УДК 681.518.22

Заяц А.М., к.т.н., проф.,
Логачев А.А., ст. преп.
ФГБОУ ВО «СПбГЛТУ им. С.М. Кирова»

АВТОМАТИЗИРОВАННЫЙ ПОИСК И ОБРАБОТКА АНОМАЛЬНЫХ ЗНАЧЕНИЙ В СИСТЕМЕ ПРОГНОЗИРОВАНИЯ ЛЕСНЫХ ПОЖАРОВ

Аннотация. *В данной работе рассматривается проблема наличия аномальных значений (выбросов) в исходных статистических данных, используемых для прогнозирования лесных пожаров и предупреждения их возникновения. Предложены робастные математические методы обработки данных и обнаружения аномальных значений в исходной выборке, а также их отбраковки или коррекции для последующего использования в статистическом анализе.*

Ключевые слова: статистический анализ, аномальные значения, выбросы, робастность, робастная обработка данных, прогнозирование, лесные пожары.

Abstract. In this article we consider the problem of the presence of abnormal values (outliers) in the initial statistical data used to forecast wildfires and to prevent their occurrence. We proposed robust mathematical methods of data processing and detection of abnormal values in the initial sample, as well as ways of their exception or correction for subsequent use in statistical analysis.

Keywords: Statistical analysis, abnormal values, outliers, robust, robust data processing, forecasting, wildfires.

Введение. Среди всех используемых мер по борьбе с лесными пожарами одной из наиболее эффективных является использование современных информационных систем и технологий для прогнозирования и поддержки принятия решений по снижению уровня лесной пожароопасности на этапе предупреждения возникновения возгораний [1].

Существующие системы поддержки принятия решений по выбору мер по профилактике лесных пожаров основываются на анализе исходных инвентаризационных, таксационных и статистических данных. Зачастую возникают проблемы, связанные с недостаточным объемом исходных данных, что в свою очередь усугубляет проблему наличия в них аномальных значений (выбросов), вызванных ошибками при фиксации результатов наблюдений или сбоями соответствующего оборудования [2]. Аномальные значения способны существенно исказить функционирование математических моделей статистического анализа данных, что может привести к снижению надежности и некорректной работе всей системы [3,4].

Целью настоящей работы является анализ методов робастной обработки исходных данных, которые могут использоваться для выявления аномальных значений и их отсеивания или исправления в рамках разрабатываемого модуля автоматизированного поиска и обработки аномальных значений в системе поддержки принятия решений по предупреждению лесных пожаров [5].

Выявление аномальных значений в исходной выборке. Большинство существующих критериев поиска аномальных значений основываются на допущении, что распределение результатов наблюдений соответствует нормальному закону распределения случайной величины [6-8]. Для нахождения выбросов среди таких значений часто используют критерий Смирнова (критерий Граббса, критерий Смирнова(Граббса)) [9].

Пусть имеется ряд значений, принадлежащих исходной выборке $\mathrm{x}_i \in \mathrm{X}$. Согласно критерию Смирнова значение x_i является аномальным, если удовлетворяется условие (1).

$$\frac{x_i - \overline{X}}{S} > K_n \quad (1)$$

В условии (1) $\overline{X}$ - среднее значение исходной выборки X, S - выборочное среднеквадратическое отклонение случайной величины, K_n - табличное значение процентной точки критерия Смирнова для n наблюдений, взятое из табл. 1.

Таблица 1.

Процентные точки критерия Смирнова–Граббса

Число наблюдений n	*Значение K_n*
5	**1.869**
6	**1.996**
7	**2.093**
8	**2.172**
9	**2.237**
10	**2.294**
15	**2.493**
20	**2.623**
25	**2.717**
30	**2.818**

Пусть по исследуемому лесному участку имеется ряд наблюдений за уровнем влажности (см. табл. 2).

Таблица 2.

