О Т З Ы В
официального оппонента на диссертационную работу
Ковалевского Артёма Павловича
«Статистические критерии апостериорного обнаружения
разладки временных рядов и их применения»,
представленную на соискание ученой степени
доктора физико-математических наук
по специальности 05.13.17 — Теоретические основы информатики

Актуальность темы

Актуальность избранной диссертантом темы не вызывает сомнений. Объектом исследования являются временные ряды, в том числе возникающие при компьютерном анализе текстов. Предмет исследования — статистические критерии анализа соответствия этих временных рядов вероятностным моделям. Следует выделить два этапа исследования: теоретико-вероятностный анализ, то есть доказательство соответствующих утверждений, и разработку алгоритмов применения статистических критериев к временным рядам. Актуальным является подход автора к анализу текста как временного ряда и использование математического аппарата теории случайных процессов. При этом разладка интерпретируется как изменение свойств временного ряда с некоторого неизвестного исследователю номера наблюдения. Изучение текста как процесса позволяет дополнить выводы, получаемые при интерпретации текста как точки в многомерном пространстве признаков. Другое приложение разработанных и изученных в диссертации статистических критериев – анализ статистической значимости отклонений многомерных временных рядов от модели линейной зависимости между компонентами. Алгоритмизация таких статистических тестов позволяет, как это проиллюстрировано на примерах, выделять разнородные фрагменты данных, что особенно актуально для анализа больших массивов данных.

Степень обоснованности научных положений, выводов и рекомендаций

В первых трех главах диссертации А. П. Ковалевский разрабатывает математический аппарат анализа временных рядов. В первой главе изучается наиболее простая вероятностная модель. Согласно основной гипотезе, наблюдаемые величины независимы и одинаково распределены. Согласно альтернативной гипотезе, их распределение меняется в некоторый момент времени. Последовательные оценки момента изменения хорошо известны из работ предшественников. Менее изучен вопрос о построении статистического критерия, позволяющего принять решение о том, была разладка или нет. А. П. Ковалевский предлагает строить такие тесты на единой основе эмпирического моста. Он проводит сравнение критериев и находит наилучший в этом достаточно широком классе. Во второй и третьей главах автор изучает модель фрактального гауссовского шума. Вторая глава посвящена построению оценок параметров, а третья – построению статистических критериев. В частности, предложена оценка бинарным знаковым методом и основанный на ней статистический критерий проверки фрактальности гауссовского шума. Полученные в первых трех главах результаты обоснованы доказательствами и результатами моделирования.

В четвертой главе диссертации А. П. Ковалевский изучает временные ряды, возникающие при компьютерном анализе текста. Используется вероятностная интерпретация моделей Ципфа и Мандельброта, известная по работам Бахадура и Карлина как бесконечная урновая схема. Автор совместно с М. Г. Чебуниным доказал функциональную центральную

предельную теорему для процесса числа разных слов в этой модели. На основании анализа текстов М. К. Щербакова и М. И. Цветаевой сделан вывод о том, что модель Мандельброта в наибольшей степени соответствует реальным текстам. Исследована зависимость параметров моделей от языка и года написания. Другой процесс, используемый для анализа текстов, – последовательность индикаторов появления служебных слов (предлогов, союзов, частиц). Слову текста сопоставляется 1, если слово является служебным, и 0 в противном случае. Полученный временной ряд анализируется методами глав 1 и 3. На основании исследования массива текстов и их конкатенаций выявлены уровни значимости статистического критерия однородности, позволяющие с довольно низкими вероятностями ошибок отличать конкатенацию двух текстов одного автора от конкатенации текстов разных авторов. Показано также, что с увеличением длины текста его однородность ухудшается, поэтому для анализа собраний сочинений авторов используется модель фрактального шума. Анализ параметров модели показывает, что фрактальность действительно присутствует, и разработанные в главе 3 методы позволяют анализировать однородность собрания сочинений.

В пятой главе изучаются регрессионные модели: регрессия на порядковые статистики и регрессия с циклическим трендом. Предложены статистические критерии обнаружения разладки, то есть изменения свойств случайной последовательности. Для циклического тренда проведено их сравнение методами математического моделирования.

В шестой главе А. П. Ковалевский применяет разработанные в главе 5 методы к медицинским и экономическим данным.

В теоретической части результаты диссертации обоснованы доказательствами, а в области приложений – корректным применением доказанных теоретически утверждений.

Оценка новизны и достоверности

В диссертационной работе А. П. Ковалевского выделены следующие новые научные результаты:

1. Широкий класс статистических критериев апостериорного обнаружения разладки в модели выборки с единых позиций: критерии построены на основе функционалов от эмпирического моста.
2. Сравнение статистических критериев апостериорного обнаружения разладки в модели выборки с точки зрения их относительной асимптотической эффективности по Питмену.
3. Алгоритм применения статистического критерия, основанного на норме эмпирического моста, к анализу однородности текста на естественном языке.
4. Модели фрактального броуновского моста и склейки фрактальных броуновских движений для временного ряда, построенного по тексту на естественном языке.
5. Центрированный знаковый метод, модифицированный знаковый метод и бинарный знаковый метод оценивания параметра Хёрста.
6. Статистический критерий проверки гипотезы фрактальности применен для проверки фрактальности текстов на естественном языке.
7. Построен статистический критерий проверки разладки фрактального гауссовского шума, основанный на разности оценок параметра Хёрста. Критерий применен к анализу однородности текста на естественном языке. Разработан алгоритм выявления склейки текстов.
8. Разработан класс критериев обнаружения разладки регрессии с циклическим трендом, основанных на значениях эмпирического моста.

