Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Алгоритм и программа для отсеивания аномальных результатов измерений

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Гарина Андрея Александровича

Научный руководитель:
д.т.н., проф. кафедры
информатики и
программирования А. С. Фалькович

(подпись, дата)

Зав. кафедрой информатики
и программирования,
к.ф.-м.н. М.В. Огнева

подпись, дата

Саратов 2017

Введение

При статистической обработке данных, полученных при измерениях, часто возникает проблема отбраковки результатов, находящихся за пределами возможной области значений случайной величины. Для нормально распределенной случайной величины часто пользуются интервалом, ограниченным полутора, двумя или тремя значениями среднего квадратичного отклонения (σ). Задача усложняется, когда распределение случайной величины не является нормальным или «выбросы» на два-три порядка превышают выборочное среднее.

Если распределение не является нормальным, алгоритм, основанный на применении среднего и среднего квадратичного отклонения, нельзя считать обоснованным. С другой стороны, аномально большое значение выброса «портит» среднее и среднее квадратичное отклонение. В случае, если таких аномально больших выбросов много, искаженные значения среднего и среднего квадратичного отклонения могут препятствовать выявлению последующих выбросов.

Поэтому актуальной является задача разработки алгоритма и программы для отсеивания аномально больших результатов измерений, критерия в которых не зависит от среднего.

Цель работы – разработать алгоритм и программу для отсеивания аномально больших результатов измерений.

Для достижения цели необходимо решить следующие задачи:

- разработать программу, реализующую алгоритм нахождения и отсеивания аномально больших чисел в выборке, основанный на применении квантилей, и позволяющую считывать из файла и обрабатывать одновременно несколько серий данных

- провести ряд расчетов и убедиться в пригодности данного алгоритма.

Методологические основы алгоритмов для отсеивания аномальных результатов измерений представлены в работах С.А. Айвазяна, И.С. Енюкова, Л.Д. Мешалкина [2], А. Афифи, С. Эйзена [3],В.Е. Гмурмана [4], Дж. Полларда [5], Дж. Тейлора [6], В. И. Марчука, С.В. Токаревой [10,11]. Обзор методов приведен в работах В. И. Марчука [10], J. Han, M. Kamber, J. Pei [8] C.C. Aggarwal [9].

Практическая значимость бакалаврской работы заключается в разработке программы, реализующей алгоритм нахождения и отсеивания аномально больших чисел в выборке, основанный на применении квантилей, и позволяющей считывать из файла и обрабатывать одновременно несколько серий данных.

Структура и объём работы. Бакалаврская работа состоит из введения, основного раздела, заключения, списка использованных источников и приложения. Общий объем работы – 59 страниц, из них 48 страниц – основное содержание, включая 21 рисунок и 4 таблицы, цифровой носитель в качестве приложения, список использованных источников информации – 20 наименований.

Основное содержание работы

Чтобы найти и исключить выбросы (ошибочные результаты измерений, слишком большие или слишком маленькие) среди значений случайной величины, часто пользуются интервалом, ограниченным полутора, двумя или тремя значениями среднего квадратичного отклонения (σ). Однако, такой подход, справедливый для нормально распределенной случайной величины, не вполне обоснован, когда распределение случайной величины не является нормальным. Критерии отбраковки, связанные со средним и средним квадратичным отклонением, не подходят также в случае, если или «выбросы» в десятки и сотни раз превышают выборочное среднее.

На рисунках 1 и 2 приведен один и тот же ранжированный ряд наблюдений, рисунки отличаются только масштабом оси ординат: на рисунке 1 максимальное значение 1,8 миллиона, на рисунке 2 – 16000.

