

Отзыв научного консультанта

о диссертации Горшенина Андрея Константиновича

«Полупараметрические методы анализа неоднородных данных и их применение в задачах математического моделирования», представленной к защите на соискание учёной степени доктора физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ (физико-математические науки).

Целью исследования А.К. Горшенина, результаты которого оформлены в виде указанной диссертации, является разработка эффективных подходов к построению математических моделей статистических закономерностей в больших объемах неоднородных данных и основанных на них методов анализа данных. Адекватная информация о таких закономерностях позволяет радикально улучшить методы статистического анализа и алгоритмы, применяемые в интеллектуальных информационных технологиях принятия решения. С учетом этих обстоятельств актуальность проблематики исследования А.К. Горшенина, нашедшего отражение в его докторской диссертации, не вызывает сомнения.

В качестве теоретического базиса разрабатываемых моделей автор использует оригинальные предельные теоремы теории вероятностей, в том числе специально доказанные новые теоремы, обобщающие классические результаты. В качестве итоговых математических моделей предлагаются возникающие в них предельные распределения вероятностей. Основным отличием рассматриваемых теорем от классических является то, что в классических постановках задач теории вероятностей асимптотика рассматривается при неограниченно возрастающем числе наблюдений (или "элементарных" случайных величин), который считается неслучайным параметром. Подход, реализованный А.К. Горшениным, подразумевает, что этот параметр является случайным. Такое предположение очень хорошо согласуется с практикой. Действительно, к примеру, в классических задачах математической статистики число доступных наблюдений всегда считается неслучайным, тогда как во многих реальных ситуациях (в телекоммуникации, экономике, страховании, медицине, биологии, и других областях) число наблюдений, которые будут доступны в итоге проведенных экспериментов, заранее не известно и само является наблюдением. Поэтому если элементы выборки трактуются как реализации случайных величин, то точно так же реализацией некоторой случайной величины следует считать и объем выборки. Если это обстоятельство принимать во внимание на этапе конструирования вероятностных моделей типа асимптотических аппроксимаций, то удастся очень хорошо объяснять возникновение во многих ситуациях так называемых "тяжелых хвостов", когда существенно выделяющихся наблюдений в выборке оказывается намного больше, нежели предписывается классическими моделями, например, нормальным распределением вероятностей.

В предельных теоремах для статистик (функций от наблюдаемых значений), построенных по выборкам случайного объема в качестве предельных законов выступают смеси вероятностных распределений, в которых ядро (смешиваемое распределение) определяется свойствами "элементарных" случайных величин, а смешивающее ("весовое") распределение определяется асимптотическим поведением случайного объема выборки. Выбор конкретного вида зависимости функции от наблюдений зависит от изучаемого явления: к примеру, если исследуется процесс типа случайного блуждания, то это сумма случайных величин; если же исследуются экстремальные характеристики процесса, то это максимальная или минимальная порядковая статистика. При этом, если о свойствах "элементарных" величин как правило заранее можно сделать более-менее реалистичные аналитические предположения, то поведение числа доступных наблюдений как правило определяется факторами, являющимися "внешними" по отношению к исследуемой системе, и редко поддается аналитическому описанию. В итоге разрабатываемые смешанные модели оказываются устроенными таким образом, что аналитический вид ядра (смешивающего распределения) оказывается известным (например, нормальное, показательное, пуассоновское распределение или одно из возможных распределений экстремальных значений) с точностью до некоторых параметров, по которым и происходит смешивание в соответствии с неким распределением вероятностей, которое далеко не всегда можно оценить с помощью параметрических статистических методов, и приходится применять непараметрические методы статистического оценивания. Указанное обстоятельство объясняет термин "полупараметрические методы", фигурирующий в названии диссертации.

Диссертация удачно структурирована. В первой главе собраны доказанные автором новые предельные теоремы для статистик, построенных по выборкам случайного объема (включая суммы случайного числа случайных величин и экстремальные порядковые статистики). Полученные здесь результаты обобщают и конкретизируют ранее известные. Параграфы 1.2 и 1.3 посвящены предельным теоремам для экстремальных порядковых статистик, построенных по выборкам случайного объема, имеющего обобщенное отрицательное биномиальное распределение – смешанное пуассоновское распределение при смешивающем обобщенным гамма-распределении. Такой класс распределений индексов очень широк и позволяет строить гибкие модели. При таких индексах предельное распределение экстремальных порядковых статистик допускает представление в виде смесей известных распределений: Фреше, Снедекора–Фишера, строго устойчивого и др. Также получены явные выражения для числовых характеристик (моментов) этих распределений и изучены их аналитические свойства. Для некоторых важных частных случаев получены оценки скорости сходимости, которые могут быть использованы при оценке степени адекватности предлагаемых моделей. Рассмотрено и асимптотическое поведение центральных порядковых статистик – выборочных