Показания датчиков уровня влажности насаждения

		Показание датчика
Датчики (X)	**x_1**	**24**
	x_2	**25**
	x_3	**23**
	x_4	**25**
	x_5	**26**
	x_6	**24**
	x_7	**52**
	x_8	**25**
	x_9	**24**
	x_{10}	**24**

Для исследуемого участка рассчитываются среднее значение $\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n} = 27{,}2$ и выборочное среднеквадратическое отклонение $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 8{,}3$. Затем, используя формулу (1), рассчитываются значения критерия Смирнова (см. табл. 3).

Таблица 3.

Рассчитанные значения критерия Смирнова для исходных данных

		Показание датчика	**Значение критерия Смирнова**

Датчики (X)	**x_1**	**24**	**0,39**
	x_2	**25**	**0,27**
	x_3	**23**	**0,51**
	x_4	**25**	**0,27**
	x_5	**26**	**0,14**
	x_6	**24**	**0,39**
	x_7	**52**	**2,99**
	x_8	**25**	**0,27**
	x_9	**24**	**0,39**
	x_{10}	**24**	**0,39**

Согласно табл. 1. для 10 наблюдений значение процентной точки критерия Смирнова равно 2,294, соответственно условию (1) удовлетворяет значение x_7, поэтому оно является аномальным.

Обработка аномальных значений. После обнаружения выбросов происходит изменение исходной выборки [10-12]. При этом может использоваться два метода:

1. Исключение выбросов. Аномальные значения отбрасываются из исходной выборки, все последующие расчеты проводятся по оставшимся данным [13]. Здесь стоит учитывать, что после исключения аномального значения x_A, следует заново проверить оставшиеся значения, поскольку для новой модифицированной выборки $X_{Mod} \notin x_A$ значения $\overline{X}_{Mod}$ и S_{Mod} изменятся, что может привести к обнаружению новых аномальных значений.

2. Модификация выбросов. Аномальные значения заменяются удовлетворяющими исходному распределению.

В зависимости от объема исходных данных, разброса значений и уровня засорения могут использоваться следующие модификации:

1. Замена выявленного аномального значения x_A на среднее значение текущей выборки $\overline{X}$. При этом стоит учитывать, что искажение, вносимое аномальным значением, может быть настолько велико, что даже после замены $x_A = \overline{X}$ данное значение может оставаться выбросом. Поэтому следует производить пересчет модифицирован-

ных значений $\overline{X}_{Mod}$ и S_{Mod} до тех пор, пока модифицированное аномальное значение x_{Mod} не перестанет удовлетворять условию (1).

2. Замена выявленного аномального значения x_A на максимальное или минимальное значение оставшейся выборки $X_{Mod} \notin x_A$. В таком случае искажения, вносимые аномальным значением x_A, не оказывают влияния на получаемый результат, поэтому последующий пересчет, который использовался при $x_A = \overline{X}$ не требуется. Исключение составляют случаи, когда в оставшейся выборке X_{Mod} обнаруживается другое аномальное значение, которое не было обнаружено при первоначальной проверке. В таком случае происходит обработка уже нового аномального значения.
3. Замена выявленного аномального значения x_A на максимально допустимое для данной выборки. Определение аномальности производится проверкой выполнения условия (1). Из него же можно вывести максимальное допустимое не аномальное значение для данной выборки $x_{\max}$, как показано в формуле (2). При этом аналогично замене $x_A = \overline{X}$ стоит учитывать, что и после $x_A = x_{\max}$ требуется производить пересчет модифицированных значений $\overline{X}_{Mod}$ и S_{Mod} до тех пор, пока модифицированное аномальное значение x_{Mod} не перестанет удовлетворять условию (1).

$$x_{Mod} = S_{Mod} * K_n + \overline{X}_{Mod} \quad (2)$$

Соответственно для значений, приведённых в таблице 2, модифицированные аномальные значения для всех перечисленных способов замен примут следующий вид: $x_{Mod} = \overline{X}_{Mod} = 24{,}72$, $x_{Mod} = X_{\max} = 26$ и $x_{Mod} = x_{Mod\max} = S_{Mod} * K_n + \overline{X}_{Mod} = 27.14$ и представлены на графике 1.

