УДК 519.254

Практическое использование методов поиска и корректировки аномалий для построения точных прогнозов

О. В. Рычка
Донецкий национальный технический университет, г. Донецк
olga_rychka@mail.ru

Аннотация
В данной статье описаны особенности и основные функциональные возможности разработанной автоматизированной системы поиска и корректировки аномальных измерений. Проанализирована работа предложенных автором методов и доказана их эффективность на реальных статистических данных. Тестирование показало, что использование новых методов для поиска и обработки аномалий даёт положительные результаты, а полученный по найденной модели прогноз практически совпадает с реальными данными. Намечены направления дальнейших исследований.

Введение

Одной из важных проблем интеллектуального анализа данных является обнаружение значений, которые не соответствуют общей модели поведения, т.е. аномалий. Существуют различные причины выявления аномальных данных. Например, для своевременного обнаружения изменений в системе. Ещё одной существенной причиной обнаружения выбросов является необходимость построения надёжной математической модели для прогнозного анализа данных. В этом случае, результаты анализа без предварительной обработки данных могут быть искажены, что негативно повлияет на качество полученной прогнозной модели.

О важности поиска аномальных измерений говорит и наличие государственных стандартов в этой области, в частности, ГОСТ 8.736-2011 [1] и ГОСТ Р ИСО 16269-4-2017 [2].

Одним из главных инструментов анализа экспериментальных данных и обнаружения закономерностей в них является регрессионный анализ. Наиболее простыми и распространёнными уравнениями регрессии являются линейные. Они применяются в различных областях науки и отраслях промышленности (например, топливной, угольной и др.).

В настоящее время существует множество методов обнаружения аномалий для различных типов данных [3-8]. Однако, большинство предлагаемых решений применимы только для временных рядов, а также зависят от объема исходных данных. Поэтому задача разработки и реализации алгоритма поиска аномальных измерений в исходных статистических данных является актуальной. В работах [9] и [10] автором были предложены улучшенные методы обнаружения и обработки аномалий, описаны критерии оценки эффективности разработанных методов. Целью данной статьи является практическая оценка, разработанных автором алгоритмом поиска и последующей обработки аномальных данных.

Функциональные возможности программного комплекса

Учитывая большой объем анализируемых данных, существует необходимость разработки автоматизированных систем анализа и обработки данных. Поэтому был разработан комплекс программ, с помощью которого было проведено множество экспериментов с различными выборками.

Комплекс программ включает набор программных модулей, состоящий из главного интерфейса, написанного на языке C# в среде Microsoft Visual Studio 2017 и макросов, разработанных с применением языка программирования Visual Basic for Applications для Microsoft Excel, поскольку при необходимости автоматизировать обработку данных в MS Excel, данный язык является наиболее удобным.

Разработка осуществлялась исходя из того, что программное приложение должно выполнять следующие основные функции:

- ввод данных вручную или с помощью их загрузки из файла формата Microsoft Excel;
- предобработка данных (проверка на правильность ввода, сортировка);
- поиск аномальных измерений;
- отбрасывание или корректировка найденных аномальных значений;
- расчет коэффициентов эффективности;
- выбор наиболее точной модели;
- вывод результатов.

Разработанный программный комплекс состоит из следующих модулей:

- модуль для обнаружения и удаления аномальных данных;
- модуль для корректировки аномальных данных;
- модуль для графического отображения обнаруженных аномальных данных;
- модули для реализации модификаций методов.

Опишем подробнее работу программного приложения, исходя из функций, представленных выше.

При вводе данных, учитывается заданное пользователем количество наблюдений. Таблица с выгружаемыми данными должна иметь следующий формат: первая строка содержит название переменных (X в первом столбце и Y во втором), последующие строки – исходные статистические данные (рис. 1).

