ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА
о диссертации КОВАЛЕВСКОГО Артема Павловича
«СТАТИСТИЧЕСКИЕ КРИТЕРИИ АПОСТЕРИОРНОГО ОБНАРУЖЕНИЯ РАЗЛАДКИ ВРЕМЕННЫХ РЯДОВ И ИХ ПРИМЕНЕНИЯ»,
представленной на соискание ученой степени доктора физико-математических наук
по специальности 05.13.17 – Теоретические основы информатики

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

Глубокое научное исследование того или иного природного, технологического, экономического или социального явления или процесса требует, помимо сбора и обработки данных и проведения экспериментов, построения адекватной математической модели. Особое развитие в настоящее время получают рандомизированные (вероятностные) модели, позволяющие учитывать стохастический характер изучаемого процесса, возможные ошибки в заданных параметрах и экспериментальных данных и т. п.

При разработке или использовании той или иной рандомизированной модели важную роль играет соответствие этой модели реальным данным. В данной работе исследуются статистические подходы к обнаружению *разладки* – изменению параметров модели – на основе статистического анализа временных рядов.

Хорошо известно, что многие содержательные и перспективные научные исследования получаются на стыках научных дисциплин, кажущихся, на первый взгляд, несовместимыми. В этом смысле в рецензируемой диссертации весьма интересными являются приложения предложенных в работе конструкций математической статистики для решения проблем из разных областей – от медицины до литературоведения. Эти актуальные приложения определяют специальность 05.13.17 – Теоретические основы информатики, по которой написана диссертация.

СОДЕРЖАНИЕ РАБОТЫ

Работа состоит из Введения, шести глав, Заключения и списка литературы.

Начальный фрагмент Введения представляет собой определенную «выжимку» из автореферата: после общего краткого обзора тематики сформулированы цели работы, описаны методы исследования и публикации (с оценкой личного вклада автора), приведен список конференций и семинаров, на которых представлялись полученные диссертантом результаты. Далее приводится подробный и достаточно объемный (7 страниц текста) обзор литературы.

Глава 1 посвящена построению, исследованию и сравнительному анализу статистических критериев для поиска момента разладки временных рядов с независимыми элементами. В разделе 1.1 кроме обзора литературы для «смешанной» (состоящей из выборок случайных величин двух сортов) выборки $X_i^{(n)}$ вводятся фундаментальные понятия *эмпирического моста* $Z_n(t)$ как ломаной, соединяющей точки $\left(\frac{k}{n}, \frac{nS_k - kS_n}{sn\sqrt{n}}\right)$ (здесь $S_n = \sum_{i=1}^{n} X_i^{(n)}$ и s – корень квадратный из выборочной дисперсии), – см. формулу (1) на стр. 22 и *приближенного бахадуровского наклона*

$$c(\theta) = 2 \lim_{n\to\infty} \left[-n^{-1} \ln\left(1 - F(J_n)\right)\right]$$

