Outliers in groundwater quality time series

ZBIGNIEW W. KUNDZEWICZ
Institute of Geophysics, Polish Academy of Sciences, 01-452 Warsaw, Poland (also at the University of Karlsruhe)

JÜRGEN IHRINGER, ERICH J. PLATE
Institute of Hydrology and Water Resources, University of Karlsruhe, D-7500 Karlsruhe 1, FR Germany

JOST GRIMM-STRELE
Landesanstalt für Umweltschutz Baden-Württemberg, Karlsruhe, FR Germany

Abstract The problem of the detection of outliers in time series of non-equidistant groundwater quality data is addressed. The operational check of data consistency, i.e. the procedure of judging if the newly acquired data element fits the temporal structure of the remainder of the data set, is presented. This includes issuing the forecast of the variable in question, assessment of the variance of the forecast error, and checking if the newly incoming data point does not substantially depart from the forecast value. The methodology used is based on a modified exponential smoothing method which, in the problem considered, was more feasible than Kalman filtering and the geostatistics of temporal fields.

Ecarts dans les séries temporelles de qualité des eaux souterraines

Résumé On aborde dans cette communication le problème de la détection des écarts dans les séries temporelles nonéquidistantes des données de qualité des eaux souterraines. On présente ici le contrôle opérationnel de consistance des données, c'est à dire la méthodologie par laquelle on estime si les éléments de données nouvellement acquises s'ajustent bien à la structure temporelle du reste de la série de données. Cette méthodologie comprend: la prévision de la variable en question, la détermination de la variance de l'erreur de prévision et la vérification que les points correspondant aux nouvelles données ne d'écartent pas substantiellement des valeurs prévues. La méthodologie employée est basée sur une méthode de lissage exponentielle modifiée, qui pour le problème considéré, est plus pratique que le filtre de Kalman et que les procédés géostatistiques des champs temporels.

INTRODUCTION

The amount of groundwater quality data collected repetitively in several
locations and pertaining to some dozens of parameters is very high. Therefore, in order to screen the data, one has to use methods of plausibility analysis that are both reliable and objective and also lend themselves well to routine applications. Such methods should render the means for identifying, interpreting and classifying suspect data points. A plausibility analysis of an incoming data point should be made before the point in question is accommodated into the data base.

The term outlier will be used herewith with reference to such observations that do not follow the pattern present in the remainder of the set of data. There are several possible sources of outlier observations, notably errors of various types (instrument malfunctioning, human mistakes) or unexpected events that have really happened. Statistical methods do not seem, in general, to offer the means for distinguishing the sources of outliers.

Most procedures for the detection of outliers given in the literature do not lend themselves well to the specific situation of structured groundwater quality data where the time series of data are often acquired with unequal time intervals (non-equidistant data) at one station and observations at different stations are not performed simultaneously. The methods offered in the literature apply to the case of equal time intervals. However, also in this case, Barnett & Lewis (1984), the authorities in the broad area of outliers, used the wording "Outliers in time series — a little explored area".

Although, theoretically, there are two kinds of outliers that can occur in time series (additive and innovative), it is additive outliers that emerge in all cases embraced by this study.

In the study of IHW (1986) reported in the paper by Hiessl et al. (1987), a method for operational detection of outliers in equidistant time series of groundwater quality data was suggested. Hiessl et al. (1987) used a particular case of a low-pass filtering method, i.e. a variant of exponential smoothing. In the present report an extension of the method of exponential smoothing to the realistic case of data acquired non-equidistantly is described. The more complete results of the study are presented in IHW (1988). Within the framework of this study, other methods of detection of outliers in groundwater quality time series were also considered, namely Kalman filtering and geostatistics that are both applicable to the case of non-equidistant data. This can be done within the framework of a continuous Kalman filter that has been applied, for instance, in medical statistics in monitoring the state of health of leukaemia patients who are not controlled at regular time intervals after they have left hospital (Jones, 1984). Use of a discrete Kalman filter with non-equidistant data is also possible, with a small time interval that is the common divisor of all the time intervals between consecutive measurements. A Kalman filter formulated for a vector state representation allows one to perform, in principle, a plausibility analysis of spatial-temporal data.

