УТВЕРЖДАЮ:

Зам. председателя КарНЦ РАН

_____ Н.В. Ильмаст

«28» апреля 2021 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию Горшенина Андрея Константиновича
«Полупараметрические методы анализа неоднородных данных и их применение в задачах математического моделирования»,
представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Актуальность темы диссертационной работы

При обработке массивов реальных данных зачастую возникают трудности, связанные с отсутствием точного закона или системы уравнений, описывающих процесс, в рамках которого эти наблюдения были получены. Также возможна ситуация, при которой известен общий вид модели, но не ее параметры или их оценки. В данных обстоятельствах ключевым становится разработка подходов, позволяющих получить значимую информацию о системе или процессе в условиях минимальных априорных предположений. Зачастую случайный характер факторов, влияющих на процесс формирования или получения данных, приводит к необходимости развития вероятностных моделей и статистических подходов к оцениванию их параметров.

Диссертация А.К. Горшенина ориентирована на развитие методов исследования и анализа неоднородных данных с точки зрения вероятностно-статистических моделей и методов. Особый интерес представляет использование полупараметрического подхода. В его рамках одна составляющая вероятностной модели, а именно семейство распределений, определяется на основе строго обоснованных предельных теорем, гарантирующих, что при выполнении достаточно естественных для практических приложений условий, может возникать только конкретный класс распределений. Фактически, речь идет о параметрическом статистическом оценивании. В тоже время, параметры у этого семейства также являются

случайными величинами – и для оценивания параметров их распределений требуются непараметрические процедуры.

Кроме того, модели и методы, развиваемые в диссертационной работе, ориентированы в том числе и на использование распределений с произвольно тяжелыми хвостами. Именно такие законы чаще всего и наблюдаются в реальных данных вместо предписываемых классической теорией семейств распределений. Таким образом, предлагаемые соискателем подходы ориентированы на более адекватное вероятностно-статистическое описание процессов, наблюдаемых в прикладных областях.

Структура и основное содержание диссертации

Диссертация состоит из введения, 7 глав, заключения, обширной библиографии из 458 источников, включая 121 работу соискателя, 28 таблиц, 175 рисунков и 30 оформленных в виде псевдокода вычислительных алгоритмов. Общий объем диссертации составляет 355 страниц.

Во *введении* отражены актуальность тематической области, ключевые направления исследований, основные полученные и выносимые на защиту результаты, их новизна, теоретическая и практическая значимость.

Первая глава посвящена обоснованию вида смешанных распределений, возникающих при рассмотрении сумм и максимальных элементов выборок, объем которых является случайной величиной с обобщенным или классическим отрицательным биномиальным распределениями. В частности, доказаны новые версии закона больших чисел и центральной предельной теоремы для случайных сумм. Данные теоретические результаты используются соискателем в дальнейших главах в качестве базиса для построения математических моделей (в терминах вероятностных распределений) для размеров частиц лунного реголита, экспериментальных рядов турбулентной плазмы, объемов и интенсивностей осадков, потоков тепла «океан-атмосфера».

Во *второй главе* рассмотрены аналитические свойства моделей на основе нормальных и гамма-распределений. В частности, продемонстрирована устойчивость в метрике Леви для конечных масштабных и сдвиговых смесей нормальных законов относительно смешивающего распределения, а также доказана теорема об устойчивости дисперсионно-сдвиговых смесей нормальных законов. В модели случайного округления данных при условии, что распределение шума имеет распределение типа конечных нормальных или гамма-смесей, получены оценки для неизвестного математического ожидания наблюдений в предположении зашумления конечными смесями нормальных и гамма-распределений, а также построены доверительные интервалы для него. Также предложено использование метода скользящего разделения смесей

(СРС) для оценивания распределений случайных коэффициентов в стохастических дифференциальных уравнениях Ланжевена.

В *третьей главе* исследуются методы анализа данных на основе скользящего разделения смесей, в частности: получены выражения для моментов для каждого положения окна, разработаны алгоритм автоматического определения числа компонент в процессе шагов СРС-анализа и метод оценивания параметров распределения сигнала при наличии конечного смешанного нормального шума.