для набора статистик J_n и предельной функции распределения F – см. формулу (2) на стр. 23, использующегося для сравнения критериев. В разделе 1.2 доказывается лемма 1.1 о форме приближенного бахадуровского наклона для «гауссовского» предельного случая (3). В разделе 1.3 доказываются: лемма 1.2 о $\mathbb{C}$-сходимости эмпирического моста $Z_n(t)$ к броуновскому мосту, лемма 1.3 о сходимости случайной функции $Z_n(t)/\sqrt{n}$ к «детерминированной» функции $z_\theta(t)$ с кусочно-линейным графиком (см. рис. 1.1). Лемма 1.3 помогла автору диссертации доказать теорему 1.1 о соотношении статистик, имеющих вид функциональных норм. В разделе 1.4 рассмотрена статистика $J_n(g) = \left|\frac{\sum_{k=1}^{n} g(k/n)X_k}{s\sqrt{n}}\right|$, где $g(t)$ – функция ограниченной вариации, такая, что $\int_0^1 g(t)dt = 0, \int_0^1 g^2(t)dt = 1$. Для этой статистики доказан аналог центральной предельной теоремы (теорема 1.2) и получен вид предельной «детерминированной» функции для последовательности $J_n(g)/\sqrt{n}$ (теорема 1.3 и следствие 1.1). В разделе 1.4 в теореме 1.4 получен вид оптимальных коэффициентов статистики $J_n(g)$ (в смысле сравнения приближенных бахадуровских наклонов, подсчитанных по формуле (4)) при известной функции распределения $F_T(t)$ момента разладки T. Получены частные выражения коэффициентов для случая известного T (следствие 1.2) и равномерного распределения T на интервале $(0,1)$ (следствие 1.3). Имеются также рассуждения о ситуации, когда функция $F_T(t)$ неизвестна. В теореме 1.5 получены выражения для приближенных бахадуровских наклонов, подсчитанных по формуле (4), для «полезных» функционалов $J_n^\infty = \sup_{t\in[0,1]}|Z_n(t)|, J_n^{|r|} = \left|\int_0^1 [Z_n(t)]^r\right|^{1/r}$, $J_n(g_{1/2}) = |Z_n(1/2)|$; при этом функционал J_n^∞ дает наилучший критерий (т.е. максимальным оказывается соответствующее значение приближенного бахадуровского наклона). Именно этот критерий используется в дальнейшем при изучении однородности временных рядов. В разделе 1.6 сформулированы основные результаты главы 1.

Если в первой главе речь идет о временных рядах с независимыми элементами, то во главе 2 рассмотрен случай, когда элементы ряда обладают долговременной зависимостью. Конкретнее, рассматривается *фрактальное броуновское движение*, т.е. гауссовский случайный процесс $X_H(t), t \geq 0$ с нулевым математическим ожиданием и корреляционной функцией

$$\mathbf{E}X_H(t)X_H(s) = \frac{\sigma^2}{2}(t^{2H} + s^{2H} - |t-s|^{2H})$$

(см. формулу (7) на стр. 43), где H – *параметр Херста* (здесь $0 < H \leq 1$). В разделе 2.1 описаны различные представления и свойства этого процесса. В разделе 2.2 последовательно рассматриваются: «наиболее точный» *метод максимального правдоподобия* для оценки параметров σ^2 и H и численный алгоритм получения этих оценок на персональной ЭВМ, а также более «грубые» *метод нормированного размаха* (в виде алгоритма 2.1) и *метод дисперсии* (в виде алгоритма 2.2); дается также краткий сравнительный анализ этих (и подобных им) алгоритмов. В разделах 2.3 и 2.4 автор предлагает рассмотреть эффективные модификации «плохого» (обладающего большой дисперсией) *метода знаков* (формула (17) на странице 52 и алгоритм 2.3) для приближения параметра Херста. Для начала в разделе 2.3 доказывается сильная состоятельность оценки (17) (теорема 2.1), а затем

предлагается и достаточно подробно обосновывается *центрированный метод знаков* (алгоритм 2.4). Далее для задачи о проверке гипотезы $H = 1/2$ представляются и обосновываются *элементарный знаковый критерий* (алгоритм 2.5). В разделе 2.4 рассматривается модификация алгоритма 2.5, основанная на свойстве самоподобия фрактального броуновского движения и, как следствие, на целесообразности разбиения данных на блоки и рассмотрения статистик V_n (см. соответствующие формулы без номеров на страницах 63 и 65); при этом автор приводит довольно внушительный теоретический материал (доказательства соответствующих вариантов центральной предельной теоремы для случаев $H = 1/2$ – теорема 2.2 и следствие 2.1 и $H \neq 1/2$ – теорема 2.3 и следствия 2.2, 2.3). В этом же разделе 2.4 имеются соображения о *методе периодограммы* и о весовой версии статистик V_n. Далее представлен экономичный *бинарный знаковый метод* и доказан соответствующий варианты центральной предельной теоремы для случая $H = 1/2$ – теоремы 2.4, 2.5, лемма 2.5 и следствия 2.4, 2.5. В разделе 2.5 сформулированы основные результаты главы 2.