The time series of groundwater quality data available in a particular location can be used in the geostatistical detection of outliers. After the temporal semivariogram has been determined, every newly incoming data can be compared with the kriged estimate and, in the case of a large discrepancy between them, it can be labelled an outlier. Application of geostatistics to the detection of outliers in structured groundwater quality data (in the case of
temporal, spatial, and spatial-temporal structures) is presented in IHW (1988), Kundzewicz et al. (1989) and Bárdossy & Kundzewicz (1989).

The finding of the study was that, for the purpose in question, a modified exponential smoothing was an acceptable operational tool. In the Kalman filter method one needs to identify the parameters of the state model and the noise covariances. This is, in principle, possible in several ways (nonlinear optimization, augmenting the system state by the parameters to be identified, and — last — manual fitting via trial and error). However, none of the methods fulfills the requirement stated in the problem formulation — operational on-line applicability with low computational time and storage requirements.

Also, geostatistical methods are, in the application considered, computationally more demanding, that is less feasible than the modified exponential smoothing. Moreover, the time series available in the case studies considered embrace too few temporal points for reliable determination of a temporal semivariogram (even if the assumption of stationarity is made). Therefore, the inherent problems of small data sets manifest themselves in the unsatisfactory semivariograms featuring oscillations with unacceptably high amplitude.

EXPONENTIAL SMOOTHING FOR EQUIDISTANT DATA

The method of exponential smoothing for equidistant data is very common in the literature on applied time series analysis, economy etc. A good presentation of exponential smoothing has been given, for instance, in Thomopoulos (1980).

Exponential smoothing yields a forecast for the time instant $t_{n+1}$ based on the last incoming data point $z(t_n)$, the last forecast value $\hat{z}(t_{n-1}|t_n)$ and one model parameter $\alpha$. The model equation reads:

$$\hat{z}(t_{n+1}|t_n) = \alpha z(t_n) + (1 - \alpha) \hat{z}(t_{n-1}|t_n)$$

(1)

The time interval between consecutive time instants of measurements does not occur in equation (1). Thus the basic variant of exponential smoothing is valid for the equidistant case only.

Although the form of equation (1) is conveniently recursive, the newest forecast value obviously depends on all points of the time series observed so far. The weights given to particular past time instances are smaller the older the data point. Equation (1) can be reformulated as:

$$\hat{z}(t_{n+1}|t_n) = \alpha z(t_n) + (1 - \alpha)[\alpha z(t_{n-1}) + (1 - \alpha) \hat{z}(t_{n-2}|t_n)]$$

$$= \alpha z(t_n) + (1 - \alpha) \alpha z(t_{n-1}) + (1 - \alpha)\alpha^2 z(t_{n-2}) + ...$$

$$+ \alpha(1 - \alpha)^{n-1} z(t_1) + (1 - \alpha)^n \hat{z}(t_0|t_1)$$

$$= \alpha \sum_{j=0}^{n-1} (1 - \alpha)^j z(t_{n-j}) + (1 - \alpha)^n \hat{z}(t_0|t_1)$$

(2)
Adjusting the parameter $\alpha$, one can determine the dependence of the forecast upon the last incoming data point. The higher the value of $\alpha$, the stronger the smoothing (i.e., the participation of the filtered part of the forecast). The sum of all weights reads:

$$\alpha \sum_{j=0}^{n-1} (1 - \alpha)^j + (1 - \alpha)^n = 1 \quad (3)$$

If the number $n$ is high, one can assume that the weights attributed to the oldest data points are negligibly small. That is:

$$\lim_{n \to \infty} (1 - \alpha)^n = 0 \quad (4)$$
$$\lim_{n \to \infty} \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j = 1 \quad (5)$$

and

$$\hat{z}(t_n, t_{n+1}) = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j z(t_{n-j}) \quad (6)$$

Thomopoulos (1980) developed the following equation for the variance of the forecast error, $e(t_{n+1}) = \hat{z}(t_n, t_{n+1}) - z(t_{n+1})$:

$$\text{var}[e(t_{n+1})] = \text{var}[\hat{z}(t_n, t_{n+1})] + \sigma^2 \quad (7)$$

where $\sigma$ is the standard deviation of $z$.

