Восканян К.Л., Иванова Т.И., Кузнецов А.Д., Сероухова О.С.

ФГБОУ ВО Российский государственный гидрометеорологический университет

Программный контроль наличия выбросов во временных рядах метеорологических величин, измеренных с помощью АМС

Аннотация. В данной работе рассматриваются два метода детектирования выбросов (аномальных наблюдений) применительно к временным рядам, полученным от автоматических метеорологических станций. Анализируются результаты численных экспериментов по детектированию выбросов во временных рядах температуры воздуха и приводятся практические рекомендации о целесообразности применения каждого из рассматриваемых методов.

Ключевые слова. Автоматические метеорологические станции, временные ряды метеорологических величин, контроль данных метеорологических измерений.

Введение

В настоящее время временные ряды, получаемые с использованием автоматических метеорологических станций (АМС), формируются в архивы данных. Временные ряды из таких архивов используются для составления прогнозов различной заблаговременности, поведения климатических исследований, принятия управленческих решений и др. Надежность результатов, при использовании архивных данных, напрямую зависит от качества исходных данных. Ведь во временных рядах метеорологических величин от АМС всегда присутствует вероятность наличия выбросов – резко выделяющиеся наблюдения, которые существенно отклоняются от распределения остальных выборочных данных. Именно поэтому возникает необходимость в предварительном контроле данных на предмет наличия в них аномальных наблюдений, т.е. значений, временная изменчивость которых значительно отличается от изменчивости соседних членов временного ряда. Своевременное выявление аномальных наблюдений, содержащихся в данных АМС, помогает снизить вероятность составления некорректных прогнозов погоды и повышает эффективность планирования.

Основная проблема выделения аномальных наблюдений заключается в том, что эти резко выделяющиеся значения могут как отражать экстремальные свойства изучаемого явления, так и быть обусловлены сбоями в работе оборудования в процессе измерений. В первом случае, выбросы представляют особый интерес, поскольку они могут быть связаны, например, с критическими значениями природного процесса и такая экстремальная аномалия обязательно должна оставаться во временно ряде, так как она отражает реально произошедшее катастрофическое событие. Во втором случае, это могут быть действительно грубые ошибки, связанные со сбоем в работе оборудования.

Таким образом, выявление и исключение грубых ошибок («выбросов») является важной задачей первичного анализа временных рядов. Они должны диагностироваться, прежде всего, путем специального анализа и, желательно, автоматически и в реальном масштабе времени. Однако конечный вывод о том, являются ли диагностированные значения временного ряда выбросами или экстремальными значения должен принимать исследователь. Рассматриваемые ниже алгоритмы призваны выявить подозрительные на выбросы данные, освобождая исследователя от необходимости «ручной» проверки всей их совокупности.

В данной работе рассматривается два из всего существующего многообразия методов детектирования выбросов.

1. Классификация методов определения аномальных наблюдений

В настоящее время контроль качества данных проводится практически везде, где происходит накопление, промежуточное хранение, использование и реализация цифровых данных, представленных в виде результатов измерений.

Существующие методы можно классифицировать как геометрические, аналитические и статистические [1 - 10]. Графические методы включают в себя визуальный анализ графического представления изменения во времени как самого временного ряда, 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, которые и используются для оценки временной изменчивости членов временного ряда. Рассматриваются линии регрессии, автокорреляционные функции и др. Аналитические методы определения выбросов включают в себя элементы математического анализа, вычислительной и аналитической геометрии, эвристические методы. Здесь так же применяются классические статистические характеристики, но, в отличие от статистических методов, вначале проводятся аналитические операции. Например, определение «подозрительных» измерений можно проводить с помощью предварительной кластеризации исследуемого ряда. Также можно отметить и наличие эвристических подходов, когда используемый алгоритм не базируется полностью на статистических методах или использования аналитических преобразований, а декларируется с помощью подтверждения его эффективности на основе проведения численных экспериментов с использование реальных временных рядов.

Возможны и комбинации этих подходов к анализу качества членов временного ряда. Однако еще раз подчеркнем, что конечное решение: считать данное изменение «выбросом», связанным со сбоем работы измерительной аппаратуры или действительно наблюдаемое критическое значение исследуемого параметра остается за исследователем.

2. Описание методов выявления аномальных наблюдений

В данном разделе остановимся на рассмотрении двух подходов для выявления значений временного ряда, которые с точки зрения использованного алгоритма можно с той или иной долей вероятности отнести к «выбросам» [10].

2.1 Сравнение двух значений временного ряда на границе контрольного участка

Данный подход можно рассматривать как комбинацию статистического и эвристического подходов. Алгоритм проверки значений членов временного ряда на наличие «выбросов» заключается в следующем.

