются реализациями одного и того же нестационарного распределения вероятностей.

По совокупности сгенерированных траекторий можно оценить, насколько значимо отклонение модельного и фактического распределений. Используем для этого расстояние между функциями распределения в норме С:

$$\rho = \left\| \tilde{F}_N(\{y\}; x, t_0+N) - F_N(x, t_0+N) \right\|. \tag{2.4.2}$$

Рассмотрим также все попарные расстояния между ВПФР для сгенерированных траекторий

$$\tilde{\rho} = \left\| \tilde{F}_N(\{y\}; x, t_0+N) - \tilde{F}_N(\{y'\}; x, t_0+N) \right\|. \tag{2.4.3}$$

В результате моделирования должен получаться ансамбль траекторий временного ряда, который обладает следующими двумя свойствами. Во-первых, СУС расстояний (2.4.2) приближенно равен СУС расстояний (2.4.3). Это необходимо для того, чтобы отклонение сгенерированной траектории от фактической находилось на уровне статистической неопределенности, характерной для данного ряда. Во-вторых, если

сравнить сгенерированные выборки в окне $[t_0+1; t_0+N]$ с исходной ВФР $F_N(x,t_0)$, то соответствующий СУС должен быть приблизительно равным $\rho^*(N)$ в соответствии с формулами (2.1.6), (2.1.7) для встык-выборок. Пример такого ансамбля как решения уравнения Фоккера-Планка для эволюции выборочного распределения приростов биржевого индекса РТС приведен на рис. 13.

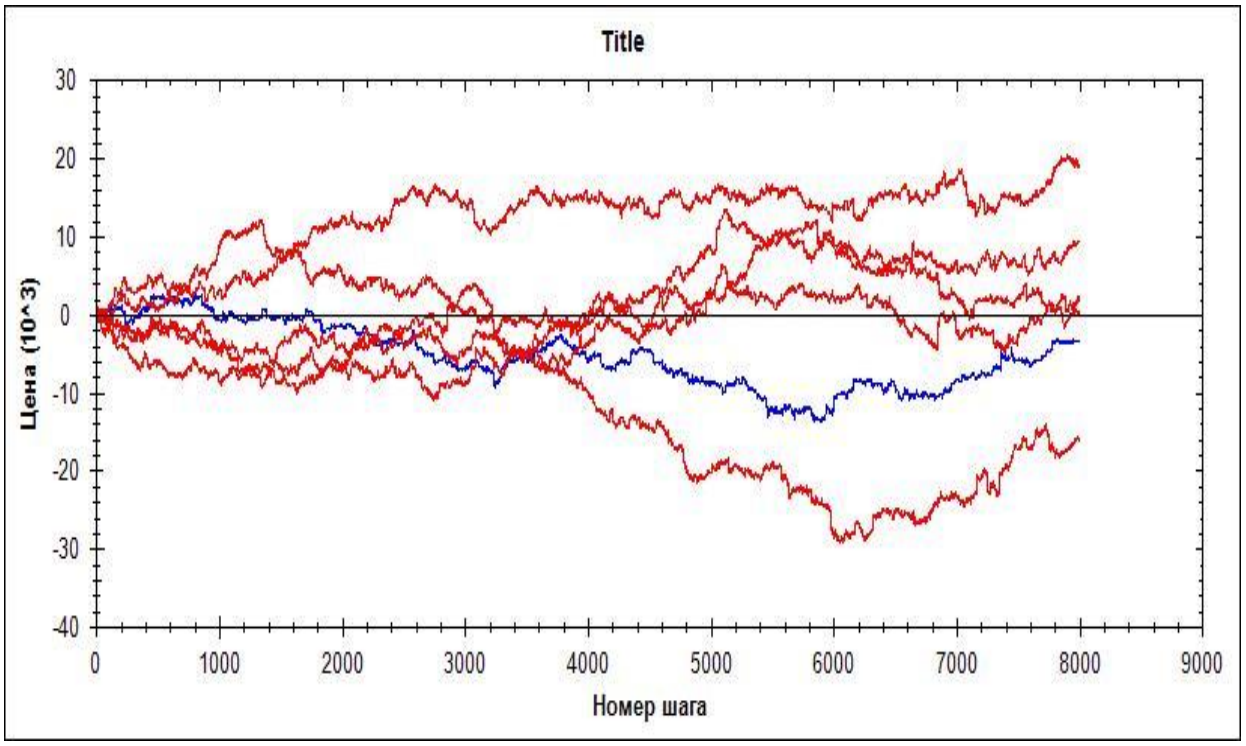

Рис. 13. Ансамбль траекторий в окне $N = 8000$
для решений уравнения Фоккера-Планка

Таким образом, использование уравнения Фоккера-Планка позволило, во-первых, сгенерировать ряд с теми статистическими свойствами его выборочного распределения, которые для него характерны, и, во-вторых, спрогнозировать выборочное распределение в среднем заметно точнее, чем наивный прогноз, в рамках которого распределение считается стационарным.

Тем самым численный эксперимент показал, что кинетическая модель эволюции распределения действительно отвечает определенным тенденциям изменения ВПФР, а не просто дает некоторое случайное ее изменение.

Описанный подход позволяет построить численный алгоритм моделирования нестационарного временного ряда с определенными непараметрическими свойствами его ВПФР, эволюционирующей в соответствии с уравнением Фоккера-Планка (или иным модельным уравнением, которое описывает эволюцию ВПФР рассматриваемого ряда).

2.5. Статистический анализ функционалов, заданных на траектории случайного процесса

В разделах 2.1-2.4 диссертации построен инструмент, позволяющий тестировать функционалы, заданные на случайной траектории, не по единственной ее реализации, а по набору траекторий, имеющих близкие статистические свойства. Процесс тестирования состоит в следующем.

Пусть на выборке длины N задан некоторый функционал $\Psi\{x(t-T+1),..., x(t)\}$. Это может быть, например, статистика в виде скользящей средней, а может быть и некоторая сложная конструкция в виде управления некоторой другой случайной траекторией. В первом случае для анализа функционала достаточно знать собственно ВПФР, а во втором случае важным является и последовательность значений СВ на траектории. Последняя задача наиболее востребована при анализе торговых стратегий, оперирующих с биржевыми рядами.

При тестировании функционала Ψ требуется определить, во-первых, его статистические свойства на выборках, отвечающих данной модели эволюции ВПФР, и, во-вторых, изучить устойчивость функционала при изменении параметров уравнения эволюции или при разладке динамики ВПФР.

Первая задача решается следующим образом. Пусть выбран интересующий нас фрагмент временного ряда и на нем построен пучок виртуальных траекторий числом N. Обозначим Ψ_j значение функционала на j-ой траектории. Его статистические свойства полностью определяются выборочным распределением $\Phi_N(\Psi)$, которое строится по имеющимся N значениям. В частности, можно определить среднее, дисперсию, нормированное среднее (коэффициент Шарпа):

$$\overline{\Psi} = \frac{1}{N}\sum_{j=1}^{N}\Psi_j, \quad \sigma_\Psi^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\Psi_j - \overline{\Psi}\right)^2, \quad S_\Psi = \frac{\overline{\Psi}}{\sigma_\Psi}, \tag{2.5.1}$$

а также моду, медиану, меньшие квантили и тому подобные величины.

Формула (2.5.1) дает корректный ответ на вопрос, какова, например, средняя доходность торговой системы на определенном промежутке времени. На практике может не быть достаточного количества данных, чтобы доходность, построенная по единственной реализации, могла быть достаточно полно протестирована на независимых встык-выборках.

Решение второй задачи требует предварительной кластеризации фрагментов временного ряда по типичным состояниям и определения характерных изменений

скорости (коэффициента сноса) с тем, чтобы можно было ввести разумные коэффициенты модуляции в формулах (2.3.6), (2.3.17). В частности, изменяя скорость в крайних значениях, можно проверить устойчивость функционала на изменение волатильности. Влияние корреляции между элементами ряда на значение функционала изучается варьированием коэффициента диффузии.

Например, если требуется смоделировать изменение вероятности срабатывания индикатора в определенной области значений случайной величины, то для этого можно провести «модуляцию» скорости с помощью произвольной подходящей функции $\alpha(x,t)$, т.е. рассматривать вместо скорости $u(x,t)$ величину $v(x,t)=u(x,t)\alpha(x,t)$. Величина $\alpha(x,t)$ только должна быть такой, чтобы на следующем шаге по времени расчетная вероятность была бы неотрицательной.

Применительно к биржевым рядам можно выделить три основных кластера, к которым относят ВПФР соответствующих фрагментов выборок: тренд вверх, тренд вниз, боковое движение цены. Если требуется протестировать поведение функционала при смене кластерной принадлежности фрагмента траектории ряда, то берутся два эталонных представителя ВПФР каждого кластера и генерируются соответствующие два фрагмента ряда, которые затем соединяются встык, и через которых в скользящем окне заданной ширины N пропускается изучаемый функционал.

К задаче смешанного типа относится задача оптимизации функционала, заданного на траектории. Предположим, что определено окно, в котором задан функционал (в примере на рис. 13 это окно длиной 8000), и имеется определенный исторический отрезок траектории, на котором этот функционал тестируется. Тестирование предполагает определение оптимального значения некоторого параметра функционала, чтобы его значение на тестируемом отрезке было максимальным. Это – типичная задача, возникающая при разработке торговых стратегий на бирже. Тогда вместо того, чтобы тестировать функционал в большом нестационарном окне, в котором есть только одна траектория, можно проводить анализ на ансамбле траекторий, имеющих одинаковые статистические свойства.

Итак, описан алгоритм тестирования функционала, заданного на случайной траектории. Он является комплексным, т.е. в нем используются выходные данные из алгоритма определения СУС, алгоритма построения оптимальной гистограммы, алгоритма кластеризации ВПФР и алгоритма прогнозирования ВПФР. Перечисленные задачи являются в определенной степени стандартными при любом нестационарном анализе временного ряда.

Глава III. Структура численного алгоритма моделирования нестационарных временных рядов

3.1. Алгоритм оптимального разбиения гистограммы

Численный алгоритм, реализующий метод оптимального равномерного разбиения гистограммы для нестационарного временного ряда, устроен следующим образом.

1. На вход подается анализируемый ряд длиной N, обозначим его $\{x_k\}\Big|_1^N$. Он разбивается на целое число $[N/T]$ непересекающихся выборок длины T, где T меняется от 10 (минимальная анализируемая выборка) до $[N/10]$, где 10 есть минимальное число выборок, по которым будет строиться распределение расстояний в той или иной норме. Если исходный ряд имеет длину, которая меньше 100, то задача оптимизации выборки через функционал индекса нестационарности не ставится. Элементы, принадлежащие k-ой выборке длины T, перенумеровываются: именно, s-ый элемент k-ой выборки y_{sk} есть элемент $x_{s+(k-1)T}$, $1 \le s \le T$.

2. Для каждой k-ой выборки длины T решается задача определения оптимального разбиения на n_k классовых интервалов по формулам (2.2.12) и (2.2.17). Уравнение в (2.2.17) решается итерационным способом с начальным приближением $n_k^0 = 2T^{1/3}$. Этот скейлинг достаточно хорошо описывает зависимость числа интервалов от длины выборки. Особенности конкретного распределения приводят к некоторому отклонению от этой зависимости, но сходимость достаточно быстрая – за 3-6 шагов. Итерационный процесс применяется к уравнению (2.2.17) и имеет вид

$$\frac{1}{n_k^i} = 2\psi\left(\frac{2\sqrt{T}}{\Sigma_T(n_k^{i-1})}\right), \quad i = 1, 2, \ldots \ . \tag{3.1.1}$$

Итерации состоят в делении пополам отрезка $[n_k^{i-1}, n_k^i]$ и продолжаются до тех пор, пока длина итерируемого промежутка не становится меньше или равной двум. Тогда следующая итерация считается последней. Итерационная запись (3.1.1) решения уравнения относительно числа классовых интервалов n в зависимости от длины выборки T представляет численное решение трансцендентного уравнения

$$\frac{\sqrt{T}}{\sum_{i=1}^{n}\sqrt{f_T(i)\left(1-f_T(i)\right)}} = nt_{1-1/(2n)}, \qquad (3.1.2)$$

где $t_{1-1/(2n)}$ есть квантиль распределения Стьюдента порядка $1-1/(2n)$ с $T-1$ степенью свободы.

3. По имеющимся непересекающимся выборками полученный набор мелкости разбиения $\varepsilon_k(T) = 1/n_k(T)$ позволяет построить функцию распределения $\Phi(\varepsilon, T)$ значений мелкости по выборкам фиксированной длины T. Далее численно ищется такое значение $\varepsilon^*(T)$, для которого выполнено равенство

$$\Phi(\varepsilon^*, T) = 1-\varepsilon^*. \qquad (3.1.2)$$

Для решения уравнения (3.1.2) отрезок [0;1], которому принадлежат значения $\varepsilon^*(T)$, равномерно разбивается на достаточно большое число промежутков (на порядок большее, чем максимальное из $n_k(T)$), после чего за решение принимается тот единственный номер промежутка j, умноженный на шаг (ширину h промежутка), для которого $\Phi(jh, T) + jh - 1 \le 0$, но $\Phi(jh+h, T) + jh + h - 1 > 0$. Решение имеет вид $\varepsilon^*(T) = jh$. Таким образом, в качестве оптимального разбиения принимается значение $n^*(T) = [1/\varepsilon*(T)] + 1$, трактуемое как величина, обратная к СУС распределения мелкостей разбиения набора независимых выборок.

3.2. Алгоритм определения длины выборки для выявления нестационарности

Численный алгоритм, реализующий описанный метод определения оптимальной выборки, устроен следующим образом.

1. Первый этап совпадает с этапом 1, описанным в разделе 3.1.

2. Входной информацией для этого расчетного алгоритма является набор значений $n^*(T)$, получаемый по алгоритму п. 3.1.

3. Выбирается норма для определения расстояния между встык-выборками. Для определенности считаем, что это гистограммная норма L1. Оптимальное разбиение гистограммы строится для каждой выборки объема T в соответствии с описанием алгоритма этого процесса в п. 3.1.

4. Численно находится СУС $\rho^*(T)$ для распределения расстояний между встык-выборками по функции распределения $G_T(\rho) = \int_0^{\rho} g_T(r)dr$, которая вычисляется с шагом 0,01. Именно, находится такой номер $k(T)$ ячейки равномерного разбиения отрезка [0; 2], при котором совместно выполняются неравенства

$$k(T): \quad \sum_{k=1}^{k(T)} g_T(k) \ge 1 - \frac{0{,}01k(T)}{2}, \quad \sum_{k=1}^{k(T)+1} g_T(k) < 1 - \frac{0{,}01(k(T)+1)}{2}. \qquad (3.2.1)$$

5. Вычисляется функционал индекса нестационарности

$$J(T) = 0{,}005 \cdot k(T) \cdot n^*(T). \qquad (3.2.2)$$

Прямым перебором определяется $\arg\max J(T)$. Это число как оптимальный объем выборки T_{opt} является выходной информацией данного алгоритма.

3.3. Алгоритм решения уравнения Фоккера – Планка для ВПФР

Численный алгоритм, реализующий схему решения эмпирического уравнения Фоккера-Планка, состоит из следующих этапов.

1. Первый этап совпадает с этапом 1, описанным в разделе 3.1.

2. Входной информацией для этого расчетного алгоритма является набор значений $n^*(T)$, получаемый по алгоритму п. 3.1. Также входной информацией является объем выборки T_{opt} из п. 3.2, на котором наиболее эффективно определять разладку и который отвечает наибольшей нестационарности относительно встык-выборок.

