Journal of Hydrology 548 (2017) 135–144

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends

Jina Jeong [a], Eungyu Park [a,*], Weon Shik Han [b], Kueyoung Kim [c], Sungwook Choung [d], Il Moon Chung [e]

[a] Department of Geology, Kyunpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea
[b] Department of Geosystem Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
[c] Korea Institute of Geoscience and Mineral Resources, 124 Gwahak-ro, Yuseong-gu, Daejeon 34132, Republic of Korea
[d] Division of Earth and Environmental Sciences, Korea Basic Science Institute, 162 Yeongudanji-ro, Ochang-eup, Choengju 28119, Republic of Korea
[e] Korea Institute of Construction Technology, 283 Goyangdae-ro, Ilsanseo-gu, Goyang-si, Gyeonggi-do 10223, Republic of Korea

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form 2 December 2016
Accepted 27 February 2017
Available online 6 March 2017
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Wolfgang Nowak, Associate Editor

Keywords:
Outlier identification
Three sigma rule
Interquartile range
Median absolute deviation
Anomaly detection

ABSTRACT

A hydrogeological dataset often includes substantial deviations that need to be inspected. In the present study, three outlier identification methods – the three sigma rule (3σ), inter quantile range (IQR), and median absolute deviation (MAD) – that take advantage of the ensemble regression method are proposed by considering non-Gaussian characteristics of groundwater data. For validation purposes, the performance of the methods is compared using simulated and actual groundwater data with a few hypothetical conditions. In the validations using simulated data, all of the proposed methods reasonably identify outliers at a 5% outlier level; whereas, only the IQR method performs well for identifying outliers at a 30% outlier level. When applying the methods to real groundwater data, the outlier identification performance of the IQR method is found to be superior to the other two methods. However, the IQR method shows limitation by identifying excessive false outliers, which may be overcome by its joint application with other methods (for example, the 3σ rule and MAD methods). The proposed methods can be also applied as potential tools for the detection of future anomalies by model training based on currently available data.

1. Introduction

A groundwater model is an abstract of a real system, which is commonly used to estimate or predict various phenomena regarding the physical and chemical states (Anderson and Woessner, 1992; Mercer and Faust, 1980; Solomatine and Ostfeld, 2008). A calibration step (in a physically based model) or training (in a data driven model) based on observations enables a groundwater model to have predictive capabilities. The measured data used in this step should capture the full variability and the frequency of the targeted system responses. However, a hydrogeological dataset often includes substantial deviations with a minor amount of points termed "outliers" (Hawkins, 1980; Barnett and Lewis, 1994; Helsel and Hirsch, 2002; William et al., 2002; Liu et al., 2004; Maimon and Rockach, 2005). In this sense, the representativeness of the measurements is an important factor for the quality of the predictions by a groundwater model.

Statistically, outliers are defined as data that have a low probability of being consistent with other data and potentially mask the actual characteristics of the data. The outliers are commonly caused by two problems with the hydrogeological data: (1) erroneous data measurements, transmission, or transcription and (2) local deviations from the major data generating processes over a limited time. Incorporating the outliers into statistical analyses without using caution may lead to incorrect interpretations of the physical or chemical state of the groundwater (William et al., 2002; Helsel and Hirsch, 2002; Gibbons and Coleman, 2001; Hodge and Austin, 2004; Zar, 2007). However, an observation that appears to be a potential outlier may provide non-negligible information that accounts for the variability of the groundwater state. Consequently, the identification and the relevant treatment of potential outliers in measurements are essential for a dataset to be representative, which, in turn, allows a groundwater model to be more accurate in the prediction of long-term state changes (Liebetrau, 1979; Hirsch et al., 1982, 1991; Hirsch and Slack, 1984; Helsel and Hirsch, 2002; Maimon and Rockach, 2005; Morton and Henderson, 2008; Khaliq et al., 2009; Moyer et al., 2012).

* Corresponding author.
E-mail address: egpark@knu.ac.kr (E. Park).

http://dx.doi.org/10.1016/j.jhydrol.2017.02.058
0022-1694/

The outliers have been evaluated through several methods that could be categorized into two groups of graphical and statistical tests (Rousseeuw, 1984; Hodge and Austin, 2004; Barnett and Lewis, 1994). In both approaches, the residuals between the data and the corresponding estimations are used to construct a criterion for making a decision about whether the data point is an outlier or a normal data point. In the relatively simple approach using graphical tests, probability plots such as P-P or Q-Q plots and box-and-whisker plots have been extensively used (Rousseeuw et al., 1999; Zar, 2007). Additionally, as another graphical test, the residual versus fit plot is constructed to identify the outliers on the basis of residual variance (Zar, 2007). However, since all graphical-based tests do not allow hypothesis testing, the interpretations and evaluations of the outliers are inevitably subjective. In addition, it has been reported that the graphical methods are difficult to apply in cases with multiple outliers (Maimon and Rockach, 2005).