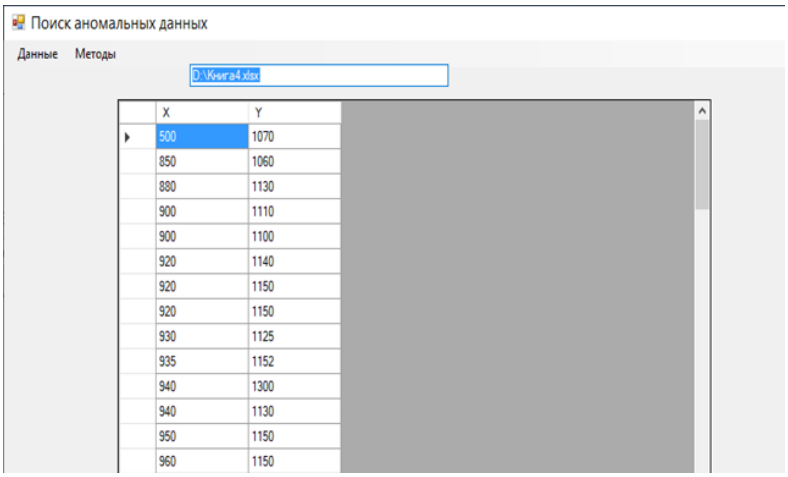

Рисунок 1 – Вид окна с исходными данными

Под функцией предобработки данных подразумевается проверка на правильность ввода и последующая сортировка статистических данных. В приложении организован контроль ошибок. Если статистические данные введены некорректно, то приложение сигнализирует об этом пользователю с помощью соответствующего сообщения, после чего возможно изменение данных.

После ввода корректных данных, пользователь определяется с тем, будет ли он использовать один из методов (метод отбрасывания аномальных данных, метод корректировки аномальных данных) или выберет одну из модификаций и нажимает соответствующую кнопку меню, что приводит к программному осуществлению всех необходимых расчетов и выводу итоговой таблицы (рис. 2), которая включает следующие параметры эффективности метода для каждого значения вероятности попадания в заданную область:

а) значения коэффициентов детерминации R^2;

б) величины доверительных интервалов;

в) величины смещений;

г) количество измерений (исходное и после отбрасывания);

г) точность;

д) коэффициенты нового линейного регрессионного уравнения.

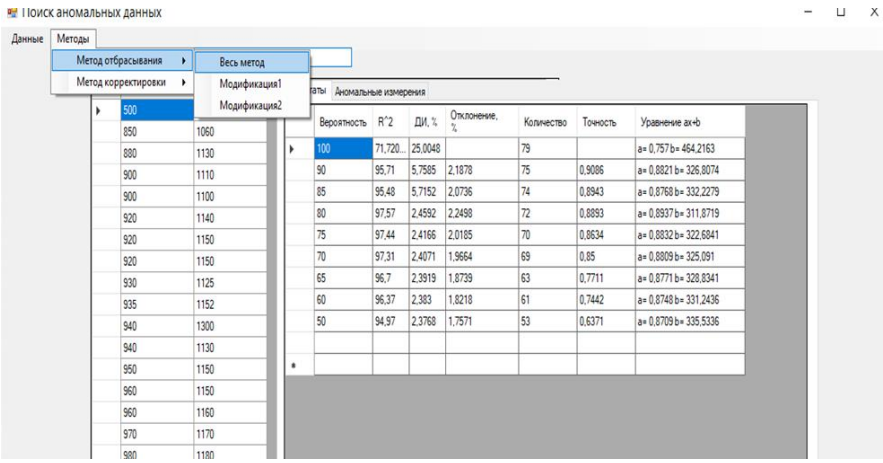

Рисунок 2 – Вид окна после выбора пункта меню «Метод отбрасывания-Весь метод»

При необходимости, можно просмотреть найденные в результате расчетов аномальные измерения для каждого значения вероятности, перейдя на соответствующую вкладку (рис. 3)

