Journal of
open research software

Huang, S et al 2016 OutlierFlag: A Tool for Scientific Data Quality Control by Outlier Data Flagging. *Journal of Open Research Software*, 4: e20, DOI: http://dx.doi.org/10.5334/jors.90

SOFTWARE METAPAPER

OutlierFlag: A Tool for Scientific Data Quality Control by Outlier Data Flagging

Shuai Huang[1,2], Yaqiang Wang[2], Yuanli Xie[1], Peng Zhao[3] and Johannes Lüers[4]

[1] Northwest University, Xi'an, China

[2] State Key Laboratory of Severe Weather & Key Laboratory of Atmospheric Chemistry of CMA, Chinese Academy of Meteorological Sciences, Beijing, China

[3] University of Bayreuth, Germany, now at University of Innsbruck, Austria

[4] University of Bayreuth, Germany

Corresponding authors: Yaqiang Wang (wangyq@cams.cma.gov.cn), Peng Zhao (peng.zhao@uibk.ac.at)

Scientific datasets collected by instruments usually include outliers which have to be flagged in data quality control process. OutlierFlag was developed to make this process accurate and simple by providing a suitable outlier data flagging algorithm and a user friendly GUI. The algorithm consists of three steps performed one by one: limitation check, error check and standard deviation check. Several parameters are configurable so the algorithm can be used for various datasets. OutlierFlag is an open source software written in Java and the MeteoInfo library was used for data plotting function.

Keywords: outlier detection; observational data; outlier data

(1) Overview

Introduction

In observed scientific datasets there are often a wide variety of outliers, which are caused by unexpected abrupt changes in the surrounding environment, instrument fluctuations, or miss-operation by observers [1]. Outliers are the error values that differ distantly from other data. According to their characteristics they can be sorted into three type errors of beyond extreme, constant value and numerical mutation [2]. Moreover they can be classified into random error, systematic error and negligence error in the light of generated reasons [3]. Data quality is one of the major concerns in scientific studies, and many kinds of analysis are sensitive to errant values and outliers [4]. On the basis of temporal and spatial variation of objective elements and scientific standards of observed file format, data quality control is an operation process associated with technical and rational testing of observation datasets for identifying abnormal values. Facing the huge amount of data and poor data quality, manually data checking is difficult and time-consuming. Taking the advantage of the high speed of modern computers, the in this study presented quality control software has made this work easy and efficient.

For automatic and accurate outlier data flagging, we developed a three-step algorithm which was first introduced by Zhao and Lüers [5] with limitation check, error check and standard deviation check. This algorithm is now implemented in a program called OutlierFlag, which was designed especially for time serials quantitative continuous data-sets rather than other data-sets such as discrete or qualitative data-sets and developed to provide a user-friendly GUI to end-users.

Although some other software packages are available for outlier detection, no single outlier detection algorithm or combination of them has the ability to detect all outliers in a simple and efficient way according to our experience. For example, the well-known R's 'outliers' package[6] provides mainly Dixon's test and Grubbs' test algorithms [7–9] for detecting only one or two outliers in a small data-set, or a general built-in unconfigurable function to detect a outlier if there is the largest difference between it and sample mean. As the performances of these algorithms normally depend on the parameter setting, OutlierFlag provides more flexible parameter setting functions and user manual outlier detection functions for improved results. Furthermore, Outlier gives a much easier, more friendly and interactive GUI than other software.

Implementation/architecture

The developed outlier data flagging algorithm includes the following three steps:

1. **Limitation Check:** The data points will be marked as outliers if their values exceed the range between minimum and maximum limitation values. The limitation values are assigned according to the features of the observational data series and the environment of the

observation station. Reasonable limitation values are the key to ensure the accuracy of the detection results.

