EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

Semiautomatic Verification of Groundwater Measured Data

Uwe Rueppel, Peter Goebel[1]

Abstract

Groundwater is an unseen yet vital resource. It is the most important resource for potable water supply in Germany. Particularly in the state of Germany Hesse, where drinking water originates to 95.2% from groundwater, there is no substitute (HLUG 2007). However, groundwater does not only play an extraordinarily decisive role in water-economical regards, but as a substantial component of the hydrological cycle, it must be seen with its great importance for the ecological system. The main objectives of groundwater monitoring are to study the variations and long-term trends in the quantitative and qualitative condition of groundwater. The provided information is intended to serve as basis for assessment of environmental quality goals and norms, to ensure compliance with regulations and to prevent excessive use of the groundwater supplies for a sustainable groundwater management. In order to give a representative picture of the groundwater status in consideration of the complexity of hydro-geological systems and heterogeneous groundwater bodies, the net of measuring points is very dense. In a large scaled catchment area, there will be easily some thousands of these groundwater objects, each of them producing groundwater data. In addition with the trend of data loggers in crucial or hard accessible rough nature environments for an automated retrieval and storage of information from one or more sensors, enormous data pools are the result. Due to various reasons, these data do not always have the required quality in order to accomplish the necessary analyses. Since the data pools are too big a manual check up is not suitable. For these reasons series of measurements are in principle to be automatically examined for plausibility – whereas a skilled worker still has the opportunity to make the final decision - to obtain high quality data.

1. Data Quality

Before we discuss possible active improvements of the data quality we first introduce the notion of data quality and then get from a broad definition to a more specific view, regarding the groundwater monitoring domain.

In information society, data quality is a concern for professionals involved with a wide range of information systems. Data quality plays a crucial role in all data processing domains. For most data-driven applications the so-called garbage-in-garbage-out-principle applies. This means that applications will unquestioningly process the most nonsensical of input data and produce nonsensical output. In order to avoid problems arising from low quality data, international research communities have been formed (e.g., the International Association for Information and Data Quality (IAIDQ) and the Deutsche Gesellschaft für Informations- und Datenqualität e.V. (DGIQ), both established in 2004).

The quality of data is often defined as the suitability of the data for a specific data-processing use, e.g., after (Juran/Blanton, 1999) we define "data to be of high quality if they are fit for their intended uses in operations, decision making and planning". This can be put short in “fitness for use” (Kahn/Strong/Wang, 2002). It must be noted, that the term data quality is often used synonymously with information quality. The quality of data is characterized - and to some extent measured - by how they keep conditions of various criteria, called *dimensions*. There are nearly 200 with data quality associated dimensions, each covering a specific aspect of data quality (Naumann, 2007). Concerning the dimensions, there is no general agreement among the data quality researchers, either on which set of dimensions defines data quality or on the exact meaning of each dimension (Batini/Scannapieco, 2006). However, these lists of dimensions commonly (according to their meaning) include: *Accessibility, Appropriate Amount of Information, Be-*

[1] Institute of Numerical Methods and Informatics in Civil Engineering, Department of Civil Engineering and Geodesy, Technische Universitaet Darmstadt, Petersenstrasse 13, Darmstadt, D-64287, PH (+49) 6151-163444, FAX (+49) 6151-165552, e-mail: rueppel, goebel@iib.tu-darmstadt.de, internet: http://www.iib.bauing.tu-darmstadt.de

lievability, Completeness, Concise Representation, Consistent Representation, Ease of Manipulation, Free-of-Error etc. In summary data quality can be defined as a set of quality dimensions The selection of the relevant dimensions and the exact definition of the dimensions is reserved to the expert of the respective application domain.

This papers focuses upon concepts to deal with the data quality dimensions Completeness and particularly Free-of-Error. Completeness can be defined by the extend to which information is not missing and is of sufficient breadth and depth for the task at hand and Free-of-Error by the extend to which information is correct and reliable (Kahn/Strong/Wang, 2002). Before we discuss how to free the data sets of errors, different types of errors will be presented.

2. Error Types and Completeness

The measurement error is understood as the difference between a measured value and the correct (true) value, whereas normally the "truth" is hard to know. This chapter gives an overview on how errors can arise during the gauging process and whether these errors endanger the data quality regarding their suitability for the groundwater management.

From the collection to the storage into a data base a measured value is exposed to various possibilities of errors. Problems can arise from low equipment quality, inappropriate measurement methods, mix-up of identifiers, measuring, and writing or data transfer errors. Measure points are very often placed in unobserved rough nature environments. Hence, the most common sources for measurement errors are external impacts. The reason for these impacts are widespread, e.g., vandalism, damage caused by animals, drifted measurement points, frost and heat.

