**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

На правах рукописи

Рычка Ольга Валентиновна

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ВЫЯВЛЕНИЯ И КОРРЕКТИРОВКИ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Донецк – 2021

Работа выполнена в ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Министерства образования и науки Донецкой Народной Республики, г. Донецк.

Научный руководитель: кандидат технических наук, доцент
Григорьев Александр Владимирович,
ГОУВПО «ДОННТУ» (г. Донецк),
профессор кафедры «Программная инженерия» им. Л.П. Фельдмана

Официальные оппоненты:
Кобак Валерий Григорьевич
доктор технических наук, доцент,
ФГБОУВО «Донской государственный технический университет» (г. Ростов-на-Дону),
профессор кафедры «Программное обеспечение вычислительной техники и автоматизированных систем», профессор кафедры «Вычислительные системы и информационная безопасность»

Чернышева Оксана Александровна
кандидат технических наук,
ГОУВПО «ДОНБАССКАЯ НАЦИОНАЛЬНАЯ АКАДЕМИЯ СТРОИТЕЛЬСТВА И АРХИТЕКТУРЫ» (г. Макеевка),
доцент кафедры «Специализированные информационные технологии и системы»

Ведущая организация: **Государственное учреждение «Институт прикладной математики и механики (ГУ «ИПММ») (г. Донецк)**

Защита состоится «29» марта 2022 г. в 12.00 часов на заседании диссертационного совета Д 01.024.04 при ГОУВПО «ДОННТУ» и ГОУВПО «ДОННУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 1, ауд. 203 Тел./факс: 380(62) 304-30-55, e-mail: uchensovet@donntu.org.

С диссертацией можно ознакомиться в библиотеке ГОУВПО «ДОННТУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 2. Адрес сайта университета: http://donntu.org

Автореферат разослан «___» _____ 20____ г.

Ученый секретарь
диссертационного совета Д 01.024.04
кандидат технических наук, доцент Т.В. Завадская

ОБЩАЯ ХАРАКТЕРИС ТИКА РАБОТЫ

Актуальность темы исследования. В условиях использования большого объёма данных, а также высокой значимости результатов их анализа важной задачей является обнаружение аномальных данных (выбросов, несогласованных наблюдений, исключений). Появление аномальных измерений может быть обусловлено «человеческим фактором» или влиянием на систему внешних факторов. Игнорирование их наличия может привести к существенному ухудшению качества модели, поэтому важное значение имеет предварительная обработка исходных статистических данных, которая включает этапы выявления аномальных измерений, их корректировку или удаление.

Качественной можно считать модель, которая обладает следующими свойствами: простота, максимальное соответствие реальным данным (необходимо стремиться к максимально возможному значению коэффициента детерминации R^2), высокие прогнозные качества.

Одним из главных инструментов анализа экспериментальных данных и обнаружения закономерностей в них является регрессионный анализ. С его помощью находится математическая модель влияния одной или нескольких независимых переменных X_i на зависимую переменную Y.

Частным, но важным случаем регрессионного анализа являются парные линейные модели. В пользу использования линейных моделей говорит и широкая область их практического применения. Так, на основе линейных моделей построено множество сложных методов машинного обучения, в том числе и нейронные сети. Найденное линейное уравнение может быть начальной точкой для построения более сложных моделей. Ещё одним преимуществом парных линейных регрессионных моделей является возможность приведения большинства нелинейных моделей к линейному виду. Поэтому парные линейные регрессионные модели применяются в различных областях науки.

Следует отметить, что несмотря на многообразие методов обнаружения выбросов в исходных статистических данных, существующие подходы, в большинстве своём, применимы только для одномерных выборок (чаще всего временных рядов) и на каждой итерации метода осуществляется анализ лишь одного подозрительного значения. Когда речь идёт о зависимости между несколькими переменными, методы, основанные на поиске отклонения от среднего или предыдущего (последующего) значения не дают положительных результатов. Это связано с тем, что не учитывается наклон линии регрессии. Помимо этого, описанные в литературе методы являются чувствительными к объёму исходной выборки.

Поэтому, разработка и реализация алгоритма поиска аномальных измерений в исходных данных и основанных на нём методов последующей обработки данных с целью повышения качества парных регрессионных моделей для дальнейшего их использования в прогнозировании,

проектировании, здравоохранении и других областях является актуальной научно-прикладной задачей.

Степень разработанности темы исследования. Исследования, связанные с обнаружением аномальных данных, проводились уже в XIX веке. По данной теме написано большое количество статей и книг различными учёными, например, такими как Россиув и Лерой, Барнет и Левис, Бекман и Кук, Хоакинс, Ходж и Остин, Маркоу и Синх и др. Также необходимо отметить работы Дж. Тьюки, Д. Хоглина, Ф. Мостеллера, В.М. Бухштабера, С.А. Айвазяна, П. Веллемана, И.С. Енюкова, Л.Д. Мешалкина. В современной литературе описаны десятки различных методов нахождения и устранения выбросов из исходных статистических данных. Основными из них являются: метод Граббса, метод Титьена-Мура-Бекмана, методы Эктона и Прескотта-Лунда. Их главные преимущества – простота понимания и применения. Однако, существующие методы имеют ряд общих недостатков:

– методы плохо формализованы и заключаются в поиске лишь одного аномального значения на каждом шаге;

– большинство из них применяются только для одномерных выборок, т.е. поиск аномалий осуществляется только по одной из переменных;

− нет конкретных рекомендаций о дальнейших действиях исследователя после нахождения выбросов в исходных данных;

− существующие методы обнаружения ненадежных и аномальных измерений опираются на конкретные законы распределения вероятностей, однако исследователю они априорно не известны;

− чувствительность большинства методов к объёму выборки;

– большинство существующих методов реализованы в таких специализированных программных пакетах, как MathCad, MatLab, Statistica, Mathematica и др.; однако данные программные средства, в основном, ориентированы на математиков и инженеров, поэтому являются достаточно сложными в использовании для специалистов других областей; помимо этого, в них не уделяется достаточное внимание последующему анализу и обработке аномальных значений.