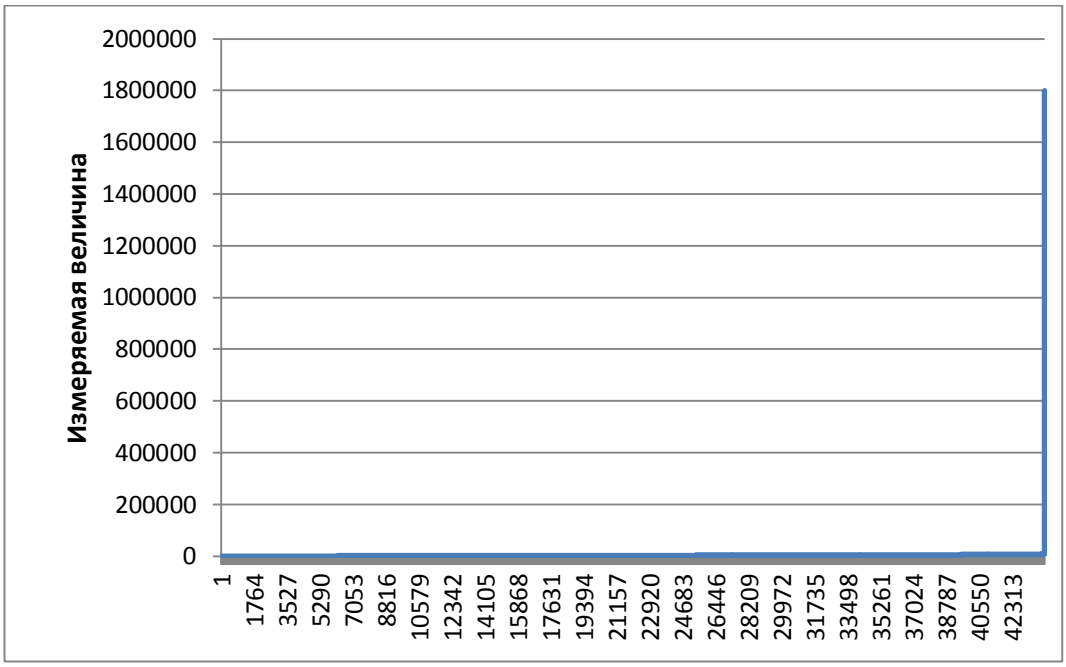

Рисунок 1 - Ранжированный ряд 44 тыс. значений измеренной величины. 24 значений аномальных выбросов не отбракованы и выглядят как несколько точек, все остальные значения сливаются с осью абсцисс