В *четвертой главе* на основе нового варианта центральной предельной теоремы из первой главы построена вероятностная модель распределений размеров частиц лунного реголита. Для оценивания параметров с учетом специфики формы представления анализируемых реальных данных, собранных миссиями «Аполлон-11, 12, 14–17» и «Луна 24», разработаны эффективные процедуры на основе имитационного моделирования (бутстрепа) и минимизации статистики хи-квадрат. Продемонстрировано, что зависимость математического ожидания и среднеквадратического отклонения является нелинейной.

В *пятой главе* конечные смешанные нормальные распределения используются как вероятностные модели процессов в турбулентной плазме. С помощью процедуры, основанной на бутстрепе, реализована аппроксимация различных одно- и двусторонних спектров. Развит вероятностно-статистический подход к анализу эволюции характеристик микротурбулентности в переходном процессе электронно-циклотронного резонансного нагрева плазмы, существенно использующий моменты смешанных моделей. Показано, что их использование в качестве дополнительных наблюдений позволяет заметно повысить точность прогнозов с помощью нейронных сетей. Кроме того, продемонстрированы результаты прогнозирования и самих этих моментов.

Шестая глава посвящена применению моделей и методов, развитых в главах 1-3 для анализа экстремальных явлений в пространственно-временных метеорологических и океанологических рядах. При этом используются различные процедуры, сочетающие как статистические подходы, так и методы машинного обучения для заполнения пропусков и прогнозирования. В частности, продемонстрировано применение метода оценивания распределений коэффициентов стохастического дифференциального уравнения Ланжевена, описывающего турбулентные потоки между океаном и атмосферой.

Седьмая глава ориентирована на описание архитектур и интерфейсов прикладных программных комплексов, разработанных соискателем для анализа данных. Результаты их работы представлены в главах 3-6 для различных временных рядов.

В *заключении* кратко сформулированы основные итоги проведенных исследований и полученные в работе результаты

Степень обоснованности научных положений и выводов, сформулированных в диссертации

К *основным результатам диссертации* относятся:

1. Смешанные вероятностные модели для выборок со случайным объемом на основе: а) нового варианта центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых (теорема 1.10); б) схемы максимума для выборок, объем которых описывается важным для прикладных задач семейством обобщенных отрицательных биномиальных распределений (теоремы 1.1 и 1.2, а также исследование свойств распределений в теоремах 1.3-1.8); в) обобщения теоремы Реньи (закона больших чисел для случайных сумм) для математического моделирования редких событий (теорема 1.9).
2. Доказательства устойчивости в метрике Леви дисперсионно-сдвиговых (теорема 2.12) и конечных сдвиговых смесей нормальных распределений относительно возмущений параметров смешивающего распределения (теоремы 2.8-2.11), обосновывающие корректность полупараметрических вычислительных процедур разделения смесей этих семейств распределений.
3. Комплекс полупараметрических методов анализа неоднородных данных (главы 2 и 3) и результаты аналитического исследования некоторых их свойств в моделях аддитивного зашумления конечными смесями и округления наблюдений (теоремы 2.13-2.16).
4. Полупараметрический подход к статистическому оцениванию распределений случайных коэффициентов стохастических дифференциальных уравнений Ланжевена (разделы 2.1 и 6.5).
5. Статистическая методология построения моделей сгруппированных скрытых наблюдений при заданных характерных точках их эмпирической функции распределения (глава 4 на основе теоретических результатов главы 1).
6. Комплекс методов и алгоритмов статистической идентификации и классификации экстремальных наблюдений, на основе обобщенных отрицательных биномиальных распределений числа наблюдений и обобщенных гамма-моделей для данных (разделы 6.1-6.4 на основе теоретических результатов главы 1).
7. Программные комплексы для автоматизации обработки массивов неоднородных данных на высокопроизводительных вычислительных ресурсах, реализующие разработанные полупараметрические методы; решение с их

помощью некоторых задач математического моделирования в физике плазмы, селенологии, метеорологии, океанологии (главы 4-7).