В разделе 3.1 описан план представления несколько разнородных материалов главы 3. В разделе 3.2 обсуждается новый *критерий «минимального спейсинга»* для определения нормальности малых выборок размера $n < 8$: для $n = 3, 4$ получен ряд аналитических результатов, а случай $n = 5$ исследован методами численного статистического моделирования, проведено сравнение с критерием Шапиро-Уилка. В разделе 3.3 несколько «запоздало» приводится алгоритм, соответствующий бинарному знаковому методу из раздела 2.4. В разделе 3.4 рассматриваются и обосновываются два алгоритма моделирования фрактального гауссовского шума: точный, основанный на специальном представлении корреляционной матрицы (алгоритм 3.2) и приближенный, основанный на построении скользящих средних (алгоритм 3.3). В разделе 3.5 формулируется алгоритм проверки соответствия модели фрактального гауссовского шума (алгоритм 3.4) и приведены специально насчитанные таблицы для этого алгоритма. В разделе 3.6 приведено аналитическое «упражнение» по поиску формул для распределения отношения компонент двумерного нормального вектора (теоремы 3.4, 3.5), которые предлагается использовать для исследования однородности фрактальных гауссовских шумов. В разделе 3.7 подходы, используемые выше для изучения статистики фрактального броуновского движения, предлагается использовать для модели (реализованной с помощью алгоритма 3.5, который известен как *метод обратной функции распределения для случайной последовательности*), в которой элементами последовательности являются величины, распределенные по симметричному устойчивому закону (более подробно рассмотрен случай *параметризации Золотарева*). В разделе 3.8 сформулированы основные результаты главы 3.

В главе 4 речь идет о статистическом анализе текстов. В разделе 4.1 приводится краткая аннотация разделов главы 4. В подразделе 4.2.1 текст рассмотрен как набор из слов, рассматриваемых как независимые одинаково распределенные случайные величины с тремя вариантами индивидуального дискретного распределения (распределение Мальденброта с бесконечным носителем, распределение Ципфа и геометрическое распределение). Доказывается теорема 4.1 о монотонности математического ожидания различных слов по параметрам рассматриваемых

распределений, а также теоремы 4.2 и 4.3 о состоятельности специально определенных оценок $\tilde{\theta}_n$ и $\widetilde{M}_n$ для всех трех случаев. В подразделе 4.2.2 приводится (без доказательства) формулировка специальной функциональной предельной теоремы, связанной с рассматриваемой моделью текста (теорема 4.4). В разделе 4.2.3 представлены результаты анализа текстов, основанного на доказанных теоремах (используются тексты М. Щербакова, М. Цветаевой и др.). Автору удалось «забраковать» модель с геометрическим распределением элементов текста и провести анализ зависимостей параметров текста от языка и от времени написания. Еще один подход к анализу текстов, основанный на изучении отклонений M_n эмпирического моста (с различными подходами к выбору уровня значимости, исследованию частот первого и второго рода и т.п.), рассмотрен в разделе 4.3: вводится словарь авторского инварианта, список исследуемых произведений и их анализ в различных сочетаниях (при этом кроме статистических результатов формулируется ряд достаточно обстоятельных и объемных литературоведческих соображений). В коротком (две страницы текста с таблицей) разделе 4.4 обозначена (без внятных объяснений) возможность проверки гипотез о фрактальности для текстов с помощью методик из главы 2. В разделе 4.5 сформулированы основные результаты главы 4.

В главе 5 снова (как и в главах 1–3) рассматриваются теоретико-статистические проблемы: на сей раз речь идет о регрессионной модели $Y_{ni} = \sum_{j=1}^{m} \theta_j g_j(i/n) + \varepsilon_i$, где θ_j – неизвестные параметры, $g_j(u), u \in [0,1]$ – заданные функции, а ε_i – независимые в совокупности одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной дисперсией σ^2. В свою очередь, в главе 6 рассматриваются приложения предложенных в главе 5 критериев.