2. **Error Check:** It is generally agreed that continuous adjacent points could be related to each other, esp. when dealing with environmental data and the distribution of them has a certain variation and characteristics within a time segment, so the sudden change of the data value or the difference between two continuous data points are likely to indicate an outlier. OutlierFlag firstly calculates the error of each data point as the difference between its value and the mean value of adjacent points within a use-defined window (11 points as the default width. Then OutlierFlag gets a sub list for the error of each data point using another user-defined window (21 continuous points as default value), and calculates the quartile value of the list with a user-assigned percentage (default value is 0.9). The data point is flagged as an outlier if its error is larger than the product of the quantile value and a user-defined multiplier (default value is 2.3)[5].
3. **Standard Deviation Check:** This step performs a further detection according to the standard deviation threshold. Firstly, OutlierFlag constructs for each data point x_i a data list by by x_i and its adjacent data points with the default data point number of 29, i.e. x_{i-14}, x_{i-13}, ..., x_i, x_{i+1},... x_{i+14}. Then the error of t x_i is calculated as the difference between x_i and the mean value of this data list. The standard deviation of the data list is calculated as well. The data point x_i is determined as an outlier if its error is bigger than 3 or other user defined times of the standard deviation.

OutlierFlag is written in Java, and the plotting functions are based on MeteoInfo library [10]. The data flow of OutlierFlag is described in **Figure 1**.

The input data file is required to be an ASCII file with identical columns in each line. The first line has to be the header containing all column names. The data will be showed in a table after they are loaded from the file. OutlierFlag provides the functions to strictly sort the data by time if a column of time stamp exists. Then the outliers can be automatically flagged following the three steps as described above with the default or user-defined parameters. The data scatter chart can be plotted for a final result check and further manual expert flagging. The flagged data points are easy to be distinguished by several colors. Flag codes are added in the data table as follows:

Flag = V0, passed all check;
Flag = V1, non- numeric value;
Flag = V2, failure in limitation check;
Flag = V3, failure in error check;
Flag = V4, failure in standard deviation check;
Flag = VU, not passed manual check.

Hourly and daily data average calculation can be performed with this the software if the data has a time column. The flagged data points are excluded in the data average calculation.