Таким образом, существующие методы обнаружения аномальных данных имеют ряд существенных недостатков, что делает актуальным новые исследования и разработки в этой области для повышения качества регрессионных моделей.

Целью диссертационной работы является совершенствование методов предварительной обработки исходных статистических данных для повышения точности линейных регрессионных моделей при построении эффективных прогнозов.

Для достижения цели поставлены и решены следующие задачи:

1. Выполнить сравнительный анализ существующих методов обнаружения аномальных и ненадёжных измерений.

2. Предложить и обосновать усовершенствованные методы обработки статистических данных, эффективность которых не должна зависеть от объёма исходной выборки, а их применение – не приводить к ухудшению качественных характеристик модели.

3. Дать рекомендации по выбору конкретного метода обработки исходных статистических данных в зависимости от вида модели.

4. На основе усовершенствованных методов поиска и корректировки исходных данных выполнить построение алгоритмов и осуществить разработку программного комплекса.

5. Провести программное моделирование с использованием разработанного комплекса программ для оценки эффективности предложенных методов.

Объект исследования. Объектом исследования является процесс анализа данных, основанный на использовании парных линейных регрессионных моделей**.**

Предмет исследования. Предметом исследования являются математические модели, методы и алгоритмы обработки статистических данных для повышения качества регрессионных моделей.

Научная новизна полученных результатов заключается в следующем.

1. Разработан новый метод поиска аномалий, основанный на построении области надёжности, которая зависит от наклона уравнения регрессии, доверительной вероятности и соответствующего коэффициента, что позволяет одновременно обнаруживать аномальные измерения как по независимой переменной (X), так и по зависимой переменной (Y). Это приводит к повышению качества прогнозов (от 10%), полученных по линейным регрессионным моделям, возрастанию коэффициента детерминации (от 10% до 30%) и уменьшение трудоёмкости (количество элементарных операций) – до $2*10^n$ раз, по сравнению с существующими методами.

2. Получил дальнейшее развитие метод корректировки аномалий, отличием которого является изменение значений аномальных статистических данных на значения, соответствующие рассчитанной области надёжности, а также отсутствие сокращения объема исходных статических данных из-за отбрасывания, что особенно важно при моделировании на выборках малого объёма.

3. Предложены два упрощения алгоритма поиска аномальных данных, позволяющие сократить трудоёмкость анализа, выбор которых зависит от надёжности исходных значений X и Y, что приводит к обнаружению аномальных данных по одной из соответствующих переменных. Спецификой первого упрощения является то, что оно может быть использовано для поиска аномальных данных при многомерной линейной зависимости.

4. Обоснована возможность применения предлагаемого в работе метода обнаружения выбросов не только для линейных регрессионных

прогнозных уравнений, но и для нелинейных моделей с внутренней линейностью.

Теоретическая и практическая значимость работы. Теоретическая значимость результатов работы заключается в том, что предлагаемые методы повышения качества регрессионных моделей, основанные на обнаружении и последующей обработке аномальных измерений в исходных статистических данных, являются эффективным инструментом для последующей разработки точных прогнозов, использующихся в различных отраслях науки и техники. В частности:

1. Показано, что предложенный в работе подход позволяет обнаружить выбросы и скорректировать вид модели без дополнительного графического отображения (на примере трёх наборов данных из «квартета Энскомба»).
2. Предложенные методы поиска и корректировки аномалий не имеют ограничений на объём выборки, в отличие от существующих.
3. Предложенные методы поиска аномальных данных и их последующей корректировки в дальнейшем могут быть дополнены и расширены для применения при построении многомерных регрессионных моделей.

Практическая значимость работы состоит в том, что результаты работы могут применяться в различных предметных областях, таких как здравоохранение, экономика и других, при решении задач прогнозирования, проектирования, оптимизации и т.д. В работе определены оптимальные параметры использования предложенных методов корректировки исходных данных, на основе которых даны практические рекомендации по выбору конкретного метода. Разработан оригинальный комплекс программ, реализующий новый алгоритм поиска аномальных данных и методы их последующей обработки, отличающийся наличием различных модулей для автоматизированной обработки исходных статистических данных, их графического отображения и построения наилучшей модели.

Методология и методы исследования. Для решения поставленных задач в работе использовались методы теории вероятности, математической статистики, математического моделирования, численные методы, регрессионный анализ.

Связь работы с научными программами, планами, темами. Работа выполнена в соответствии с тематическими планами Донецкого национального технического университета и является частью исследований, в которых автор принимала участие как исполнитель: гостемы Н-22-10 «Программное обеспечение высокопроизводительных вычислительных, интеллектуальных и моделирующих систем»; гостемы Н-1-16 «Анализ современных методов инженерии программного обеспечения для информационно-вычислительных и интеллектуальных систем»; гостемы Н-16-18 «Исследование методов, технологий и средств инженерии программного обеспечения на различных

классах приложений»; гостемы Н-2020-14 «Усовершенствование средств инженерии программного обеспечения для актуальных классов IT-приложений».