В разделе 5.1 сформулирована предельная теорема МакНила, выделен частный практически важный случай (38) *линейной регрессии с циклическим трендом* с тригонометрическими функциями $g_j(u)$ и приведена формула (39) предельной ковариационной функции для этого случая. В разделе 5.2 рассматривается «одномерная» двухпараметрическая регрессия $Y_i = a + bX_i + \varepsilon_i$, где X_i – порядковые статистики для некоторого заданного распределения. Приводятся (без доказательства) теорема 5.1 о сходимости соответствующего эмпирического моста $Z_n^0(t)$ к гауссовскому процессу (приведена также формула для корреляционной функции этого процесса), а также следствие 5.1 о соответствующей среднеквадратической сходимости моста $Z_n^0(t)$ с формулами из работы [65]. В разделе 5.3 сходимость к гауссовскому процессу доказана для эмпирического моста $Z_n(t)$, соответствующего линейной регрессии с циклическим трендом (38) – теорема 5.2. Далее конструируется семейство критериев J_d для обнаружения разладки модели (при этом особо выделены частные случаи $d = 1,2,3$); рассматривается также асимптотически наиболее мощный критерий $J = \sup_{t\in[0,1]}|Z_n(t)|$. В разделе 5.4 осуществляется визуальное (по графикам траекторий) и «монте-карловское» (с помощью моделирования набора траекторий) исследование годности рассматриваемых критериев. В разделе 5.5 сформулированы основные результаты главы 5.

В разделе 5.6 (это первый раздел главы 6) кратко формулируется план представления материалов главы 6. В разделе 5.7 изучается статистика временных рядов, представляющих собой замеры концентрации маркеров *FB MoM* и *PA MoM* в крови людей (в зависимости от массы тела). Диссертант предлагает улучшенные (по сравнению с «фирменной» моделью компании «Typolog Software LTG & Co.») модели для корректировки значений *MoM*, в которой логарифм концентрации *MoM* предполагается нормально распределенным (с точностью до линейного преобразования). Для этих моделей численно оцениваются значения параметров и исследованы эмпирические мосты. Более подробный анализ моделей «с логарифмами» (см. формулу (41) на странице 214 и формулу (42) на странице 216) и их модификаций рассматривается в разделе 5.8 для изучения цен на жилую недвижимость. В разделе 5.9 техника исследования «одномерной» двухпараметрической регрессии с порядковыми статистиками из раздела 5.2 применяется для исследования модели «с логарифмами» (51) (в частности, вторично формулируется теорема 5.1; на сей раз она названа «теорема 6.1»). Важными (и едва ли не основными во всей диссертационной работе) видятся описанные в разделе 5.10 приложения разработанных автором статистических критериев к анализу дефектов строительных конструкций. Наконец, в разделе 5.11 сформулированы основные результаты главы 6.

Текст Заключения диссертации (как и начало Введения) представляет собой «выжимку» из автореферата: имеются разделы «Теоретическая значимость», «Практическая значимость», «Достоверность», «На защиту выносятся…» + благодарности коллегам.

Список литературы содержит 210 названий.

НОВИЗНА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Новыми являются как предложенные в диссертации статистические критерии апостериорного обнаружения разладки для модели выборки с независимыми компонентами, для фрактального броуновского движения и для регрессионных моделей, так и общие подходы к их сравнительному анализу, подразумевающие применение специальных функциональных и асимптотических методов. В связи с новизной самих статистических технологий, новыми являются и их приложения к проблемам сравнения литературных текстов, обработки медицинских и экономических данных, исследования надежности строительных конструкций.

СТЕПЕНЬ ОБОСНОВАННОСТИ И ДОСТОВЕРНОСТЬ НАУЧНЫХ ПОЛОЖЕНИЙ, ВЫВОДОВ И РЕКОМЕНДАЦИЙ, СФОРМУЛИРОВАННЫХ В ДИССЕРТАЦИИ

Требующиеся автору факты из литературы сформулированы верно (с соответствующими ссылками). Для новых лемм, теорем и их следствий приведены корректные доказательства. Объективными и достоверными являются выводы автора, сформулированные на основании его прикладных статистических исследований.

ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ ЗНАЧЕНИЕ РАБОТЫ

Предложенные автором критерии вносят значимый вклад в соответствующие разделы теории вероятностей и математической статистики. Особую роль в работе играют приложения построенных критериев для статистической обработки текстов,

медицинских, экономических и инженерных данных. Отметим также, что упомянутые новые критерии (как и подавляющее большинство других математических конструкций) обладают определенным «универсализмом» с точки зрения расширения сферы их возможных применений, и поэтому можно сформулировать весьма длинный список новых приложений (и в любом случае этот список будет неполным).