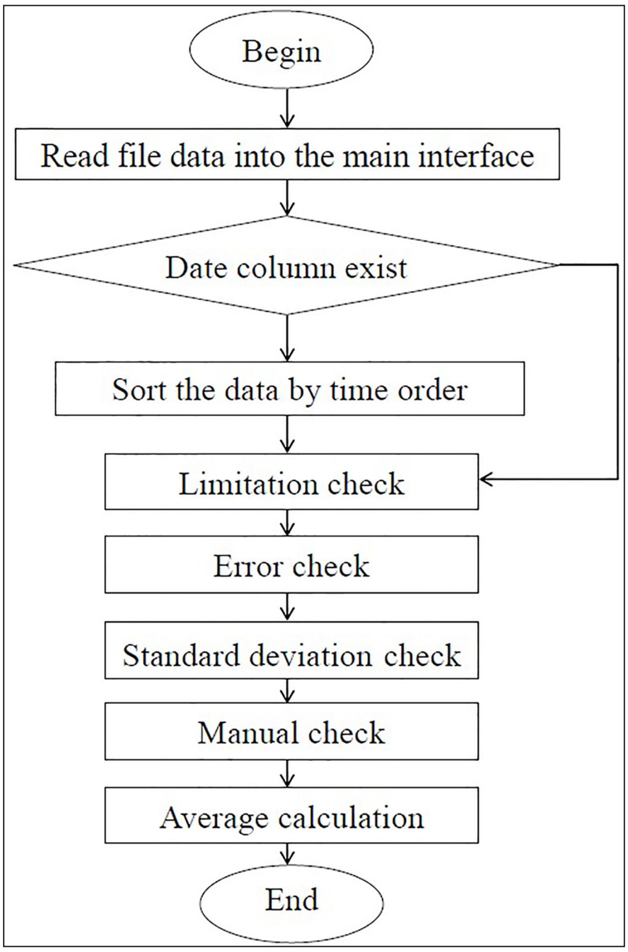


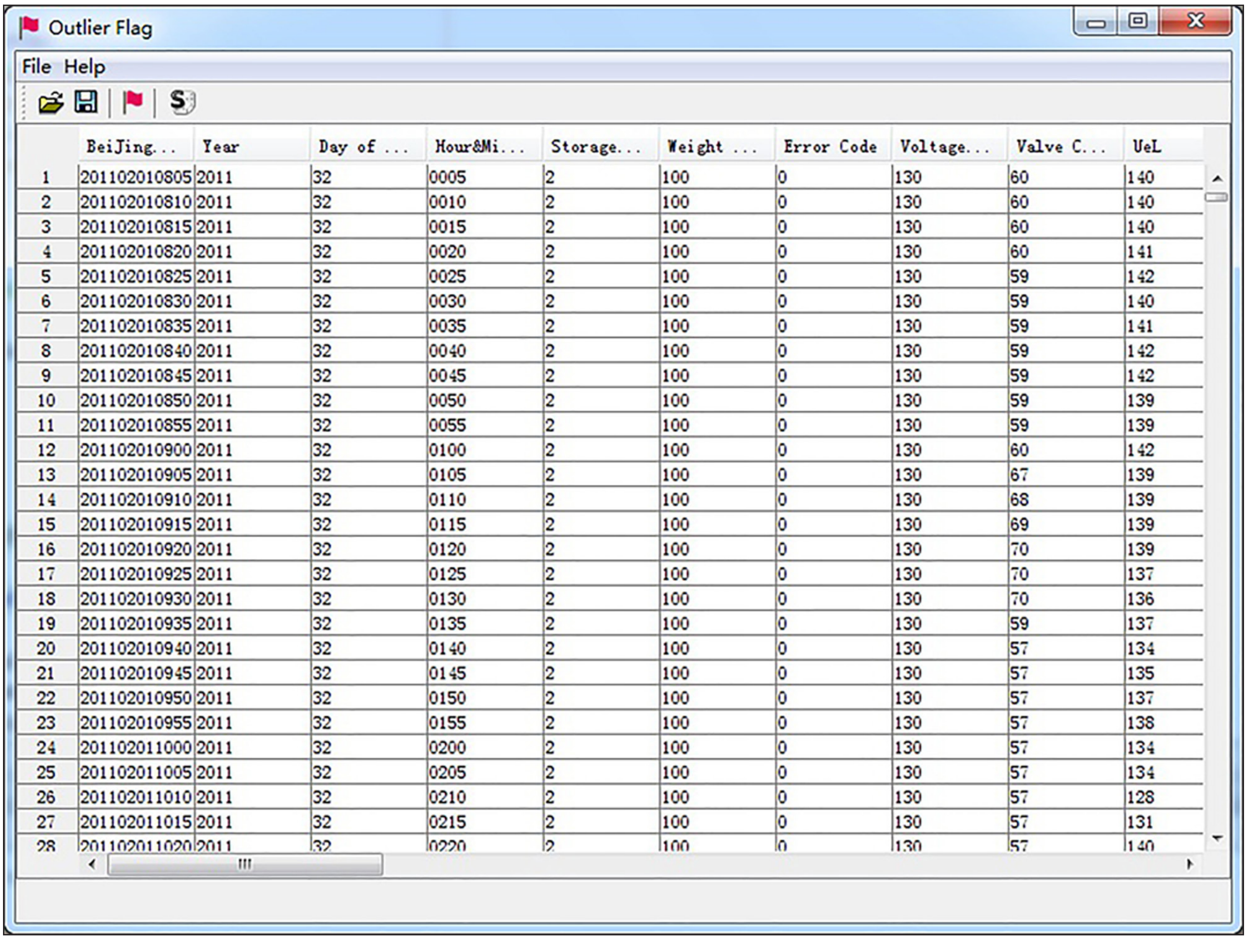
Figure 1: Data flow of OutlierFlag.

The existing "larger" outliers with huge sudden change values may influence the "smaller" outliers' detection using any kind of algorithm. So this 3-step algorithm is used to detect "larger" outliers firstly and then the "smaller" outliers. The point is the "larger" outliers detected by previous steps will not take part in the next step.