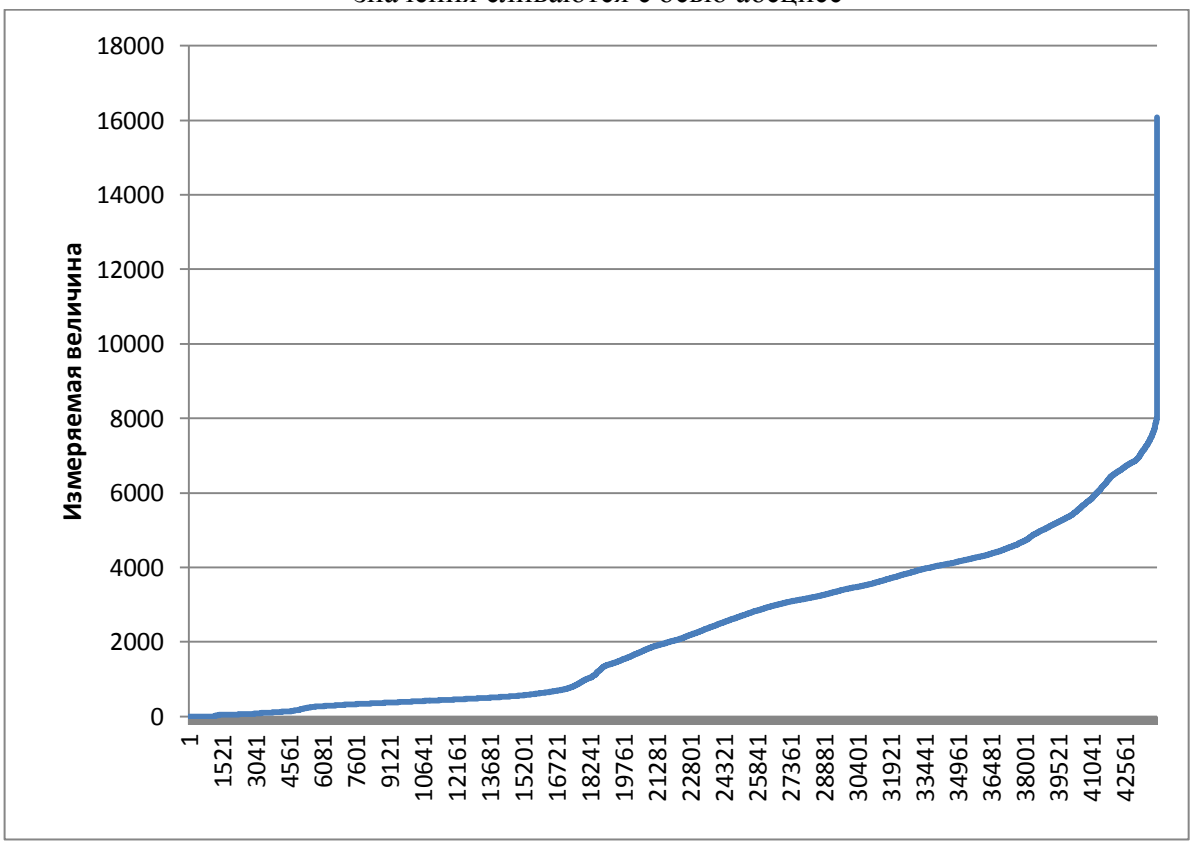

Рисунок 2 - Ранжированный ряд 44 тыс. значений измеренной величины. Аномальные выбросы исключены.

Поскольку очень большие значения «выбросов» влияют на величины среднего

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

и среднего квадратичного отклонения, ограничение приемлемых данных интервалом

$$(\bar{x} - 3\sigma; \bar{x} + 3\sigma) \qquad (2)$$

где $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$ (3)

непригоден, так как огромные величины выбросов влияют на среднее и среднее квадратичное отклонение. На практике в этих случаях предлагают применять этот критерий несколько раз, последовательно исключая выбросы и уменьшая интервал, пока не прекратится отбраковка информации.

Однако если закон распределения случайной величины отличается от нормального, применение среднего и среднего квадратичного отклонения для отсеивания выбросов недостаточно обоснованы. В этом случае более подходящими являются оценки, не связанные с параметрами распределения (непараметрические).

Алгоритм отсеивания аномальных выбросов основан на применении оценки зависимости величины квантилей K_x от их уровней x.

Данный метод предназначен только для исключения аномальных выбросов, и не избавляет от необходимости дальнейшей подготовки выборки с помощью традиционных методов сглаживания данных.

Квантиль уровня x (x-квантиль) случайной величины y с функцией распределения $F(y)$ – это число K_x, удовлетворяющее условию

$F(K_x) = x$.

Эмпирическое значение квантиля K_x для некоторой выборки – это число, стоящее на $(x \cdot n)$ месте в ранжированной, то есть упорядоченной по возрастанию, выборке, где $0 < x < 1$ - уровень квантиля, n - объем выборки.

Несмотря на то, что распределение случайной величины в наших выборках не является нормальным, зависимость $K_x(x)$ практически во всех случаях при отсутствии в выборке аномальных выбросов является экспоненциальной.

Программа фильтрации аномальных выбросов в экспериментальных данных

В настоящей работе исходный код программы написан на языке программирования C# платформы .NET.

Решение состоит из трех проектов:

- BL (Business Level) основная логика программы, этот проект отвечает за построение регрессионных зависимостей и всех необходимых вычислений;
- DAL(Data Access Level) проект отвечает за загрузку информации с файла (excel) в определенном формате;
- WPF(Windows Presentation Foundation) - этот проект отвечает за отображение результирующей информации.