We classify three types of errors: Firstly, *accidental errors* (or *random errors*) are always irregularly distributed and are the result of a multiplicity of influences and coincidental combinations of diverse changes in the operational procedure. However, not all the measurement affecting conditions can be defined and the accuracy of the measuring instruments is limited as well. Accidental errors are unavoidable. They can not be prevented by a plausibility check up. These types of errors are, however, limited and do not endanger the quality of the data. A number of accidental errors are usually a priori excluded by the use of data logger. To this kind of excluded accidental errors belong the mix-up of two measuring points or reading off and writing errors (e.g., input of the comma in the wrong place, transposed digits, meter error, slipping in the line when copying values from tables etc.).

Secondly, *biased errors* (or *systematic errors*) are based on inaccurate measuring methods and incorrect measuring instruments. One recognises a biased error by the fact that all measured values are falsified on one side. They also can hardly be recognised by a plausibility check up (Figure 1).

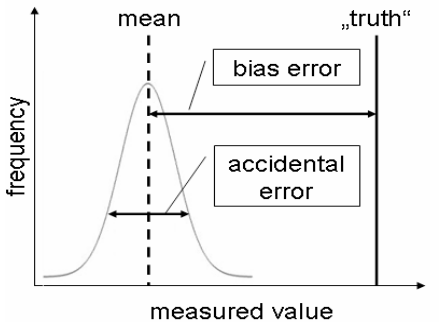

Figure 1: Bias and accidental error

Thirdly, *gross errors* result mostly from gross malpractice of the user or from significant external impacts. The resulting error is thereby so large that a plausibility check up can help to identify and mark suspected outliers. Unlike methods of labelling outliers in time-invariant normal distributed data sets, it is not of interest to spot data points with the most extreme deviation from the arithmetic mean.
What is rather requested is an approach which takes into account the dynamics and time-variance of groundwater levels. In this article, an outlier is referred to as an observation that lies an abnormal distance from the surrounding data points.

The crucial point is to decide what will be considered as abnormal or normal. Furthermore, one can differentiate between two types of outliers (Fox 1972). The simplest case is the additive outlier (AO), where a single observation in the time series is affected, being unreasonably high or low. This kind of outlier affects only the current observation. An innovation outlier (IO) affects future values of the series. IO's have an after-effect and produce a perturbation (Figure 2).

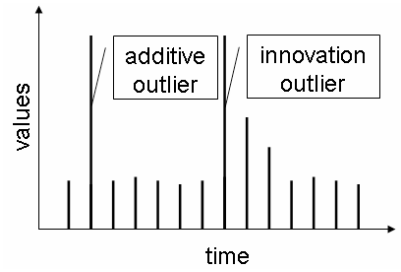

Figure 2: Additive and innovation outlier

Another reason for low quality data sets in groundwater monitoring are missing values (low completeness). This can be sporadic missing values or blocks of values. In order to perform a 'normal' analysis in this situation the gaps need to be closed. Even though the analysis should ideally take into account that there is a greater degree of uncertainty than if the calculated values had actually been observed. Methods to improve completeness are addressed briefly in the outlook.

3. Test Criteria

There are various ways of controlling the quality of groundwater measured data in practise. Trivial check ups are the basis for plausibility tests. They check that the data is not trivially invalid, e.g., the measurement date can not be after 'today' and a percentage measurement must be in the range 0 to 100. Another popular method for the identification of gross errors is bound checking with boundary values. E.g., a groundwater engineer knows about a measuring point, that its highest level ever gauged was 90m (upper boundary value) and the lowest 86m (lower boundary value). If he then puts these values into this simple set of laws, the regulation will be violated with a measured groundwater level of 90.2m. These tests already provide useful results.
But these rules don't take into account the unsteady dynamics of groundwater levels (e.g., during different seasons) and don't regard the distance between the value to be checked and the surrounding data points (Figure 3).

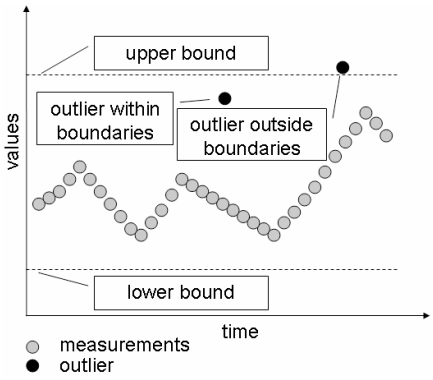

Figure 3: Bounds checking

To meet this requirement advanced methods are common.