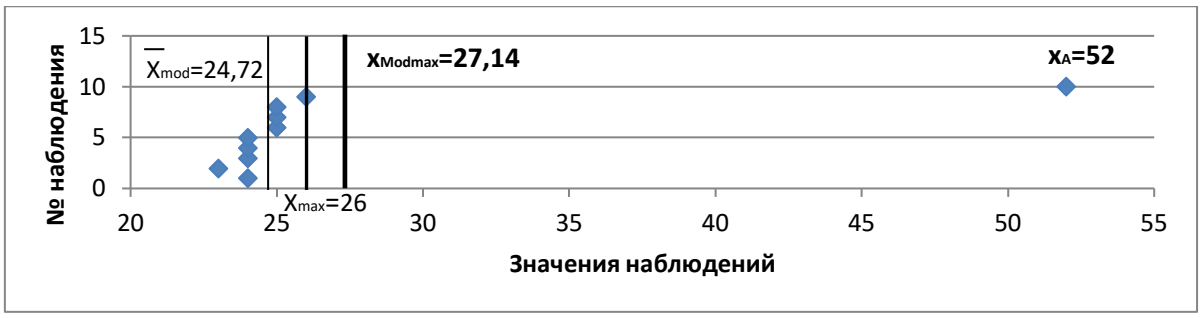

График 1. Сравнение модифицированных аномальных значений

Заключение. В результате проведенного анализа и применения представленных математических методов обработки данных были получены модифицированные множества значений X_{Mod}. Результат сравнения ошибки при оценке уровня лесной пожарной опасности по исходному множеству X и по модифицированному множеству X_{Mod} показан на графике 1.