Examples

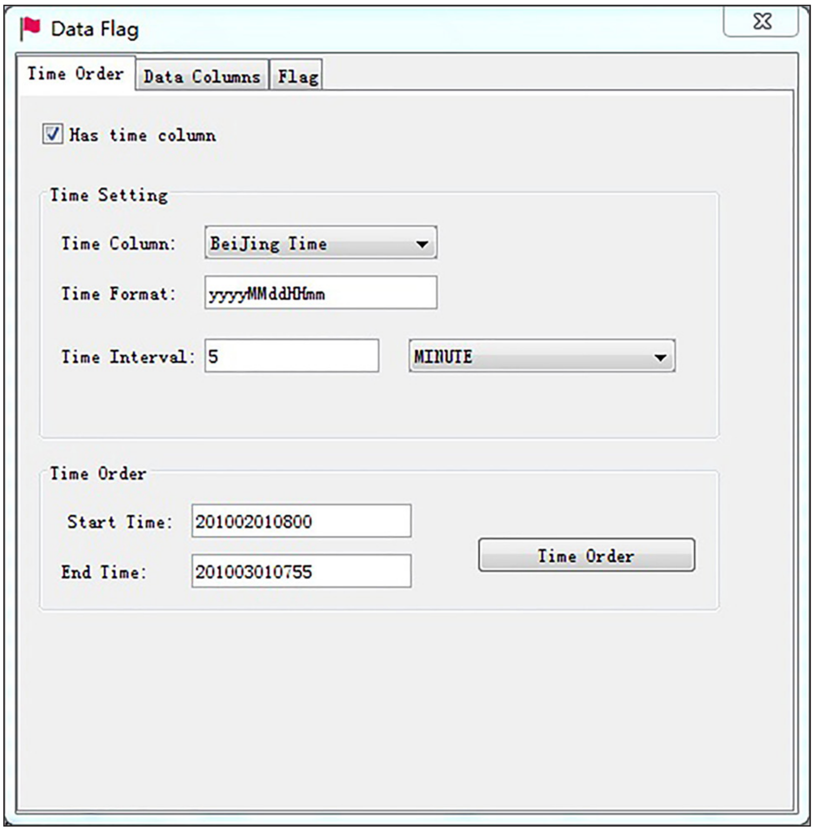
To demonstrate the usage of the OutlierFlag, an example data file "54826PMMUL201102_T.txt", located in sample folder of OutlierFlag directory, was provided with the software, and step-by-step instructions can be found in the help documentation of the software. The data file includes aerosol mass concentration data (PM1, PM2.5 and PM10) collected by a GRIMM 180 dust monitor (Magee Scientific Co, USA) at Mount Tai meteorological station in February, 2011. A time column is included in the data with 5 minute interval. The data processing steps were described in **Figure 2** to **Figure 6**.

According to the operation steps of the software, we show the use of the software sequentially with the example data to give readers a detailed understanding of OutlierFlag. Firstly, click the "Open File" button in the toolbox and set file path, column separator and title line in the "Open File" dialog, then load data of the destination file into the main table (**Figure 2**). Click the "Data



	BeiJing...	Year	Day of ...	Hour&Mi...	Storage...	Weight ...	Error Code	Voltage...	Valve C...	UeL
1	201102010805	2011	32	0005	2	100	0	130	60	140
2	201102010810	2011	32	0010	2	100	0	130	60	140
3	201102010815	2011	32	0015	2	100	0	130	60	140
4	201102010820	2011	32	0020	2	100	0	130	60	141
5	201102010825	2011	32	0025	2	100	0	130	59	142
6	201102010830	2011	32	0030	2	100	0	130	59	140
7	201102010835	2011	32	0035	2	100	0	130	59	141
8	201102010840	2011	32	0040	2	100	0	130	59	142
9	201102010845	2011	32	0045	2	100	0	130	59	142
10	201102010850	2011	32	0050	2	100	0	130	59	139
11	201102010855	2011	32	0055	2	100	0	130	59	139
12	201102010900	2011	32	0100	2	100	0	130	60	142
13	201102010905	2011	32	0105	2	100	0	130	67	139
14	201102010910	2011	32	0110	2	100	0	130	68	139
15	201102010915	2011	32	0115	2	100	0	130	69	139
16	201102010920	2011	32	0120	2	100	0	130	70	139
17	201102010925	2011	32	0125	2	100	0	130	70	137
18	201102010930	2011	32	0130	2	100	0	130	70	136
19	201102010935	2011	32	0135	2	100	0	130	59	137
20	201102010940	2011	32	0140	2	100	0	130	57	134
21	201102010945	2011	32	0145	2	100	0	130	57	135
22	201102010950	2011	32	0150	2	100	0	130	57	137
23	201102010955	2011	32	0155	2	100	0	130	57	138
24	201102011000	2011	32	0200	2	100	0	130	57	134
25	201102011005	2011	32	0205	2	100	0	130	57	134
26	201102011010	2011	32	0210	2	100	0	130	57	128
27	201102011015	2011	32	0215	2	100	0	130	57	131
28	201102011020	2011	32	0220	2	100	0	130	57	140