3. Анализ индекса нестационарности в п. 3.2 позволил определить длину $T = T_{opt}$ выборки, по которой строится ВПФР $f_T(x,t)$, так что далее эта длина фиксирована и для краткости опущена. Представляющая ВПФР гистограмма имеет вид

$$f(x) = f_j, \ x \in [(j-1)/n;\ j/n], \quad j = 1 \div n, \qquad (3.3.1)$$

где число классовых интервалов определяется по формуле $n = n^*(T_{opt})$.

В том же окне длины T строим совместную плотность распределений $\Phi(x,v,t)$ значений временного ряда $x(t)$ и его приращений $v(t) = x(t+1) - x(t)$. При этом справедлива формула

$$f(x,t) = \int_{-1}^{1} \Phi(x,v,t)dv,$$

где в пределах интегрирования учтено, что $x \in [0;1]$, так что $v \in [-1;1]$. Скоростной интервал разбивается на то же число классовых интервалов, что и одномерная гистограмма. В качестве модельного кинетического уравнения для эволюции ВПФР (3.3.1) используем уравнение Фоккера-Планка:

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x}(uf) - \frac{\lambda}{2}\frac{\partial^2 f}{\partial x^2} = 0,$$

где параметры сноса (средняя скорость $u(x,t)$) и диффузии $\lambda(t)$ определяются формулами

$$u(x,t) = \frac{1}{f(x,t)}\int v\Phi(x,v,t)dv,$$

$$\lambda(t) = \frac{d\sigma^2}{dt} - 2\,\mathrm{cov}_{x,u}(t), \qquad .$$

$$\sigma^2(t) = \int_0^1 (x-\bar{x})^2 f(x,t)dx.$$

Их разностный вид следующий. Скорость $u(i,t)$ в i-ой ячейке на шаге t определяется по модели эволюции ВПФР с предыдущего шага по времени. Считаем, что на границах промежутка, где принимает значения случайная величина, ВПФР равна нулю. Для равномерно ограниченного ряда $x(t)$ этому условию всегда можно удовлетворить, введя нужное количество дополнительных граничных ячеек разбиения, в которых плотность распределения равна нулю во все моменты времени.

В каждой ячейке разбиения, где существуют значения временного ряда, эмпирическая скорость определяется по формуле

$$u(i+1,t) = \frac{u(i,t)f(i,t) - f(i,t) + f(i,t+1)}{f(i+1,t)}, \quad u(0,t) = 0. \qquad (3.3.2)$$

Если же в i-ой ячейке плотность распределения $f(i,t)$ оказалась равной нулю, то формально скорость в такой ячейке не определена. В фиктивных граничных ячейках скорость естественно положить равной нулю. Во внутренних ячейках с нулевой плотностью скорость определяем по линейной интерполяции между окаймляющими ячейками, плотность в которых отлична от нуля.

Видно, что скорость изменения ВПФР в данный момент времени определяется значениями распределения в этот и последующий моменты. Это означает, что в текущий момент времени t эмпирическая скорость известна для предыдущего момента $t-1$, что следует учитывать при составлении эволюционной модели. Поэтому вместо (3.3.2) скорость $u(i,t)$ изменения ВПФР в i-ой ячейке в момент t определяем формулой

$$u(i+1,t)\,f(i+1,t-1)=\sum_{k=1}^{i}\big(f(k,t)-f(k,t-1)\big)=F(i,t)-F(i,t-1)\,. \qquad (3.3.3)$$

Таким образом, правая часть (3.3.3) представляет собой изменение по времени выборочной функции распределения в ячейке, предшествующей той, в которой вычисляется значение эмпирической скорости.

Оценка величины $\lambda(t)$ по выборке в соответствии с дискретным аналогом дисперсии имеет вид

$$\lambda(t)=\frac{1}{T}\sum_{k=t-T+1}^{t}\big(x(k)-x(k+1)\big)^2-\frac{1}{T^2}\big(x(t+1)-x(t-T+1)\big)^2 \qquad (3.3.4)$$

и при $T>1$ эта величина строго положительна.

4. Для повышения устойчивости решения уравнения Фоккера-Планка далее была использована численная схема, в которой, во-первых, каждый классовый интервал разбит еще на 100 ячеек (с учетом равномерности в них, по построению, выборочной плотности), и, во-вторых, аппроксимация второй производной делается в лево-разностном шаблоне, в котором значение функции в ячейке x берется со следующего шага по времени. Описанная процедура приводит к разностному уравнению:

$$f(x,t+1)=f(x,t)+\frac{f(x,t)u(x,t-1)-f(x+1,t)u(x+1,t-1)}{h}+$$
$$+\frac{\lambda(t-1)}{2h^2}\big(2f(x-1,t)-f(x,t+1)-f(x-2,t)\big).$$

Разрешая его относительно $f(x,t+1)$, получаем схему расчета:

$$f(x,t+1)=f_T(x,t)\frac{f(x,t)}{1+\lambda(t-1)/2h^2}+\frac{f(x,t)u(x,t)-f_T(x+1,t)u(x+1,t)}{h+\lambda(t-1)/2h}+$$
$$+\frac{\lambda(t-1)}{\lambda(t-1)+2h^2}\big(2f(x-1,t)-f(x,t+1)-f(x-2,t)\big). \qquad (3.3.5)$$

5. Решение уравнения Фоккера-Планка по формуле (3.3.5) осуществляется на такой горизонт τ, внутри которого отклонение прогнозной ВПФР от опорной (известной в начальный момент времени) не превосходит величины $\tau\rho^*/T$. Если предельное отклонение достигнуто, то в этот момент вычисленная ВПФР принимается за новое начальное распределение, для которого по выборке длины T находится новая скорость $u(x,t)$ и параметр диффузии $\lambda(t)$, вычисляемые уже не по элементам ряда, а непосредственно по имеющимся функциям распределения, включая фрагмент прогнозной ВПФР в скользящем окне ширины T.

6. После завершения численного решения уравнения Фоккера-Планка полученная прогнозная ВПФР вновь кластеризуется в найденное ранее оптимальное количество классовых интервалов.

3.4. Алгоритм генерации пучка нестационарных траекторий

После того, как найдена прогнозная ВПФР на заданном промежутке T, известная в каждый момент времени от t_0 (исходная ВПФР) до $t_0 + T$, можно в каждый момент времени $t_k = t_0 + k,\ 0 < k \le T$, сгенерировать случайное число из функции распределения $F(x, t_k)$, считая $F(x, t_k) = \int_0^x f(y, t_k)dy$. Алгоритм генерации таких чисел, образующих в совокупности траекторию временного ряда на этом промежутке времени, состоит в следующем.

1. Генерируются S выборок длины T из равномерного распределения на [0;1]. Обозначим их $\{y_i^s\},\ i = 1,...,T;\ s = 1,...,S$.

2. Из решения уравнения Фоккера-Планка, известного из п. 3.3, строится прогнозная ВФР в том же классовом разбиении, что и ВПФР. Она имеет вид

$$F(x,t) = (nx - j)\cdot f_{j+1}(t) + \sum_{k=1}^{j} f_k(t),\ \ x \in [(j-1)/n;\ j/n],\quad j = 1 \div n. \qquad (3.4.1)$$

3. В каждый момент времени $t_k = t_0 + k,\ 0 < k \le T$ решается уравнение относительно случайной величины x:

$$F(x, t_k) = y_k^s. \qquad (3.4.2)$$

В силу строгой монотонности ВФР $F(x, t_k)$ классовый интервал Δx_j, содержащий соответствующее решение для x_k^s, единственный. Он определяется из условий

$$F\left(\frac{j-1}{n}, t_k\right) \le y_k^s,\quad F\left(\frac{j}{n}, t_k\right) > y_k^s. \qquad (3.4.3)$$

Для найденного интервала Δx_j решение уравнения (3.4.2) имеет вид

$$x_k^s = \frac{1}{n}\left(\frac{y_k^s - \sum_{i=1}^{j} f_i(t_k)}{f_{j+1}(t_k)} + j\right), \qquad (3.4.4)$$

если $f_{j+1}(t_k) > 0$. Если же $f_{j+1}(t_k) = 0$, то в качестве x_k^s выбирается центр классового интервала Δx_j.

3.5. Блок-схема программного комплекса

Итак, в работе представлен программный комплекс, позволяющий анализировать и прогнозировать нестационарные временные ряды и связанные с ними различные стратегии распознавания образов. Он состоит из четырех последовательных блоков, структура и наполнение которых представлены на рис. 14.

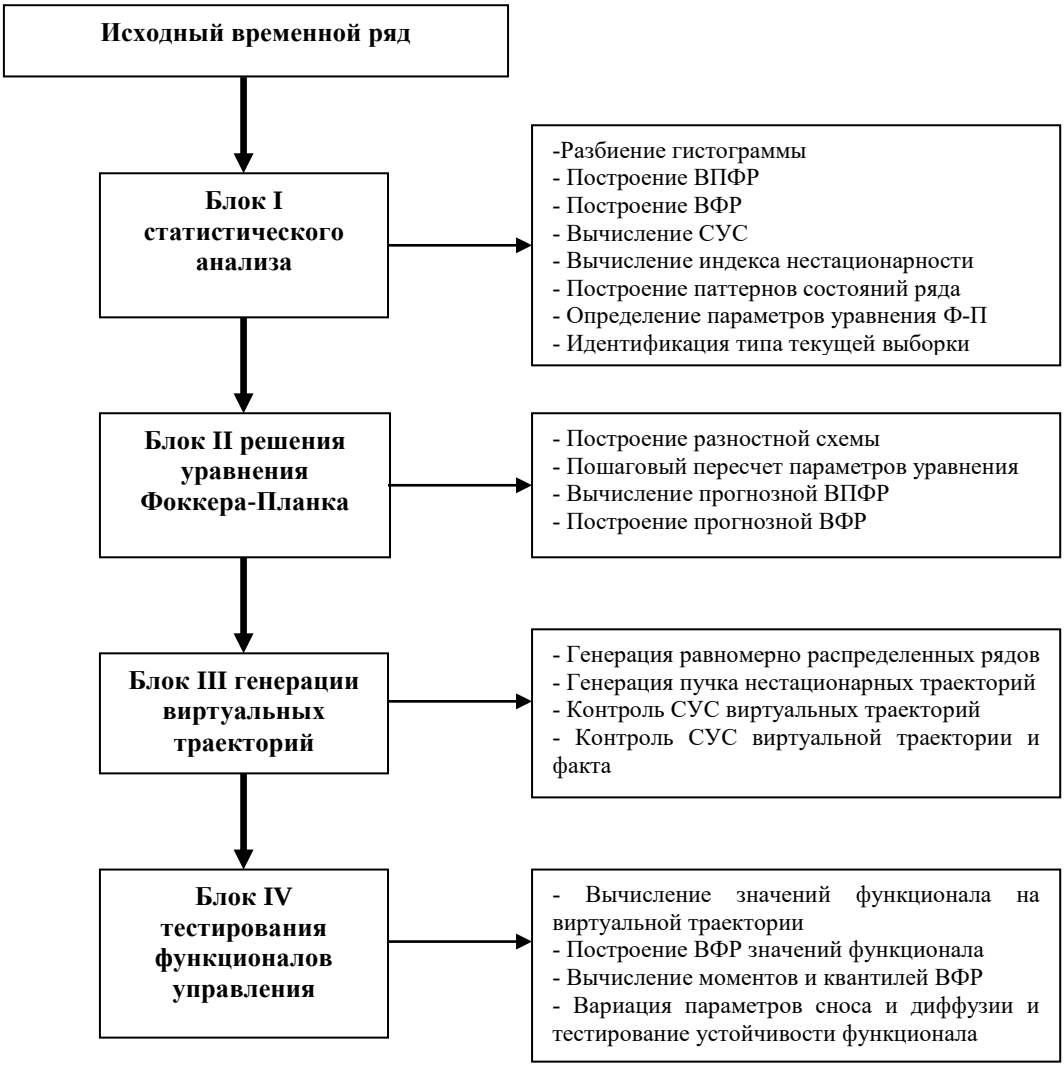

Рис. 14. Блок-схема программного комплекса

На рис. 15-20 приведены структурные схемы отдельных блоков.