1. Выбирается величина контрольного участка временного ряда *ndl*, например, *ndl* = 12. Длина такого участка задается исследователем как один из параметров алгоритма и может быть получена в процессе проведения численных экспериментов. На этом начальном участке временного ряда (и только на нем) значения временного ряда алгоритмом не проверяются. Это можно сделать, например, графическим способом, построив и проанализировав график этого участка. Соответствующим значениям вспомогательного временного ряда присваивается 0.

2. На начальном контрольном участке от $n1 = 1$ до $n = ndl$ рассчитывается величина среднеквадратического отклонения: $СКО_{ndl}$. Для исключения возможной ложной идентификации на этом этапе реализации алгоритма после расчета $СКО_{ndl}$ для контрольного участка производится его сравнение со средним квадратическим отклонением для всего временного ряда: $СКО_r$ (для заданного большого участка проверяемого ряда, например, от 1 до 2000). Если вычисленное на контрольном участке значение $СКО_{ndl}$ оказывается в *Кско* раз меньше $СКО_r$ для всего ряда (величина *Кско* задается исследователем как один из параметров

алгоритма, например, *Кско* = 10), то считается, что на контрольном участке изменчивость слишком мала и не может служить критерием наличия или отсутствия выброса.

3. Определяется абсолютная разность между последним значением контрольного участка с порядковым номером $i = ndl$ и соседним с ним значением временного ряда под номером $j = ndl + 1$: Δ_j. По этой абсолютной разности определяется, во сколько раз она превосходит среднее квадратическое отклонение контрольного участка $СКО_{ndl}$. Эта величина присваивается вспомогательному временному ряду под номером $j = ndl + 1$.

4. Контрольный участок сдвигается на единицу: от $n1 = 2$ до $n2 = ndl + 1$ и производится выполнение пп. 2 и 3 для проверки значения временного ряда под номером $j = ndl + 2$: вычисляется новое значение Δ_j для вспомогательного ряда.

Таким образом, путем последовательного сдвига контрольного участка на 1 производится контроль всех значений временного ряда (кроме первых его значений в количестве, равных длине контрольного участка).

2.1 Метод Ирвина

Этот метод можно отнести к комбинированному, поскольку здесь для проведения контроля с начало производится преобразование временного ряда, а затем статистическое оценивание. Рассмотрим реализацию этого алгоритма.

1. Как и в предыдущем методе, на начальном этапе выбирается величина контрольного участка временного ряда *ndl*, например, *ndl* = 12.

2. Выбранный отрезок временного ряда преобразуется в вариационный (например, в порядке возрастания).

3. На начальном контрольном участке от $n1 = 1$ до $n = ndl$ рассчитывается величина среднеквадратического отклонения: $СКО_{ndl}$ (для исключения ложных срабатываний алгоритма производится сравнение $СКО_{ndl}$ с $СКО_r$ так, как это описано в п. 3 предыдущего метода).

4. В вариационном ряду контрольного участка контролируются на наличие «выбросов» два первых и два последних значения. Для этого вычисляют расчётное значение критерия Ирвина $\boldsymbol{I}_{расч}$ для максимального значения вариационного ряда:

$$\boldsymbol{I}_{расч} = Abs(x_{ndl} - x_{ndl-1})/СКО_{ndl},$$

и для минимального значения вариационного ряда:

$$\boldsymbol{I}_{расч} = Abs(x_1 - x_2)/СКО_{ndl},$$

где x_1 и x_{ndl} – проверяемые на наличие «выбросов» значения, x_2 и x_{ndl} - *1* –значения в вариационного ряда, соседствующие с проверяемыми значениями; $СКО_{ndl}$ - выборочное среднеквадратическое отклонение, рассчитываемое для контрольного участка протяженностью *ndl* (с учётом проверяемых значений).

Полученное расчётное значение сравнивают с табличным $I_{табл}$, которое зависит от объёма выборки $\boldsymbol{n}$ и принятого уровня значимости $\boldsymbol{\alpha}$. Если $I_{расч} > I_{табл}$, то проверяемое значение считают грубой ошибкой («выбросом») и значению вспомогательного ряда с соответствующим номеров присваивается 1 (при проведении ранжирования контрольных участков запоминались фактические номера, соответствующие началу и концу ранжированного ряда).

При автоматизированной обработке данных удобно рассчитывать $I_{табл}$ с приемлемой точностью по соотношениям, показанным в табл. 1 для случая использования выборочного СКО при изменении объёма выборки $\boldsymbol{n}$ в пределах от 3 до 1000.