ЗАМЕЧАНИЯ ПО РАБОТЕ

1. При чтении текста диссертации (начиная с ее названия) только с главы 4 становится понятно, что диссертация содержит-таки материалы, соответствующие специальности 05.13.17 – Теоретические основы информатики (точнее - пункту 5 паспорта этой специальности: «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текстов, устной речи и изображений»). До этого момента кажется, что речь идет о работе по пункту 6 «Методы статистического анализа и вывода. Оценивание параметров. Проверка статистических гипотез» паспорта специальности 01.01.05 Теория вероятностей и математическая статистика. По-видимому, уже начиная со Введения (а может, даже с названия диссертации) стоило говорить не о «широком классе вероятностных моделей», а о важных задачах теории информатики типа проблем анализа текстов, исследования последовательностей сигналов и измерений в различных прикладных областях и т.п. (а уж потом конструировать и исследовать требуемые статистические критерии).

2. По не вполне понятным причинам автор диссертации не слишком выделяет и почти никак не описывает полученные им (в соавторстве) свидетельства и патенты [206–210] и связанные с ними теоретическо-статистические разработки. Кроме того, в работе (в частности, на страницах 154 и 160) упоминаются написанные автором программные продукты, но параметры этих программ (объем кода, особенности реализации на компьютерах различного типа) не описываются в тексте диссертации.

3. Чтение текста диссертации затруднено в связи со следующими обстоятельствами.

А). Неудачной видится организация ссылок на формулы в тексте диссертации: на весь текст диссертации (271 страниц) имеется всего 51 формул со ссылками. Это приводит, в частности, как к трудностям поиска занумерованных формул в тексте, так и к многочисленным повторам описаний обозначений (сравните, например, пары страниц 22, 25, или 63, 65, или 163, 182, или 195, 229 и др.).

Б). Помимо дублирования описаний обозначений в тексте встречаются случаи отсутствия описания обозначений. Соответствующих примеров достаточно много: на странице 22 не описано обозначение $\overline{X^2}$; на странице 34 не описано обозначение σ_θ; по-видимому, это одно и то же, что σ_1 из доказательства теоремы 1.2, т.е. предел величины s при $n \to \infty$ (кстати, обозначение s неудачное, нужно было подчеркнуть зависимость от числа выборочных значений n); не сразу понятны: обозначение $g_{1/2}$ на странице 38, обозначение S на странице 141, обозначение θ_j на странице 191 и т.д.

В). В разделах 2.2 и 2.3 автор весьма удачно представил построение обсуждаемых статистических оценок параметра Херста и их модификаций в виде алгоритмов (см. алгоритмы 2.1–2.5), а в разделе 2.4 (где речь идет о главных находках автора, связанных с возможностью разбиения данных на блоки) эта стилистика нарушена, и получается определенная «каша» из обосновывающих утверждений и их доказательств (а «лучший» алгоритм и вовсе вынесен в раздел 3.3 следующей главы). В том же разделе 2.4 основные утверждения (соответствующие варианты центральной предельной теоремы для случаев $H = 1/2$ и $H \neq 1/2$) названы не «теоремами», а «следствиями», а вспомогательные утверждения (типа теорем 2.2–2.5) названы не «леммами», а «теоремами». В целом раздел 2.4 видится весьма перегруженным.

Г). Так же, как и раздел 2.4, плохо структурирована и глава 3. Здесь «свалены в кучу» весьма разнородные материалы, связь между которыми просматривается с трудом.

Д). Отсутствуют подписи под рисунками 4.1–4.5.

Е). Материал раздела 4.3 по не слишком понятным причинам разбит на две части с анализом двух разных списков литературных текстов, при этом зачем-то меняются обозначения одних и тех же объектов (например, для отклонений эмпирического моста есть аж три обозначения: M_n, M и $|M_n|$).