Научные положения, выносимые на защиту.

1. Показано, что построение области надёжности («коридора»), представляющей собой прямоугольную область, размер которой зависит от заданного значения вероятности и величин среднеквадратических отклонений, позволяет эффективно обнаруживать ненадежные данные, как данные, которые не попали в эту область, в результате чего достигается повышение качества исходной модели (значение коэффициента детерминации R^2 может увеличиваться до 30%). При этом, при отбрасывании данных не происходит ухудшения качественных характеристик модели, поскольку число отброшенных наблюдений не является критическим и, как правило, составляет от 5% до 20% исходных данных.

2. Показано, что предлагаемый алгоритм применим также к нелинейным регрессионным моделям с внутренней линейностью (экспоненциальная, логарифмическая, степенная и др.).

3. Доказано, что применение алгоритма построения области надежности, а также соответствующей стратегии исключения/изменения данных, исходя из объема имеющейся выборки, позволяет сократить временные затраты за счет уменьшения количества вычислительных операций (при этом сокращение может быть от $4*n$ до $2*10^s$, где n – количество исходных данных, s – число аномальных измерений) и получить регрессионные модели более высокой точности.

Степень достоверности и апробация результатов. Достоверность результатов исследования обеспечивается достаточным количеством проведенных экспериментальных вычислений с использованием реальных и модельных данных. Подготовка, анализ исходных данных и интерпретация итоговых результатов базируются на современных методах обработки информации и статистического анализа.

Практическая ценность исследований подтверждается внедрением в ООО НПО «Интермет» (справка о внедрении от 23 июня 2021 г.), в научно-исследовательские работы ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» (справка о внедрении № 29-13/15 от 05 июля 2021 г.), в учебный процесс ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» при чтении лекций и проведении лабораторных работ на кафедре «Программная инженерия» им. Л.П. Фельдмана по дисциплинам: «Эмпирические методы программной инженерии», «Численные методы в информатике» (справка о внедрении № 29-12/15 от 05 июля 2021 г.).

Апробация результатов диссертации.

Основные положения диссертационной работы докладывались и обсуждались на: VIII Международной научно-практической конференции

«Математическое и программное обеспечение интеллектуальных систем (MPZIS-2010)», г. Днепропетровск: ДНУ им. Олеся Гончара, 2010 г.; «Донбас-2020 перспективи розвитку очами молодих вчених», г. Донецк, ДонНТУ, 2010 г; XIX Международной научно-практической конференции MicroCAD-2011, г. Харьков, 2011 г.; Одиннадцатой международной научно-технической конференции «Проблемы информатики и моделирования» г. Харьков: НТУ «ХПИ», 2011 г.; IV Международной научно-технической конференции «Моделирование и компьютерная графика - 2011», г. Донецк, 2011 г., VII Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2016)», г. Донецк, 2016 г.; II Международной научно-практическая конференция «Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2018)», г. Донецк, 2018 г.; VIII Международной научно-практической конференции «Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике», г. Азов, 2021 г.

Личный вклад соискателя. Все результаты и положения, составляющие основное содержание диссертации, вынесенные на защиту, получены соискателем самостоятельно в процессе научных исследований. Личный вклад автора заключается в обосновании идеи и цели работы, её реализации, а также в проведении теоретических и экспериментальных исследований, разработке вычислительных алгоритмов и комплекса программ для их компьютерной реализации, разработке рекомендаций по практическому применению результатов.

Публикации. Основные научные результаты диссертации опубликованы в 17 научных работах, из них 2 статьи в специализированных изданиях, рекомендованных ВАК ДНР, 4 – в изданиях, входящих в перечень научных изданий, утверждённых ВАК Украины, 2 – в других научных изданиях (в том числе 1 монография), 9 – в материалах международных научных конференций.

Соответствие темы и содержания диссертации паспорту специальности.

По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки), в частности: п.3 «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»; п.5 «Комплексные исследования научных и технических проблем с применением современных технологий математического моделирования и вычислительного эксперимента»; п.6 «Разработка новых математических методов и алгоритмов проверки адекватности математических моделей объектов на основе данных натурного эксперимента».

Структура и объём работы. Диссертация состоит из введения, четырех разделов, заключения, списка литературы и 4 приложений. Изложена на 162 страницах машинописного текста, включая 44 рисунка, 36 таблиц, список литературы из 98 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение содержит общую характеристику работы. Обоснована актуальность темы диссертации и её научная новизна, поставлены цели и задачи исследований, определены практическая и теоретическая значимость результатов работы, представлены сведения об их апробации, сформулированы выносимые на защиту основные положения.

В **первом разделе** «Постановка задачи и анализ известных методов обнаружения аномальных измерений» описаны основные этапы решения сложных задач с использованием математического моделирования, определена постановка задачи прогнозирования на основе линейных регрессионных моделей, выделены основные цели и задачи регрессионного анализа, как одного из методов построения математических моделей. Рассмотрены и проанализированы существующие методы нахождения и корректировки ненадёжных измерений в исходных статистических данных. В результате анализа данных методов выявлены имеющиеся недостатки и определены перспективы их устранения путём разработки новых подходов, сформулированы цели и задачи исследований.