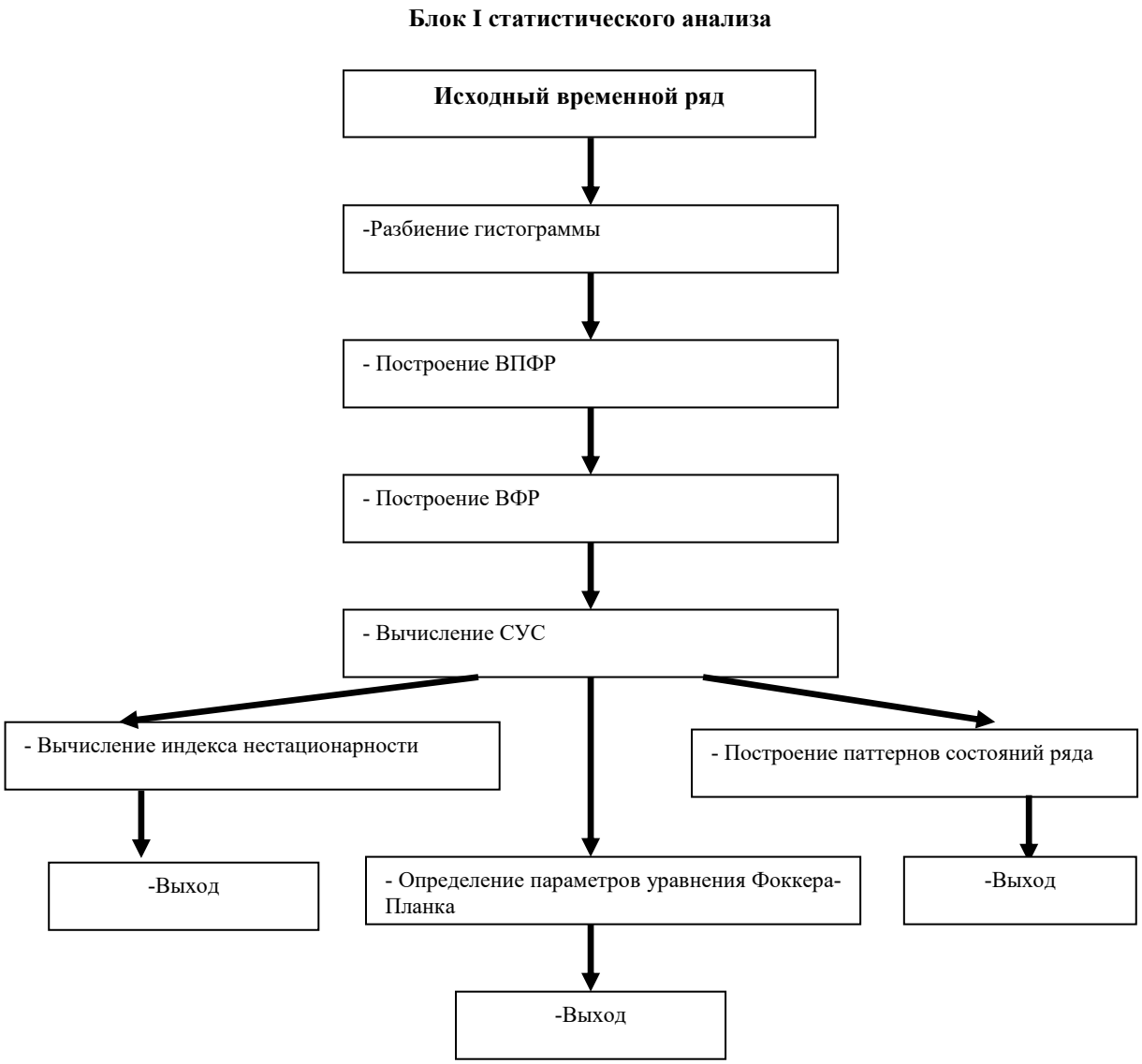

Рис. 15. Структура блока статистического анализа временного ряда

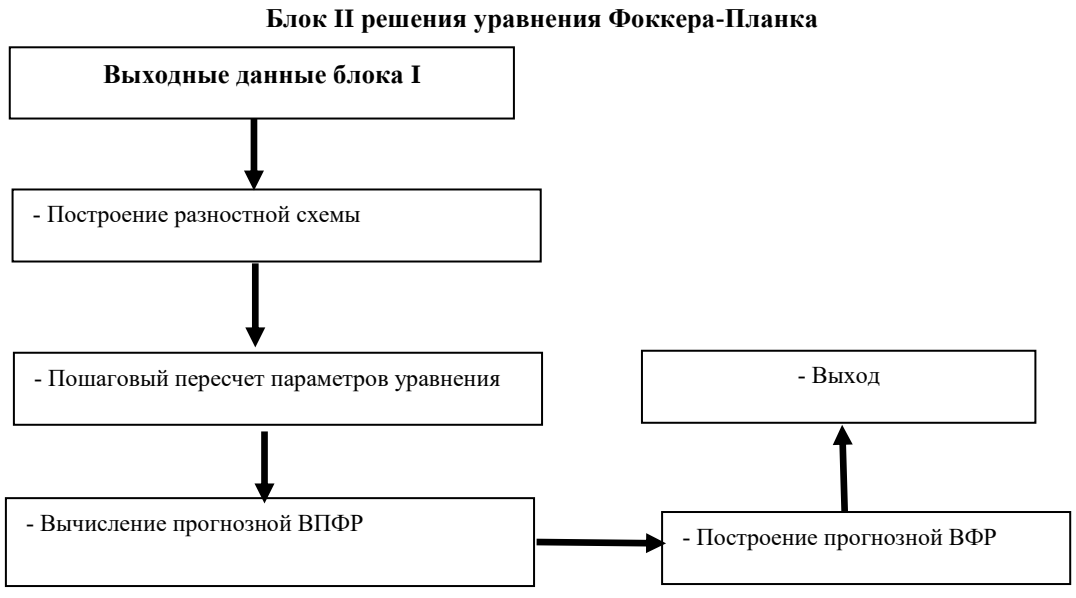

Рис. 16. Структура блока решения уравнения Фоккера-Планка

Рис. 17. Структура блока генерации пучка нестационарных траекторий

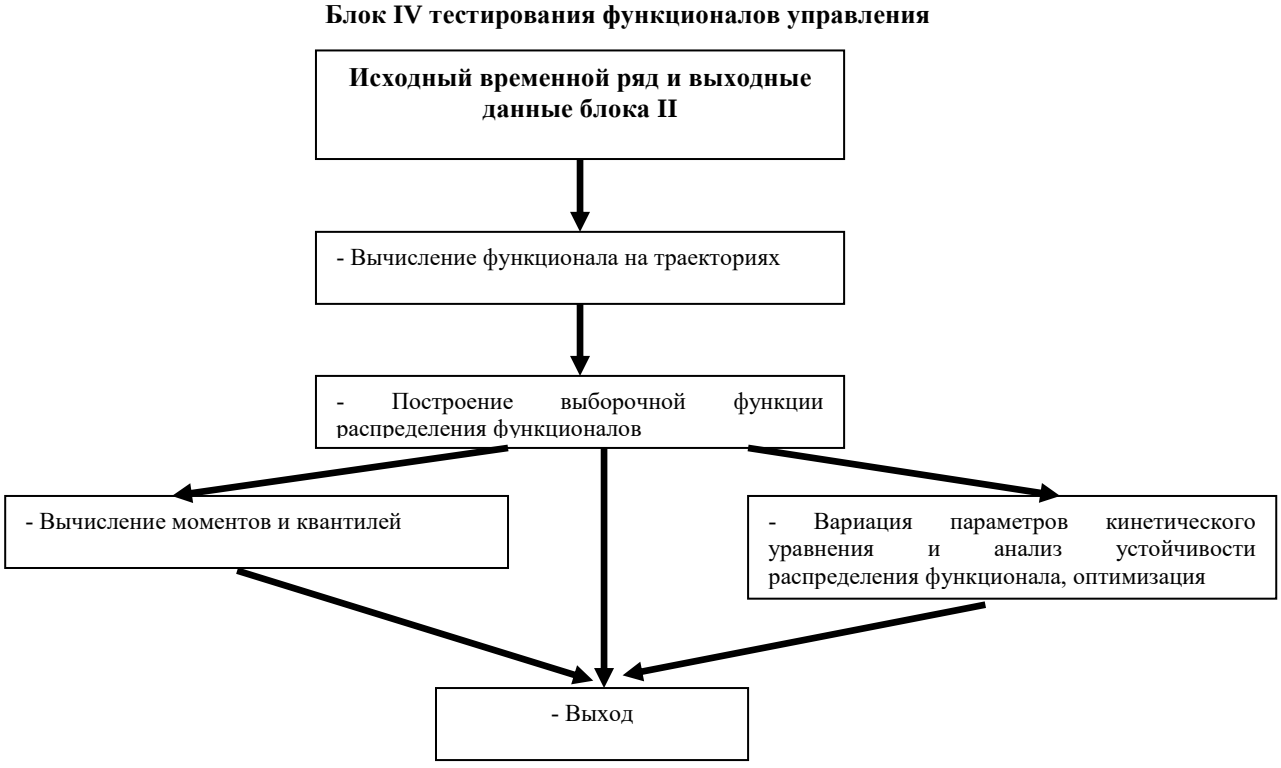

Рис. 18. Структура блока тестирования функционалов управления

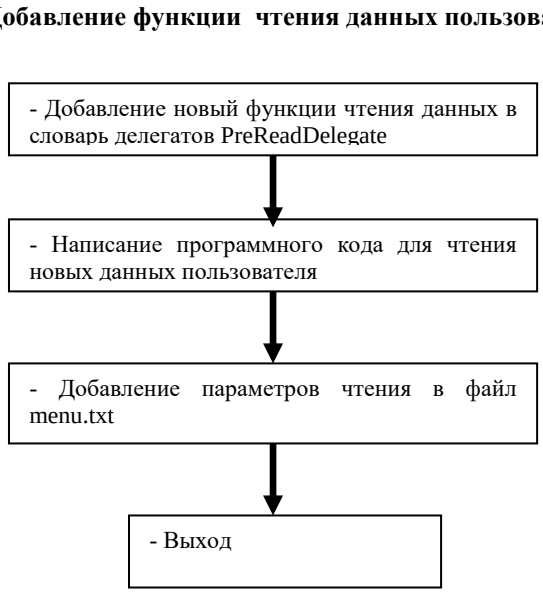

Рис. 19. Структура подсистемы добавления данных пользователя

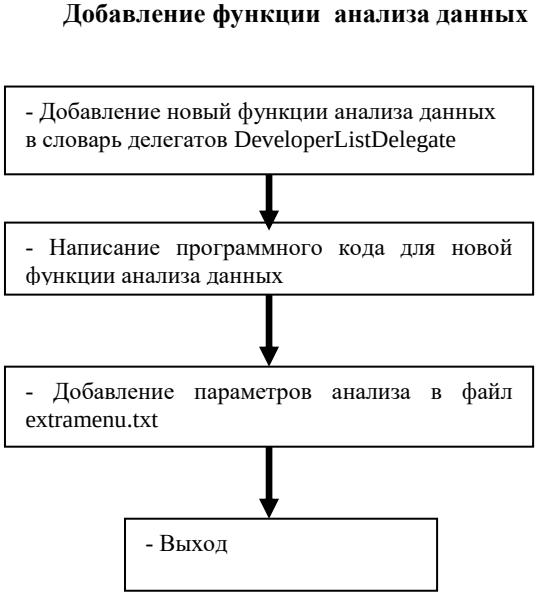

Рис. 20. Структура подсистемы добавления функции анализа данных

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. 21.

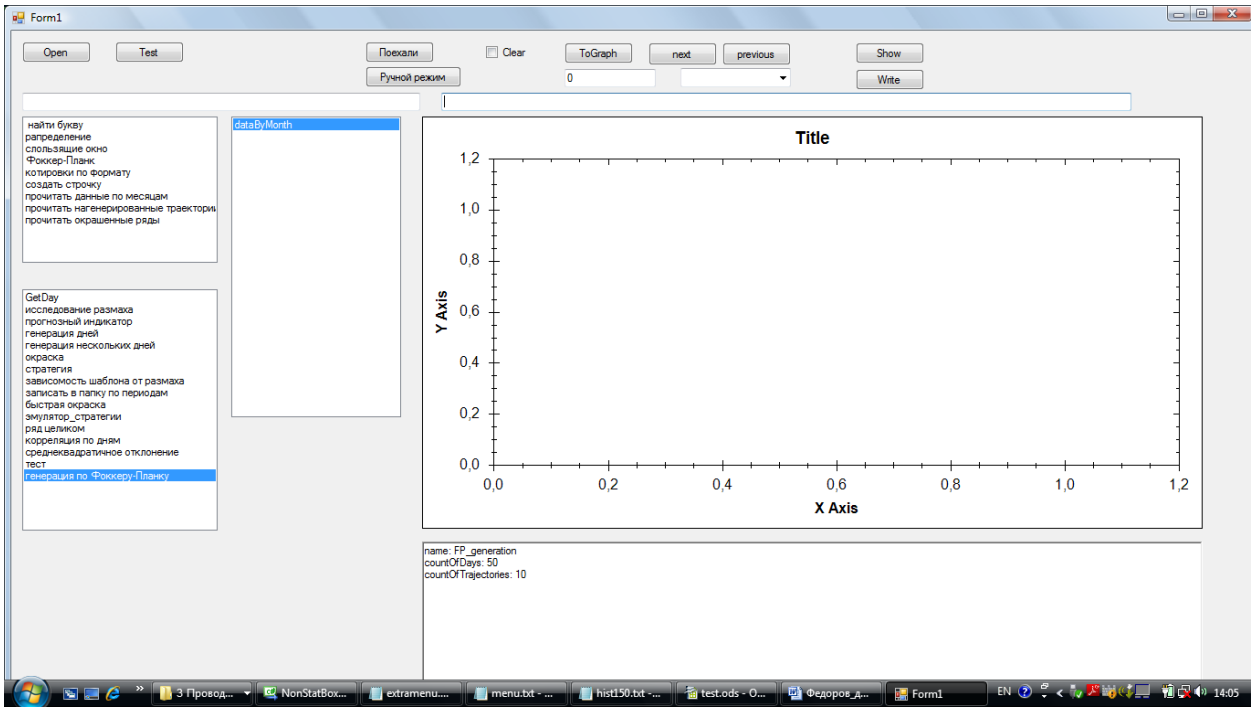

Рис. 21. Скрин-шот рабочего окна программы

Кнопка «Open» открывает файл в текстовом формате из папки, путь к которой прописывается в строке окна, расположенного под указанной кнопкой.

Кнопка «Test» обеспечивает быстрый доступ к файлу – либо последнему, который использовался, либо заданному специальной командой в коде программы.

Кнопка «Ручной режим» осуществляет запуск расчета для конкретного исследования – последнего или специально прописанного в коде.

Верхнее левое поле рабочего окна содержит первичные функции, которые можно выполнять с выбранным файлом данных.

Нижнее левое поле рабочего окна содержит функции, которые можно выполнять с результатами расчетов.

В поле окна справа отображается список проведенных исследований.

Поле с разметкой осей графика предназначено для визуализации результатов расчетов.

Нижнее поле под графиком предназначено для ввода-вывода числовых или текстовых параметров исследования, которое выбирается в поле первичных функций.

Кнопка «Поехали» осуществляет запуск расчета для задания, выбранного в любом из двух левых полей окна (поля не конфликтуют между собой), и после выбора параметра в третьем (нижнем) поле.

Кнопка «Clear» - флаг очистки поля графика при необходимости визуализации других графиков.

Кнопка «ToGraph» дает команду выведения графика или группы графиков на экран в поле визуализации.

Кнопки «next/previous» обеспечивают переход к визуализации следующего или предыдущего графика в выбранной серии графиков.

Кнопка «Show» является командой вывода в нижнее поле под графиком числовых результатов исследования.

Кнопка «Write» осуществляет запись результатов исследования в папку, путь к которой указан в строке над полем графика, а название файла совпадает с названием исследования, выбранного в поле списка проведенных исследований.

Два окна над полем графика отвечают номеру (справа) и цвету (слева) выбираемого графика в серии.

Главным объектом для работы с программой является абстрактный класс Developer. Каждый потомок класса Developer реализует ряд методов для визуализации и записи информации. Для обращения к объектам Developer используется шаблон Singleton. Класс OS содержит в себе список всех объектов Developer, доступных пользователю. Возможность визуализации обеспечивается переопределением абстрактных функций класса Developer: ToGraph() , Write(), Show().

Вычислительный алгоритм написан на языке C#.

Системные требования.

Операционная система Windows XP или более новые версии

30 Mb свободного места на диске

Для написания собственных функционалов необходима среда разработки Visual Studio 2010 или более новая версия.

Код визуализации.

Возможность визуализации обеспечивается переопределением абстрактных функций класса Developer: ToGraph() , Write(), Show().

Функция Show() позволяет выводить текстовую информацию в окно вывода.

Функция Write() позволяет записывать текстовую информацию в файл.

Функция ToGraph() обеспечивает возможность вывода графической информации. Если необходимо вывести по датам сложный графический объект, содержащий несколько графиков разного цвета, то для решения этой задачи используются интерфейс iDayGraph. Переход от одной даты к другой осуществляется по нажатие кнопок Previous и Next.

Доступные функции.

- Котировки по формату.

Позволяет прочитать текстовый файл с биржевыми котировками или же с любым численным рядом меняющимся во времени.

- Прочитать данные с заданным шагом по длине окна.

Читает котировки, разбитые по длинам заданной длины из указанной папки.

- Прочитать нагенерированные траектории

Читает ранее нагениреванные пользователем числовые ряды из текстового файла.

- Прочитать окрашенные ряды

Читает ранее классифицированные по близости к паттернам числовые ряды.

- Создать строчку

Превращает текстовый файл в строчку символов, убирая пробелы.

Дополнительные функции.

- Cut

  Вырезает кусок ряда по начальному и конечному индексу для дальнейших исследований.

- Гистограмма дельт

  Строит гистограмму попарных разностей ряда.

- Нормированная гистограмма дельт

  Строит гистограмму попарных разностей ряда, нормированных на некоторый отрезок.

- Список гистограмм приращений

  Создает список гистограмм попарных разностей ряда, построенных на встык-выборках определенной длины, нормированных на некоторый отрезок.

- Генерация

  Генерирует пучок траекторий, используя алгоритм непараметрической генерации.

- Разбить на периоды

  Разбивает ряд котировок на временные интервалы.

- Среднеквадратичное отклонение

  Считает среднеквадратичное отклонение.

- Функция распределения

  Считает функцию распределения данной гистограммы.

- GetDay

  Выбирает конкретный день из списка

- Исследование размаха

  Считает параметры максимальных дельт дня.

- Генерация дней

Генерирует каждый день заданное количество траекторий, используя алгоритм непараметрической генерации.

- Генерация несколько дней.

Генерирует заданное количество траекторий длиной в несколько дней без учета гэпов.

- Зависимость шаблона от размаха

Генерирует ряд шаблонов в зависимости от максимальной дельты дня.