Взаимодействие с программой

Дальнейшая работа с программой осуществляется следующим образом. При нажатии на кнопку «Начать работу» появляется кнопка «обзор», а при нажатии на кнопку «обзор» появляется стандартное окно выбора файла. После выбора файла, программа достает из файла данные и делает все последующие необходимые вычисления с данными. В результате программа записывает в отдельный компонент TabControl (представляет из себя вкладку

с название вычисления и результатом) одно из вычисленных значений, а именно – фактические квантили, квантили, вычисленные на основе регрессии, уровень критического квантиля, самое большое не отброшенное, самое маленькое отброшенное, количество отброшенных. Также после вычисления очищенные данные от выбросов записываются в текстовый файл с именем соответствующим названию вычисления. Результат обработки данных предоставлен в соответствии рисунком 3.

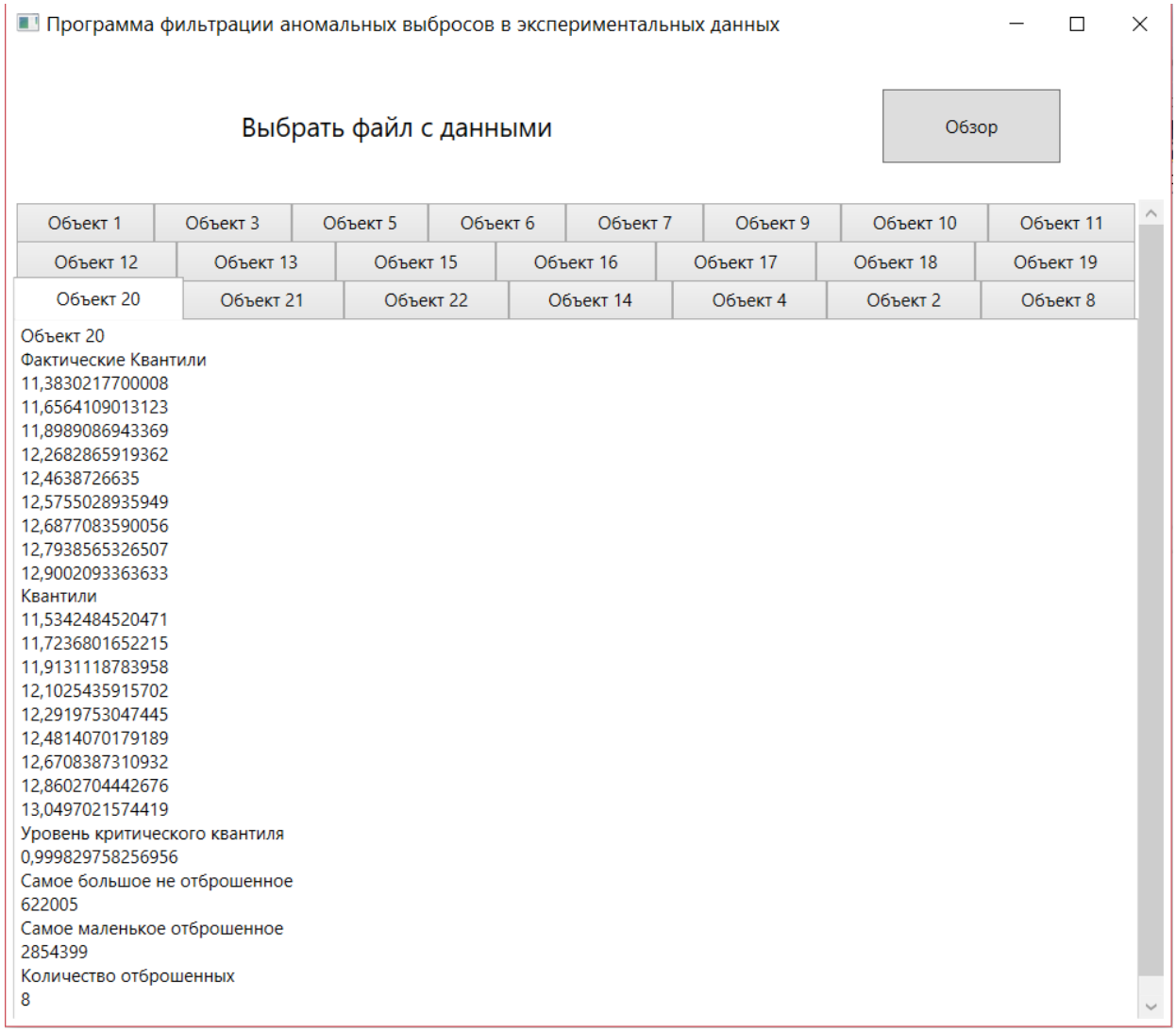

Рисунок 3 - Результат обработки данных

Результаты применения программы

Обработка нескольких рядов экспериментальных данных с помощью разработанной программы показала, что метод обладает достаточно высокой точностью.