Во **втором разделе** «Совершенствование методов выявления аномальных измерений и их последующей обработки на основе линейных регрессионных моделей» описана сущность разрабатываемых методов повышения качества парных линейных регрессионных моделей.

Основная идея методов заключается в построении области надёжности, исходя из доверительной вероятности и соответствующего коэффициента с учётом угла наклона уравнения регрессии. Первый метод заключается в обнаружении и отбрасывании аномальных и ненадёжных измерений, а второй – в корректировке таких данных. Предложены модификации этих методов и дан их сравнительный анализ, основанный на экспериментальных данных.

Строится область со сторонами $2k\cdot\sigma_e$ и $2k\cdot\sigma'_e$. Значение k определяется из формулы (1), а среднеквадратические отклонения σ_e и σ'_e по формулам (2) и (3) соответственно.

$$P_0 = 2\cdot\frac{1}{\sqrt{2\pi}}\int_0^k e^{-t^2/2}dt, \qquad (1)$$

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (a\cdot X_i + b))^2}{n-2}}, \qquad (2)$$

$$\sigma'_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (a' \cdot X_i + b'))^2}{n-2}} \qquad (3)$$

где a и b – коэффициенты исходного линейного регрессионного уравнения; a' и b' − коэффициенты перпендикуляра, построенного к исходному уравнению; Y_i – фактические значения; X_i – исходные значения по независимой переменной; n – количество исходных данных в выборке.

После того, как аномальные данные найдены, требуется выбрать последующий метод их обработки. Это может быть либо удаление, либо корректировка. Корректировка осуществляется таким образом, чтобы данные вышедшие за границы области надёжности попали в заданную прямоугольную область.

При переносе данных на уровень $A \cdot x_i + B \pm k \cdot \sigma_e$ значение независимой переменной x_i остается неизменной, а зависимая переменная меняется на соответствующее значение.

Для переноса данных на уровень $A' \cdot x_i + B' \pm k \cdot \sigma'_e$ меняются значения независимой переменной x_i.

Для нахождения новых значений x_i, при перемещении на уровень $A' \cdot x_i + B' + k \cdot \sigma'_e$ используется формула:

$$X'_i = \frac{(A \cdot X_i + B) - k \cdot \sigma' - B'}{A'}. \qquad (4)$$

Значение x_i при переносе на уровень $A' \cdot x_i + B' - k \cdot \sigma'_e$ находится по формуле:

$$X'_i = \frac{(A \cdot X_i + B) + k \cdot \sigma' - B'}{A'}. \qquad (5)$$

Графическое отображение описанных выше методов представлено на рисунках 1 и 2 соответственно.

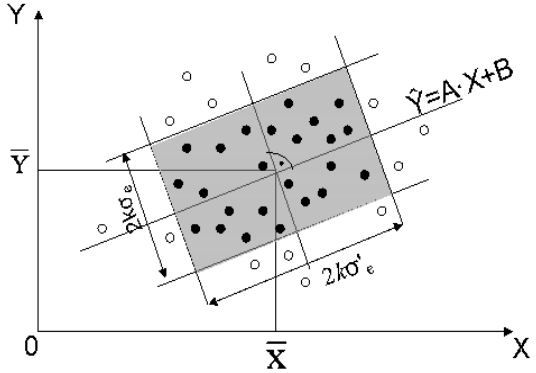

Рисунок 1 – Графическое представление метода, основанного на отбрасывании аномальных данных

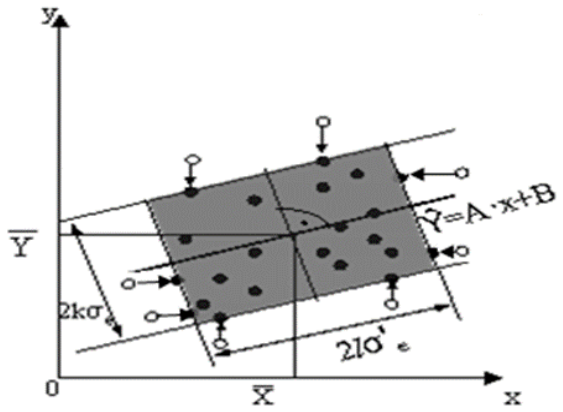

Рисунок 2 – Графическое представление метода, основанного на переносе аномальных данных

Для оценки эффективности, разработанных методов в работе используются следующие критерии:

1. Коэффициент детерминации R^2.
2. Модуль величины смещения результата прогноза. Смещение возникает из-за изменения положения нового регрессионного уравнения.

$$\Delta_{прогн}=|(A\cdot X_{прогн} + B) – (A_н\cdot X_{прогн} + B_н)| \quad (6)$$

где: первое и второе слагаемое, соответственно, линейное регрессионное уравнение до отбрасывания части статистики и линейное регрессионное уравнение после отбрасывания или корректировки части статистики.

3. Доверительный интервал прогнозных значений.
4. Точность, которая рассчитывается по формуле (7):

$$T = R^2 \cdot \frac{m}{n} \quad (7)$$

где n – исходное количество данных;

m – количество данных оставшихся после отбрасывания.

5. Количество элементарных операций ЭВМ (К), необходимых для реализации методов.

Для метода, основанного на отбрасывании данных, требуется:

$$K=26\cdot n+ 9\cdot m + 7 \quad (8)$$

где m − количество точек, оставшихся после отбрасывания.