Since the variance of the forecast can be approximated as:

$$\text{var}[\hat{z}(t_n, t_{n+1})] = \alpha^2 \sum_{j=0}^{\infty} (1 - \alpha)^j \sigma^2 = \frac{\alpha}{2 - \alpha} \sigma^2 \quad (8)$$

the following simple result can be obtained:

$$\text{var}[e(t_{n+1})] = \frac{\alpha}{2 - \alpha} \sigma^2 + \sigma^2 = \frac{2}{2 - \alpha} \sigma^2 \quad (9)$$

One can assess the variance participating in equation (9) recursively by exponentially smoothing the squares of the data points (see IHW, 1986; Hiessl et al., 1987).

In order to develop a variant of exponential smoothing applicable for the non-equidistant case, it is necessary to distinguish the concepts of the smoothed estimator $\tilde{z}(t_n)$ and the forecast $\hat{z}(t_{n-1}, t_n)$. The new value of the smoothed estimator $\tilde{z}(t_n)$ depends on the last value of the smoothed estimator $\tilde{z}(t_{n-1})$ and the newly incoming data point $z(t_n)$. This is the posterior (updated) estimate after the data point $z(t_n)$. In the present formulation the forecast is based on the last value of the smoothed estimator $\tilde{z}(t_{n-1})$. This is the prior estimate, before the data point $z(t_n)$ is known.

Assume that both the smoothness estimator $\tilde{z}(t_n)$ and the forecast
Outliers in groundwater quality time series

\( \tilde{z}(t_n, t_{n+1}) \) are of the exponential smoothing form:

\[
\tilde{z}(t_n) = \alpha z(t_n) + (1 - \alpha) \tilde{z}(t_{n-1})
\]

(10)

and

\[
\tilde{z}(t_n, t_{n+1}) = \alpha z(t_n) + (1 - \alpha) \tilde{z}(t_n)
\]

(11)

That is:

\[
\tilde{z}(t_n) = \tilde{z}(t_{n-1}, t_n) + \alpha [z(t_n) - z(t_{n-1})]
\]

(12)

EXPONENTIAL SMOOTHING FOR NON-EQUIDISTANT DATA

Although the original data are not equidistant, the details of timing play a role that decreases with time (cf. the name of the method — exponential smoothing). Therefore, assume for the time being that all time intervals between consecutive past observation points are the same. They can be, for instance, considered equal to the interval between the last and the next-to-last observation, or — as taken in this work — to the recursively updated mean period \( \bar{T} \). The recursive updating formula reads:

\[
T_{n+1} = \frac{n}{n+1} T_n + \frac{T_{n+1}}{n+1}
\]

(13)

where \( T_n \) is the mean interval up to the \( n \)th observation and \( T_{n+1} \) is the interval between the \( n \)th and the \( (n + 1) \)th observation. Now one can apply the basic exponential smoothing methodology to produce the forecast for the time instant \( t + \bar{T} \), where \( t \) is the time point where the last observation is made and \( \bar{T} \) is the updated mean interval between consecutive observations. That is, for the time instant \( t + \bar{T} \) one could write:

\[
\tilde{z}(t + \bar{T}) = \alpha z(t) + (1 - \alpha) \tilde{z}(t)
\]

(14)

\[
\text{var}[e(t + \bar{T})] = \frac{2}{2 - \alpha} \sigma^2(t)
\]

(15)

\[
e(t + T) = \tilde{z}(t, t + T) - z(t + T)
\]

(16)

However, the time instant for which a forecast is needed is not \( t + \bar{T} \), but \( t + T' \), where in general \( T' \neq \bar{T} \). Therefore interpolation is used, if \( T' < \bar{T} \) and extrapolation if \( T' > \bar{T} \) for both the forecast value and the estimate of the variance of the forecast:

\[
\tilde{z}(t, t + T') = \tilde{z}(t) + [\tilde{z}(t, t + \bar{T}) - \tilde{z}(t)] \frac{T'}{\bar{T}}
\]