После подсчета фактических квантилей с уровнями 0,1; 0,2; … 0,9 программа вычисляет логарифмы квантилей (на рис. 4 изображены точками). По этим величинам строится линейная регрессионная зависимость. На рис. 4 эта зависимость отображается сплошной линией.

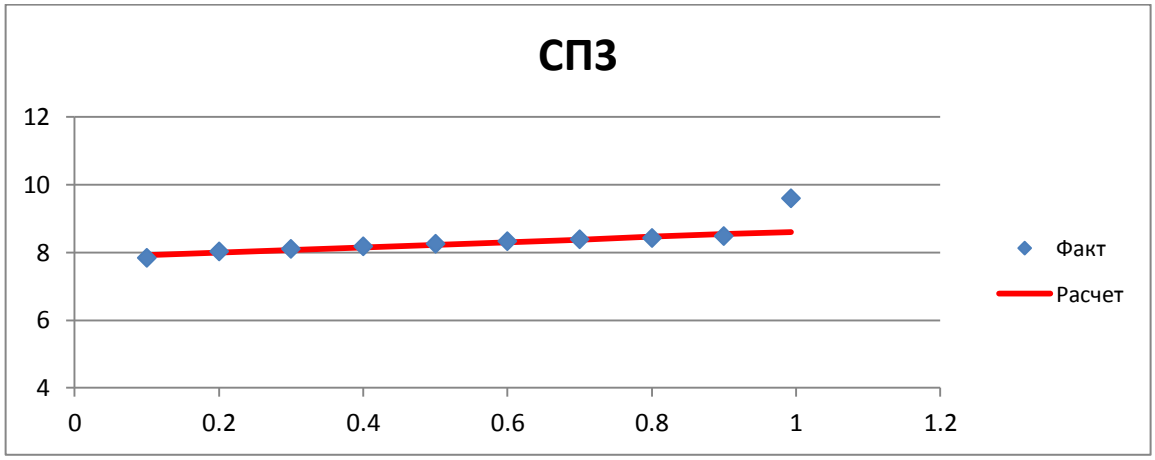

Рисунок 4 - Фактические и расчетные значения логарифмов квантилей для объекта СП3.

Далее эта зависимость рассчитывалась для уровней, больших чем 0,9 до тех пор, пока разность между логарифмом квантиля соответствующего уровня и расчетным значением становилась больше 0,1. Для объекта СП3 таким уровнем был 0,992833. Логарифм фактического квантиля этого уровня равен 9,60606, а расчетное значение – 8,605005. Разность составила 1,001. В выборке объекта СП3 было 1462 значения. Квантиль уровня 0,992833 соответствует $1462 \times 0{,}992833 \cong 1451$ значению ранжированного ряда. Таким образом, должны быть отброшены последние 10 значений ранжированного ряда наблюдений по объекту СП 3. Результаты расчетов приведены в таблице 1, отброшенные значения – в таблице 2.

Таблица 1 - результат расчета по исходному файлу, содержащему данные четырех объектов.