- Ряд целиком

Соединяет дни в один временной ряд.

- Корреляция по дням

Считает корреляцию по дням с заданным рядом из списка исследований.

- Тест

Выполняет тест, заданный пользователем в коде программы.

- Генерация по Фоккеру-Планку.

Генерирует заданное количество траекторий, используя эволюционное уравнение Фоккера-Планка.

Для работы программы необходима операционная система не ниже Windows XP, установка программы требует 30 Mb свободного места на диске.

Для написания собственных функционалов необходима среда разработки Visual Studio 2010 или более новая версия.

Примеры численных расчетов для некоторых практических задач представлены далее в главе IV.

Глава IV. Результаты численных расчетов

4.1. Тестирование корректности модели прогнозирования ВПФР по уравнению Фоккера - Планка

Для проверки корректности работы вычислительных алгоритмов, описанных в главе III, в данном разделе приводятся примеры построения прогнозной ВПФР по уравнению Фоккера-Планка, коэффициенты сноса и диффузии в котором определяются с предыдущих шагов по времени по выборке, длина которой равна горизонту прогнозирования в соответствии с методикой, изложенной выше.

Прогнозная ВПФР, обозначаемая $\tilde{f}_T(x, t_0 + T)$, должна обладать тем свойством, что отклонение $\tilde{f}_T(x, t_0 + T)$ от ВПФР начального состояния $f_T(x, t_0)$ в выбранной норме должно быть порядка СУС между встык-выборками. При этом желательно, чтобы ее отклонение от фактически состоявшейся ВПФР $f_T(x, t_0 + T)$ не превосходило бы СУС.

На рис. 22 показаны две типовые выборочные плотности распределения для встык-выборок ряда приростов цен закрытия 1-минутных интервалов для индекса РТС за неделю (5 торговых дней). Описание их эволюции в этом временном промежутке является практически актуальной задачей. Рассматриваемому промежутку отвечает длина выборки 3000, ВПФР имеет индекс нестационарности, равный в рассматриваемом примере 3,2. Анализировались данные за 2015 г., всего 400 тыс. значений.

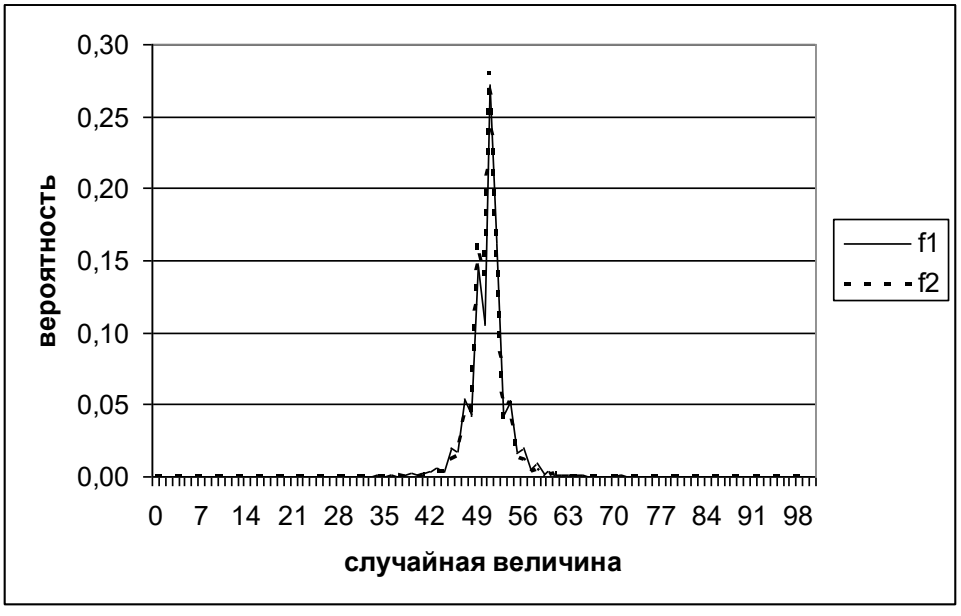

Рис. 22. Пример выборочных плотностей в соседних окнах длиной $T = 3000$

Область изменения случайной величины (нормированных на [0;1] приростов цен) на рис. 16 разбита на 100 классовых интервалов, по оси абсцисс отложен номер соответствующего интервала. Несмотря на то, что визуально распределения представляются весьма близкими, расстояние между ними в норме C равно $\rho = \|F_T(x,t_0) - F_T(x,t_0+T)\| = 0{,}13$ при том, что в стационарном случае для выборок представленных длин это расстояние должно быть порядка 0,04.

На рис. 23 показаны две ВПФР для того же ряда, только вместо фактической $f_T(x,t_0) \equiv f_1$ представлена прогнозная $\tilde{f}_T(x,t_0+T)$, которая сравнивается с фактической $f_T(x,t_0+T) \equiv f_2$. Расстояние между ними равно 0,06, что примерно в 1,5 раза больше, чем стационарный уровень шума и в два раза меньше, чем расстояние между встык-выборками. Это показывает, что кинетическое уравнение, рассматриваемое как модель изменения ВПФР, дает качественно адекватные результаты прогнозирования.

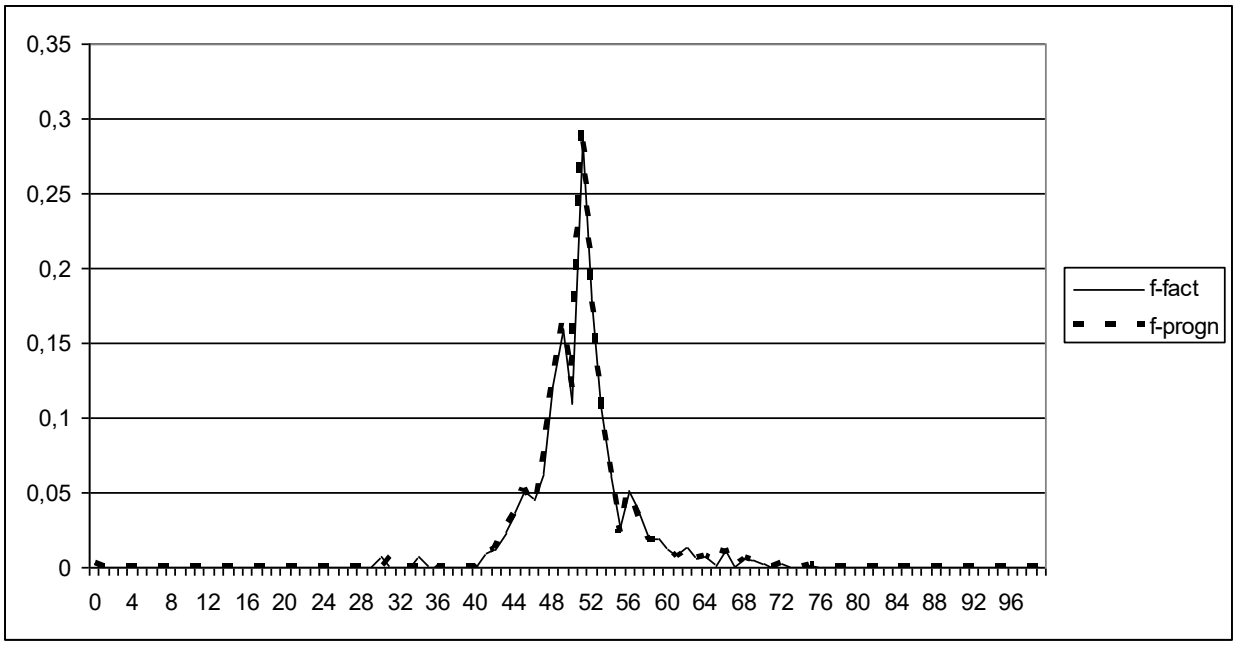

Рис. 23. Пример сравнения фактической и прогнозной ВПФР в окне длиной $T = 3000$

Интересно также оценить влияние сноса и диффузионного члена в прогнозной модели. На рис. 18 показаны фактическая и прогнозная ВПФР в модели, когда диффузионный член отсутствует. При этом расстояние между прогнозом и фактом оказалось равным 0,10. Основные отличия обусловлены различиями ВПФР в точках локальных экстремумов, когда прогнозные пики вверх по уравнению Лиувилля оказываются выше, а прогнозные

минимумы – напротив, ниже, чем сглаженные положения экстремумов с учетом диффузионного члена.

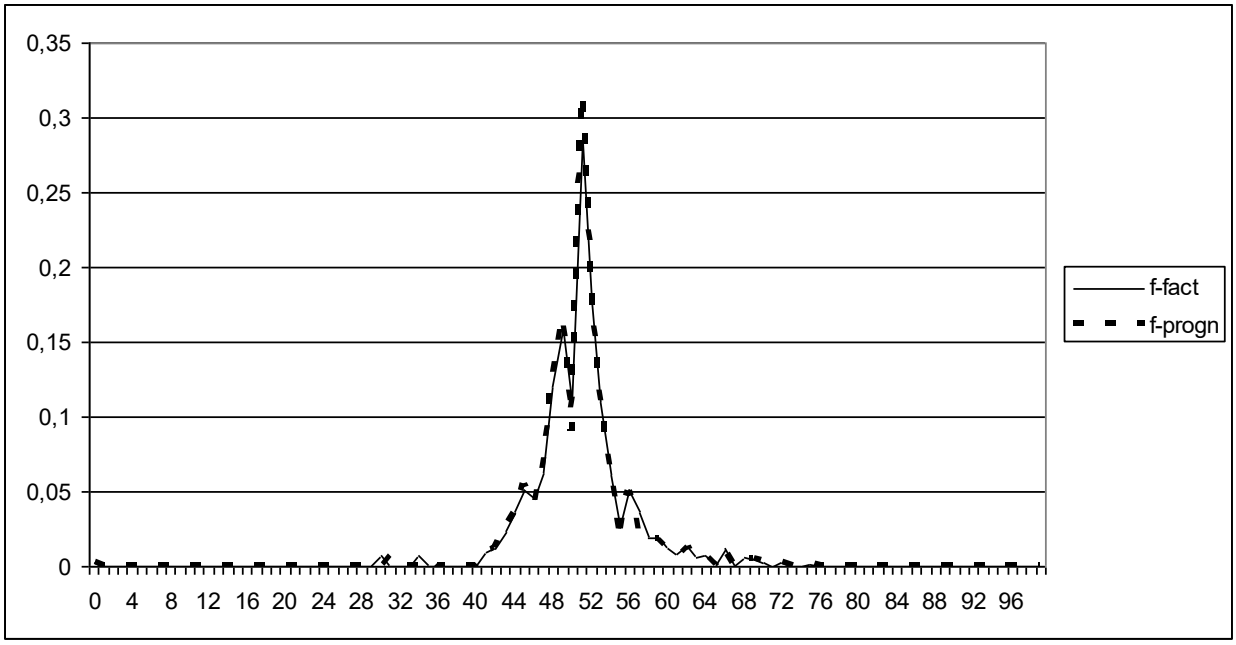

Рис. 24. Пример сравнения фактической и прогнозной ВПФР в окне длиной $T = 3000$ без учета диффузионного члена

На рис. 23-24 по оси абсцисс отложен номер классового интервала при равномерном разбиении на 100 ячеек, а по вертикальной оси – эмпирическая частота попадания в него.

Таким образом, кинетическая модель прогнозирования ВПФР по уравнению Фоккера-Планка дает более точные результаты, чем чисто механическая модель сноса по уравнению Лиувилля, а также значительно более точные, чем модель стационарного приближения.

4.2. Тестирование корректности модели генерации нестационарного временного ряда

Проверка корректности алгоритма генерации пучка траекторий из временного ряда, ВПФР которого эволюционирует в соответствии с заданным модельным кинетическим уравнением, состоит в следующем.

Во-первых, следует убедиться в том, что построенные траектории в определенном смысле близки: именно, СУС попарных расстояний между ними должен быть больше стационарного расстояния (если, конечно, ряд нестационарный) и меньше, чем СУС встык-выборок этого ряда.

Во-вторых, надо убедиться в том, что сгенерированы выборки именно из этого эволюционирующего временного ряда. Для этого надо проверить, что СУС попарных расстояний между сгенерированными выборками равен СУС расстояний между фактической траекторией в прогнозном окне и сгенерированными выборками.

В-третьих, проверка корректности моделирования эволюции ВПФР состоит в том, что СУС расстояний между опорной ВПФР $f_T(x,t_0)$ и сгенерированными выборками $\tilde{f}_T(\{y\};x,t_0+T)$ должен быть равен СУС встык-выборок данного временного ряда.

На рис. 25 приведен пример пучка траекторий (приростов цен в фактических единицах), сгенерированных описанным в работе методом на основе анализа предыдущих встык-выборок. Жирной линией выделена фактически состоявшаяся траектория.

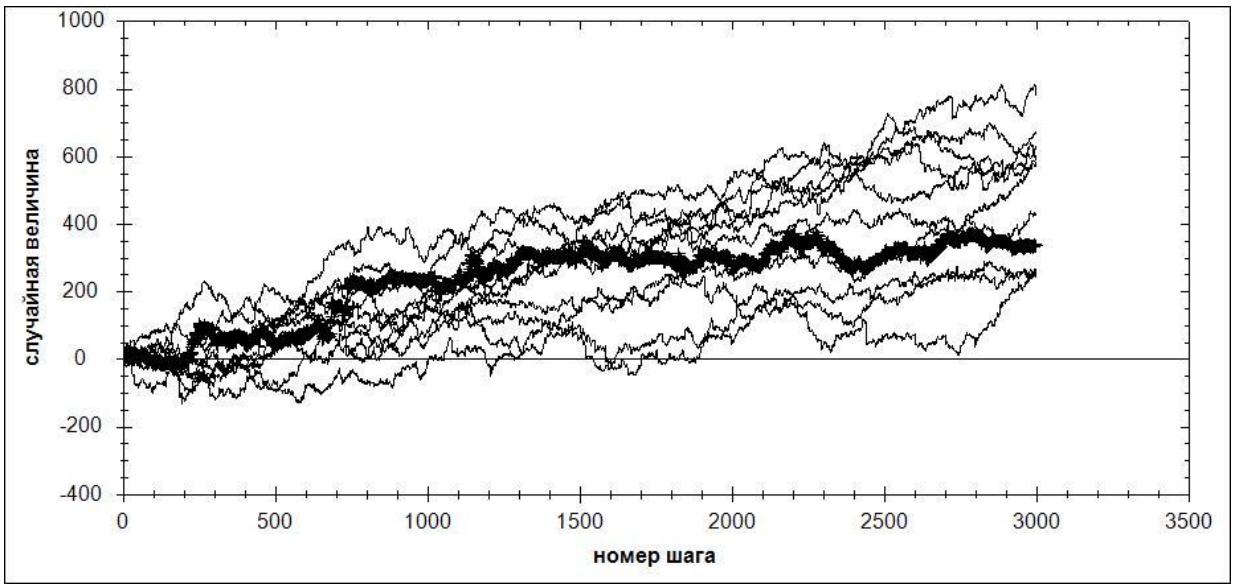

Рис. 25. Пучок траекторий в окне $T=3000$ для уравнения Фоккера-Планка

Для анализа нестационарной статистики было сгенерировано 1000 различных траекторий. Укажем основные характеристики построенного пучка. Как уже говорилось, для рассматриваемого ряда СУС встык-выборок длины 3000 равен 0,13. Далее, характерное расстояние между сгенерированными распределениями по выборкам длины T на промежутке $[t_0, t_0+T]$, оказалось равным $\rho^* = 0{,}06$. Оно также значимо больше, чем $\varepsilon = 0{,}04$, так что сгенерированы именно эволюционирующие нестационарные траектории. Этой же величине ρ^* оказалось равным и расстояние между сгенерированной $\tilde{f}_T(\{y\};x,t_0+T)$ и фактичекой $f_T(x,t_0+T)$ выборками. Поэтому

можно считать, что эксперимент генерации нестационарного ряда в заданном классе распределений проведен корректно: сгенерированные траектории имеют согласованный уровень стационарности, значимо меньший, чем расстояние между встык-выборками, и в то же время больший, чем для стационарного процесса.