(17)

\[
\text{var}[e(t + T')] = \sigma^2(t) + \{\text{var}[e(t + \bar{T})] - \sigma^2(t)\} \frac{T'}{\bar{T}}
\]

(18)
EXPONENTIAL SMOOTHING FOR DETECTION OF OUTLIERS

In addition to the newly incoming information, once one disposes with the forecast based on the information present in older data and also the assessment of the variance of the forecast error, one can perform a plausibility analysis by evaluating the following ratio:

$$\text{AND}(t) = \frac{|z(t) - \hat{z}(t - T_1, t)|}{\sigma}$$

(19)

where:
- AND = absolute normalized deviation,
- $z(t)$ = measured value at the time instant $t$,
- $\hat{z}(t - T_1, t)$ = forecast made by modified exponential smoothing, based on information present in all the older data,
- $T_1$ = time interval between the last observation and the time point at which the new observation was made,
- $\sigma$ = standard deviation of the forecast error.

The operational check of consistency, i.e. the procedure of judging if the newly incoming data point fits the temporal structure that already exists in the data, contains three steps:
- production of the forecast for the time instant at which the newly incoming observation is taken, based on the earlier data;
- assessment of the variance of the forecast error;
- checking if the newly incoming data point does not substantially differ from the forecast value.

A data point is considered to be an outlier if the variable AND attains a value higher than some parameter $C$. The meaning of $C$ is intuitively straightforward. The probability that a normally distributed variable departs

![Fig. 1 Plausibility analysis of a time series of chloride concentration (case study no. 1, well no. 926, parameter values $\alpha = 0.5$, $C = 2.0$).](image)

from the mean by more than $2\sigma$ (this corresponds to $C = 2$) is equal to $p = 4.55\%$. For $C = 1$ it is $p = 31.73\%$ and for $C = 3$ it is $p = 0.27\%$. For non-normal distributions the values of $C$ that yield the above probabilities may be substantially higher, as results from the Tschebyscheff inequality.

**CASE STUDIES**

The operational method for the detection of outliers in groundwater quality time series has been tested on two case studies from the Upper Rhine Valley, FR Germany. The analysis was made for two water quality characteristics — chloride concentration and total hardness.

The first system consists of 24 wells where data have been collected
non-equidistantly over a period of 11 years with time intervals between consecutive observations in a single well ranging from one month to over two years. An example of the performance of the algorithm is shown in Fig. 1, where upper and lower boundaries of plausible (i.e. non-outlying) values of chloride concentration and the observed time series are presented.

Some difficulties arose for several wells in the initial phases of the plausibility analyses. A compromise was found necessary between severity and gentleness in the run-up phase of the algorithm, where the assessment of the variance of the prediction error may be poor. Severity means the detection
of outliers that only departed slightly from a weakly stated norm. Gentleness is understood as treating "obvious" outliers as plausible values, based on the largely uncertain initial assessment of the estimation variance.

The second system analysed consists of two wells located very close to one another. The time series of groundwater quality characteristics are longer (over 40 values) but still non-equidistant with intervals between consecutive observations varying from less than one month to over two years. The results of the outliers detection are shown in Fig. 2(a) and (b).

There are two parameters to be chosen in the procedure:
\( \alpha \) the exponential smoothness parameter, and
\( C \) the maximum plausible value of AND (absolute normalized deviation).
By manipulating the values of these parameters, one can influence the number of data points falling outside the plausible limits. The influence of both parameters on the plausibility analysis for chloride concentration (i.e. the time series of Fig. 2(b)) is shown in Fig. 3(a) and (b).

CONCLUSIONS

It has been found that a simple modification of exponential smoothing allows crude screening of groundwater quality data acquired at non-equal time intervals. The methods with a better reputation, like geostatistics and Kalman filtering, are more costly and more demanding. Exponential smoothing was found promising in operational applications, where thousands of wells and millions of data points may be of concern.

Plausibility analysis of groundwater quality data is an important step in the building of data bases necessary for decision making on water resources.

REFERENCES