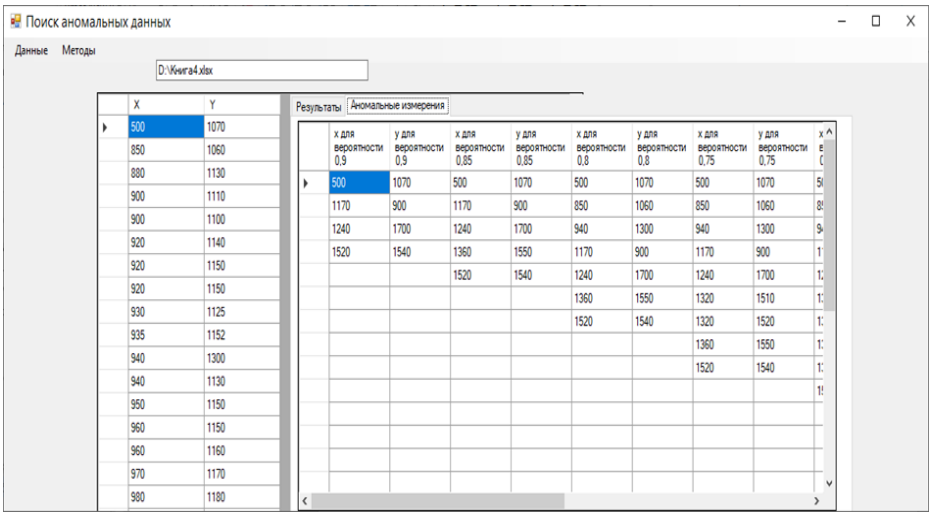

Рисунок 3 – Вид окна при выборе вкладки «Аномальные измерения»

Таким образом, все результаты отображаются в соответствующем окне программного комплекса, помимо этого результаты, а также промежуточные расчеты могут быть записаны в Excel-файл, который автоматически создаётся средствами C#. Кроме того, пользователь может не только рассчитывать показатели, но и заходить в редактор среды Microsoft Excel и работать непосредственно с ним.

Результаты применения методов поиска и обработки аномалий

Для верификации разработанных и описанных в работах [9], [10] методов рассмотрим пример на реальных данных – зависимость оборота розничной торговли непродовольственными товарами, млн. руб. от среднедушевого денежного дохода населения в РФ, руб./месяц. Значения взяты поквартально с 2013 по 2019 гг. на официальном сайте Федеральной службы государственной статистики. Величина коэффициента детерминации R^2 составляет 0,85, что говорит о достаточно тесной линейной зависимости между переменными.

Чтобы провести эксперимент и проверить работу методов, изменим 4 значения. После замены, величина коэффициента детерминации значительно уменьшилась и составила всего 0,55. Это доказывает, что аномальные данные оказывают большое влияние на исходную модель.

Все расчеты проводились с использованием разработанного программного комплекса. Результаты использования метода повышения качества парных линейных регрессионных моделей за счёт отбрасывания данных и его модификаций представлены в табл. 1.

Таблица 1 – Результаты

Вероятность попадания в область	Количество отброшенных точек			Метод, R^2	1-я модиф. R^2	2-я модиф. R^2	Метод. T	1-я модиф. T	2-я модиф. T
	Метод	1-я модиф.	2-я модиф.						
100	0	0	0	0,55	0,55	0,55			
90	2	2	2	0,69	0,69	0,61	0,64	0,64	0,561
85	4	4	2	0,789	0,81	0,61	0,67	0,68	0,561
80	4	5	2	0,789	0,85	0,61	0,67	0,69	0,561
70	5	6	3	0,79	0,854	0,57	0,64	0,66	0,51
65	7	6	5	0,83	0,854	0,5	0,61	0,65	0,44
60	8	7	7	0,8	0,876	0,4	0,56	0,65	0,3
50	9	8	8	0,77	0,87	0,36	0,51	0,61	0,24

Как видно из таблицы, при использовании предложенного метода были получены положительные результаты. Построение области надёжности при доверительной вероятности попадания данных в эту область равной 0,9 позволило обнаружить аномальные измерения. При применении самого метода, в этом случае, было выявлено два аномальных наблюдения. Их отбрасывание позволило увеличить коэффициент детерминации с 0,55 до 0,69. Отбрасывание всего 4-х значений позволило увеличить величину коэффициента детерминации до 0,79, что также свидетельствует об адекватности, найденной при этом линейной регрессионной модели. Хочется особо отметить, что данные 4 наблюдения – это те наблюдения, которые были изменены в исходной выборке для проведения эксперимента. При этом соотношение коэффициента детерминации и значения точности в данном случае являются наилучшими, поэтому дальнейшее отбрасывание нецелесообразно.