The simplest method is the comparison of the spacing to the last measured value (in Figure 4 distance [a]). The groundwater measured variables regarded in this work describe the condition of a groundwater body whose characteristics usually only slowly change. For example, a change of groundwater levels takes place relatively slowly. Therefore exceptional large distances to the value measured before can refer to outliers.

If the measured values within a temporal range vary around an (moving) average value, then the distance of the value to an average value of the last n values (in Figure 4 distance [b] with n=3) can be computed and compared with a preset boundary value.

As soon as the series of measurements show a linear trend, the distance of the value to be checked to a computed value by means of linear involution provide a better informative basis (in Figure 4 distance [C]).

If the trend of the actual measured value pattern is non-linear and is reproduced better by the graph of a parabolic function, the distance from the value to be checked to a computed value on the basis of polynomial regression allows better results (in Figure 4 distance [D]) (Greifenhagen, 2000).

Even though these tests already provide good results, they have the disadvantage to require a high level of manpower. These methods are based upon existing experience about the related measuring point or region, e.g., the maximum difference of levels in one month was 0.5m (spacing to the value measured before). A high level of manpower is needed to build up this set of laws. For each of the possibly thousands groundwater objects suitable boundary values have to be considered and decided. On the other hand, the boundary values are steady and not dynamic. Hence, the same boundary values apply for springtime as much as for autumn, unless the temporal scope is split into several values (e.g., separate preset boundary values

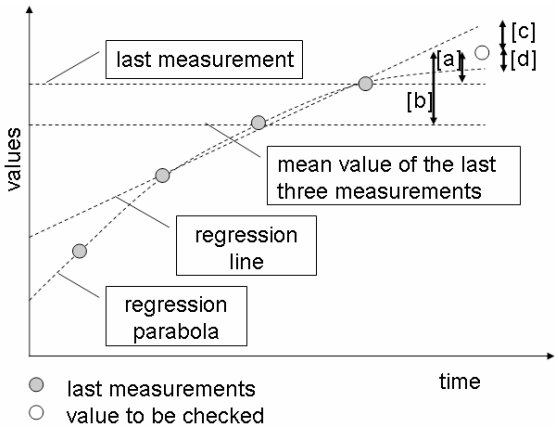

Figure 4: Test criteria

for springtime and for autumn), which would involve at the same time a proportional higher level of manpower. With these methods, the test criteria will be a trade-off between accuracy and manpower.

What is requested is an approach which takes into account the dynamics of groundwater levels during different seasons without manually adjusting rules for each groundwater object. For this reason statistical methods and methods of time series analysis have been investigated upon groundwater data sets.

4. Statistical Methods and Groundwater Measured Data

Several statistical methods are available for outlier detection in univariate data. In groundwater measured data the outlier can be either too small or too large and there is no reason to assume a specific number of outliers, hence the attention is focused on two-tailed outlier detection procedures for an undefined number of outliers in rather big sample sizes. Gibbons (Gibbons 1994) describes several methods for detecting outliers in groundwater quality data, measured near waste disposal facilities to detect release from the facilities. Most of these tests are also suitable for detecting outliers in groundwater levels. All these tests have their advantages and disadvantages.

The Rosner test is a generalisation of the Grubbs test, which is for testing a single outlier, to use for multiple outliers. To use the Rosner test, an upper limit of outliers must be specified. Skewness Test, Kurtosis Test and Shapiro-Wilk Test are outlier tests for normally distributed variables. These tests are based on a sample statistic, whereas the value of the statistic is compared to the critical value provided in corresponding tables. They can be repeatedly applied until the sample value of the statistic is less than the critical value, but the accuracy decreases with the number of repeated evaluations. The E_m-Statistic and the Dixon's test can be used for outlier detection in data sets in which a small number of outliers are suspected. Using these test one has to assume a number of outliers first. Setting the number too large, valid measurements may be marked as outliers too, whereas setting the number too small, not all outliers may be detected.

All these tests work reasonably well as long there are only a few outliers are present (Skewness Test, Kurtosis Test and Shapiro-Wilk Test) or there is a reason to assume a specific number of outliers (Rosner test, E_m-Statistic and the Dixon's test). To analyse groundwater levels it is useful to divide long time series into smaller time periods to be checked separately.