Figure 2: OutlierFlag main GUI.



☑ Has time column

Time Setting

Time Column: BeiJing Time

Time Format: yyyyMMddHHmm

Time Interval: 5 MINUTE

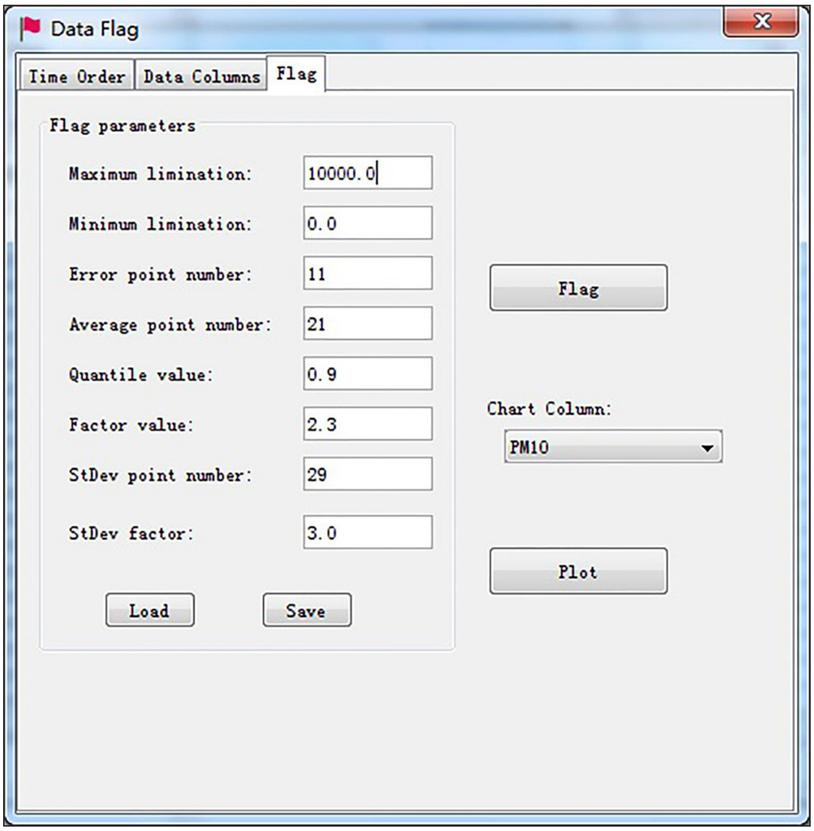
Time Order

Start Time: 201002010800

End Time: 201003010755

Figure 3: Sorting data rows by time in OutlierFlag GUI.

flag" button, then a dialog with three pages named "Time Order", "Data columns" and "Flag" will be opened. In the "Time Order" page the data rows can be sorted by time if a time column exists (**Figure 3**). Afterward choose the destination data columns by column titles in the "Data Columns" page for further outlier flag analysis. Open the "Flag" page to set the algorithm parameters of maximum limitation, minimum limitation, error point number, average point number, quartile value, factor value, standard deviation point number and standard deviation factor (**Figure 4**). Then click the "Flag" button to run the outliers detection and the result chart form will be shown (**Figure 5**). In this case, three data columns of PM_{10}, $PM_{2.5}$ and PM_1 are checked and flagged simultaneously. We can choose one of them in "Chart column" and click the "Plot" button to visualize the data in a figure, where users can manually check and mark or unmark the outliers by clicking the "Flag selected



Flag parameters

Maximum limination: 10000.0

Minimum limination: 0.0

Error point number: 11

Average point number: 21

Quantile value: 0.9

Factor value: 2.3

StDev point number: 29

StDev factor: 3.0

Chart Column: PM10

Figure 4: Algorithm parameters setting step in OutlierFlag GUI.