Результаты применения метода на основе переноса данных, а также его модификаций представлены в табл. 2.

Таблица 2 – Результаты применения метода, основанного на корректировке данных

Вероятность попадания в область	Метод, R^2	1-я модиф. R^2	2-я модиф. R^2
0	0,55	-	-
90	0,71	0,61	0,69
85	0,75	0,65	0,7
80	0,77	0,69	0,69
70	0,79	0,74	0,66
65	0,791	0,77	0,62
60	0,792	0,79	0,6
50	0,78	0,82	0,5

В результате использования данного метода, коэффициент детерминации, также вырос до 0,79, при этом были изменены пять значений статистических данных. Максимальное и минимальное значение независимой переменной X изменились с 15800 и 58848 на 22228,43, и 39907,18 соответственно. Если сравнить их с реальными данными (21800 и 38848), то можно заметить, что они стали очень близки.

Таким образом, использование предложенных автором методов привело к:

- обнаружению всех аномальных и ненадёжных измерений;
- росту коэффициента детерминации R^2, который достигает 15-30%;
- величина доверительного интервала уменьшается до 2 раз;
- построению нового линейного регрессионного уравнения, которое является надёжнее, чем исходное, что позволяет построить более точный прогноз;
- число элементарных операций менее 1000, для исходного объема выборки равного 27.

На рис. 4 представлены сравнения линий регрессии, построенных по уравнениям с реальными данными (при этом R^2=0,85) и аномальными (R^2=0,55). Как видно, из рисунка есть значительное отклонение между двумя линиями.

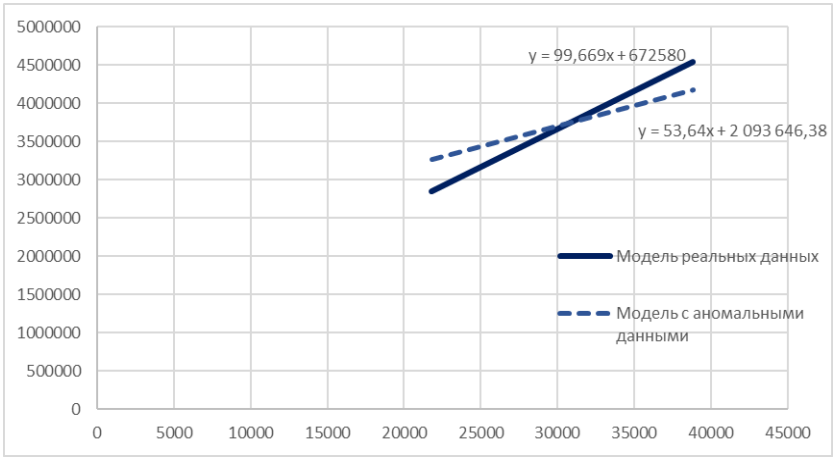

Рисунок 4 – Сравнения линий регрессии, построенных по реальным данным и с аномальными

На рис. 5 представлено сравнение линии регрессии, полученной по реальным данным и по данным после отбрасывания четырёх значений. Смещение уравнения, построенного по реальным данным при максимальном значении переменной X, равном 38848 от уравнения, полученного после отбрасывания данных, составило всего 1.14%.