Проведенные исследования описаны в диссертации ясно и подробно с соответствующей научной строгостью. Теоретические результаты подтверждаются строгими доказательствами с использованием современного математического аппарата. Для описания вычислительных алгоритмов используются формы представления в виде блок-схем и псевдокода. Приведены многочисленные наглядные иллюстрации применения разработанных соискателем методов для смоделированных и реальных данных. Полученные результаты являются новыми и оригинальными. Они представлялись автором в виде докладов на российских и международных конференциях и научных семинарах в 2012-2020 гг.

Основные результаты корректно отражены в виде научных публикаций: 82 печатные работы, включая 31 статью в журналах, рекомендованных ВАК, а также 51 работа в изданиях, индексируемых Web of Science Core Collection, Scopus. Зарегистрированы 39 программ для ЭВМ. В большинстве работ (75%) соискатель является первым или единственным автором, при этом в работах, выполненных в соавторстве, ключевая роль также принадлежит соискателю. Данное обстоятельство корректно отражено в тексте диссертации и автореферате.

Также отмечено, что основные результаты диссертации получены соискателем в рамках научных проектов, поддержанных грантами Президента России для молодых кандидатов наук, Российского научного фонда, Российского фонда фундаментальных исследований, Научного центра мирового уровня «Московский центр фундаментальной и прикладной математики» и стипендиями Президента России.

Теоретическая и практическая значимость полученных автором диссертации результатов для развития соответствующей отрасли науки

Результаты диссертации являются как фундаментальными, поскольку существенным образом теоретически развивают подходы к построению математических вероятностных моделей, так и прикладными – в силу эффективности продемонстрированных соискателем путей решения реальных задач на основе созданных подходов в широком спектре прикладных областей.

Доказанные в диссертации теоремы показывают, что распределения с произвольно тяжелыми хвостами возникают и для выборок, в которых наблюдения имеют конечные дисперсии, то есть подобные строго теоретически обоснованные модели возможно корректно использовать для анализа реальных массивов данных.

Полученные соискателем результаты в настоящий момент уже используются в ряде научно-исследовательских организаций – Институте общей физики им. А.М. Прохорова Российской академии наук, Институте океанологии им. П.П. Ширшова Российской академии наук и Центре компетенций Национальной технологической инициативы по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова. Прикладная значимость результатов работы не вызывает сомнений.

Рекомендации по использованию результатов

В качестве рекомендаций по дальнейшему использованию можно указать два направления. Во-первых, расширение исследовательских областей, в которых подобные модели и подходы к их построению и анализу могут быть эффективно применены: трафик в телекоммуникационных системах, системы мониторинга экологического загрязнения окружающей среды, медицинские приложения и др. Во-вторых, поскольку многие программные методы были апробированы на оборудовании центра коллективного пользования «Информатика» Федерального исследовательского центра «Информатика и управление» Российской академии наук, это позволяет в дальнейшем использовать их в качестве компонентов цифровых научных сервисов. Подобные ресурсы могут быть весьма полезными в рамках проведения совместных научных исследований междисциплинарными коллективами.

Замечания

1. В работе не приведены примеры применения результатов теоремы 1.8 (стр. 43) о соответствии асимптотического распределения выборочных квантилей распределению Стьюдента для каких-либо реальных данных.
2. В разделе 4.3 на стр. 149 упоминается статистическая проверка качества аппроксимации распределений, оценки параметров которых получены методом минимизации статистики хи-квадрат. Однако не указано, какие выборки используются для вычисления Р-значений, как при этом выбираются параметры соответствующего распределения хи-квадрат.
3. В разделе 5.3.2 на стр. 189 не указана разница в скорости обучения нейронных сетей прямого распространения и рекуррентных архитектур в абсолютных единицах.
4. В диссертации есть незначительное количество опечаток, например, на стр. 288 в подписи к рисунку 7.7 должно быть указано «динамическая компонентА».

Указанные замечания носят характер уточнений по изложению отдельных вопросов в диссертационном исследовании и не влияют на высокую оценку работы в целом.