Таким образом, использование уравнения Фоккера-Планка позволило, во-первых, сгенерировать ряд с теми статистическими свойствами его выборочного распределения, которые для него характерны, и, во-вторых, спрогнозировать выборочное распределение в существенно точнее, чем наивный прогноз, в рамках которого распределение считается стационарным. Тем самым численный эксперимент показал, что кинетическая модель эволюции распределения применительно к рассматриваемым временным рядам действительно отвечает определенным тенденциям изменения ВПФР, а не просто дает некоторое случайное ее изменение. Подчеркнем, что модель эволюции ВПФР использует непараметрические методы сравнения выборочных распределений.

4.3. Формирование паттернов и распознавание фрагментов траекторий

Одной из основных задач в проблеме идентификации образов является создание библиотеки паттернов. Разработанный в диссертации алгоритм позволяет:

- формировать средневзвешенную функцию распределения по экспертно отобранным фрагментам случайной траектории;

- формировать средневзвешенную функцию распределения по алгоритмически отобранным фрагментам случайной траектории в соответствии с параметрами регрессионной аппроксимации в окне заданной произвольной длины;

- генерировать случайные траектории, обладающие статистическими свойствами паттернов;

- идентифицировать фрагменты новой траектории по степени их близости набору базисных паттернов в рамках байесовского метода распознавания в скользящем окне произвольно заданной длины.

Выяснилось, что для рассматриваемых примеров набор паттернов инвариантен к перенормировке фрагмента траектории на величину выборочного среднеквадратичного отклонения в окне, выбранном в качестве окна идентификации скользящей выборки. Тем самым оказалось возможным построить практически важные эталонные распределения, характеризующие типовые состояния той гипотетической динамической системы, которая генерирует наблюдаемый нестационарный временной ряд.

Опишем алгоритм формирования паттернов на примере временного ряда курса рубль/доллар по данным за 2015 г. (рис. 26).

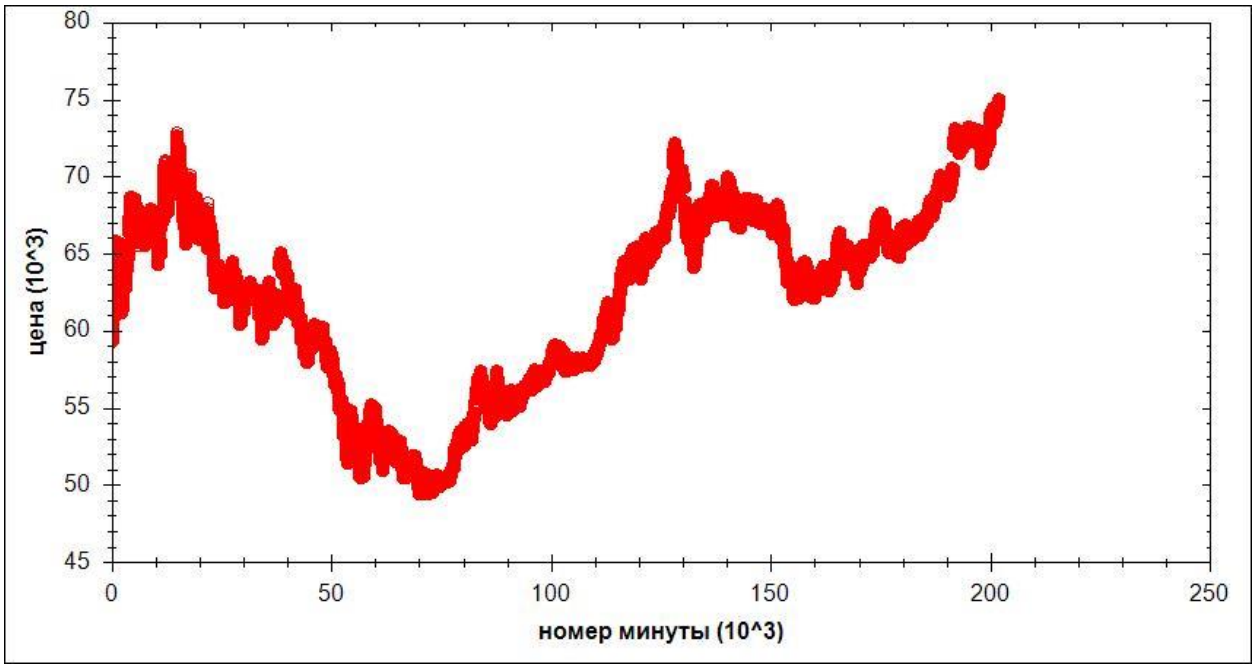

Рис. 26. График движения инструмента (курс рубля к тысяче долларов США) за 2015 г. с шагом 1 минута (цена закрытия)

Экспертно отобранные фрагменты показаны на рис. 27-28.

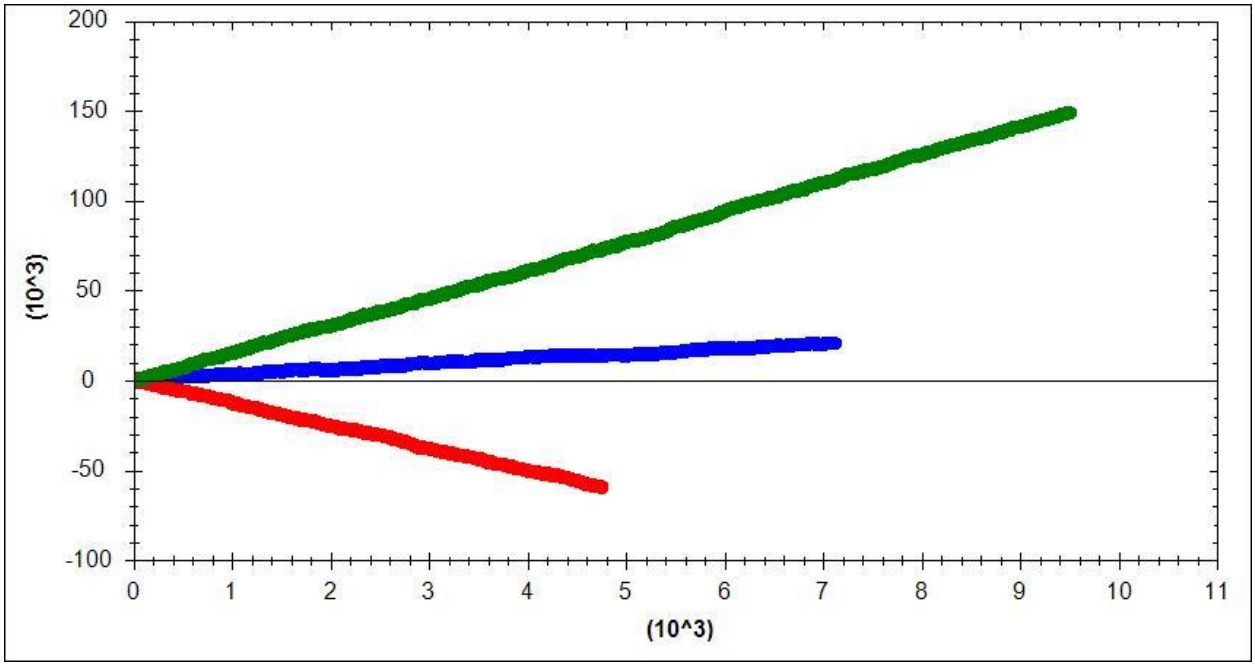

Рис. 27. Отобранные фрагменты траекторий для создания паттернов

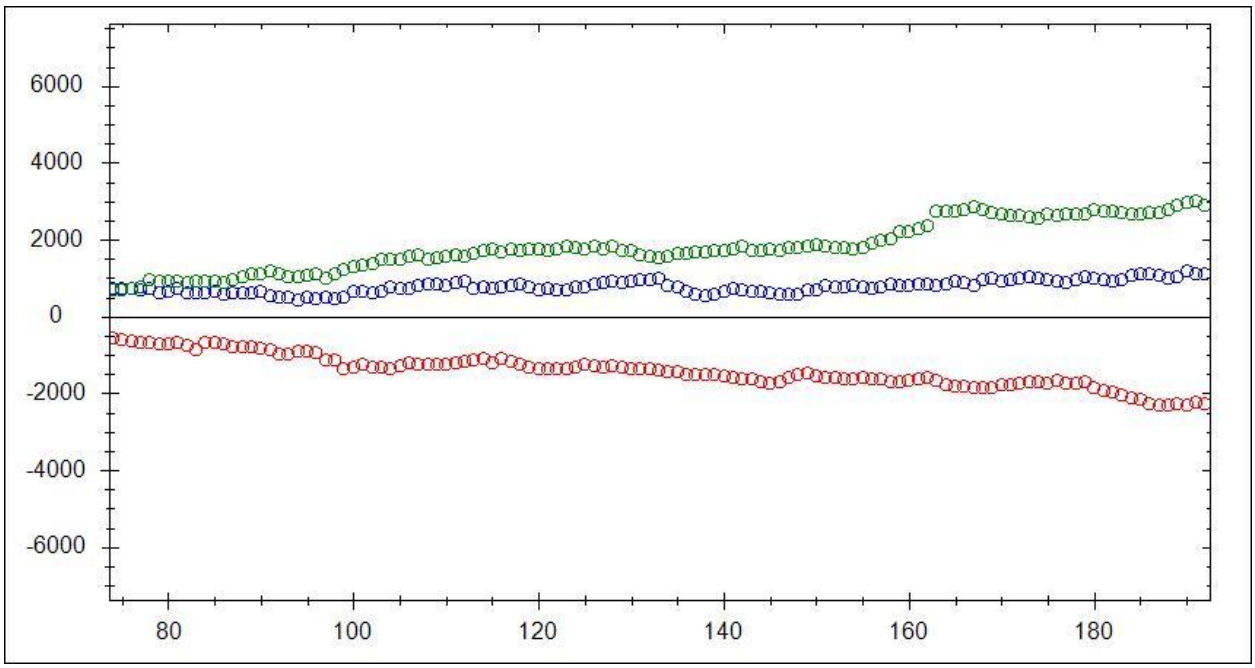

Рис. 28. Отобранные фрагменты траекторий для создания паттернов

(детальное разрешение)

Эти фрагменты отвечают трем типам состояния временного ряда, имеющим практическую важность для торговых систем: тренд вверх, тренд вниз и «боковик». Соответствующие средневзвешенные функции распределения приведены на рис. 29 для некоторого размаха суточной выборки.

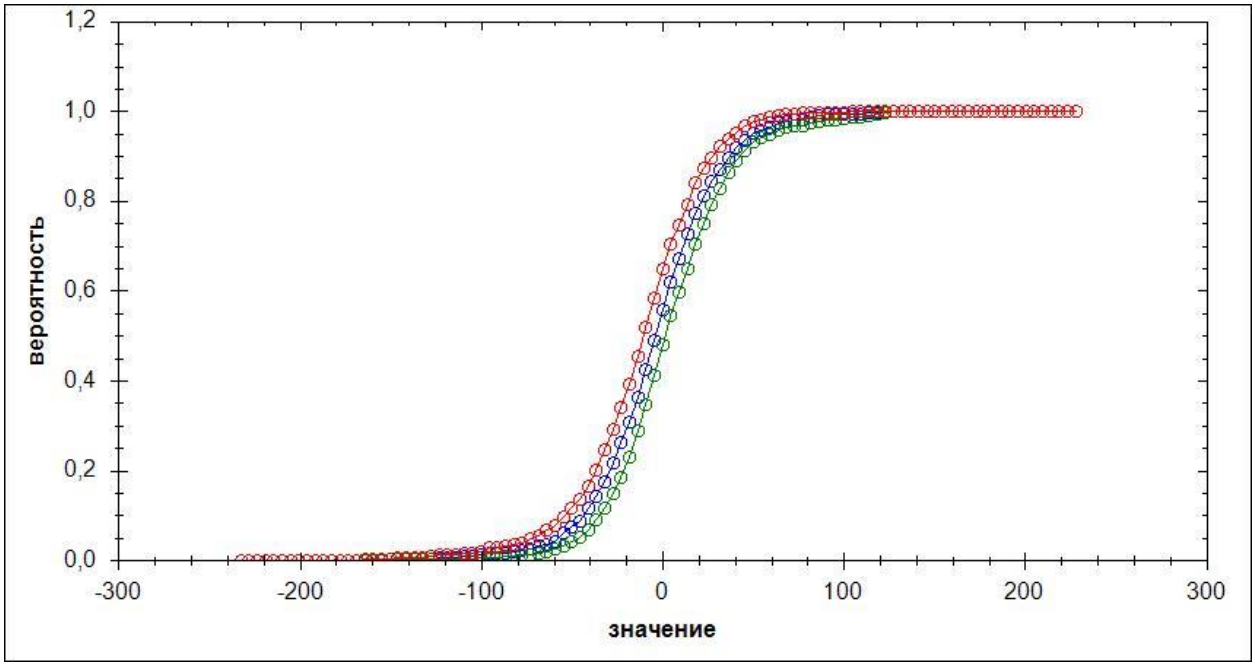

Рис. 29. Пример паттернов для заданного размаха суточной выборки

Паттерн, у которого левая часть расположена выше всех, отвечает тренду вниз, когда отрицательные приросты цены встречаются чаще, чем положительные. Нижний в этом смысле паттерн отвечает тренду вверх, а промежуточный паттерн – переменчивому боковому движению с равновероятным движением цены вверх или вниз. Далее по методике, описанной в главе III, идентифицируются фрагменты реальной траектории. Примеры идентификации приведены на рис. 30-32.