Ж). Странной видится сквозная нумерация разделов глав 5 и 6.

З). Имеется также довольно много мелких редакционных замечаний.

З1). Замечание из раздела 1.3 на странице 26 о том, что понятие эмпирического моста было введено при исследовании статистики энергопотребления в работе [174] следовало привести в разделе 1.1 при определении этого самого эмпирического моста.

З2). В тексте имеются две таблицы 3.1, две таблицы 3.2, две таблицы 3.3. Неудачно оформлена первая из таблиц 3.1: название таблицы расположено на странице 99, а сама таблица – на странице 100. Такая же ситуация с таблицей 4.4 (название на странице 172, а сама таблица – на странице 173), с таблицей 4.5 (название на странице 174, а сама таблица – на странице 175; да еще между названием и таблицей вставлен рисунок 4.8) и с таблицей 5.1 (часть названия на странице 200, а сама таблица – на странице 201).

З3). На странице имеются ссылки на теоремы 2 и 3; теорем с такими номерами в тексте нет. По-видимому, речь идет о теоремах 3.2 и 3.3. Аналогичная проблема на странице 195, где упоминается несуществующая теорема 1 (очевидно, речь идет о теореме 5.1).

З4). На странице 119 есть ссылка на несуществующие таблицы 2.2–2.10 (по-видимому, имеются в виду таблицы 3.2–3.10), а вместо обозначения τ неверно используется обозначение tau.

З5). Совпадают формулировки теорем 5.1 (страница 195) и 6.1 (страница 229). Кроме того, на странице 229 имеется «пустая» ссылка «[?]» (по-видимому, имеется в виду работа [170]).

Вообще говоря, текст диссертации скомпонован неаккуратно, без особого уважения к читателю и выглядит не как единый труд, а как формальное соединение ранее написанных статей.

4. Определенную досаду у оппонента (как у специалиста по численному статистическому моделированию) вызывает использование «плохой» (неэкономичной, неадекватной) формулы моделирования стандартного гауссовского распределения на странице 99. Здесь безусловно лучшими являются *формулы Бокса-Мюллера* $X_1 = \sqrt{-2\ln U_1}\sin 2\pi U_2, X_2 = \sqrt{-2\ln U_1}\cos 2\pi U_2$, где $U_1, U_2 \in U(0,1)$. О качестве численного моделирования линейной регрессии с циклическим трендом из раздела 5.4 судить трудно из-за скудости описания этого моделирования.

ОБЩЕЕ ЗАКЛЮЧЕНИЕ

Научный уровень диссертации А. П. Ковалевского «Статистические критерии апостериорного обнаружения разладки временных рядов и их применения» соответствует требованиям, предъявляемым нормативными актами Российской Федерации к диссертациям на соискание ученой степени доктора наук. Работа формально соответствует пункту 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текстов, устной речи и изображений» паспорта специальности 05.13.17 – Теоретические основы информатики. Полученные результаты опубликованы (42 научные статьи и одна монография) и представлены на научных семинарах и конференциях; кроме того, А.П.Ковалевский имеет 3 свидетельства на программы на ЭВМ и 2 патента (с соавторами). Автореферат правильно отражает содержание диссертации.

Автор диссертации Артем Павлович Ковалевский заслуживает присуждения ему ученой степени доктора физико-математических наук по специальности 05.13.17 – Теоретические основы информатики.

Официальный оппонент
ведущий научный сотрудник лаборатории стохастических задач
Федерального государственного бюджетного учреждения
науки Института вычислительной математики и математической
геофизики Сибирского отделения Российской Академии наук,
доктор физико-математических наук (01.01.07 – Вычислительная
математика), профессор

Войтишек Антон Вацлавович

11 февраля 2019 года

Подпись А. В. Войтишека удостоверяю.
Ученый секретарь ИВМиМГ СО РАН, к.ф.-м.н.
Л. В. Вшивкова

Сведения об организации: Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук; адрес: 630090, Новосибирск, проспект Академика М. А. Лаврентьева, 6; телефон: +7 (383) 330 83 53; адрес электронной почты: contacts@sscc.ru; сайт: https://icmmg.nsc.ru