	СП1	СП2	СП3	СП4
Уровни	Логарифмы фактических квантилей			
0,1		4,488636	7,838738	4,736198
0,2	5,075174	4,60517	8,029433	4,969813
0,3	5,298317	5,298317	8,101678	5,075174
0,4	5,811141	5,811141	8,188689	6,35437
0,5	7,17012	6,336826	8,263075	7,096721
0,6	7,862882	6,590301	8,333511	7,31322
0,7	7,972466	6,784457	8,386629	7,536364
0,8	8,039157	6,907755	8,433812	7,772753
0,9	8,082711	7,069874	8,495152	7,956827
Уровни	Логарифмы расчетных квантилей			
0,1	2,904252	4,580392	7,924865	4,722549
0,2	3,714633	4,931699	8,001043	5,175692
0,3	4,525013	5,283007	8,07722	5,628834
0,4	5,335394	5,634315	8,153398	6,081976
0,5	6,145774	5,985622	8,229576	6,535118
0,6	6,956155	6,33693	8,305754	6,98826
0,7	7,766535	6,688237	8,381932	7,441403
0,8	8,576916	7,039545	8,458109	7,894545
0,9	9,387296	7,390853	8,534287	8,347687
Уровень критического квантиля	0	0,992833	0,992833	0,998294
Максимальное неотброшенное значение	0	4200	5977	5000
Минимальное отброшенное значение	0	7400	23000	100711
Количество отброшенных значений	0	10	10	2
		3,513076	0,761778	4,531422
		7,716981	8,605005	8,793096

По объекту СП1 выбросов не было. В этом случае в файле output.txt на соответствующих местах записываются нули.

Таблица 2. Последние 20 значений ранжированного ряда наблюдений по объекту СП 3.

СП 3
…
5728
5728
5769
5787
5794
5805
5849
5856
5954
5977
23000
37500
45600
47600
47700
48300
49200
49400
49600
53000

Граница отбрасываемых наблюдений проходит между числами 5977 и 23000.

Программа была также протестирована на нехарактерной выборке (объект ПВ 2), в которой есть и аномально большие выбросы, и «обыкновенные», не аномально большие, то есть превышающие реальные значения всего в 2-3 раза (Рисунки 19, 20). Объем этой выборки 44056 значений.

Для объекта ПВ 2 программа определила уровень критического квантиля 0,995623. Логарифм фактического квантиля этого уровня равен 6,906581, а расчетное значение – 6,277216. (таблица 3, рисунок 21).

Таблица 3. Результат расчета по объекту ПВ2.

	ПВ2
Уровни	Логарифмы фактических квантилей
0,1	4,962845
0,2	5,081404
0,3	5,214936
0,4	5,407172
0,5	5,605802
0,6	5,717028
0,7	5,866468
0,8	5,993961
0,9	6,082219
Уровни	Логарифмы расчетных квантилей
0,1	4,959442
0,2	5,106577
0,3	5,253712
0,4	5,400847
0,5	5,547982
0,6	5,695116
0,7	5,842251
0,8	5,989386
0,9	6,136521
Уровень критического квантиля	0,995623
Максимальное неотброшенное значение	994
Минимальное отброшенное значение	1000
Количество отброшенных значений	136

Соответственно, было отброшено 136 значений. При этом максимальное не отброшенное значение – 994, минимальное отброшенное значение – 1000. То есть алгоритм и программа отсекли все аномально большие выбросы и значительную часть «не аномальных». Об этом можно судить по рисунку 21 и таблице 4. Оставшуюся часть «не аномальных» выбросов, в 2-3 раза превышающих значения измеряемого параметра, можно убрать обычными методами, например, сглаживая величины, не попадающие

в интервал «среднее плюс-минус три средних квадратичных отклонения». Таким образом, программа показала свою пригодность и на этом примере.

Работа алгоритма и программы тестировалась на выборках объема несколько десятков тысяч (от 23 тыс. до 44 тыс.) значений и на выборках объема 1460 значений. Большие выборки показали более устойчивый результат. При работе с выборками в несколько десятков тысяч значений отсутствовали ошибки как первого рода – когда отбрасывалось значение, не являющееся аномальным выбросом, так и второго рода – когда не отбрасывался аномальный выброс.