Количество элементарных операций, необходимых для реализации метода переноса данных:

$$K = 35\cdot n + m' + 10\cdot l+7 \quad (9)$$

где m' − количество перенесенных точек на уровень $A\cdot xi + B \pm k\cdot\sigma_e$; l − количество перенесенных данных на уровень $A'\cdot xi + B' \pm k\cdot\sigma'_e$.

Часто при решении практических задач достаточно обрабатывать ненадёжные значения, возникающие только по зависимой переменной Y. В данном случае исследователь абсолютно уверен в значениях независимой переменной X и крайние величины x_i незначительно удалены от остальных.

Суть этой модификации заключается в том, что достаточно только определить так называемый коридор, верхняя и нижняя границы которого будут равноотстоять от линии исходного регрессионного уравнения. Расстояние между этими границами будет равным $2k\sigma_e$. Также данная модификация применима для поиска и отбрасывания аномальных данных при многомерных линейных регрессионных моделях.

В данной модификации при отбрасывании данных количество элементарных операций будет значительно ниже, чем в исходном методе:

$$K = 16 \cdot n + 9 \cdot m + 1. \quad (10)$$

А при корректировке данных:

$$K = 25 \cdot n + m' + 1. \quad (11)$$

Помимо первой модификации исходного метода, может быть применена и вторая модификация. При использовании второй модификации отсекаются данные, которые являются аномальными по независимой переменной X. В отличие от первой модификации метода повышения качества прогнозной модели с помощью отбрасывания данных, во второй модификации находится коридор, границы которого параллельны линии, построенной по найденному уравнению перпендикуляра. Расстояние между ними составляет $2k\sigma'_e$.

Данную модификацию рекомендуется использовать в случаях, когда в исходных данных регрессора X существует достаточно большой размах между крайними его значениями и они удалены на значительное расстояние от остальных точек. В этом случае исследователю следует обратить особое внимание на эти значения, поскольку они могут оказаться выбросами.

При использовании данной модификации значение коэффициента детерминации R^2 в отдельных случаях уменьшается. Однако это связано с тем, что крайние точки являются определяющими для уравнения регрессии, и если они удалены на значительное расстояние от остальных точек, то они оказывают большое влияние на исходное уравнение. Рекомендуется отбрасывать не более 10-15% исходных статистических данных.

Одним из примеров положительного эффекта от применения данной модификации может служить один из наборов данных квартета Энскомба (Рисунок 3).

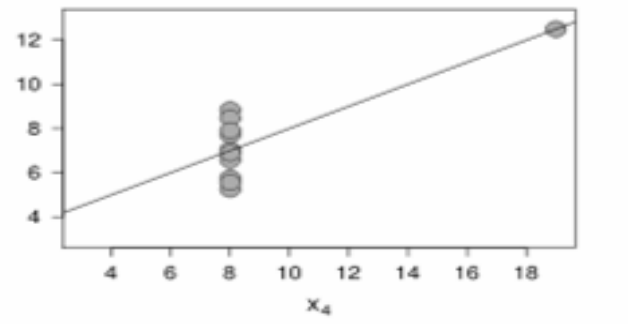

Рисунок 3 – Набор данных из квартета Энскомба

Исходное уравнение регрессии имеет следующий вид: $y = 3 + 0{,}5\,x$. Коэффициент детерминации R^2 при этом составляет 0,67. Однако, использование предложенного в данной работе подхода или второй его модификации позволило выявить аномальное значение по независимой переменной X. Поскольку, данное значение было значительно удалено от остальных и являлось определяющим для уравнения регрессии, то его удаление значительно изменило все основные параметры. Так коэффициент детерминации вместо 0,67 стал 0,09, тем самым давая исследователю понять, что исходное предположение о виде модели было ошибочным. Таким образом, данный пример демонстрирует, что применение предложенного в работе метода поиска и последующего устранения аномалий позволяет получить скорректированную модель данных, которая наиболее точно соответствует исходным измерениям. При этом не происходит искусственного сокращения области анализа, поскольку отбрасывается небольшой процент (в общем случае, от 5% до 25%) исходных данных.

Количество элементарных операций для данной модификации, при отбрасывании данных составляет:

$$K = 19 \cdot n + 9 \cdot m + 4. \quad (12)$$

А при корректировке измерений:

$$K = 28 \cdot n + 10 \cdot l + 4. \quad (13)$$

В ходе проведения экспериментов на различных наборах данных были получены следующие результаты:

- коэффициент детерминации R^2 растет и достигает максимального значения при вероятности попадания в заданную область равную 65%;
- величина доверительного интервала уменьшается до 2.6 раз;
- величина смещения прогнозного значения $Y_{прогн}$ не превышает 2.5%;
- число элементарных операций составляет в среднем 800 и 1550 для исходного объема выборки 24 и 47 значений соответственно при отбрасывании данных, и 1110, 1700 при корректировке данных. Однако, требующееся для предложенных методов количество элементарных операций является меньшим, чем при применении известных методов.

Таким образом, можно сказать, что применение предложенных методов даёт хорошие результаты.

Третий раздел «Разработка программных модулей для решения задачи обнаружения и обработки аномальных измерений» посвящен разработке архитектуры оригинального комплекса программ, предложенных во втором разделе методов. Архитектура комплекса включает в себя четыре программных модуля (Рисунок 4). Каждый модуль является универсальной компонентой, способной интегрироваться с другими программными модулями.