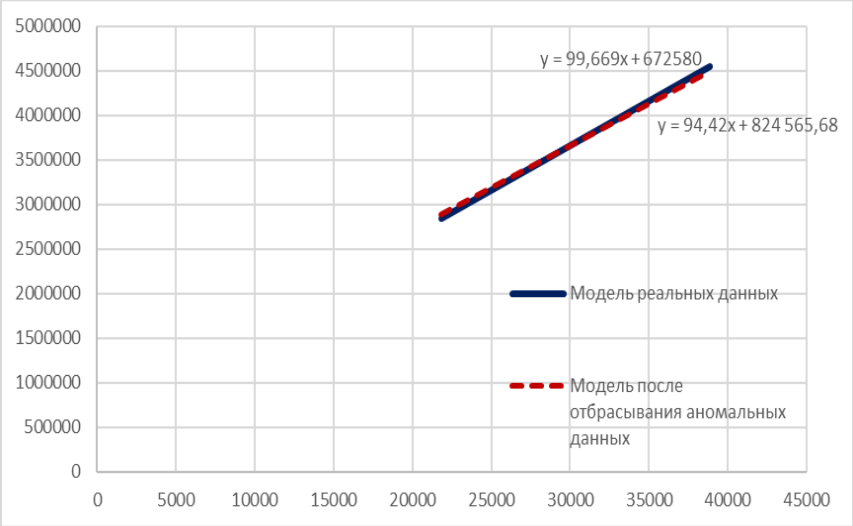

Рисунок 5 – Модель реальных данных и модель после отбрасывания аномальных данных

Таким образом, можно сделать вывод, что применение метода, основанного на отбрасывании данных, привело к получению модели идентичной реальной регрессионной модели, а значит она может быть использована для прогнозирования, поскольку результаты, полученные по данной модели будут точными и надёжными.

Сравнивая модель, полученную по реальным данным и модель по данным после корректировки. Можно увидеть, что линии регрессии практически совпадают. Отклонение составляет 0,41% (рис. 6).

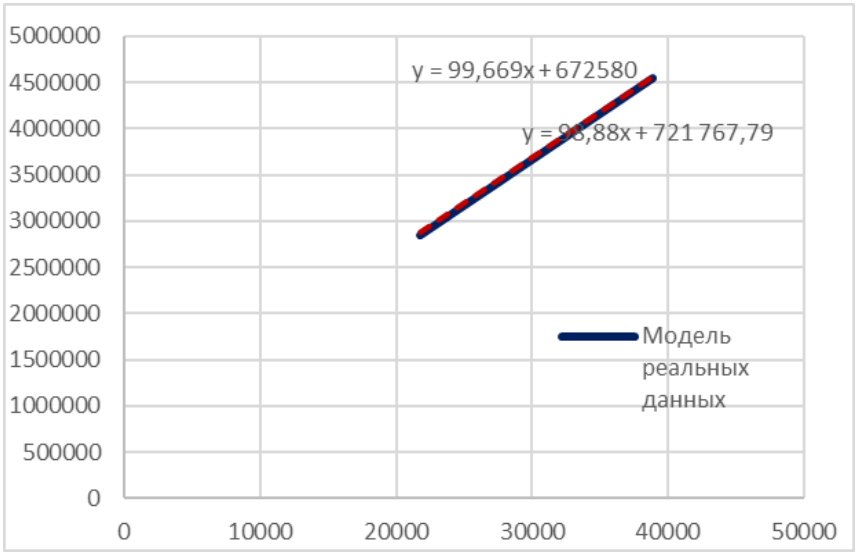

Рисунок 6 – Модель реальных данных и модель после корректировки аномальных данных

Для проверки адекватности полученных моделей по ним был осуществлён прогноз на 4 квартал 2019 г. и 2020 г., и сравнение с реальными данными с сайта Федеральной службы статистики. В таблице 3 представлен прогноз:

- по исходной модели, полученной по реальным данным;
- по модели с аномальными данными;
- по модели после отбрасывания аномальных данных;
- по модели после корректировки аномальных данных.

Таблица 3 – Прогнозные значения оборотов розничной торговли непродовольственными товарами

Период прогноза	X	Прогноз по исходной модели	Прогноз по модели с аномалиями	Прогноз по модели после отбрасывания аномалий	Прогноз по модели после корректировки аномалий	Реальные данные
4 квартал 2019г.	41328	4791741,41	4310480,3	4726755,44	4808280,44	4873283,6
4 квартал 2020г.	42543	4912840,46	4375652,9	4841475,74	4928419,64	5048843,7
Отклонение, % для 2019 г.	-	1,67	11,55	3	1,33	-
Отклонение, % для 2020 г.	-	2,7	13,33	4	2,4	-

Как видно из таблицы, по моделям, построенным после обработки аномальных данных получены достаточно точные прогнозы. При использовании модели, полученной после корректировки 5 значений отклонение от прогноза, получилось даже меньше, чем при использовании модели, построенной по реальным данным (1,33% для 2019 г. и 2,4% для 2020 г.).