5. Time Series and Groundwater Measured Data

A time series is understood as a row which consists of observed values in chronological order (Stier/2001). The discrete data are typically measured at successive times and spaced at (often equidistant) time intervals. In this manner groundwater data over the time can be interpreted as time series. Time series analysis comprises methods that attempt to understand such time series by detecting structures and regularities. They try to explain sequences of data points from themselves. With adequate methods, no more input is necessary apart from the actual time series. The application range of time series analyses covers all scientific fields which are concerned with questions of time-dependent developments, e.g., share prices, sales figures, demographical rows, manufacturing processes, and meteorology. Thus it is also suitable for the investigation of groundwater hydrograph curves as it offers broad mechanisms for investigation and evaluation.

If time series include outliers, then the square deviation is, e.g., strongly affected with the method of the least squares. Therefore, for example 100 deviations of the size of '1' result in the same weight as 1 distance of the size of '10'. Due to the fact, that the groundwater time series can include outliers, the approach must be robust against the influence of outliers.

Within this work a prototypical software-tool is implemented with a robust approach with autoregressive models using a generalized m-estimator as proposed by Schlittgen (Schlittgen 2001).This means that for each time series an autoregressive (AR) model is build with a maximum likelihood estimator (MLE).

Then an approximative "conditional mean type" (ACM) robust filter is applied to the time series (Martin 1979) to free the row from noise. This filtered row can then be compared with the measured row. If the difference is still very great and many alleged outliers are identified – which are due to the user experience not all outliers in fact – the row computed by the stochastic model can be put closer to the measured row by smoothing. The user can choose between two different smoothing algorithms to find the best fit. In most cases the best results are achieved with not using the smoothing algorithms and they are off-state by default. Data which can not pass these verification tests are piped with its computed models to graphical processing modules to provide an adjustment.

The computed model (continuous line) and the hydrograph curve (dotted line) are presented to the user by the implementation of an interactive chart. The continuous line overlies the dotted line and only if computed values diverge from original values, the dotted line is shown as an indication of a suspected out-lier (Figure 5).

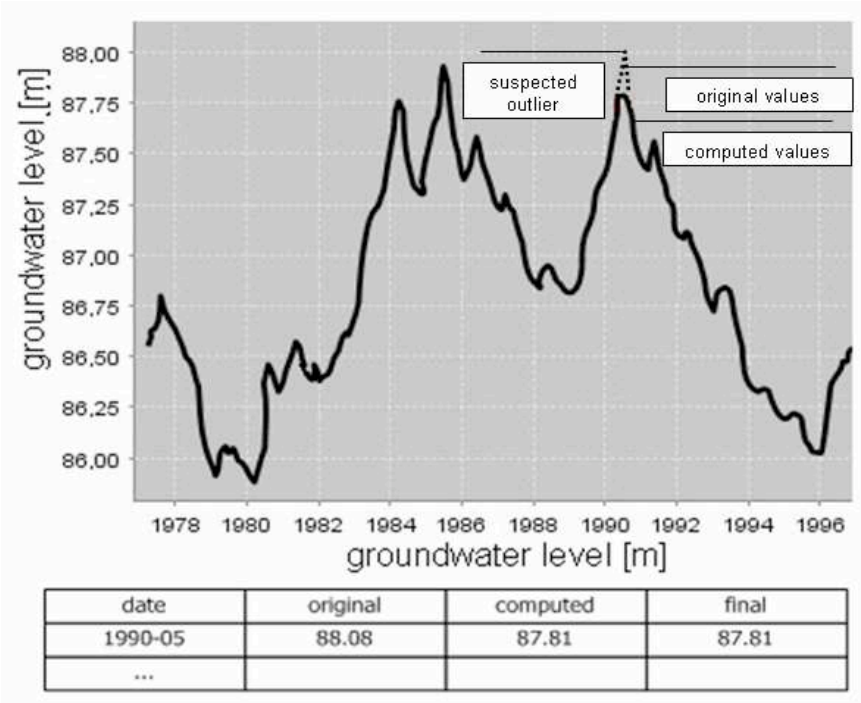

Figure 5: Graphical representation of a hydrograph curve

According to the recommendation of the 'Working Group of the German Laender on Water Issues (LAWA)' the measurement and the computed model are graphically prepared in a hydrograph curve. These can be verified by a skilled worker [LAWA 1999].Thus a fast and suitable survey of the development can be achieved. When a point in the time series is selected, the data for the chosen measurement is shown in a table next to the graph. Alongside the measured value and the value computed by the model, the date of the measurement and the difference to the previous and following measured values is shown for a better orientation. The interactive graph opens the corresponding data set in tabular representation

by double-clicking on it. Here the skilled user finally decides with his experience about the correctness of the data. Unlike most domains concerned with time series, groundwater management deals with many time series measured at the same time (e.g., gauged groundwater levels at different places over a specific period of time). Some of theses time series are similar. Information obtained from similar time series can be consulted for decision making. In doubtful situations, the user can compare, e.g., the current hydrograph curve with a hydrograph curve of a gauging station close to it and in the same groundwater body. To find a good gauging station to compare with the time series, the user gets support by the correlation analysis.