Рисунок 4 – Состав модулей программного комплекса

Помимо модулей можно выделить восемь логических блоков, представленных на рисунке 5.

Рисунок 5 – Логические блоки комплекса

Разработанный программный комплекс можно представить в виде UML диаграммы активности (деятельности) (Рисунок 6).

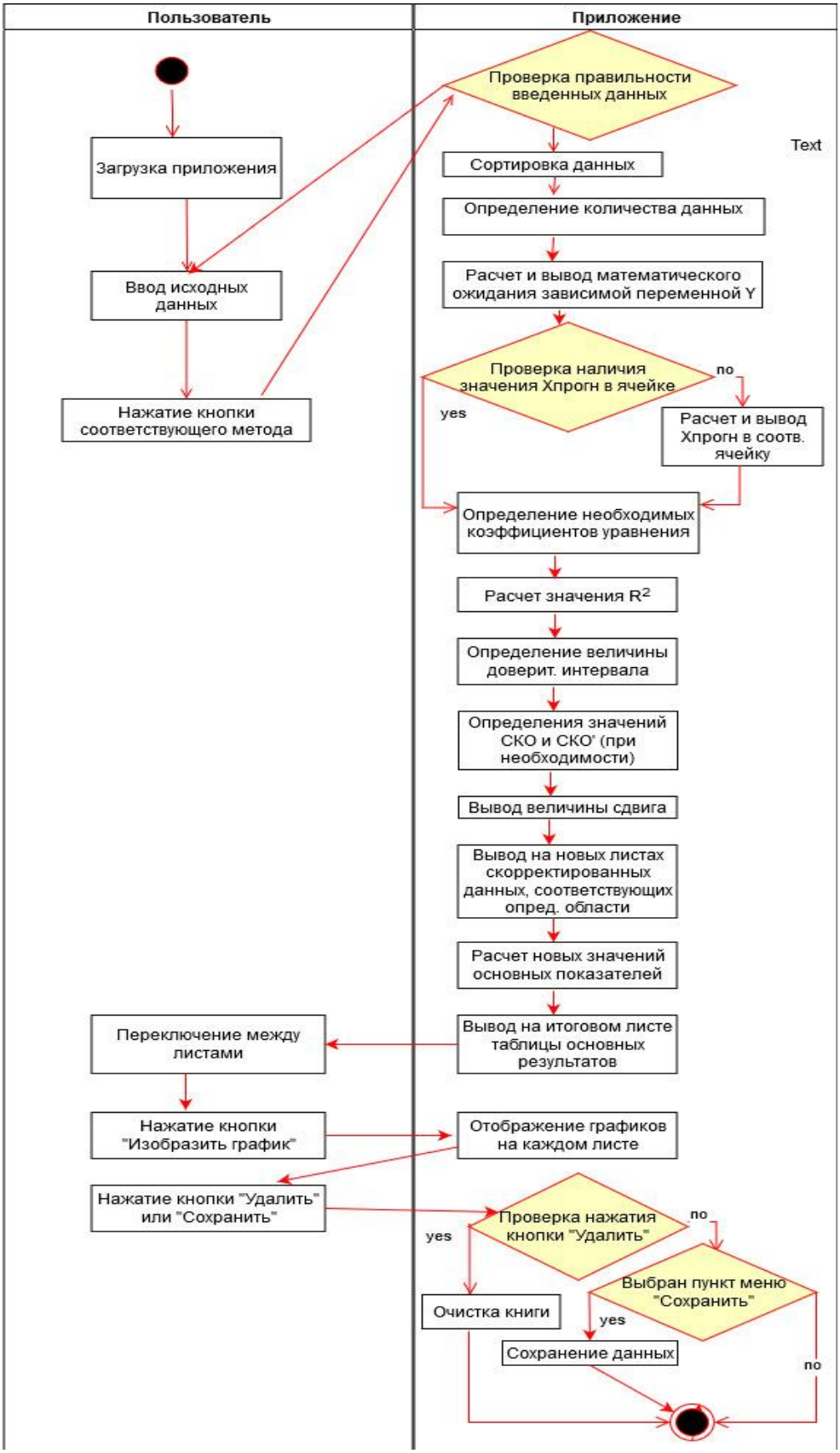

Рисунок 6 – Диаграмма активности

В **четвёртом разделе** «Сравнительный анализ предложенных методов и рекомендации по их применению» даны основные рекомендации по выбору одного из двух методов повышения качества регрессионной прогнозной модели.

На основе проведенных исследований можно дать следующие рекомендации по выбору одного или другого метода:

1. В случае большого объёма предпочтительнее метод основанный на отбрасывании данных, в противном случае – метод корректировки аномальных данных.

2. В случае уверенности в исходных данных по одной из переменных можно использовать одно из упрощений методов.

3. При применении методов следует учитывать критерий точности, поскольку данный критерий может являться индикатором для прекращения обработки исходных данных. Как только величина T, становится меньше порогового значения 0,5, следует остановить применение, выбранного метода и выбрать наилучший вариант. Также, следует отметить, что не рекомендуется отбрасывать более 20% исходных измерений.

ЗАКЛЮЧЕНИЕ

В диссертационной работе дано теоретическое обоснование и приведено решение актуальной научно-технической задачи совершенствования методов обработки исходных статистических данных с целью выявления и дальнейшей корректировки аномальных измерений, что позволяет повысить точность парных регрессионных моделей, используемых для прогнозирования в различных областях науки и техники.