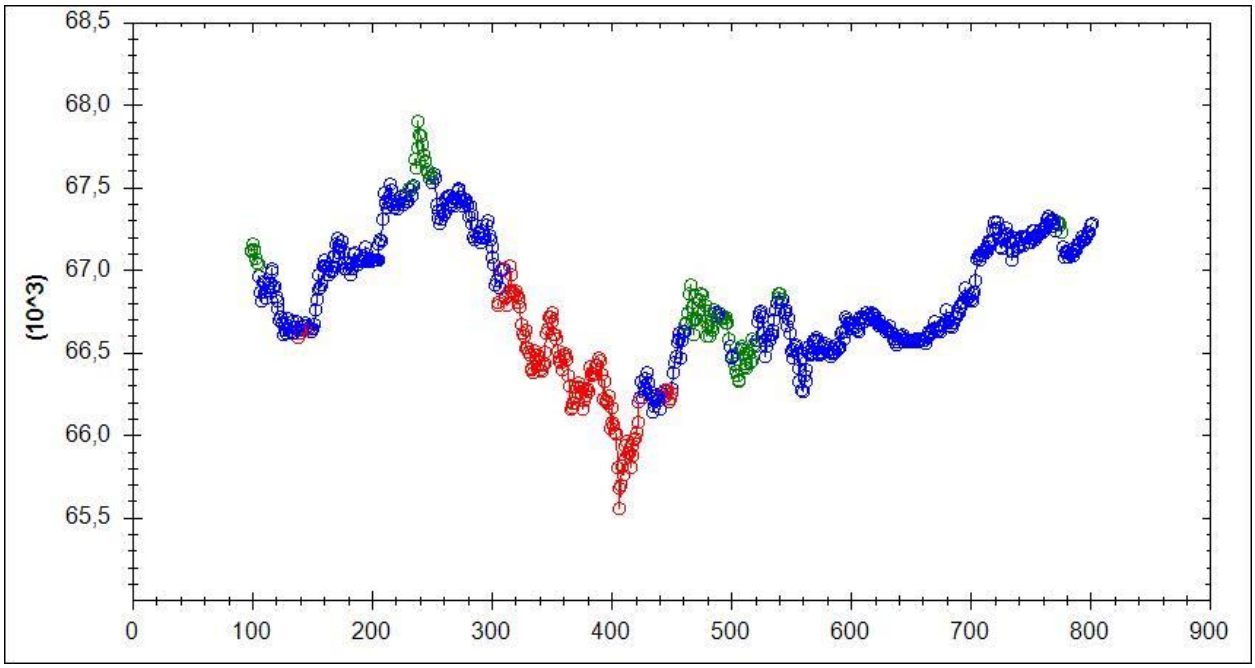

Рис. 30. Распознавание типа траектории в скользящем окне 100 мин, пример 1

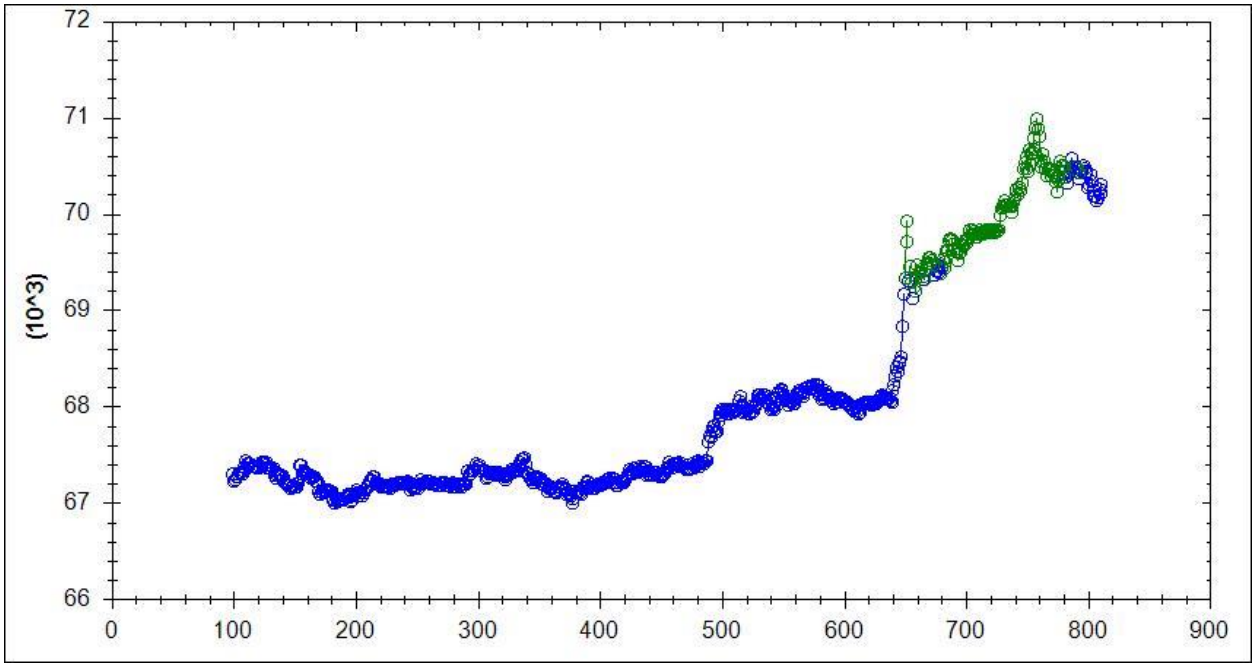

Рис. 31. Распознавание типа траектории в скользящем окне 100 мин, пример 2

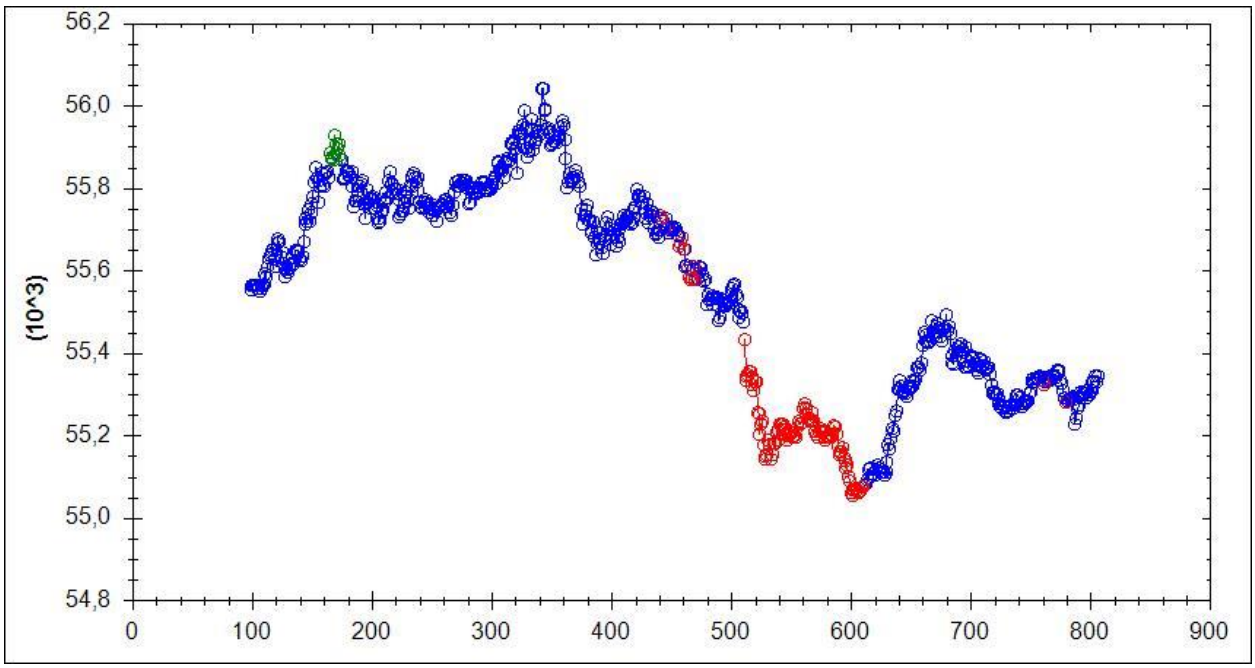

Рис. 32. Распознавание типа траектории в скользящем окне 100 мин, пример 3

На рис. 30-32 синим цветом показана идентификация «боковика», зеленый цвет отвечает тренду вверх, а красный – тренду вниз. Видно, что существуют очевидные ошибки идентификации типа «пропуска цели» (т.е. ошибки второго рода), обусловленные заданным окном сканирования ряда, когда явное движение вверх или вниз – правда, внутри окна меньшей длины, чем выбранное окно сканирования – идентифицируется как «боковик». Ошибки же первого рода (ложная тревога) не встречаются. Эта ситуация типична для выбранного способа распознавания. Уменьшение длины окна сканирования приводит к появлению ошибок первого рода, а увеличение – к росту ошибок второго рода. Возможности оптимизации длины выборки с целью уменьшения совокупной ошибки идентификации ограничены, поскольку требуется, чтобы число сигналов было достаточно велико на заданном конечном промежутке времени.

4.4. Пример статистического анализа функционала доходности торговой системы

Рассмотрим теперь практически важную задачу, когда на траектории случайного процесса задан некоторый функционал, управляющий другим процессом – уже не случайным. Пусть траектория – это ряд движения цены на рис. 26, а процесс, который требует управления – это работа торговой системы, реализующей некоторую стратегию купли-продажи финансового инструмента. Предположим, что разработана трендовая стратегия покупки по текущей цене, если идентифицировано движение цены вверх, и

продажи, если цена предположительно движется вниз. Функционал управления – это собственно команды, которые генерирует торговая система, исходя из анализа движения цены в скользящем окне определенной длины. Если провести оптимизацию этой длины скользящей выборки по имеющейся исторической траектории, то на участке оптимизации будет наблюдаться картина устойчивого роста накопленной доходности (рис. 33).

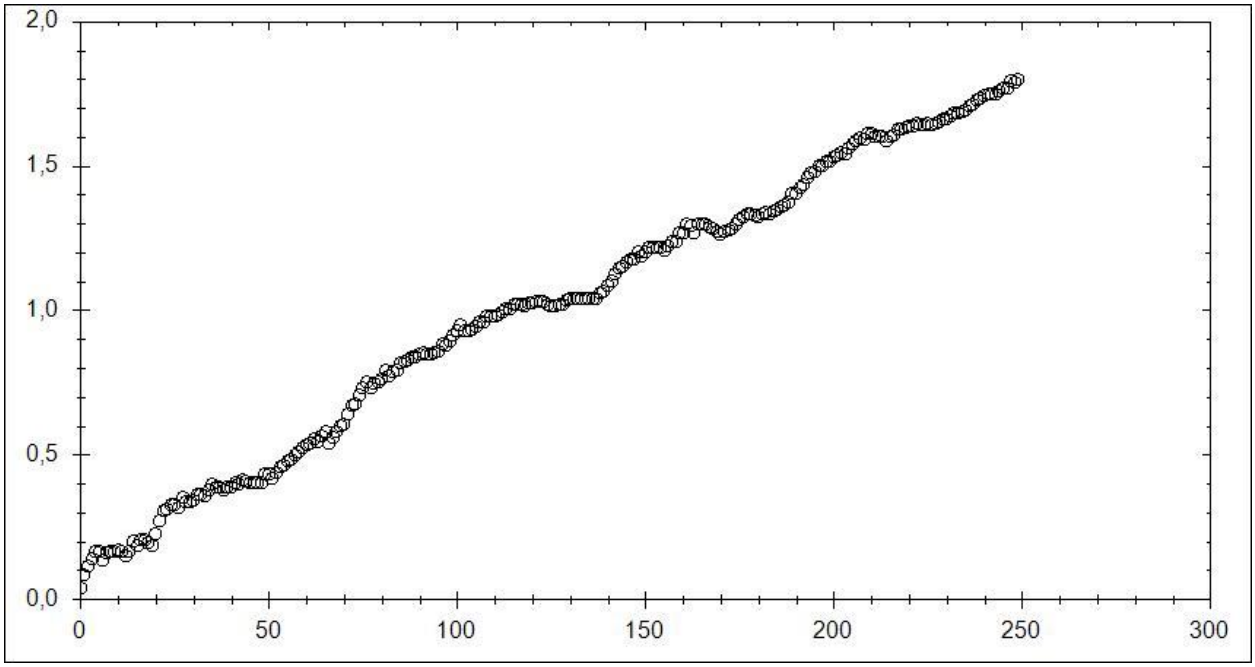

Рис. 33. График кумулятивной доходности торговой системы по реальной траектории на участке оптимизации параметров

Основная проблема состоит в том, что если использовать эту торговую систему со «старыми» оптимальными параметрами на новом участке траектории, т.е. вне области оптимизации, то показатели доходности обычно бывают существенно ниже, что связано с нестационарностью исходного процесса. Дело в том, что управляющий функционал зависит как от последовательности значений случайной величины на выборочной траектории, так и от вида распределения, поэтому интерес представляет не поиск оптимального управления на одной (пусть и достаточно длинной) исторической траектории, а оптимизация на ансамбле эволюционирующих траекторий. Это позволит более правильно найти оптимальное окно сканирования в области тестирования и снизит ошибку работы системы на новом участке.

Построив для ВПФР приростов цены эмпирическое уравнение Фоккера-Планка, как описано в главе III, получаем на участке тестирования ансамбль траекторий, аналогичный пучку на рис. 25. На каждой траектории пучка при фиксированном значении тестируемого

параметра строится функционал управления торговой системой, так что получается набор траекторий кумулятивных доходностей (рис. 34).

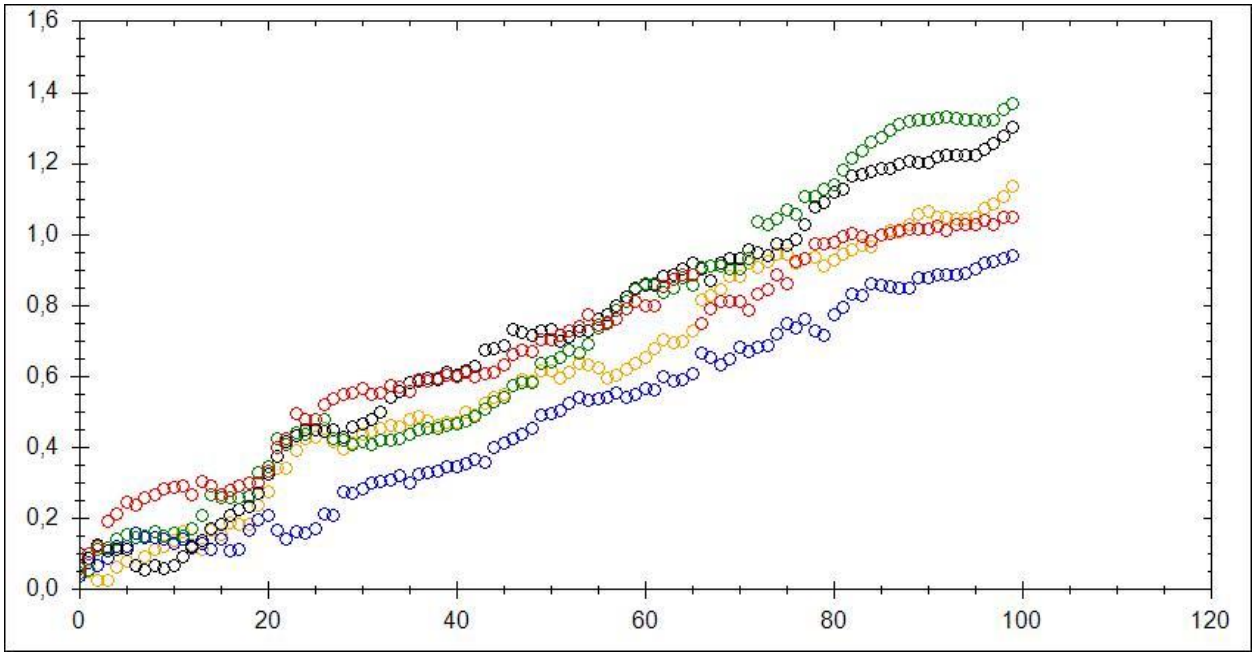

Рис. 34. Графики кумулятивной доходности торговой системы

по виртуальным траекториям

Для каждой выборки длины T строится функционал управления $\Psi\{x(t-T+1),..., x(t)\}$. При тестировании функционала Ψ требуется определить, во-первых, его статистические свойства на выборках, отвечающих данной модели эволюции ВПФР, и, во-вторых, изучить устойчивость функционала при изменении параметров уравнения эволюции или при разладке динамики ВПФР.

Метод решения этих задач был описан выше в разделе 2.5. Применительно к рассматриваемому примеру он состоит в следующем. Пусть выбран интересующий нас фрагмент временного ряда и на нем построен пучок виртуальных траекторий числом *N*. Обозначим Ψ_j значение функционала на j-ой траектории. Определяем средние по ансамблю величины: первый момент дисперсию, нормированное среднее:

$$\overline{\Psi} = \frac{1}{N}\sum_{j=1}^{N}\Psi_j, \quad \sigma_\Psi^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\Psi_j - \overline{\Psi}\right)^2, \quad S_\Psi = \frac{\overline{\Psi}}{\sigma_\Psi}, \tag{4.4.1}$$

Оптимальным следует считать то значение тестируемого параметра, при котором нормированное среднее максимально: $S_\Psi \to \max$.