Выводы

С использованием языков программирования C# и Visual Basic for Application была разработана автоматизированная система поиска и корректировки аномальных данных по предложенным автором методам, которая позволяет сократить трудоёмкость проводимых испытаний.

В работе был выполнен анализ эффективности рассматриваемых методов на реальных статистических данных с применением автоматизированной системы. Тестирование показало, что использование новых методов для поиска и обработки аномалий даёт положительные результаты, поскольку были обнаружены все аномальные данные в рассматриваемом примере, а полученный по найденной модели прогноз практически совпадает с реальными данными (отклонение менее 2,5%).

В дальнейших исследованиях будет проводиться оценка эффективности результатов предложенных методов на практических предметных областях с целью построения точных прогнозов. Помимо этого, планируется расширение функциональности автоматизированной системы для многомерных регрессионных моделей.

Литература

1. ГОСТ 8.736-2011 "Государственная система обеспечения единства измерений. Измерения прямые многократные. Методы обработки результатов измерений. Основные положения".

2. ГОСТ Р ИСО 16269-4-2017 Статистические методы. Статистическое представление данных Часть 4. Выявление и обработка выбросов.

3. Кириченко, А. В. Математические модели и методы анализа и прогнозирования: предварительная обработка результатов эксперимента, проверка статистических гипотез, корреляционный анализ, парный регрессионный анализ: учебное пособие / А. В. Кириченко и др. - Саратов: КУБиК, 2019. - 259 с.

4. Chandola, V. Anomaly detection: A survey / V. Chandola, A. Banerjee, V. Kumar // ACM Comput. Surv., 2009. – № 41, 3, Article 15. – 58 p.

5. Попукайло, В. С. Обнаружение аномальных измерений при обработке данных малого объема // Технология и конструирование в электронной аппаратуре, 2016. – № 4-5. – С. 42-46.

6. Wilson, J. Holton. Regression Analysis: Understanding and Building Business and Economic Models Using Excel, 2nd Edition / J. Holton Wilson, Barry P. Keating, Mary Beal. — New York, USA, Business Expert Press, LLC, 2016. — 205 p.

7. Кузовлев, В. И. Выявление аномалий при прогнозном анализе данных / В. И. Кузовлев, А. О. Орлов // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение, 2016. – № 5. – С.75-85.

8. Кузовлев, В. И. Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в

системах поддержки принятия решений / В. И. Кузовлев, А. О. Орлов // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. Журн, 2012. - № 9. - URL: https://cyberleninka.ru/article/n/metod-vyyavleniya anomaliy-v-ishodnyh-dannyh-pri-postroenii-prognoznoy-modeli-reshayuschego-dereva-v-sistemah-podderzhki-prinyatiya/viewer.

9. Рычка, О. В. Разработка алгоритма реализации методов повышения качества регрессионных моделей, используемых при проектировании технических систем // Информатика и кибернетика. – Донецк: ДонНТУ, 2020. - № 3 (21). - С.42-48.

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

Рычка О.В. Практическое использование методов поиска и корректировки аномалий для построения точных прогнозов. *В данной статье описаны особенности и основные функциональные возможности разработанной автоматизированной системы поиска и корректировки аномальных измерений. Проанализирована работа предложенных автором методов и доказана их эффективность на реальных статистических данных. Тестирование показало, что использование новых методов для поиска и обработки аномалий даёт положительные результаты, а полученный по найденной модели прогноз практически совпадает с реальными данными. Намечены направления дальнейших исследований.*

Ключевые слова: *аномальные измерения, автоматизированная система, поиск аномалий, корректировка аномалий, эффективность, прогноз.*

Rychka O.V. Practical use of methods for finding and correcting anomalies to build accurate forecasts. *This article describes the features and main functionality of the developed automated system for searching and correcting anomalous measurements. The work of the methods proposed by the author is analyzed and their effectiveness is proved on real statistical data. Testing has shown that the use of new methods for finding and processing anomalies gives positive results, and the forecast obtained from the found model practically coincides with real data. The directions of further research are outlined.*

Key words: *anomalous measurements, automated system, anomaly search, anomaly correction, efficiency, forecast.*

Статья поступила в редакцию 23.11.2022
Рекомендуется к публикации профессором Зори С. А.