Результаты диссертационного исследования могут быть сформулированы следующим образом:

1. Выполненный анализ наиболее часто используемых в рамках регрессионного анализа методов обнаружения и устранения аномальных данных показал наличие ряда недостатков, основным из которых является большая трудоёмкость вычислений. Исходя из этого, разработка новых методов обнаружении и устранении аномальных измерений для повышения качества парных регрессионных моделей, основанных на использовании доверительной вероятности, соответствующего коэффициента и учитывающих наклон уравнения регрессии, является актуальной научно-практической задачей.

2. Разработаны алгоритмы функционирования методов обнаружения и корректировки экспериментальных данных, базирующиеся на основных математических и статистических принципах, позволяющие улучшить качество регрессионных моделей, для дальнейшего их использования при прогнозировании и проектировании. Использование предложенных в работе методов ведёт к повышению качества прогнозов, полученных по линейным регрессионным моделям (от 10%), за счет предварительной обработки

исходных данных. Возрастание коэффициента детерминации при этом может достигать от 10% до 30%.

3. Разработана архитектура оригинального комплекса программ для реализация предложенного алгоритма поиска и обработки аномальных данных, включающая следующие модули: программные модули на языке C# и Visual Basic for Application, встроенном в MS Excel для обнаружения и удаления аномальных данных, программный модуль для корректировки выбросов, программный модуль для графического отображения, обнаруженных аномальных данных, модули для модификаций методов.

4. Проведённые численные эксперименты позволили сформулировать рекомендации по выбору одного из методов, наиболее подходящих в определенных ситуациях.

5. Обосновано, что предложенные в работе методы повышения качества парных регрессионных моделей одинаково эффективно используются как для линейных, так и для нелинейных регрессионных прогнозных уравнений с внутренней линейностью. Для этого исходное нелинейное регрессионное уравнение путём специальных преобразований приводится к линейному виду.

6. Показано, что предложенные методы обнаружения аномальных значений в исходных статистических данных позволяют наиболее быстро и точно проводить процедуру анализа данных на наличие грубых выбросов. Сокращение временных затрат на поиск и обработку аномальных данных достигается за счет уменьшения количества вычислительных операций (при этом сокращение может быть от $4*n$ до $2*10^s$, где n – количество исходных данных, s – число аномальных измерений).

7. Применение метода, состоящего в обнаружении и дальнейшем изменении значений аномальных измерений, позволяет обеспечить его реализацию для выборок малого объёма, поскольку в отличие от метода, где аномальные данные отбрасываются, в данном случае сохраняется исходное количество данных.

8. Представлены результаты применения методов для построения модели зависимости оборота розничной торговли непродовольственными товарами от среднедушевого денежного дохода населения в РФ, а также зависимостей из других предметных областей, подтверждающие эффективность предложенных в работе методов.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

- публикации в рецензируемых научных изданиях, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени кандидата и доктора наук ДНР:

1. **Рычка, О.В.** Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью / **О.В. Рычка** // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта», Донецк. – 2020. – №3(18). – С. 101-110.

2. **Рычка, О.В.** Разработка алгоритма реализации методов повышения качества регрессионных моделей, используемых при проектировании технических систем / **О.В. Рычка** // Научный журнал «Информатика и кибернетика», Донецк. – 2020. – № 3 (21). - С.13-19.

- публикации в рецензируемых научных журналах и изданиях, рекомендуемых ВАК Украины:

3. **Рычка, О.В.** Метод повышения качества прогнозных регрессионных моделей / **О.В. Рычка,** А.В. Смирнов // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. – 2010. – № 12. – С. 141-147.

4. Смирнов А.В. Новый метод улучшения качества прогнозных регрессионных моделей / А.В. Смирнов, **О.В. Рычка** // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. – 2011. – № 13. – С. 168-172.

5. **Рычка, О.В.** Разработка и анализ метода повышения точности прогнозных регрессионных моделей и его модификаций / **О.В. Рычка** // Питання прикладної математики і математичного моделювання: зб. наук. пр. / Вид-во Дніпропетр. нац. ун-ту, 2011. – С. 200-212

6. **Рычка, О.В.** Исследование эффективности применения метода повышения качества прогнозных регрессионных моделей и его модификаций / **О.В. Рычка** // Наукові праці Донецького національного технічного університету. Серія «Проблеми моделювання та автоматизації проектування»(МАП-11). Випуск 9 (179). – 2011. – С. 72-78.

- монографии и публикации в других изданиях:

7. Новые методы повышения точности прогнозных регрессионных моделей: монография / **О.В. Рычка** – LAP Lambert Academic Publishing, 2014. – 61 с.

8. **Рычка, О.В.** Описание и программная реализации методов обработки данных для повышения точности прогнозирования / **О.В. Рычка** // Научный журнал «Информатика и кибернетика», Донецк. – 2016. – № 1(3). – С.92-97.

– публикации по материалам научных конференций:

9. **Рычка, О.В.** Разработка и анализ метода повышения точности прогнозных регрессионных моделей и его модификаций / **О.В. Рычка** // Математичне та програмне забезпечення інтелектуальніх систем: матеріали VIII Міжнародної науково-практичної конференції – м. Дніпропетровськ – 10-12 листопада 2010 – С.196-197.