The correlation analysis represents one of the most important components of statistical methods in the hydrogeology. The correlation analysis examinees the stochastic interrelations between equivalent variables and estimates the strength of dependence and confidence regions (Greifenhagen 2000). A correlation of groundwater time series is 'good', if it is beyond 0.9 and is calculated by monthly values of at least 3-4 years (LAWA 1999). The correlation coefficient does not change as long as the master data – which determine the groundwater level in both groundwater gauging stations – do not vary. So the correlation is computed once and needs only to be checked in appropriate time intervals. In terms of clearness to the user an optical comparison of two time series in form of a difference hydrograph curve is a good aid and is presented to the user in combination with the correlation coefficient. It is important to notice that a good correlation does not necessarily imply causation. But measuring points may only be compared with each other if there is causation. Hence, the program provides a list of gauging stations in the closer periphery with good correlation but the user eventually has to decide about the usability of each correlation.

After the data is checked by the user the data can then be edited and corrected or rejected. Thus suitable treatment is offered for, e.g., incorrect long term trends of a gauging station and temporary mismeasurements.

6. Conclusion and Outlook

It is common to find outliers and to deal with missing data in groundwater measured data. Low quality data often hinder 'normal' analysis work. Outliers usually demand special attention, since they may indicate problems in sampling or data collection or transcription. In this paper a variety of methods to detect and free the data sets of errors are presented. Most offered test criteria are implemented in the software-system 'Grundwasser-Online" (GWO 2007). The statistical and time series methods presented in this paper are still subject of our research. They have been implemented with Java© and R (R 2007) and show encouraging results. R, a free language and software environment for statistical computing and graphics, provides methods for time series analyses and has got an outlier package for most presented statistical methods.

There are different concepts to deal with missing data we experiment with at the moment, in which time series analysis and statistical methods are involved. One possible approach is to adapt a special model, e.g., an AR model to such a missing values time series. The results can be improved; by prognosis of the missing values by the model adapted so far. With the so assessed values the estimation of the autocorrelation function is improved and thus the AR model will be better (Schlittgen 2001). Other techniques use the least squares methods to compute missing values by means of linear combinations of the existing values or statistical methods of multiple imputation. Another approach to close gaps in time series in groundwater management uses the information of time series of correlating measuring points.

Methods of computer science can help to discover and correct many of these errors, but the struggle for high data quality can only be successful if it is paired with an abatement of the error causes.

Bibliography

Batini, C. and Scannapieco, M. (2006): Data Quality, Concepts, Methodologies and Techniques

Fox, A.J. (1972): Outliers in time series. Journal of the Royal Statistical Society, Series B, pp.350-363.

Gibbons, R.D. (1994): Statistical Methods for Groundwater Monitoring; Wiley-Interscience; 1st edition

Greifenhagen, G. (2000): Untersuchungen zur Hydrogeologie des Stadtgebietes Darmstadt mit Hilfe eines Grundwasserinformationssystems - unter Verwendung von einer Datenbank, Datenmodellierung und ausgewählten statistischen Methoden.

GWO (2007): www.grundwasser-online.de

HLUG (2007): Hessian national office for environment and geology, www.hlug.de

Juran, J. M., Blanton, A. G. (1999): Juran's Quality Handbook, Fifth Edition, p. 2.2, McGraw-Hill, 1999.

Kahn, B., Strong, D., Wang, R. (2002): Information Quality Benchmarks, Product and Service Performance, Communications of the ACM, April 2002. pp. 184-192.

LAWA (1999): Empfehlungen zur Optimierung des Grundwasserdienstes (quantitativ), Länderarbeitsgemeinschaft Wasser.

Martin, R.D. (1979): Approximate conditional-mean type smoothers and interpolators, in: Smoothing techniques for curve estimation, Proc. Workshop, 1979

Naumann, F. (2007): Datenqualität, in: Informatik Spektrum, S. 27-32

R (2007): http://www.r-project.org

Ruckdeschel, P. (2001): Ansätze zur Robustifizierung des Kalman-Filters.

Schlittgen, R. (2001): Angewandte Zeitreihenanalyse.

Stier, W. (2001): Methoden der Zeitreihenanalyse.