Заключение

Результаты диссертации А.К. Горшенина являются значимым научным достижением в области математического моделирования на основе развития современных полупараметрических подходов теории вероятностей и математической статистики и создания программных комплексов анализа неоднородных данных. Они соответствуют следующим пяти пунктам паспорта специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»:

1. «Разработка новых математических методов моделирования объектов и явлений»;
2. «Развитие качественных и приближенных аналитических методов исследования математических моделей»;
3. «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»;
4. «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента»;
5. «Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента».

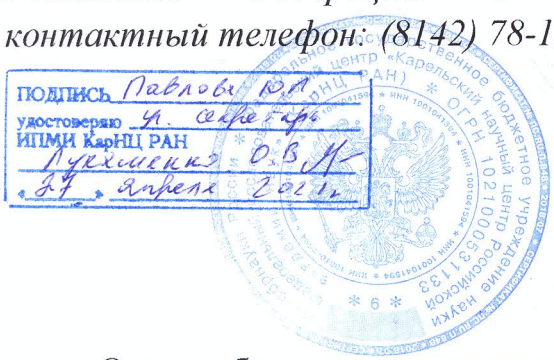
Автореферат корректно отражает основное содержание проведенных исследований и в полной мере соответствует тексту диссертации.

Представляются несомненными вклад соискателя в развитие теоретических подходов в области математического моделирования и существенный потенциал использования разработанных методов и подходов для исследования данных в широком спектре прикладных областей, не ограничивающихся рассмотренными соискателем.

Диссертационная работа А.К. Горшенина является научно-квалификационной работой, выполненной на высоком научном уровне, и полностью удовлетворяет требованиям к докторским диссертациям, установленным Положением о присуждении ученых степеней (утверждено Постановлением Правительства РФ от 24 сентября 2013 г. №842). Горшенин Андрей Константинович **заслуживает присуждения степени доктора физико-математических наук** по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ».

Отзыв подготовил:

Павлов Юрий Леонидович, главный научный сотрудник, руководитель лаборатории теории вероятностей и компьютерной статистики Института прикладных математических исследований КарНЦ РАН, доктор физико-математических наук по специальности 01.01.05 – «Теория вероятностей и математическая статистика», профессор, заслуженный деятель науки Российской Федерации. Электронная почта: pavlov@krc.karelia.ru, контактный телефон: (8142) 78-12-18.

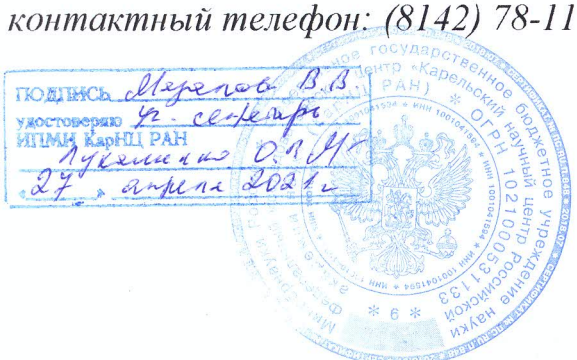


Ю.Л. Павлов

27.04.2021

Отзыв обсужден и единогласно утвержден на заседании Ученого совета Института прикладных математических исследований КарНЦ РАН: протокол №3 от 23.04.2021; присутствовало на заседании: 12 чел.; результаты голосования: за – 12, против – 0.

Мазалов Владимир Викторович, директор Института прикладных математических исследований КарНЦ РАН, председатель Ученого совета Института прикладных математических исследований КарНЦ РАН, доктор физико-математических наук, профессор, заслуженный деятель науки Российской Федерации. Электронная почта: vmazalov@krc.karelia.ru, контактный телефон: (8142) 78-11-08.



В.В. Мазалов

27.04.2021

Сведения о ведущей организации

Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук» (КарНЦ РАН). Адрес: 185910, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11 http://www.krc.karelia.ru/. Электронная почта: krcras@krc.karelia.ru, контактный телефон: (8142) 76-60-40, 76-97-10.