Формула (4.4.1) дает корректный ответ на вопрос, какова, например, средняя доходность торговой системы на определенном промежутке времени. На практике может не быть достаточного количества данных, чтобы доходность, построенная по единственной реализации, могла быть достаточно полно протестирована на независимых встык-выборках. Управляющий функционал следует оптимизировать не на одной фактической длинной траектории, уходящей в прошлое, которое в силу нестационарности процесса потеряло актуальность в настоящем, а на пучке относительно небольших выборок, которые отвечают текущим свойствам ряда.

Вторая задача устойчивости оптимального значения параметра решается посредством вариации параметров уравнения Фоккера-Планка, в результате которой тренд $u(x,t)$ и диффузия $\lambda(t)$ меняются определенным образом. Вычисляя статистику (4.4.1) функционала управления на новых траекториях, можно определить допустимые пределы, внутри которых управление устойчиво. Чувствительность функционала определяется как его логарифмическая производная по параметру, например:

$$\Lambda_{\Psi} = \frac{\partial \ln \overline{\Psi}}{\partial \ln \lambda}. \qquad (4.4.2)$$

Задавая допустимые границы вариации (4.4.2), можно в численном эксперименте получить допустимые границы вариации параметров уравнения Фоккера-Планка, т.е. выяснить, предположим, при какой предельной волатильности торговая стратегия на бирже имеет положительное математическое ожидание (4.4.1).

В результате построен инструмент, позволяющий тестировать функционалы, заданные на случайной траектории, не по единственной ее реализации, а по набору траекторий, имеющих близкие статистические свойства.

4.5. Пример распознавания языка фрагмента текста

Приведем пример статистической идентификации текущего фрагмента выборки применительно к другой области – к математической лингвистике. В работе [4] возникла задача идентификации выборочных распределений применительно к вопросу о том, на каком языке могла быть написана рукопись, известная как Манускрипт Войнича. Были высказаны определенные предположения, которые позволили отобрать паттерны типовых частотно-упорядоченных распределений букв, используемых в текстах на разных языках.

Распределения упорядоченных частот букв в литературных текстах на одном и том же языке отличаются в норме L1 в пределах 0,08-0,13 безотносительно к тому, какой именно это язык. При этом 90%-ый доверительный интервал составляет $[0{,}085;\ 0{,}115]$. На рис. 35

приведены эталонные распределения, построенные по текстам на некоторых европейских языках без учета гласных букв.

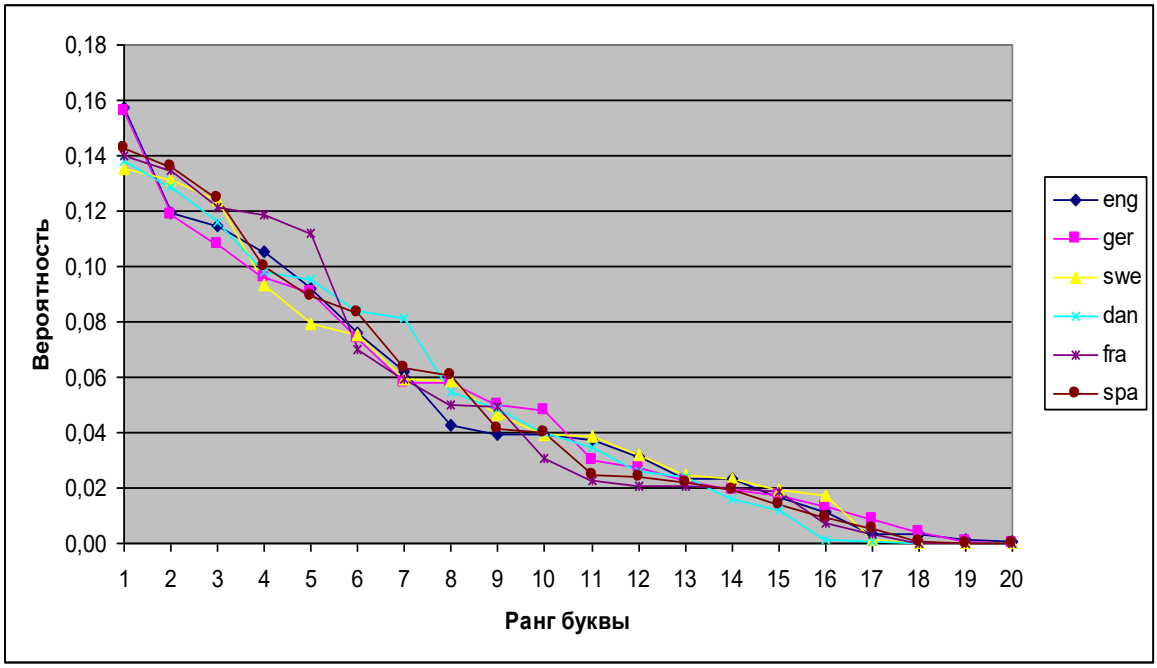

Рис. 35. Упорядоченные частоты букв в текстах на европейских языках без огласовки

Разработанный в диссертации инструмент сравнения и идентификации классовой принадлежности текущей выборки был применен к существующей транскрипции Манускрипта в латинский алфавит. Оказалось, что весь текст как таковой наиболее близок к смеси латыни и датского языка. Расстояние до эталона смеси составило 0,09, что является минимальным среди всех возможных сочетаний пар современных европейских языков, включая латынь.

Если разделить рукопись на четыре части приблизительно по 45 тыс. знаков каждая, то первые две части оказались ближе всего к эталону датского языка, расстояние до которого составило 0,08, третья часть близка латыни с расстоянием 0,10, а четвертая часть опять-таки близка смести латыни и датского языка с расстоянием 0,07.

Последующая детализация фрагментов текста позволяет выяснить более точно, на каком преимущественно языке написан тот или иной фрагмент. Рассмотрим фрагменты длиной 1000 знаков (примерно 2,5 страницы рукописи). Эталон, расстояние до которого оказывается наименьшим, указывает на наиболее вероятный язык данного фрагмента текста. Соответствующая языковая «раскраска» MB дана на рис. 36.

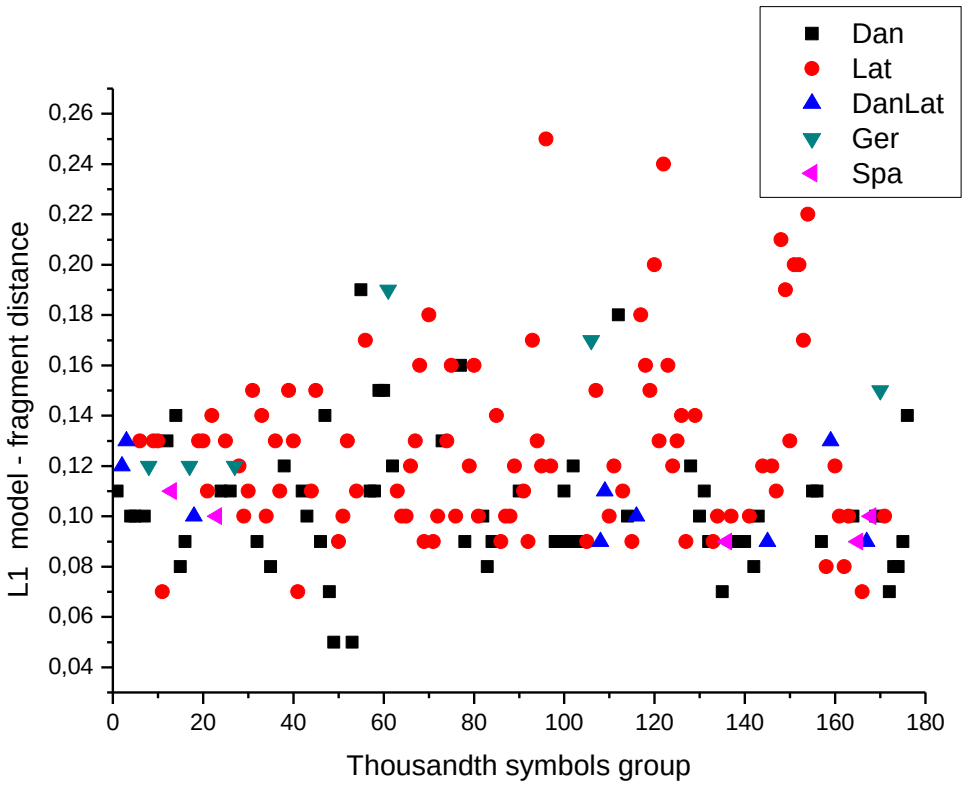

Рис. 36. Ближайшие распределения к фрагментам длиной 1000 символов

Однако следует подчеркнуть, что около 15 % фрагментов идентифицированы на низком уровне доверия, поскольку расстояния до ближайших эталонов оказались весьма большими (более 0,15). Это может быть обусловлено тем, что эталонные распределения были построены по современным текстам, но также и тем, что в библиотеке эталонов не нашлось правильного. Тем не менее, большая часть текста идентифицирована на высоком уровне доверия.

Аналогичные задачи могут быть поставлены и решены и в других областях, где требуется статистическое распознавание выборки по близости ее к эталонному распределению.

4.6. Анализ уровня стационарности сейсмограмм

При анализе распределения частоты землетрясений, происходящих в определенном регионе, в зависимости от магнитуды, актуальным является вопрос о том, насколько стационарны данные наблюдений. Практическая необходимость такого знания состоит, например, в том, что если распределение магнитуд стационарно, то с заданным уровнем значимости можно определить сейсмическую опасность и установить СНИП при строительстве тех или иных сооружений. Если же ряд магнитуд нестационарный, то, во-

первых, следует определить, на каком фактически уровне значимости следует принимать соответствующие решения, а также на каком горизонте прогнозирования будут справедливы прошлые статистики.

На рис. 37-38 представлены данные Мирового каталога землетрясений по Японии, начиная с 1900 г. Это магнитуды событий не менее 4 баллов и промежутки времени (в натуральных логарифмах секунд) между последовательными событиями.

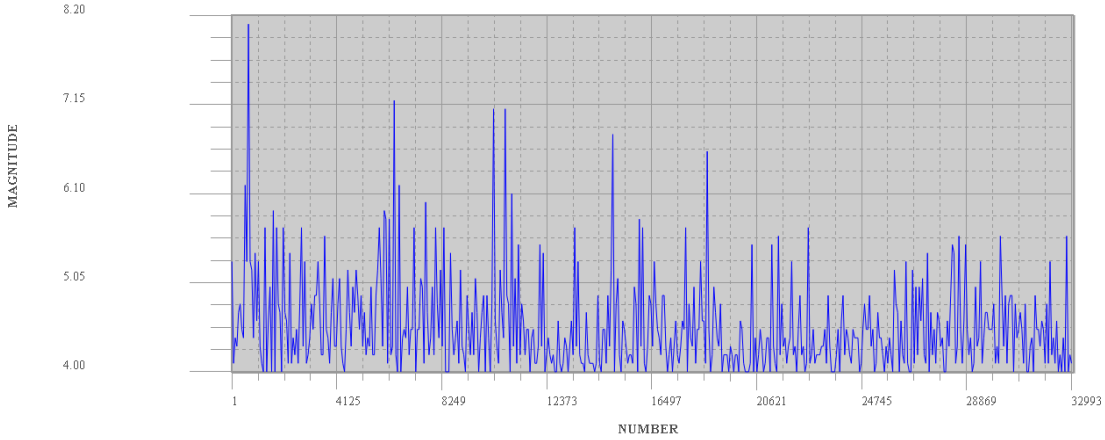

Рис. 37. Временной ряд магнитуд землетрясений

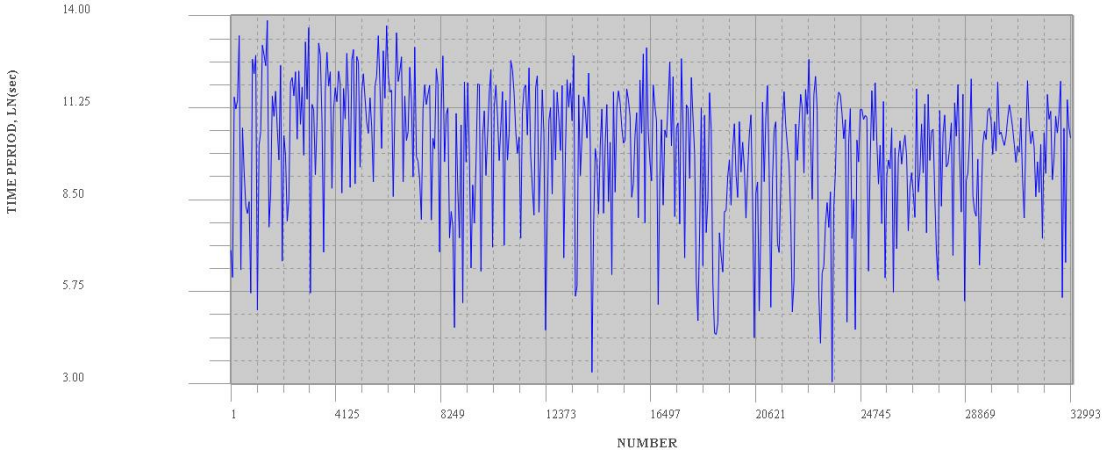

Рис. 38. Временной ряд промежутков времени между землетрясениями

На рис. 39 показано эмпирическое распределение промежутков времени между всеми событиями, а на рис. 40 – эмпирическое распределение событий по магнитуде с указанием точности аппроксимации предположительно стационарного распределения.

Рис. 39. Распределение вероятностей промежутков времени между землетрясениями

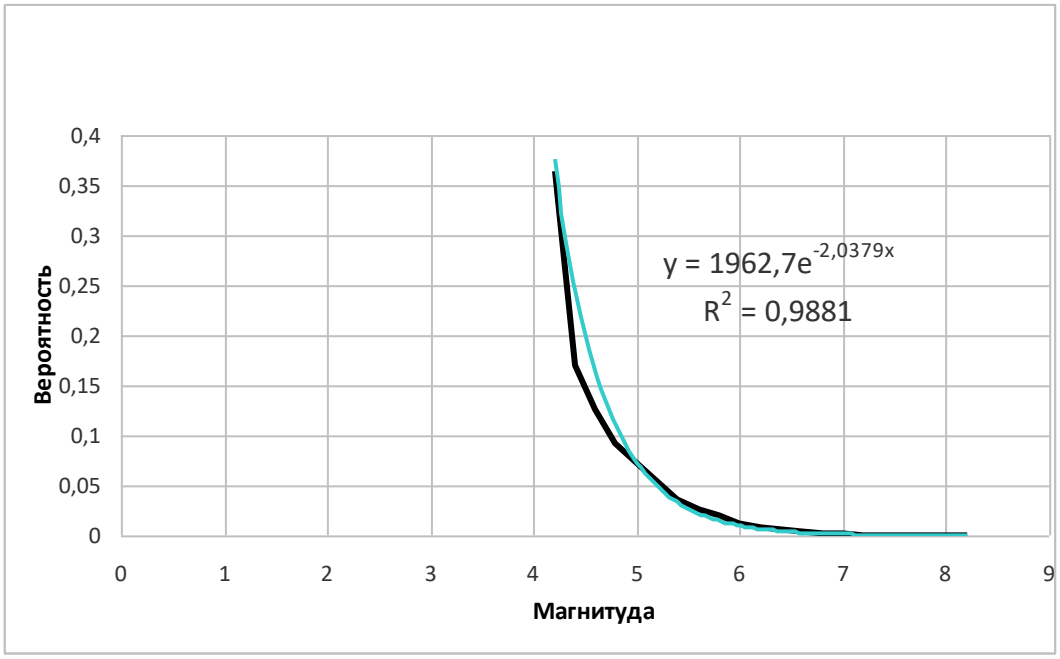

Рис. 40. Распределение вероятностей магнитуд землетрясений

Поскольку экспоненциальная зависимость считается теоретически верной, в указанном Каталоге можно обнаружить утверждение о том, что вероятности магнитуд определены с точностью не хуже 0,025. Однако, построив зависимость СУС выборочных распределений указанных случайных величин в зависимости от длины выборки, обнаруживаем (рис. 41), что наилучшая точность по магнитуде равна 0,10 на выборках примерно 800-1200 данных (2,5-3,5 года), а по промежуткам времени точность составляет 0,15 на выборках 5-7 тыс. данных (16-24 года). На рис. 41 показан также согласованный уровень значимости для стационарных распределений.

