ИССЛЕДОВАНИЕ АЛГОРИТМОВ И МЕТОДОВ ОПРЕДЕЛЕНИЯ АНОМАЛЬНЫХ НАБЛЮДЕНИЙ ВО ВРЕМЕННЫХ РЯДАХ, ПОЛУЧЕННЫХ С ПОМОЩЬЮ АВТОМАТИЧЕСКИХ МЕТЕОРОЛОГИЧЕСКИХ СТАНЦИЙ

Кузнецов А.Д.[1], Восканян К.Л.[1], Ефременко Д.С.[1], Сероухова О.С.[1], Солонин А.С.[1]

[1] – ФГБОУ ВПО Российский государственный гидрометеорологический университет, Санкт-Петербург, Россия, kvosia@mail.ru

Аннотация. Рассмотрена возможность использования различных алгоритмов для поиска аномальных наблюдений во временных рядах, полученных с помощью автоматических метеорологических станций, и проведена оценка параметров таких алгоритмов, обеспечивающих их эффективную работу.

Ключевые слова: временные ряды метеорологических величин, поиска аномальных наблюдений во временных рядах.

В процессе работы с метеорологическими данными от автоматических метеорологических станций часто возникает потребность в их предварительной обработке [1-3]. В данной работе мы рассматриваем этап, который относится к контролю качества данных, который подразумевает различные проверки исходных метеорологических параметров. В качестве данных мы рассматриваем временные ряды температуры, атмосферного давления и относительной влажности, полученные от автоматической метеорологической станции РГГМУ.

Классификация методов определения аномальных наблюдений. В настоящее время, контроль качества данных проводится практически во всех отраслях и для всех видов человеческой жизнедеятельности, где происходит накопление, промежуточное хранение, использование и реализация цифровых данных, представленных в виде результатов измерений. В течении последних нескольких десятилетий эта область науки развивалась значительными темпами. На текущий момент существует вполне определенная классификация методов определения выбросов, среди которых можно выделить статистические, data mining - аналитические методы определения выбросов и графические методы, которые, в свою очередь, присутствуют, как по отдельности, так и в каждом из предыдущих классов методов определения аномальных значений.

Графические методы включают в себя непосредственный визуальный анализ распределения как самого временного ряда во времени, так и его характеристик.

Статистические методы основаны на предположении, что исследуемый временной ряд подчиняется тому или иному закону распределения и имеет соответствующие ожидаемые величины статистических характеристик. Для таких распределений справедливы основные теоремы, леммы и неравенства теории вероятностей. В статистических методах также применяется и графический метод определения выбросов, исходя из графика исследуемого временного ряда вместе с нанесенным на него теоретическим распределением, линией регрессии, автокорреляционной функцией и другими параметрами.

Аналитические методы определения выбросов, включают в себя элементы математического анализа, вычислительной и аналитической геометрии, эвристические методы. Как правило, в качестве определяющего конечного критерия — относить ли наблюдение к выбросу или нет, применяются классические статистические характеристики. Но, в отличие от статистических методов, вначале проводятся определенные аналитические операции.

Исследуемые методы определения аномальных наблюдений. В данной работе в качестве методов выявления аномальных наблюдений исследовались следующие [4-12]: критерий Ирвина, метод сигм или метод Z-Score, метод MAD (median absolute deviation - абсолютное отклонение медианы), метод Boxplot, метод LOCI (Local correlation integral) с использованием Евклидовой метрики, метод, ABOD (Angle-Based outlier detection), метод LDOF (Local Distance-Based Outlier Detection), метод DBSCAN (Density-based spatial clustering of applications with noise).

В данной работе производились исследования методов определения выбросов на временных рядах температуры, атмосферного давления и относительной влажности, для дискретностей рядов - 15* минут, за четыре сезона 2015 и 2016 года. А именно - за январь, апрель, июль и октябрь. Длительность каждой выборки составляет 7 суток. Методы применялись на рядах с искусственными ошибками. Искусственные ошибки моделировались в различных участках временных рядов - на "подъемах", "спусках", в экстремумах и медианах, а также, на стабильных отрезках временного ряда, где нет сильных перепадов величин. Величины ошибок моделировались в зависимости от статистических и средних соседних значений от точки, где моделировался искусственный выброс. На основе анализа наибольшего числа верно детектированных выбросов в рассматриваемых выборках определялись оптимальные параметры (если это возможно) или диапазоны параметров для рассматриваемых методов, в зависимости от рассматриваемой метеорологической величины, времени года (сезона) измерения и места появления выброса. Эффективность метода рассчитывается для оптимальных коэффициентов, при которых детектировано наибольшее число присутствующих выбросов, по следующей формуле:

$$Э_{факт} = \frac{N_{true}}{N} 100\%,$$

где N_{true} – число верно определённых выбросов; N – число попыток определения выбросов.

Литература

1. Восканян К.Л., Кузнецов А.Д., Сероухова О.С. Автоматические метеорологические станции. Часть 1. Тактико-технические характеристики. Учебное пособие. – СПб.: РГГМУ, 2016. – 195 с., ISBN 978-5-86813-421-0

2. Восканян К.Л., Кузнецов А.Д., Сероухова О.С. Автоматические метеорологические станции. Часть 2. Цифровая обработка данных автоматических метеорологических станций. Учебное пособие. Санкт-Петербург, РГГМУ, 2015. – 80 с., ISBN 978-5-86813-423-4

3. Малинин В.Н. Статистически методы анализа гидрометеорологической информации. – СПб.: Изд. РГГМУ, 2008. – 407 с.

4. IEEE 19th International Conference on Data Engineering (ICDE'03), Bangalore, India, March 5-8, 2003, LOCI: Fast Outlier Detection, Using the Local Correlation Integral.

5. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World, from arxiv, 0903.3257.pdf.

6. M. Ester, H. Kriegel, J. Sander, X. Xu A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Minchen, Germany, @informatik.uni-muenchen.de).

7. http://d-scholarship.pitt.edu/7948/1/Seo.pdf (A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets).

8. H. Kriegel, M. Schubert, A. Zimek. Angle-Based Outlier Detection in High-dimensional Data. Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany, http://www.dbs.ifi.lmu.de.

9. D. Hawkins. Identification of Outliers. Chapman and Hall, London, 1980.

10. V. Barnett and T. Lewis. Outliers in Statistical Data. John. Wiley&Sons, 3rd edition, 1994.

11. https://elki-project.github.io ELKI Data Mining Framework.

12. https://cran.r-project.org/ - The Comprehensive R Archive Network.

INVESTIGATION OF ALGORITHMS AND METHODS FOR DETERMINING ANOMALOUS OBSERVATIONS IN TIME SERIES OBTAINED USING AUTOMATIC METEOROLOGICAL STATIONS

Kuznetsov A.D.[1], Voskanyan K.L.[1], Efremenko D.S.[1], Seroukhova O.S.[1], Solonin A.S.[1]

[1] – *Russian State Hydrometeorological University, Saint-Petersburg, Russia, kvosia@mail.ru*

Abstract. The possibility of using various algorithms to search for anomalous observations in the time series obtained using automatic meteorological stations was considered, and the parameters of such algorithms ensuring their efficient operation were evaluated.

Keywords: time series of meteorological variables, search for anomalous observations in time series.