ЗАКЛЮЧЕНИЕ

В результате выполнения работы были разработаны алгоритм и программа для отсеивания аномально больших результатов измерений.

Разработана программа, реализующая алгоритм нахождения и отсеивания аномально больших чисел в выборке, основанный на применении квантилей. Программа позволяет считывать из файла и обрабатывать одновременно несколько серий данных. Результаты расчетов и исходные данные с скорректированными значениями выбросов выводятся в отдельные файлы. С помощью разработанной программы проведены расчеты, свидетельствующие о пригодности данного алгоритма.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Справочник по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана. – М. : Финансы и статистика, 1990. Т. 2. 526 с.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М. : Финансы и статистика, 1983. 472 с.
3. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. – М. : Мир, 1982. 488 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. Пособие для вузов / В.Е. Гмурман. – 9-е изд., стер. – М.: Высш. Шк., 2003. – 479с.
5. Поллард Дж. Справочник по вычислительным методам статистики / Пер. с англ. – М.: Финансы и статистика, 1982. 344 с.
6. Тейлор Дж. Введение в теорию ошибок. Пер. с англ. – М.: Мир, 2009. – 272 с.
7. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2011. – 816 с.
8. Han J., Kamber M., Pei J. Data Mining Concepts and Techniques. Third Edition. Waltham, MA, USA : Morgan Kaufmann Publishers is an imprint of Elsevier, 2012. 703 p.
9. Aggarwal C.C. Outlier Analysis. New York : Springer Science; Business Media 2013. 446 p.
10. Марчук, В. И. Способы обнаружения аномальных значений при анализе нестационарных случайных процессов / В. И. Марчук, С.В. Токарева – Шахты: ЮРГУЭС, 2009. – 210 с.
11. Пат. № 2302655 МПК G06F15/00 (2006.01) Способ обнаружения аномальных измерений без оценки функции тренда и устройство, его реализующее// Марчук В.И., Шерстобитов А.И., Воронин В.В., Токарева С.В. (Россия). - 2005132148/09; 17.10.2005 Заявл. 17.10.2005; опубл. 10.07.2007, Бюл. №19. 14 с.

12. Корн Г. Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. - М.: Наука, 1974. – 832 с.
13. Справочник по специальным функциям // под ред. М. Абрамовица, И. Стиган. М. : Наука. Главная редакция физико-математической литературы, 1979. 832 с.
14. Фалькович, А. С Непараметрический метод определения и отсеивания аномальных результатов измерений // Компьютерные науки и информационные технологии. Мат-лы междунар. научн. конф. - Саратов, 2016 - С. 429-431
15. Гутер, Р.С. Элементы численного анализа и математической обработки результатов опыта // Р.С. Гутер, Б.В. Овчинский. - М.: Гос. изд-во физ-мат. лит-ры, 1962. – 356 с.
16. Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. — СПб: Питер, 2001. — 368 с.: ил. (Серия «Библиотека программиста») ISBN 5-272-00355-1
17. Нейгел К. C# 4.0 и платформа .NET 4 для профессионалов: Пер. с англ. / К. Нейгел, Б. Ивьен, Д. Глинн, К. Уотсон. – М.: ООО “И.Д. Вильямс”, 2011. – 1440 с.
18. Натан А. WPF 4. Подробное руководство. - Пер. с англ. - СПб.: Символ-Плюс, 2011. - 880 с., ил.
19. Сеппа Д. Программирование на Microsoft* ADO.NET 2.0. Мастер-класс. / Пер. с англ. — М.: Издательство «Русская Редакция»; СПб.: Питер, 2007. — 784 стр.: ил.
20. Рихтер Д. CLR via C#. Программирование на платформе Microsoft .NET Framework 4.5 на языке C#. 4-е изд. – Питер, 2016 , – 896 с.