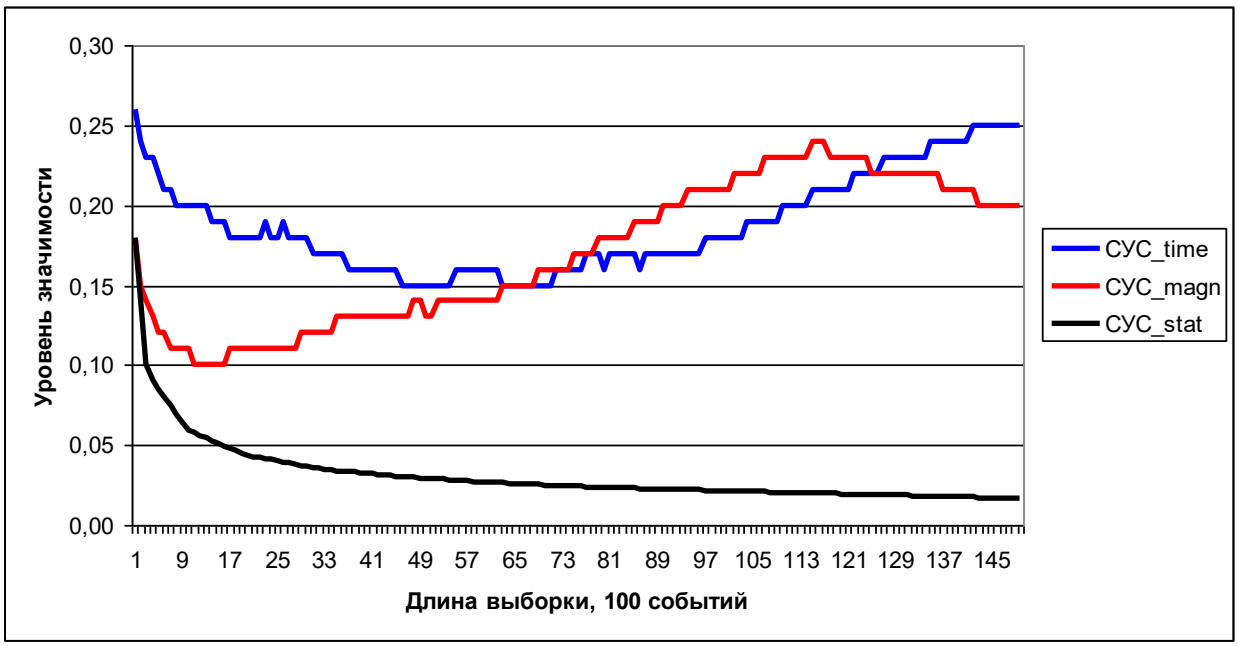

Рис. 41. СУС распределений в зависимости от длины выборки

Результаты рис. 41 показывают, что анализируемые ряды являются сильно нестационарными, по магнитуде наилучшая статистика собирается 2,5 года и существует примерно год, после чего должна быть пересчитана. Если исходить из более реальных требований частоты пересчета, связанных со строительством зданий (например, один раз за 25 лет), то наилучшим уровнем значимости будет 0,15. Эта величина существенно отличается от декларируемой точности 0,025 Каталога.

Заключение

В настоящей диссертации разработан новый метод непараметрического моделирования нестационарных временных рядов и построен программный комплекс, реализующий соответствующие численные алгоритмы.

Моделирование ряда основано на решении эмпирического уравнения Фоккера-Планка, что позволяет задавать скользящее среднее, дисперсию, размах и аналогичные величины изменяющимися во времени в соответствии с требованиями статистического эксперимента, причем выборочные распределения остаются в классе распределений, характерных для данного ряда.

Разработанный в диссертации метод позволяет тестировать индикаторы-предикторы изменения какого-либо свойства временного ряда и функционалы распознавания состояний ряда в широком диапазоне изменения его выборочных статистик. На практике для такого тестирования обычно используются исторические данные за достаточно большой промежуток времени, но надо помнить, что при этом берется единственная реализация случайного процесса. Однако подчеркнем, что в будущем реализуется произвольная траектория из ансамбля возможных путей, а в прошлом есть только одна.

К достоинствам метода следует отнести и то, что он позволяет провести стресс-тест на работоспособность индикатора в пределах, контролируемых исследователем. Исторический же ряд данных не предоставляет таких возможностей. Кроме того, для квалифицированного тестирования ряд прошлых данных требует предварительного выявления интересных ситуаций, кластеризации их, определения ошибок при кластеризации, что весьма трудоемко и не дает полного представления об имеющихся локальных паттернах ряда. Таким образом, численный код, генерирующий по фрагменту траектории нестационарного ряда ансамбль его нестационарных же реализаций, представляет практическую важность.

Теоретический аспект исследования состоит в использовании согласованного уровня стационарности ряда, что позволило формализовать точность оценки нестационарной ВФР и выбрать адекватный уровень значимости критерия близости распределений. Если задать уровень значимости априори, то может оказаться, что ряд на выборке данного объема не способен в среднем удовлетворить критерию, что приведет к ошибочному отклонению верной гипотезы.

С помощью построенного метода оказалось возможным провести оптимизацию решающих правил, распознающих локальные состояния (паттерны) временных рядов. Основные результаты исследования состоят в следующем.

1. Разработана математическая модель нестационарного временного ряда на основе численного решения эмпирического кинетического уравнения, которым описывается эволюция его выборочной функции распределения, и построена система индикаторов для идентификации уровня нестационарности в задачах статистического анализа нестационарных временных рядов.
2. Построен численный алгоритм генерации ансамбля траекторий нестационарного временного ряда и выборок из него в пределах заданного горизонта прогнозирования на основе решения уравнения Фоккера – Планка относительно выборочной плотности функции распределения, отвечающей данному временному ряду.
3. Разработан метод статистического анализа функционалов, заданных на траектории нестационарного случайного процесса, реализованный в виде программного комплекса с интерфейсом.

Список литературы

1. Айвазян С.А. Программное обеспечение персональных ЭВМ по статистическому анализу данных: проблемы, тенденции, перспективы отечественных разработок. // Заводская лаборатория. Диагностика материалов. 1991, т. 57, № 1, с. 54-58.

2. Айвазян С.А., Степанов В.С. Инструменты статистического анализа данных. // Мир ПК, 1997, № 8, с. 32-41.

3. Anderson T.W. An introduction to Multivariate Statistical Analysis. – Wiley-Interscience, 2003. – 752 p.

4. Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А. Статистические закономерности европейских языков и анализ рукописи Войнича // Препринты ИПМ им. М.В. Келдыша. 2016. № 52. 36 с.

5. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. (пер. с англ.) – М.: Мир, 1982. – 488 с.

6. Бартлетт С. Введение в теорию случайных процессов. (пер. с англ.) – М.: ИЛ, 1958. – 384 с.

7. Боголюбов Н.Н., Крылов Н.М. Общая теория меры в нелинейной механике. – Н.Н. Боголюбов, Избранные труды, т.1. – Киев: «Наукова Думка», 1969. – с. 411-464.

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

9. Боровков А.А. Математическая статистика. – М.: Физматлит, 2007. – 704 с.

10. Босов А.Д., Орлов Ю.Н. Моделирование нестационарных временных рядов с помощью эмпирического уравнения Лиувилля и уравнений эволюции моментов // Препринты ИПМ им. М.В. Келдыша. 2011. № 52. 28 с.

11. Босов А.Д., Орлов Ю.Н. Кинетико-гидродинамический подход к прогнозированию нестационарных временных рядов на основе уравнения Фоккера-Планка // Труды МФТИ, 2012. Т. 3. № 4. С. 134-140.

12. 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.

13. Босов А.Д., Кальметьев Р.Ш., Орлов Ю.Н. Моделирование нестационарного временного ряда с заданными свойствами выборочного распределения // Математическое моделирование, 2014. № 3. С. 97-107.

14. Бэстенс Д.Э., ван дер Берт В.М., Вуд Д. Нейронные сети и финансовые рынки: принятие решений в торговых операциях. - М.: ТВП, 1998.

15. Wass J.A. How Statistical Software Can Be Assessed. // Scientific Computing & Automation, 1996, (October), p. 14–24.

16. Векслер Л.С. Статистический анализ на персональном компьютере. // Мир ПК, 1992, № 2, с. 89-97.

17. Виленкин С.Я. Статистические методы исследования систем автоматического регулирования. – М.: Советское радио, 1967.

18. Власюк А.А., Орлов Ю.Н. Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки // Препринты ИПМ им. М.В. Келдыша. 2015. № 17. 25 с.

19. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.

20. Главные компоненты временных рядов. Сб. статей / Ред. Д.Л. Данилов и А.А. Жиглявский. СПбГУ, 1997.

21. Дубровин Б.А., Новиков С.П., Фоменко А.Т. Современная геометрия. – М.: Наука, 1986. – 759 с.

22. Ермаков С.М. Метод Монте-Карло и смежные вопросы. – М.: Наука, 1975. – 471 с.

23. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. – М.: Наука, 1982. – 296 с.

24. Заславский Г.М. Стохастичность динамических систем. – М.: Наука, 1984. – 270 с.

25. Ивченко А.Ю., Орлов Ю.Н. Практические аспекты задачи распознавания образов // Препринты ИПМ им. М.В. Келдыша. 2016. № 17. 20 с.

26. Каган А.М., Линник Ю.В., Рао С.Р. Характеризационные задачи математической статистики. – М.: Наука, 1972.

27. Калиткин Н.Н. Численные методы. – М.: Наука. 1978. – 512 с.

28. Кендалл М., Стюарт А. Статистические выводы и связи. (пер. с англ.) – М.: Наука, 1973. – 900 с.

29. Кендалл М., Стюарт А. Многомерный статистический анализ и временные ряды. (пер. с англ.) – М.: Наука, 1976. – 736 с.

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

31. Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л. Метод базисных паттернов в анализе нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2016. № 7. 20 с.

32. Клочкова Л.В., Орлов Ю.Н., Федоров С.Л. Моделирование ансамбля нестационарных траекторий с помощью уравнения Фоккера-Планка // Журнал Средневолжского математического общества, 2016. – Т.18. № 1.

33. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.
34. Кожевников А.С. Программное обеспечение для статистического моделирования и анализа случайных процессов со скачками, описывающих динамику цен акций предприятий авиационной отрасли. // ЭЖ «Труды МАИ», 2012, № 59.
35. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
36. Кремер Н.Ш., Путко Б.А. Эконометрика. – М.: ЮНИТИ-ДАНА, 2002. – 311 с.
37. Кузнецов С.П. Динамический хаос. – М.: Физматлит, 2001. – 296 с.
38. Лемешко Б.Ю., Постовалов С.Н. О распределении статистик непараметрических критериев согласия при оценивании по выборкам параметров наблюдаемых законов. // Заводская лаборатория. Диагностика материалов. 1998, т. 64, № 3, с. 61-72.
39. Лемешко Б.Ю., Помадин С.С. Проверка гипотез о математических ожиданиях и дисперсиях в задачах метрологии и контроля качества при вероятностных законах, отличающихся от нормального. // Метрология, 2004, № 4, с. 3-15.
40. Лемешко Б.Ю., Чимитова Е.В. Построение оптимальных L-оценок параметров сдвига и масштаба распределений по выборочным квантилям. // Сибирский журнал индустриальной математики, 2001, т.4, № 2, с. 166-183.
41. Лившиц М.Е., Иванов-Муромский К.А., Заславский С.Я., Войтинский Е.Я., Лернер В.А., Ромм Б.И. Численные методы анализа случайных процессов.– М.: Наука, 1976. – 128 с.
42. Лоскутов А.Ю., Котляров О.Л., Истомин И.А., Журавлев Д.И. Проблемы нелинейной динамики. Локальные методы прогнозирования временных рядов // Вестник МГУ, Сер 3. Физика и Астрономия. 2002. № 6. С. 3-21.
43. Лукашин Ю.П. Адаптивные методы прогнозирования экономических показателей. М.: «Статистика», 1979.
44. Льюис К.Д. Методы прогнозирования экономических показателей. М.: Финансы и статистика, 1986.
45. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. Диагностика материалов. 1991, т.57, № 7, с.64-66.
46. Орлов А.И. О современных проблемах внедрения прикладной статистики и других статистических методов. // Заводская лаборатория. Диагностика материалов. 1992, т. 58, № 1, с. 67-74.
47. Орлов Ю.Н., Осминин К.П. Анализ нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2007. № 36. 24 с.

48. Орлов Ю.Н., Осминин К.П. Кинетические уравнения для прогнозирования нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2008. № 47. 28 с.

49. Орлов Ю.Н., Осминин К.П. Методика определения оптимального объема выборки для прогнозирования нестационарного временного ряда. // ИТВС, 2008, № 3, с. 3-13.

50. Орлов Ю.Н., Осминин К.П. Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Мат. Мод., 2008, № 9, с. 23-33.

51. Орлов Ю.Н., Суслин В.М. Кинетические уравнения для некоторых моделей демографии // Математическое моделирование, 2003. Т.15. №3. С.43-54.

52. Орлов Ю.Н., Федоров С.Л. Моделирование и статистический анализ функционалов, заданных на выборках из нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша. 2014. № 43. 26 с.

53. 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.

54. 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.

55. 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.

56. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. 2-е изд. М.: Инфра-М, 1997.

57. Уилкс С. Математическая статистика. (пер. с англ.) – М.: Наука, 1967. – 632 с.

58. Федоров С.Л. Анализ функционалов, заданных на выборках из нестационарного временного ряда. // Труды II Международной научно-практической конференции Теоретические и прикладные аспекты современной науки, Белгород, август 2014. С. 9-16.

59. Yu. Orlov, S. Fedorov, A. Samuylov, Yu. Gaidamaka, D. Molchanov. Simulation of Devices Mobility to Estimate Wireless Channel Quality Metrics in 5G Network // Proc. ICNAAM, September 19-25, 2016, Rhodes, Greece.

60. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа (пер. с нем.) – М.: Финансы и статистика, 1982.

61. Хайкин С. Нейронные сети. Полный курс (пер. с англ.) – Москва, С-Петербург, Киев, «Вильямс», 2006.

62. Цветков Э.И. Нестационарные случайные процессы и их анализ. – М.: «Энергия», 1973.

63. Шугай Ю.С. Нейросетевые алгоритмы прогнозирования событий и поиска предвестников в многомерных временных рядах. // Искусственный интеллект. Донецк, 2004, № 2, с. 211-215.

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

65. Классификация и снижение размерности: Справ. изд. – М.: Прикладная статистика, 1989. – 607 с.

66. Система ЭВРИСТА. Электронное издание Центра статистических исследований, 1997, № 114-97.1.0.RUS Серия Б.