10. **Рычка, О.В.** Методы повышения точности прогнозирования при использовании регрессионных моделей / **О.В. Рычка** // Матеріали доповідей IV Міжнародної науково-практичної конференції «Сучасна інформаційна Україна: інформатика, економіка, філософія». Випуск Інформатика. – м. Донецьк – 13-14 травня 2010 – С.410-414.

11. **Рычка, О.В.** Метод повышения точности прогнозных регрессионных моделей с возможностью использования в современных компьютерных технологиях / **О.В. Рычка** // Донбас-2020: перспективи розвитку очима молодих вчених: матеріали V науково-практичної конференції, Донецьк – 25-27 травня 2010 р. – С.476-479.

12. **Рычка, О.В.** Методы обработки данных для повышения качества прогнозирования при использовании регрессионных моделей / **О.В. Рычка** // Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: Тези доповідей XIX міжнародної науково-практичної конференції, Ч.IV – м.Харків – 01-03 червня 2011 р. – С.74.

13. **Рычка, О.В.** Анализ эффективности новых методов для повышения точности регрессионных прогнозных моделей / **О.В. Рычка** // Проблеми інформатики і моделювання. Тезиси одинадцятої міжнародної науково-технічної конференції. – м.Харків, НТУ «ХПІ». – 2011. – С.66.

14. **Рычка, О.В.** Повышение эффективности прогнозирования при использовании линейных регрессионных моделей // Моделирование и компьютерная графика – 2011: материалы 4-й международной научно-технической конференции. – г. Донецк – 5-8 октября 2011 г. – С. 213-217.

15. **Рычка, О.В.** Описание и программная реализация методов обработки данных для повышения точности прогнозирования / **О.В. Рычка** // Информатика, управляющие системы, математическое и компьютерное моделирование в рамках II форума «Инновационные перспективы Донбасса» (ИУСМКМ-2016): VII Международная научно-техническая конференция – г.Донецк, ДонНТУ. – 26 мая 2016 г. – С.117-125.

16. **Рычка, О.В.** Улучшение прогнозных значений с использованием метода отбрасывания данных. / **О.В. Рычка** //Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2018): сборник научных трудов II Международной научно-практической конференции, Донецк, ДонНТУ. – Том. 1. – 14-18 ноября 2018 г. – С. 44-50.

17. **Рычка, О.В.** Программная реализация алгоритмов методов поиска и обработки аномальных измерений. / **О.В. Рычка**, А.В. Григорьев // Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике: материалы и

доклады VIII Международной научно-практической конференции, г. Азов. – Т7. № 1 (6). – 26-29 мая 2021 г. – С. 111-116.

Личный вклад соискателя в публикациях: [3] – обоснована актуальность разработки метода поиска аномальных измерений для регрессионных моделей; [4] – определены достоинства и недостатки существующих методов поиска аномалий, проведён сравнительный анализ наиболее известных методов с предлагаемым; [17] – обоснованы алгоритмы и предложена программная реализация методов поиска и обработки аномальных измерений;

АННОТАЦИЯ

Рычка О.В. **Совершенствование методов выявления и корректировки аномальных измерений для повышения качества линейных регрессионных моделей.** – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки) – ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Донецк, 2021 г.

В диссертационной работе дано теоретическое обоснование и приведено решение актуальной научно-практической задачи совершенствования методов поиска аномальных наблюдений и последующей их обработки, что позволяет повысить точность парных регрессионных моделей, используемых для прогнозирования в различных областях науки и техники.

Предложены и обоснованы усовершенствованные методы обработки статистических данных, эффективность которых не зависит от объема исходной выборки. Рассмотрены критерии оценки эффективности методов. Разработаны вычислительные алгоритмы предложенных методов на основе которых был создан комплекс программ, обеспечивающий решение таких задач, как нахождение аномальных данных, построение графиков для отображения обнаруженных выбросов, корректировка данных, определение наиболее подходящей зоны надёжности при которой достигается наибольшая эффективность, вывод результатов, подбор наилучшей модели.

Ключевые слова: аномальные данные, алгоритмы, выборка, корректировка данных, регрессионная модель, метод, анализ, эффективность, комплекс программ

ANNOTATION

Rychka O.V. **Improvement of methods for detecting and correcting anomalous measurements to improve the quality of linear regression models.** – Manuscript.

Candidate’s Thesis in Engineering Science by specialty 05.13.18 – Mathematical modeling, numerical methods and software complexes (technical

sciences) – STATE HIGHER EDUCATIONAL ESTABLISHMENT «DONETSK NATIONAL TECHNICAL UNIVERSITY», Donetsk, 2021

In the dissertation work, a theoretical justification is given and a solution of urgent scientific and practical problem of improving methods for searching of anomalous observations and their subsequent processing is given, which makes it possible to increase the accuracy of paired regression models used for forecasting in various fields of science and technology.

The improved methods of statistical data processing are proposed and substantiated, the efficiency of which does not depend on the size of the initial sample. Criteria for evaluating the effectiveness of the methods are considered. The computational algorithms of the proposed methods were developed, on the basis of which a complex of programs was created that provides the solution of such problems as finding anomalous data, plotting graphs for displaying detected outliers, correcting data, determining the most suitable reliability zone at which the greatest efficiency is achieved.

Keywords: anomalous data, algorithms, sampling, data correction, regression model, method, analysis, efficiency, complex of programs, output of results, selection of the best model.