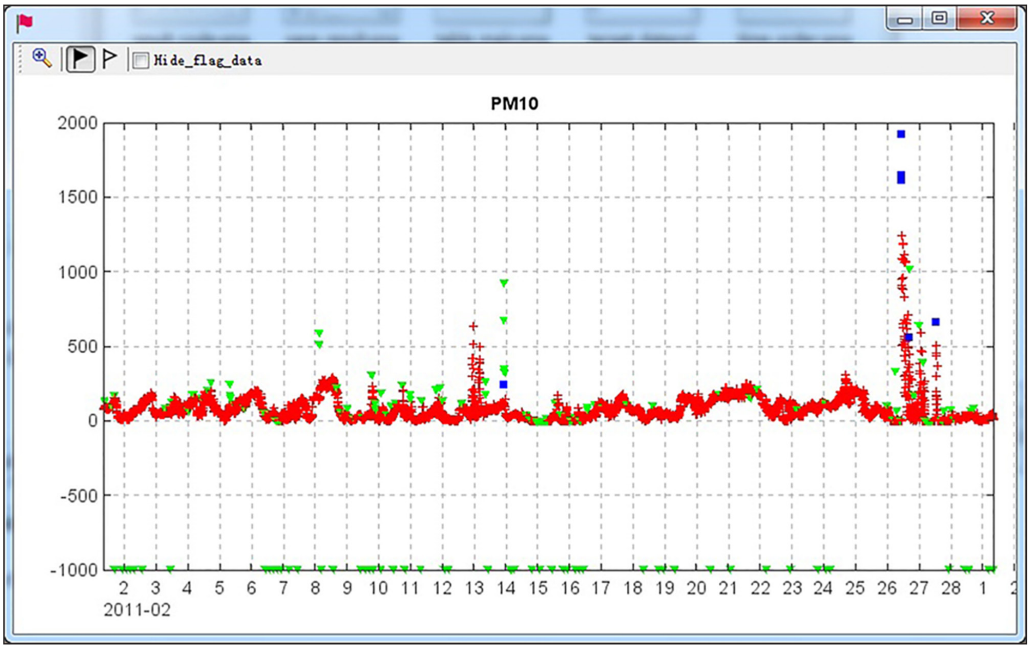


Figure 5: Results of outlier detection in OutlierFlag GUI.

points" or "Unflag selected points" above the chart form. In order to facilitate the user to watch the chart without flagged data, software will eliminate the outliers when choosing "Hide flag data" check box. A dialog of daily and hourly average values of data sets will be displayed by clicking the "Data average" button in the main menu, and the data can be averaged by ignoring flagged outliers (**Figure 6**).

Quality control

OutlierFlag has been tested with many observation data files, such as measurements of the aerosol mass concentration of PM_{10}, $PM_{2.5}$ and PM_1 collected by a GRIMM 180 dust monitor, or atmospheric visibility data observed by a FD-12 visibility meter at Mount Tai meteorological station from June 2010 to March 2012. One of the outlier detection results is shown in **Figure 7**.

At the same time, the algorithm of OutlierFlag has been widely used in the data quality control processing of atmospheric composition data observed by stations of China Meteorological Administration [11]. The results show that OutlierFlag has ability to identify outliers easily and accurately.