Результаты, полученные автором, являются новым научным знанием. Результаты, касающиеся анализа текстов, согласуются, в частности, с результатами наших с О. В. Кукушкиной, А. А. Поликарповым и О. Г. Шевелевым исследований. Основные результаты диссертации опубликованы в 48 печатных работах, они неоднократно обсуждались на международных и всероссийских конференциях и симпозиумах и получили одобрение ведущих специалистов.

Общие замечания по диссертационной работе

1) В целом диссертационная работа А. П. Ковалевского производит очень хорошее впечатление. По первым трём главам, посвящённым разработке математического аппарата исследования, формулировкам и доказательствам теорем, построению алгоритмов статистической обработки данных применительно к анализу **скалярных** числовых временных рядов существенных замечаний нет. К сожалению, в диссертации **отсутствует обобщение результатов на случай многомерных (векторных) временных рядов наблюдений**. Исходные данные в виде многомерных временных рядов возникают в различных практических приложениях (в сейсморазведке, в рекомендательных системах и др.).

2) По главе 4, посвящённой анализу временных рядов **нечисловой** природы (текстов на естественном языке), необходимо отметить, что для применения здесь методов, разработанных для анализа временных рядов **числовой** природы, требуется корректная однозначная числовая идентификация (числовое представление) элементов нечисловых рядов (в том числе текстов). Однако в диссертации **не уделено достаточного внимания вопросам перехода от номинальной шкалы представления текстовых исходных данных к метрической.**

3) В диссертации используются два подхода к числовому представлению текстов: **индексный** и **индикаторный**. Оба подхода не противоречат условиям теорем, сформулированных и доказанных в диссертации. При **индексном** подходе автор сопоставляет элементам текста (словам) их индексы в упорядоченном множестве слов (словнике, словаре) и успешно использует получающееся таким образом числовое представление текстового ряда при исследовании адекватности частотно-ранговых распределений Ципфа-Мандельброта. **Неясно, почему индексный подход к числовому представлению элементов текстовых рядов не использован автором диссертации при решении задач апостериорного обнаружения и локализации разладок текстов.**

4) При решении проблем апостериорного обнаружения и локализации разладок в диссертации используется **индикаторный** подход к числовому представлению, сопоставляющий служебным словам 1, а остальным словам 0. Этот подход хорошо работает, если суммарные частоты употребления служебных слов в сравниваемых текстах существенно различаются. Поскольку **при индикаторном подходе по сравнению с индексным подходом используется только суммарная частота появления служебных слов в тексте и полностью теряется информация о частотах появления отдельных служебных слов, это снижает эффективность алгоритмов обнаружения и локализации разладки в тексте.** Для компенсации потерь информации в этом случае следовало бы использовать либо **индексно-индикаторный** подход к числовому представлению, когда служебным словам сопоставляются их индексы в упорядоченном наборе служебных слов, а остальным

словам сопоставляются нули, либо **частотный** подход, сопоставляющий словам частоты их появления в тексте, либо, наконец, **векторный индикаторный** подход, приводящий к необходимости **совместного** анализа рядов индикаторов появления в тексте каждого служебного слова. Но в последнем случае возникает **вопрос, как обрабатывать векторные временные ряды, поскольку вся теория и алгоритмы построены в диссертации для скалярных числовых временных рядов** (см. замечание 1).

Отмеченные в замечаниях недостатки не влияют на значимость главных теоретических и практических результатов диссертации.

Заключение

Диссертация А. П. Ковалевского является законченным научно-исследовательским трудом, выполненным автором самостоятельно на высоком научном уровне. В работе приведены научные результаты, позволяющие квалифицировать их как научное достижение. Полученные автором результаты достоверны, выводы и заключения обоснованы. Работа написана доходчиво, грамотно и аккуратно оформлена. По каждой главе и работе в целом сделаны четкие выводы.

Автореферат соответствует основному содержанию диссертации.

Диссертационная работа отвечает критериям Положения о порядке присуждения ученых степеней, а ее автор Ковалевский Артём Павлович заслуживает присуждения ученой степени доктора физико-математических наук по специальности 05.13.17 — Теоретические основы информатики.

Официальный оппонент

Поддубный Василий Васильевич,
доктор технических наук по специальности 05.13.16 - применение вычислительной техники, математического моделирования и математических методов в научных исследованиях, профессор по кафедре прикладной информатики,
главный научный сотрудник лаборатории когнитивных исследований языка
Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет»
(Национальный исследовательский Томский государственный университет; Томский государственный университет; НИ ТГУ; ТГУ)
634050, г. Томск, пр. Ленина, 36, http://www.tsu.ru, тел. 8 (382-2) 529-852, rector@tsu.ru.

Дата 25.02.19

Подпись д.т.н., профессора В. В. Поддубного заверяю

Подпись
УДОСТОВЕРЯЮ
Ученый секретарь ТГУ
Н.А. Саломатова