α	*Iтабл*

0,01	-205,06(n – 3) + 424,26(n - 2,5) - 352,483(n – 2) +143,747(n - 1,5) - 33,401(n -1) + 6,381(n - 0,5) + 1,049
0,05	-229,21(n – 3) + 422,39(n - 2,5) - 320,96(n -2) + 124,594(n -1,5) - 26,15(n-1)+4,799(n-0,5) + 0,7029
0,1	-132,78(n-3) + 224,24(n-2,5) - 165,27(n-2) +68,614(n-1,5) - 16,109(n-1)+3,693(n-0,5) + 0,549

3. Апробация методов выявления аномальных наблюдений

Для иллюстрации работы рассмотренных алгоритмов использовался временной ряд приземной температуры воздуха (рис. 1а) и тот же временной ряд после внесения искусственных выбросов (рис. 1б). Значения временного ряда с номерами 80 и 341 уменьшены на 2 ^{0}C, а с номером 157 – увеличено на 2 ^{0}C.

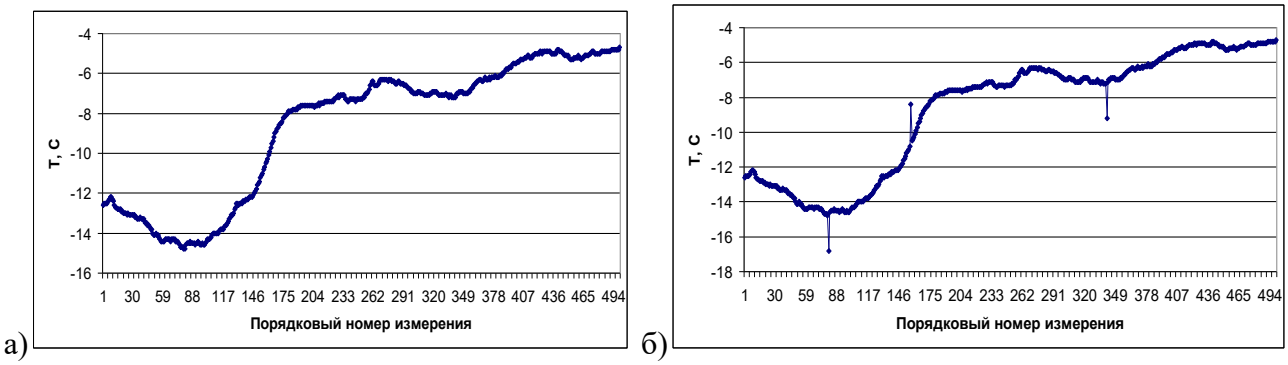

Рис. 1 Временной ряд приземной температуры воздуха [^{0}C] с дискретностью 15 мин
а) исходный ряд; б) исходный ряд после искусственного внесения «выбросов»

После реализации рассмотренных методов график вспомогательных рядов для исходного ряда выглядит так, как это представлено на рис. 2 – значения временного ряда, которые можно было бы рассматривать как «выбросы» не обнаружены.

Рис. 2 Вспомогательные ряды при использовании для контроля качества двух методов

Результаты, полученные при применении двух рассмотренных методов в форме графиков вспомогательных временных рядов для исходного ряда с искусственно внесенными «выбросами» соответственно представлены на рис. 3.

а) 

б) 

Рис. 3 Вспомогательные ряды для контроля качества при использовании
а) двух методов; б) метода Ирвина

В первом случае алгоритм четко указывает на номер подозрительного на выброс измерения. Во втором случае, с учетом специфики метода Ирвина «выброс» держится в вариационном ряде до тех пор, пока он не покинет либо весь контрольный участок, либо его часть.

Заключение.

1. В работе отмечена важность контроля качества временных рядов как в процессе формирования архивов данных, так и при использовании данных из уже имеющихся архивов.
2. Указано, что существующие алгоритмы в подавляющем своем количестве только помогают исследователю находить подозрительные на «выбросы» данные, конечный контроль, но уже только этих данных, а не всего временного ряда остается за исследователем.
3. В работе приведены два алгоритма реализации контроля качества временных рядов.
4. На конкретном примере показано работоспособность приведенных в работе алгоритмов.

Список литературы

1. Spiros Papadimitriou Carnegie Mellon University, Hiroyuki Kitagawa University of Tsukuba, Christos Faloutsos Carnegie Mellon University, Phillip B. Gibbons, Intel Research Pittsburgh; IEEE 19th International Conference on Data Engineering (ICDE'03), Bangalore, India, March 5-8, 2003, LOCI: Fast Outlier Detection, Using the Local Correlation Integral

2. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data , Ke Zhang1 and Marcus Hutter1,2 and Huidong Jin1,2,3, from arxiv, 0903.3257.pdf

3. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databaseswith Noise, Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Miinchen, Germany

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

5. Angle-Based Outlier Detection in High-dimensional Data. Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek. Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany, http://www.dbs.ifi.lmu.de

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

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

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

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

10. 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.