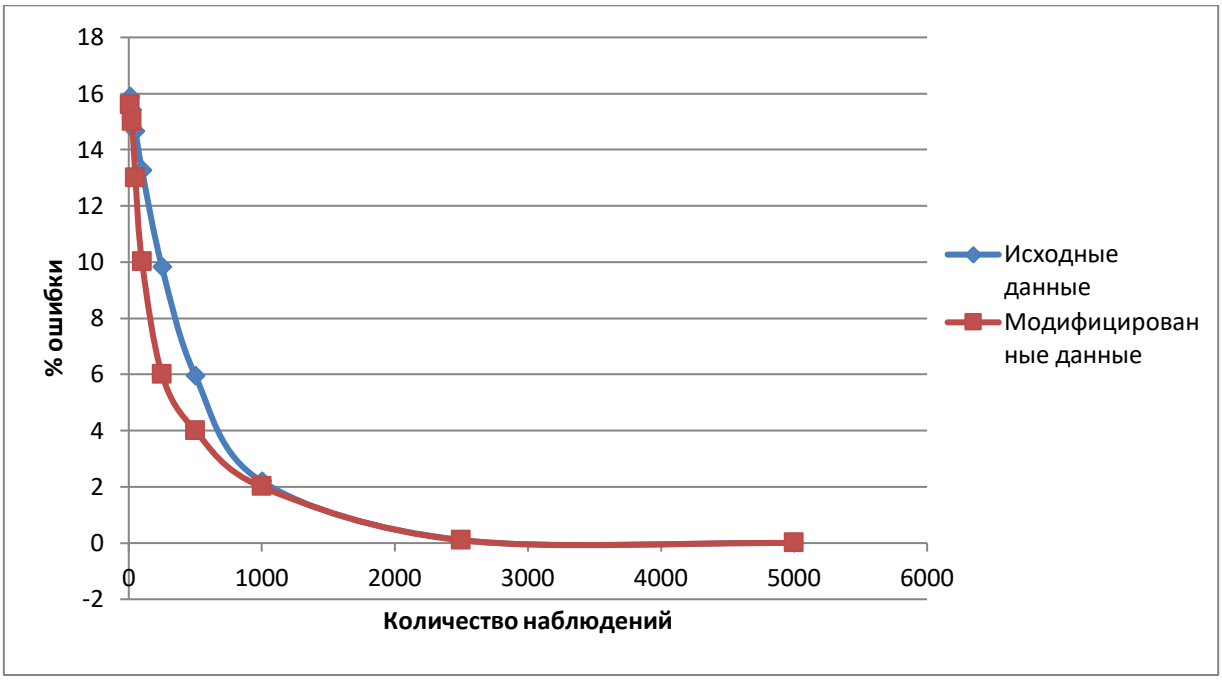

График 2. Сравнение ошибки при оценке уровня лесной пожарной опасности по исходному и модифицированному множествам

Из графика видно, что при большом числе исходных данных искажения, вносимые аномальными значениями, практически полностью отсутствуют и проявляют себя лишь при малом объеме статистики. Это подтверждает необходимость использования робастных методов обработки данных при небольшом количестве исходного статистического материала, что часто встречается при прогнозировании лесных пожаров.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Заяц А.М., Логачев А.А. Математические модели для поддержки принятия решений по предупреждению лесных пожаров при ограниченном объеме исходных данных // Известия высших учебных заведений. Приборостроение. 2016. Т. 59. № 5. С. 342-347.
2. Заяц А.М., Думов М.И. Обзор беспроводных сенсорных сетей и технологий информационных систем оценки лесной пожароопасности и мониторинга лесов // Информационные системы и технологии: теория и практика. Сборник научных трудов. отв. ред. А. М. Заяц. 2016. С. 9-21.
3. Богатырев В.А., Богатырев С.В. Надежность мультикластерных систем с перераспределением потоков запросов // Известия высших учебных заведений. Приборостроение. 2017. Т. 60. № 2. С. 171-177.
4. Богатырев В.А., Винокурова М.С., Петров П.А., Назарова М.Л., Шабаков Р.В. Контроль и безопасность функционирования дублированных компьютерных систем // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 2. С. 368-372.
5. Заяц А.М., Логачев А.А. Информационная система мониторинга лесов и лесных пожаров с использованием беспроводных сенсорных сетей // Известия Санкт-Петербургской лесотехнической академии. 2016. № 216. С. 241-254.
6. Уткин Л.В., Жук Ю.А. Робастная модель обнаружения аномалий с использованием модели засорения // Вестник компьютерных и информационных технологий. 2013. № 7 (109). С. 47-51.
7. Уткин Л.В., Жук Ю.А., Коолен Ф. Робастная модификация метода лассо для полногеномного поиска ассоциаций с учетом целевых значений фенотипа // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 1. С. 150-160.

8. Гоголевский А.С., Уткин Л.В., Хабаров С.П. Метод обнаружения аномальных измерений в системах реального времени на основе алгоритмов машинного обучения // Известия Санкт-Петербургской лесотехнической академии. 2014. № 206. С. 173-180.
9. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерный статистический анализ. М.: «Финансы и статистика», 2000. 352 с.
10. Иванова В.В., Челошкина К.С. Подготовка исходных данных для бизнес-анализа // В сборнике: Science, society, progress proceedings of articles the international scientific conference. 2016. С. 85-98.
11. Ярушкина Н.Г., Эгов Е.Н. Алгоритм выявления новых аномалий в диагностике технических временных рядов // Автоматизация процессов управления. 2016. № 2 (44). С. 24-34.
12. Харин Ю.С., Малюгин В.И. Вероятностно-статистическое прогнозирование: оптимальность, робастность, применения // Вестник БГУ. Серия 1, Физика. Математика. Информатика. 2009. № 1. С. 72-84.
13. Никифоров С.К., Степченко А.А., Алаев Н.П. Методика отбраковки измерений с аномальными значениями среднеквадратической ошибки // Известия Волгоградского государственного технического университета. 2008. Т. 5. № 8 (46). С. 45-48.

В сборнике: Информационные системы и технологии в моделировании и управлении Материалы всероссийской научно-практической конференции. Ответственных редактор Н.Н. Олейников. 2017. С. 29-35.