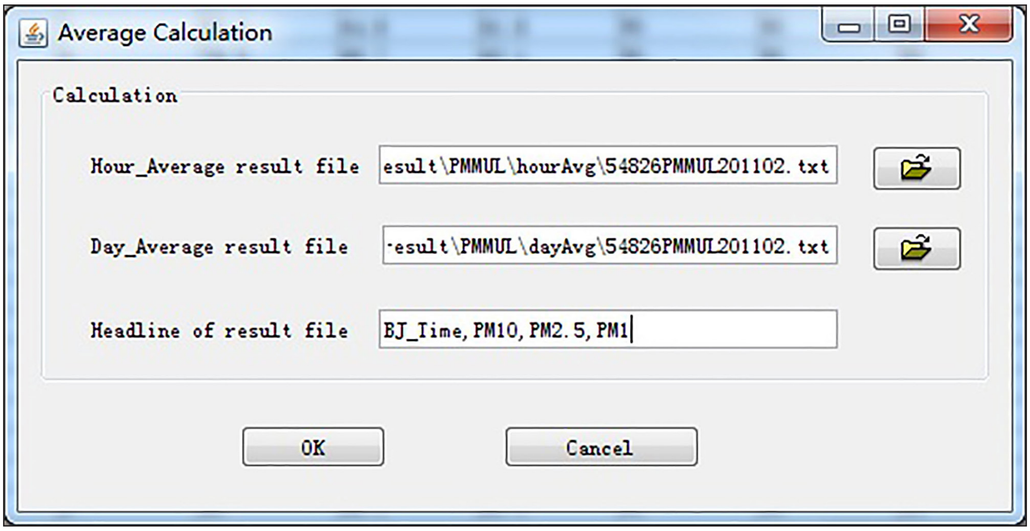
(2) Availability

Operation system

As it is implemented in Java, OutlierFlag runs on any operation system that runs Java.

Programming language

Java 1.7



Hour_Average result file: esult\PMMUL\hourAvg\54826PMMUL201102.txt

Day_Average result file: ·esult\PMMUL\dayAvg\54826PMMUL201102.txt

Headline of result file: BJ_Time,PM10,PM2.5,PM1

Figure 6: Average calculation in OutlierFlag GUI.

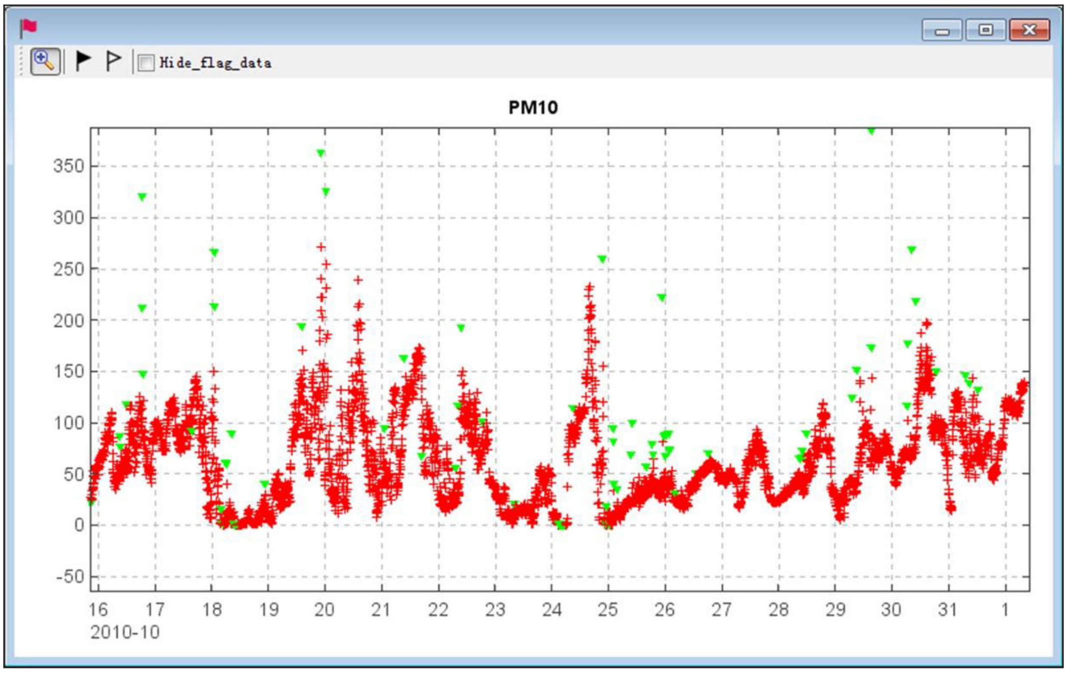


Figure 7: Outlier detection results by OutlierFlag for observation data files, which is aerosol mass concentration (PM_{10}, October 2010).