The approach based on statistical inference is an objective alternative to the aforementioned graphical approach by allowing hypothesis testing on outlier distributions. For the purpose of outlier identifications on the basis of measurements, Chauvenet's criterion, Grubbs' test, Peirce's criterion, Dixon's test, and Rosner's test have been generally used when the tests commonly rely on an assumption of Gaussianity for diagnostics (Dixon, 1953; Grubbs, 1969; Hawkins, 1980; Gibbons, 1994). When there are discernible trends and outliers as in groundwater measurements, robust regression methods can be employed as trend estimators to obtain residuals less affected by outliers. The Theil-Sen (TS), least median square (LMS), and least trimmed square (LTS) estimators are the frequent options (Rousseeuw, 1984; Rousseeuw and Leroy, 1987; Gibbons, 1994; Helsel and Hirsch, 2002). In the outlier identification methods, it is commonly assumed that the residuals are from independent and identical Gaussian distributions (Hodge and Austin, 2004), which may not be appropriate in accounting for the frequency and variability of groundwater data (Hirsch et al., 1982).

The majority of conventional outlier identification methods are proposed to be compatible with Gaussian groundwater state data, and it is difficult to be effectively extended to account for the variability of data, which often are non-Gaussian. This shortcoming justifies the development of alternative outlier identification methods for groundwater state data. In the current study, outlier identification methods are proposed based on the common methods of interquartile range (IQR), median absolute deviation (MAD), and three sigma (3σ) rules, which address data showing non-Gaussianity where the ranges of the normal data are defined from ensemble estimations. It is frequently reported that an ensemble of multiple estimations can improve prediction qualities (Brown et al., 2005). In the validations, simulated groundwater level data showing non-Gaussian statistical distributions are employed to evaluate the proposed methods, and the applications of the methods are discussed based on the results. In addition, to test the practicality of the developed methods, groundwater level measurements from an actual monitoring well are acquired and analyzed.

2. Method development

As criteria for the separation of normal data from outlying data, three different improved methods are proposed that are based on the 3σ rule, IQR rule, and MAD. Comparing the methods in their conventional form, the 3σ method is not fully sufficient in the outlier identification of the groundwater state data because the method is limited by assuming a Gaussian data distribution while the data is often non-Gaussian. For the purposes of outlier identification in the groundwater state data, the conventional IQR and MAD methods are more appropriate as they take advantage of non-parametric statistics (i.e., the quartile and median) instead of parametric measures (e.g., the mean, variance, etc.).

The proposed modification allows the prediction to be based on multiple statistical estimations from the resampling of observed data rather than from the observations only. Therefore, the proposed outlier identification methods in this study must be combined with the ensemble method based on random resampling of observed data, such as bagging (bootstrap aggregating) or subagging (subsampled bootstrap aggregating). In the conventional outlier identification methods listed above (3σ, IQR, and MAD), increasing or decreasing trends in time-series data are not explicitly considered because statistical tests for outliers are conducted based on observed data that are often limited in number. In the modified methods proposed in the present study, a statistical analysis based on a large number of trend estimations from resampled observations are carried out, and a general trend in the time-series data can be effectively drawn. Additionally, based on the proposed methods, the asymmetric statistical dispersion of the observed data along major trends can be addressed by adopting the non-parametric statistics from the ensemble of the estimated trend from resampling, which enables the proposed methods to identify outliers even for highly skewed non-Gaussian data.

In the following sections, an example of the ensemble regression method employed in this study is briefly explained, and then three modified outlier identification methods based on the ensemble estimations are described in detail.

2.1. Ensemble regression estimator

The groundwater state time-series $D = [(t_1, y_1), \cdots, (t_n, y_n), \cdots, (t_N, y_N)]$ is set as the training information observed at a monitoring well where t_n is an explanatory (or independent) variable of a measured time, y_n is a response (or dependent) variable of either a qualitative or quantitative groundwater state observed at t_n, and N is the total number of observations. Linear regression has been often adopted to explain and estimate temporal changes in the groundwater state over long periods of time. In the present study, linear regression is also adopted as a key estimator of the groundwater state change while the application to the problem is different from the conventional methods.

It is well understood that diversity among the estimations promotes the prediction accuracy in ensemble approaches (Brown et al., 2005). Therefore, in this study, a limited number (N_s) of data are resampled (i.e., subsampled, $N_s < N$) from the training data (D) without replacement, and estimations are made repeatedly from the resampling based on the concept of subagging (Büchlmann and Yu, 2002), where a small number of subsamples compared to the sample size (N) is used to maximize the diversity of the estimations. The number of Monte Carlo samplings of estimations, M, is decided based on the consistency of the statistics, which have more or less asymptotic characteristics with an increase in M. For consistency purposes, a sufficiently large number, $M = 5000$, is used in this study. In the estimation of the weight vector of every subsample ($\boldsymbol{\omega}^{\mathrm{MLE}}$, maximum likelihood estimation of weights), the ordinary least squares (OLS) method is employed by assuming residuals of subsamples around their optimal trend with Gaussian distribution as

$$\boldsymbol{\omega}^{\mathrm{MLE}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}, \tag{1}$$

where $\mathbf{A}$ is the design matrix, superscript T indicates the transposed matrix, and $\mathbf{y}$ is a set of response variables. Notice that the assumption of a Gaussian distribution of individual subsamples does not necessarily result in a Gaussian distribution in the ensemble of the Monte Carlo results.