квантилей, получены условия сходимости их распределений к распределению Стьюдента. Доказан закон больших чисел для сумм с обобщенным отрицательным биномиальным распределением – обобщение теоремы Реньи, являющейся своего рода законом больших чисел для геометрических случайных сумм – в котором для слагаемых не предполагается независимость и одинаковая распределенность и используется оригинальная нетрадиционная нормировка, что в итоге позволило получить нетривиальные выводы о долговременных трендах в метеорологических и океанологических процессах. Параграф 1.4 содержит новый вариант центральной предельной теоремы для сумм случайного числа независимых, но необязательно одинаково распределенных слагаемых для схемы серий, обосновывающий сближение распределение таких сумм с классом смесей нормальных законов. Этот результат стал основой при построении моделей распределения размеров частиц лунного реголита.

Предлагаемый подход построения смешанных вероятностных моделей может быть эффективно применен к повышению точности прогнозирования случайных процессов, поведение которых может быть описано стохастическим дифференциальным уравнением в условиях недостатка информации об аналитическом виде его (случайных) коэффициентов. В параграфе 2.1 описан метод скользящего разделения смесей и предложено его использование в качестве базовой процедуры статистического оценивания распределений случайных коэффициентов стохастического дифференциального уравнения с произвольными коэффициентами (уравнения Ланжевена). "Изюминкой" данного подхода является то, что он позволяет существенно расширить размерность признакового пространства на этапе обучения в методах интеллектуального анализа и прогнозирования процессов.

Алгоритмы анализа данных, основанные на методе скользящего разделения смесей, и их свойства описаны в главе 3.

Глава 4 посвящена методам построения смешанных моделей размеров частиц по сгруппированным (скрытым) данным. Здесь предложены оригинальные методы построения смешанных логнормальных моделей для распределений размеров частиц во всех 317 пробах лунного реголита, доставленных миссиями «Аполлон-11, 12, 14–17» и «Луна 24», которые представлены в каталоге NASA в весьма сжатой специфической форме. Для статистической обработки таких данных автором разработаны две процедуры: на основе бутстреп-метода и минимизации статистики хи-квадрат.

Глава 5 посвящена описанию различных методов анализа данных, основанных на конечных смесях нормальных распределений и их применению в скользящем режиме к изучению тонкой структуры процессов, наблюдаемых в экспериментах с турбулентной плазмой.

Глава 6 посвящена разработке вероятностных моделей и методов исследования метеорологических и океанологических данных. Особое внимание уделяется вопросам выявления экстремальных наблюдений в рассматриваемых пространственно-временных рядах. Здесь используется широкий набор

различных статистических процедур, алгоритмов машинного обучения и нейронных сетей для решения задач заполнения пропусков и прогнозирования. Особо следует обратить внимание на то, что предложенный в диссертации подход позволил существенно расширить представления о том, в каком смысле можно понимать "аномальную экстремальность" того или иного выделяющегося наблюдения. Еще одной "изюминкой" этой главы является демонстрация возможностей применения обобщенной теоремы Реньи, доказанной автором, к выделению глобальных климатических трендов и оцениванию их числовых характеристик.

В главе 7 рассматриваются программные решения и комплексы, которые использовались для анализа неоднородных данных и визуализации результатов в главах 3–6.

Основные результаты диссертационного исследования своевременно опубликованы в 82 печатных работах, из них 31 статья – в журналах, рекомендованных ВАК РФ, и 51 работа – в изданиях, индексируемых в Web of Science Core Collection и/или Scopus. Получены 39 свидетельств о государственной регистрации программ для ЭВМ, зарегистрированные в Федеральной службе по интеллектуальной собственности (Роспатент).

В ходе работы А.К. Горшенин проявил себя как самостоятельный, вдумчивый и инициативный исследователь, стремящийся найти нетривиальные подходы к решению важных задач. Результаты, представленные в диссертации А.К. Горшенина, свидетельствуют о его высокой квалификации в области разработки вероятностно-статистических моделей и их применения к изучению реальных процессов и явлений.

Диссертация А.К. Горшенина является крупным научным достижением и значительным вкладом в теорию и методологию построения вероятностно-статистических моделей, а также их применение к анализу неоднородных данных. Считаю, что диссертация соответствует всем критериям Положения о присуждении ученых степеней, а ее автор Горшенин Андрей Константинович заслуживает присуждения степени доктора физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ.

Научный консультант –
доктор физико-математических наук,
профессор, заведующий кафедрой
математической статистики факультета
ВМК МГУ имени М.В. Ломоносова



В.Ю. Королев

28.01. 2021 г.

В.Ю. Решетов