Dependencies

MeteoInfo Libraries

Archive

Name: OutlierFlag
Persistent identifier: https://dx.doi.org/10.6084/m9.figshare.3175630.v3
License: GNU Lesser General Public License

Code Respository

Name: OutlierFlag
Identifier: https://bitbucket.org/yaqiang/outlierflag
License: GNU Lesser General Public License
Date published: 07/15/14

Language

English

(3) Reuse potential

The typical usage of OutlierFlag is to perform a quality control for time serials quantitative continuous data-sets. It was generally designed to be useful for many kinds of data such as atmosphere data, soil data, hydrology geographical data and so on. The outlier flag algorithm implementation code can be used in other software which needs these functions.

Support for OutlierFlag

When users or developers run into problems or discover bugs we encourage them to either open an issue on Bitbucket page or to contact us via email (yaqiang.wang@gmail.com) directly.

Acknowledgement

This work was supported by grants from the National Key Project of Basic Research (2014CB441201), the National Natural Science Foundation of China (41275167) and the Chinese Academy of Meteorological Sciences (2013Z007).

Competing Interests

The authors declare that they have no competing interests.

References

1. **Ren, Z** 2007 The quality control of surface monthly climate data in China. *Journal of Applied Meteorological Science*, 18(4): 516–523. (In Chinese).
2. **Yang, P, Liu, W, Zhong, J** and **Yang, J** 2011 Evaluating the quality of temperature measured at automatic weather stations in Beijing. *Journal of Applied Meteorological Science*, 22(6): 706–715. (In Chinese).
3. **Liu, X** and **Ren, Z** 2005 Progress in quality control of surface meteorological data. *Meteorological Science and Technology*, 33(3): 109–203. (In Chinese).
4. **Eischeid, J K, Baker, C B, Karl, T R** and **Diaz, H F** 1995 The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology*, 34: 2787–2795. DOI: http://dx.doi.org/10.1175/1520-0450(1995)034<2787:TQCOLT>2.0.CO;2
5. **Zhao, P, Lüers, J** and **Foken, T** 2014 GaFiR: a gap-filling package for ecosystem-atmosphere carbon dioxide flux and evapotranspiration data, University of Bayreuth, *Dept. of Micrometeorology*, Work Report Vol. 59, ISSN 1614-8916, 19 pp.
6. **Lukasz Komsta** 2011 outliers: Tests for outliers. R package version 0.14 http://CRAN.R-project.org/package=outliers.
7. **Dixon, W J** 1950 Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4): 488–506. DOI: http://dx.doi.org/10.1214/aoms/1177729747
8. **Dixon, W J** 1951 Ratios involving extreme values. *The Annals of Mathematical Statistics.* 22(1): 68–78. DOI: http://dx.doi.org/10.1214/aoms/1177729693
9. **Grubbs, F E** 1950 Sample Criteria for testing outlying observations. *The Annals of Mathematical Statistics.* 21(1): 27–58.
10. **Wang, Y Q** 2014 MeteoInfo: GIS software for meteorological data visualization and analysis. *Meteorological Applications*, 21(2): 360–368.
11. **Wang, Y Q, Zhang, X Y, Sun, J Y, Zhang, X C, Che, H Z** and **Li, Y** 2015 Spatial and temporal variations of the concentrations of PM_{10}, $PM_{2.5}$ and PM_1 in China. *Atmospheric Chemistry and Physics*. 15: 13585–13598. DOI: http://dx.doi.org/10.5194/acp-15-13585-2015

How to cite this article: Huang, S, Wang, Y, Xie, Y, Zhao, P and Lüers, J 2016 OutlierFlag: A Tool for Scientific Data Quality Control by Outlier Data Flagging. *Journal of Open Research Software*, 4: e20, DOI: http://dx.doi.org/10.5334/jors.90

Submitted: 24 August 2015 **Accepted:** 04 May 2016 **Published:** 31 May 2016

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS