Prepared
enabling change

Literature Review of Data Validation Methods

SEVENTH FRAMEWORK
PROGRAMME

Prepared
enabling change

Literature Review of Data Validation Methods

SEVENTH FRAMEWORK
PROGRAMME

COLOPHON

Title
Literature review for data validation methods

Report number
PREPARED 2011.019

Deliverable number
D3.3.1

Author(s)
Siao Sun; Jean-luc Bertrand-krajewski ;
Anders Lynggaard-Jensen;
Joep van den Broeke; Florian Edthofer;
Maria do Céu Almeida; Àlvaro Silva Ribeiro; José Menaia

Quality Assurance
By Christopher Huton

Document history

Version	Team member	Status	Date update	Comments
1.0	Siao Sun	Draft	2011.3.28	
1.1	Anders Lynggaard-Jensen; Joep van den Broeke; Florian Edthofer; Maria do Céu Almeida; Àlvaro Silva Ribeiro; José Menaia.	Draft	2011.6.8	Review/additions
1.2	Christopher Huton; Jean-luc Bertrand-krajewski.	Draft	2012.2.27	Review/comments

This report is:
PU = Public

Contents

1 Concept, scope and background

Increased efficiency of existing water systems will more than ever require reliable data. Reliable data play a key role in the analysis, monitor, and forecast of water system behaviors as bad quality data may result in an erroneous decision scheme. These data are provided in a large part by the measuring system, in which a sensor is an important element. The sensor measures a physical quantity and converts it into a signal that can be read by an observer or by an instrument. The measuring system then converts the sensor signals to values aiming to represent certain "real" physical quantity. These values, known as raw data, needs to be validated before other use in order to assure the reliability of results from data application.
Generally raw data may include errors such as noise, drift, outliers, malfunctions, etc. In addition to the possible measurement deviations related to the sensor performance itself, the errors can occur due to various reasons, e.g., the sensor installation problem and the measurement assumption violation. Thus, it is important to equip the data system with procedures that can detect the related problems and assist the user in monitoring and processing the incoming data. The data validation is an essential step to improve data reliability.
Data validation is a subject associated to various domains and thus mechanisms and techniques have been developed in different fields such as nuclear power engineering, chemical engineering, air pollution control, aerospatial industry, industry control, medicine, etc. (Conejo *et al*., 2007). Some data validation methods can be widely applied in many areas, while some methods are restricted to certain special scope due to some assumptions being made within the methods. In this report, data validation methods in water systems are of concern. One of the most important aims of this report is to identify potentially useful methods that could be developed or applied in water systems. Therefore, in addition to methods that are being used in water systems at the moment, methods that are used in other areas and may have an application potentially in the water area are also within the scope of this review.
Referring to the measurements in water systems, the commonly measured variables include rainfall, water depth, water flow rate, pH, conductivity, turbidity and pollutant concentrations. For each measurement obtained from one sensor, the data (signals) are most usually represented in the form of one dimensional time series. Exceptions come from recently developed new technologies applied in hydrological measurements. Two typical examples are the radar image for rainfall measurements and the spectra measurement (e.g., Pedersen *et al*., 2010) for water quality indicators (e.g., Winkler *et al*., 2008). Both

of them are represented by multiple dimensional data. Most methods for data validation described in the literature concern either one or several variables, each of which is represented by a one-dimension time series, while multiple dimensional data are seldom referred to due to their recent application in the area and probably due to their complexity.

Over the last decade, more and more affordable on line sensors have become available, leading to ever increasing acceptance of on line water monitoring (Hasan, 2005). These on line systems allow control mechanisms that are optimized for and respond to the actual process conditions. However, it accordingly calls for data validation that is valid for the real time coming data. The major difference between on line and off line data validation lies in the available information and the required execution time. Generally speaking, on line data validation is performed based on the past time series at a certain time point, e.g. without any information about data later than that time point, while off line data validation has the whole time series of data. Moreover, on line data validation is usually required by some real time control or management and thus the data are used for decision support or decision making as soon as possible after being obtained. Consequently the on line data validation process should be executed fast, whereas the off line data validation does not have such high requirement for this aspect. Another case calling for on line data validation is when the data read from a measuring system are with very small time steps or are multiple dimensional but the practice merely requires data at relatively larger time steps or in a reduced dimension form. Hence if the raw data are not required to be kept as the evidence to auditing for later usage, it is not necessary to keep large amounts of data all the time due to the limited storage capacity and thus most of the data can be removed at the on line stage. In this sense, by on line data validation, the validated data represent measurements of the variables in the required form where unnecessary information from raw data has been removed.

In traditional cases, data validation is mostly carried out manually by experienced observers, with the assistance of basic data and graphic visualization tools (Jorgensen *et al*., 1998). However, such a process can only be applied for a limited amount of data (Mourad and Bertrand-Krajewski, 2002). The complete reliance on human judgments to cope with abnormal signals has become increasingly difficult due to a variety of malfunctions (Venkatasubramanian *et al*., 2003). Further more, in real time application it is difficult to executed data validation manually due to time constraints. In data collection platforms, automatic data validation techniques or software can increase the level of data reliability and decision support systems. For example, NIKLAS (Lempio *et al*., 2010) has been developed as a module for real-time and

non-real-time evaluation of meteorological data. This module validates time series for the parameters of precipitation, global radiation, sunshine duration, air temperature, dew point temperature, relative humidity, wind speed and air pressure; NDBC (2009) used the software that can automatically validate observations for measurement taken from network of buoys and Coastal-Marine Automated Network stations; Edthoger *et al*. (2010) developed a system for fault tolerant control in on line drinking water quality monitoring, in which a module of data validation is incorporated.

The remainder of this reprot is organised as follows: Section 2 reviews methods used for data validation. Various approaches for faulty data detection and faulty data correction, which are the two main steps in data validation, are given and discussed; based on the review of state-of-the-art methods in this area, the potetial research directions are proposed in Section 3. Referrences are listed lastly.

2 Literature review on methods for data validation

Generally, data validation process consists of two main steps: faulty data detection and faulty data correction. Faulty data detection identifies doubtful values or errors in data and the correction process provide methods to deal with problematic data. In each category, a number of different tools and methods exist. This section reviews methods of these two categories respectively.
The measuring system usually does not provide the interested measurements directly but by transferring from signals from sensors with certain relationship. Thus some pre-processing tools such as calibration are required. It is also a very important component that can significantly affect the data validation result. In addition, other techniques such as checking the maintenance duration of sensors and data context classification (Branisavljevic *et al*., 2011) also support data validation. Hence in the third part of this section, techniques assisting data validation process beyond data error detection and correction are reviewed.

2.1 Faulty data detection methods

Faulty data detection methods generally classify the data into two classes: class of correct data and class of faulty or doubtful data (Bertrand-Krajewski *et al*., 2003; Branisqvljevic *et al*., 2011). Olsson *et al*. (2005) performed detection methods to assign each datum a confidence value between 0 and 100. Such a way of quantifying the data quality provides more refined information rather than only indicating data as correct, faulty and doubtful. There is no perfect or universal tool for this task and the success of the tool's application depends on a number of factors such as the type of monitored variable, the overall measurement conditions, the sensor used, the characteristics of the phenomenon being captured, etc. (Branisavljevic *et al*, 2011). Branisavljevic *et al*. (2011) also pointed out that the faulty data detection should not rely on just one method and some of the selected methods have to be applied successively according to a predefined order. The order of methods for faulty data detection used by Branisavljevic *et al*. (2011) is given as an example in the following:
Step 1: Zero value detection;
Step 2: Flat line detection;
Step 3: Minimum and maximum values detection based on geometric, hydraulic and data quality constraints;
Step 4: Minimum and maximum thresholds consideration based on historical values;

Step 5: Statistical test of variables that follow certain distributions;
Step 6: Mlutivariate statistical test using principal component analysis (PCA);
Step 7: Artificial neural network (ANN) non linear regression model for modeling one of the measured variables;
Step 8: One-class support vector machine classification;
Step 9: Physical modles such as Manning's equation.

2.1.1 Simple test based methods

Many faulty data detection methods are test-based methods. Some methods are based on very simple principle and are quick to execute. This group of methods are reviewed in this section.

2.1.1.1 Physical range check

The physical range of a variable is formed according to physical rules. The measured values can not exceed a physical range with a given sensor in a given location. It is both sensor and site specific. It is usually equal to the sensor's measuring range and/or to the physical conditions (Mourad and Bertrand-Krajewski, 2002). For example, with a temperature sensor that is designed for a range from 0ºC to 100ºC, a negative reading is erroneous.

2.1.1.2 Local realistic range detection

The local realistic range gives values that are usually observed in a specific measurement site (Mourad and Bertrand-Krajewski, 2002). It ensures that the measurements fall within established limits. The limits of the range are set and adjusted gradually using available information and former knowledge. The local range can be evaluated according to local geometric, hydraulic and data quality constraints. Another method to determine the range is to use statistical tools. For example, the range can be simply provided by historical minimum and maximum. It can also be defined as the 95% or 99% confidence interval of the observed values. This statistical approach is similar to the approach of extreme values detection (see Section 2.1.2.1).

2.1.1.3 Detection of gaps in the data

Gap detection mainly serves to find and exclude the gap interval from a sensor. It can also serves as an indicator of the reliability of the data obtained from this sensor: experience shows that well maintained sensors with a good data quality rarely have gaps in the data series (Olsson, 2005).

2.1.1.4 *Constant value detection*

Constant value identifies the time period when the measured variable always has the same value for a predefined threshold period (Ingleby and Huddleston, 2005; Branisavljevic et al., 2011). Constant values over a certain period of time are an indicator of unusable data which may be either due to bad digitization or to missing values in case of digital registration. Constant values can also be detected by running a variance check because the measurement values obtained from a normally functioning sensor always have a small variation even if it measures in a standard solution (Olsson *et al*., 2005). Therefore the need for a validation method allowing certain variation as normal is obvious. The running variance of a number of previous measurement values can serve as a measure of constant values.

2.1.1.5 *Signal's gradient test*

The signal gradient test detects sudden or erratic increase or decrease of values, or unrealistic gradients (Mourad and Bertrand-Krajewski, 2002). The accepted gradient can be given by a threshold on absolute or relative variations according to the physical process and local environment and conditions. NDBC (2009) used a time-continuity check to define maximum allowable difference between measurements within certain time period. This check, in essence, is similar to the detection of the unrealistic gradient.

2.1.1.6 *Tolerance band method*

In order to detect "outliers", which are observations that appear to deviate markedly from the majority of observations, a tolerance band method was developed (Edthofer et al., 2010). A smoothed curve is firstly generated based on the last measurement values (refer to 2.2.2 for the smoothing method). The deviations of the actual readings from the smoothed curve are then identified. Using this information, a tolerance band is calculated which follows the changes in the measurement values. The tolerance band becomes narrower with higher measurement accuracy. The tolerance band is centered on the smoothed curve and its local width is equal to the average deviation of the measurements from the smoothed curve multiplied by a tolerance factor which is usually in the range from 3 to 5. When a measured value lies outside of the tolerance band, it is classified as an outlier.

2.1.1.7 *Material redundancy detection*

When two sensors are redundant in a given site, the measured values obtained from them and the dynamics of their signals can be compared in order to detect unusual trends or abnormal gaps (Mourad and Bertrand-Krajewski, 2002). The difference between the two signals is used as an indicator to detect diverging values. The analysis of the

difference can be carried out by means of threshold values: when the difference between the two signals exceeds a predefined limit, the corresponding data are declared as doubtful. Other detection methods can also be used, like the Page-Hinkley algorithm (Barnett and Lewis, 1990). The use of only two sensors can lead to the detection of doubtful values. Additional data analysis should be made afterwards in order to decide which value between the two should be rejected. Martinez *et al*., (2001) believed that three redundant sensors would compose a more efficient system in case that no additional and contextual information is available to judge between measurements from two sensors.

2.1.2 *Physical or mathematical model based methods*

More complicated faulty data detection methods involve physical or mathematical models. Physical models provide analytical redundancy for data validation while statistical models are often used for detecting statistically rare values. Further, mathematical models making use of artificial intelligence are also applied in this area.

2.1.2.1 Extreme value check using statistics

This test determines "outlier" of the measurand. In this context, an outlier is simply viewed as an unusually extreme value for a variable, given the statistical model in use and a confidence interval. An outlier is detected if a value is out of the range of certain statistical frequency. However, an extreme value does not certainly lead to a faulty value as a statistical rare event could actually happen. Other methods such as analytical/material redundancy or expert opinion should be considered afterwards to assist further decision.

This test can also make use of some statistical distributions that are assumed or observed. For example, the observed velocity and water depth follow the student's t probability distribution (Branisavljevic *et al*., 2011). Assumed distributions, however,should be considered with caution.

2.1.2.2 Drift detection by exponentially weighted moving averages

Drift is a long term continuous increase or decrease of the readings from a measurement device. Edthofer *et al*. (2010) detected a drift in their software using exponentially weighted moving averages (Holt, 2004). This method produces a slope component and when it is significantly higher or lower than zero, a drift is detected. The time window for drift detection is long in comparison with other general faulty data detection. For instance, in the case of Edthoger *et al*. (2010), drift is only detected after increase or decrease of the values in a data stream over a period of several days. The statistical significance of the detected drift is then evaluated against the variations in the result and only drift substantially larger than the variations is considered as such.

After a drift is detected, it is then necessary to find out the driving force of it, that is, whether the drift is caused by measurement influence effects or by the real shift as time goes by.

2.1.2.3 *Spatial consistency method*

Correlation between measurements obtained from sensors with locations with certain spatial relationships may be used to evaluate measurement consistency (Olisson, 2003). For example, the measured water flow or water quality indexes at flow upstream should be related to that at downstream with some time delay. This class of methods can also be viewed as a special case for analytical redundancy method. For instance, a technique of spatial consistency check based on the idea that an observation at a given rain station is consistent with the neighboring observations is developed to detect outliers of rainfall records (Kondragunta, 2001). Pelczer and Cisneros-Iturbe (2007) made use of this spatial constancy check for rainfall data validation.

2.1.2.4 *Analytical redundancy method*

The essence of analytical redundancy is to check the actual system behavior against the system model for consistency (Venkatasubramanian *et al*., 2003). In measurement, most performance parameters are not directly measured, but are evaluated by modelling based on one or several measured quantities. Faulty data detection can be designed by making use of physical models. The pairs of correlated quantities may be either pairs of measured values or calculated/simulated values. It is therefore a method dealing with multiple data. Similar to the material redundancy method, the pairs of correlated values should be compared and the doubtful values can then be detected. For example, Branisavljevic *et al*, (2011) used the Manning's equation to link flow velocity and water depth.

The analytical redundancy method can provide reliable estimates even when extrapolating is needed. However, this type of methods is not widely applied in data validation nowadays probably due to the big effort that are usually required for building the physical models. But it is a potentially useful means to be used in this area particularly with the advances of computer programs in physical and mathematical models. However, model uncertainty must also be considered in such an evaluation approach (Hutton *et al*., 2011).

2.1.2.5 *Gross error detection*

Like analytical redundancy, gross error detection can be performed only if constraints are present. All gross error detection methods use, either directly or indirectly, the fact that gross errors in measurements cause the violation of the model constraints. All measurements contain

random errors. Thus the violation of constraints due to random errors is allowed.

The basic principle in gross error detection is derived from the detection in statistical applications. The random error is assumed to follow a normal distribution. Any normalized error (the difference between the measured value and the expected mean value divided by its standard deviation) which falls outside a confidence interval is declared an outlier or a gross error. The method utilizing model residuals is also referred to as residual analysis (Qin and Li, 1999).

2.1.2.6 *Multivariate statistical test using principal component analysis (PCA)* ***or kernel PCA***

One of the most popular outlier detection methods with multiple variables is the Principal Component Analysis (PCA) (Dunia *et al*., 1996; Pranatyasto and Qin, 2001; Branisavljevic *et al*., 2011). The PCA is based on linear relationship between data and transforms the data according to its variability. Data is transformed according to the correlation matrix to new coordinate system that is oriented to the direction of greatest data variability. When the PCA model is developed, its loading matrix represents the transformation matrix of the data. Using loading matrix any examined data value can be transformed and its statistical value can be compared with a threshold value.

Furthermore, the Kernel PCA can be used in nonlinear situations with a Kernel function that projects variables with a nonlinear transformation (Lee et al., 2004; Sun *et al*., 2007; Ren *et al*., 2011).

2.1.2.7 *Data mining technology*

Data mining is a process of extracting patterns from data. Data features are extracted from history data and are then used to allow later data diagnosis. One category of methodology constructs models through data mining. Instead of using physical models, they can be viewed as a black box, or data-driven models. It is another choice to model the relationship between variables instead of using a physical model. The main difference between physical and black box model is that the parameters in the former bear certain physical meaning, whereas in the latter they typically have no physical meaning, but are adjusted to produce a predictive relationship between model inputs and ouptuts. A typical and commonly used example of data mining method applied to data validation is to use ANN regression model to represent measurement function (Qin and Li, 1999; Branisavljevic *et al*., 2011). The absolute difference between modelled and measured data is compared to the threshold value. The one class support vector machine is another example to use data mining methods for building black box

models in data validation (Branisavljevic *et al.*, 2011). It is a special case of data classification method where the training data is classified into just one class forming a minimum radius sphere around the data. Data mining technology for faulty data detection can also borrow from other fields based on signal processing and pattern recognition. This is based on the assumption that the time series exhibit temporal autocorrelation such that the most recently measured values provide information that can be exploited to predict future values. A significant difference between the predicted and the measured values indicates an anomalous event. McKenna *et al.* (2008) used time series increments, multivariate distance algorithm and linear filter to detect changes in water quality data of water distribution systems.

It is worth mentioning that the methods considered in section 2.1 provides a list of possible faulty data detection methods; referring to a specific data validation case, the expert has to select the most suitable group of methods.

2.2 Data correction methods

Fauty data detection identifies doubtful or missing data. One identified data may be replaced by estimates. Possible data correction methods are reviewed in this section.

2.2.1 Interpolation

The faulty/doubtful data can simply be replaced by interpolation if no further information is available. For on line data validation, the simplest method is to use the value from the last measurement or use the trend from previous sets of measurements (Olsson *et al.*, 2005).

2.2.2 Smoothing method

When sudden change in the behavior of a system due to a sensor fault or due to a non-representative phenomenon happens, it usually generates high gradients. The signal can be filtered by using a moving average of n surrounding elements, where n is the smoothing window. Thus, the high gradients can be smoothed (Mourad and Bertrand-Krajewski, 2002). The moving average can also be a weighted average. Medians can be used instead of mean values. The main advantage of using the median as compared to moving average smoothing is that its results are less biased by outliers (within the smoothing window). Thus, if there are outliers in the data (e.g., due to measurement errors), median smoothing typically produces smoother or at least more "reliable" curves than moving average based on the same window width. The main disadvantage of median smoothing is that in the absence of clear outliers it may produce more "jagged" curves than moving average and it does not allow for weighting.

In the relatively less common cases (in time series data), when the measurement error is very large, the distance weighted least squares smoothing or negative exponentially weighted smoothing techniques can be used. All those methods will filter out the random noise and convert the data into a smooth curve that is relatively unbiased by outliers. Series with relatively few and systematically distributed points can be smoothed with bicubic splines (CS&IE).

2.2.3 Data mining technology

Data-driven models used for faulty data detection can also be applied to obtain appropriate values as the replacement of the faulty values. ANN is widely applied to this task. For instance, Kramer (1991) used ANN to estimate the replacement of faulty readings in process engineering; Elshorbagy *et al*. (2002) estimated missing stream flow data using ANN with principles of chaos theory. Genetic Programming (GP) is another artificial intelligence computation algorithm that can be potentially useful in this area. Ustoorikar and Deo (2007) used GP to fill up the gaps in a given time series of wave height data.

2.2.4 Data reconciliation

Data reconciliation is applied widely in industrial plants since 1980s (Heyen, 2006). This technique makes use of model equations to adjust the measurements. It can obtain estimates that are consistent with model constraints. Data reconciliation is a procedure to calculate a minimal correction to measured variables, to make them verify a set of model constraints.

Data reconciliation depends crucially on the assumption that only random errors are present in the data and systematic errors either in the measurements or the model equations are not present. If this assumption is invalid, reconciliation can lead to large adjustments being made to the measured values; and the resulting estimates can be very inaccurate and even infeasible. Yoo *et al* (2006) used principal component analysis (PCA) in a sensor reconciliation based on the redundancy of the measurements.

2.3 Other assisting techniques or tools

In addition to the above listed methods, a number of other tools and methods exist for assisting data validation. It is necessary to apply the adequate method, tool or idea that could give the expected results and improve data validation. Some pre-processing or supporting tools for data validation include:

2.3.1 Check of status of sensor

This test checks if the sensor connected to the data logger is in the on/off mode (Mourad and Bertrand-Krajewski, 2002).

2.3.2 *Check the duration after sensor maintenance*

The sensors should be regularly cleaned and maintained in order to ensure the reliability of the measurements (Mourad and Bertrand-Krajewski, 2002). It can be reasonably assumed that the functionality of sensors has an inverse relationship with the duration after the last maintenance.

2.3.3 *Data context classification*

If the measurement has different patterns under different conditions, the data can be classified in various classes according to the context. This approach will benefit data validation process. Branisavljevic *et al*. (2011) demonstrated that the classification of sewer flows according to wet/dry weather and day/night time improves the result of faulty data detection.

2.3.4 *Calibration of measuring systems*

In general, a measurand can not be detected by a sensor directly, but needs to be transformed from a signal (e.g. the electric current) directly measured from the sensor. This transformation process, which aims to build the relationship between the actual value of the interested measurand and the signal, is called calibration. The relationship between the sensor signals and the measurands should be calibrated in the laboratory before measuring in practical applications. However, the laboratory environment and conditions may differ from that of field. Uncertainty rises as the field calibration is quite often not carried out due to practical or technical reasons. Besides, even with field calibration, uncertainty can only be minimized but not avoided. As a result, calibration can make significant contributions to quatifying the uncertainties in some measurements such as turbidity and spectrometry for water quality (Bertrand-Krajewski *et al*., 2007; Joannis *et al*., 2008). Hence it is essential to use appropriate methods and techniques to execute the calibration in order to have reliable measurement result (Bertrand-Krajewski *et al*., 2003). If it is necessary, the calibration uncertainty should be taken into account and provided in the validated measurement result.

2.3.5 *Uncertainty consideration*

Uncertainty is an intrinsic feature of the hydrologic data, which complicates the data validation problem. Since the absolute accuracy in measurement is unattainable, a data validation system should take uncertainty into account and be able to handle the inaccurate data if required. JCGM 100 (2008) provided general rules for evaluating and expressing uncertainty in measurement that can be followed at various levels of accuracy. As many meansurands are not measured directly but are computed by a mathematical model from one or several measured values, JCGM 101 (2008) is concerned with the propagation of uncertainty through a

mathematical model as a basis for the evaluation of uncertainty of measurement, and its implementation by a Monte Carlo method. Furthermore, uncertainty presents when different models (including physical and data-driven models) are applied to data validation (Hutton *et al*., 2001) and it shouldnot be ignored.

3 Potential research directions

Data validation in water system is a field under development with certain challenges remaining. This report examines related concepts and provides a review of state-of-the art methods in this field. Typical data validation includes fautly data identifcation and faulty data correction. Numerious methods exist for both sub-processes and some methods can be used for both of fautly data detection and correction. A suiatble combination of different methods should be considered depending on the specific context of application. Also, selected methods have to be applied successively according to a predefined order. In addtion, uncertainty is an important factor that may affect the results and should be always kept in mind when performaing data validation.

Based on the litterature review, the following potential research points are identified:

1) *To study* on *data validation methods for recent technologies for hydrological measurements such as radar image for rainfall measurements and spectrometer for water quality measurements.*

There are two possibilities for the validation of multiple dimensional data. One is to work directly on the multiple dimensional data, and the other one is to transform the multiple dimensional data to single time series of corresponding variable (rainfall series or water quality index) using calibration function and then to validate the transformed one dimensional data.

2) *To use material/analytical redundancy for data validation*

The question of how to effectively dispose redundancy sensors can be explored. And the physical model and correlations between variables can be utilized to detect the data fault and to replace faulty data or complete missing data.

3) *To use data mining technology for data validation*

The currently applied method in this area such as ANN and GP can be tested and further developed. Other data mining technologies can also be explored, for example, the auto-regressive moving average (ARMA) which is considered as an effective method for hydrological series prediction that can be utilized for faulty data detection and correction.

4) *To take calibration of the measuring system into account in data validation*
5) *To include uncertainty analysis into data validation*
6) *To develop procedure that makes good use of both automatic and manual validation*

On one hand, the computer based automatic process can perform some repetitive checks efficiently without any missing element, and it also makes the application of complicate physical or mathematical models on data validation possible; on the other hand, computer can never

fully substitute expert's thought. Thus it is important to balance the use of computer and human judgment in data validation. To identify the balance between automated and manual methods will be specific to a particular applciation.

4 References:

Barnett, V. and Lewis, T. (1990). Outliers in statistical data. New York (USA): John Wiley & Sons, 3rd edition.

Bertrand-Krajewski J.-L., Bardin J.-P., Mourad M., Beranger Y. (2003) Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems. Water Science and Technology. 47(2): 95-102.

Bertrand-Krajewski J.-L., Winkler S., Torres A., Schaar H. (2007) Comparison of and uncertainties in raw sewage COD measurements by laboratory techniques and field UV-visible spectrometry. Water Science and Technology. 56(11): 17-25.

Branisqvljevic N., Kapelan Z, Prodanovic D. (2011) Improved real-time data anomaly detection using context classification. Hydroinformatics (in press)

CS &IE. Time Series Analysis. http://csie-data.com/time_series_analysis

Conejo R., Guwman E., Perez-de-la- Cruz J.-L. (2007) Knowledge-based validation for hydrological information systems. Applied Artificial Intelligence. 21: 803-830.

Dunia R., Qin S.J., Edgar T.F., McAvoy T.J. (1996) Identification of faulty sensors using principal component analysis. AIChE Journal. 42:2797-2812.

Edthofer F., Broeke. J., Ettl J., Lettl W., Weingartner A. (2010) Reliable online water quality monitoring as basis for fault tolerant control. Procedings of the Conference on Control and Fault-Tolerant Systems, Nice, France, 6-8 October.

Elshorbagy A., Simonovic S.P., Panu U.S. (2002) Estimation of missing streamflow data using principles of chaos theory. Journal of Hydrology. 255: 123-133

Hasan J. (2005) Technologies and techniques for early warning systems to monitor and evaluate drinking water: State of the Art Review. US EPA Office of Science and Technology.

Heyen Georges. (2006) Data reconciliation http://www.lassc.ulg.ac.be/webCheng00/meca0468-1/Validation_intro.pdf
Holt C.C. (2004) Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20(1): 5-10.

Hutton C.J., Vamvakeridou-Lyroudia L.S., Kapelan Z., and Savic D.A. (2011) Uncertainty quatification and reduction in urban water systems (UWS) Modeling. Evaluation report for European Project: Prepared.

Ingleby B. and Huddleston M. (2005) Quality control of ocean temperature and salinity profiles—historical and real-time data. Journal of Marine Systems. 65:158-175

Joannis C., Ruban G., Gromaire M.-C., Bertran-Krajewski J.-L.(2008) Reproducibility and uncertainty of wastewater turbidity measurements. Water Science and Technology. 57(10): 1667-1673

Jorgensen H.K., Rosenorn S., Madsen H., Mikkelsen S. (1998) Quality control of rain data used for urban runoff systems. Water Science and Technology. 37(11): 113-120.

JCGM 100 (2008) Evaluation of measurement data—guide to the expression of uncertainty in measurement. BIPM report. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

JCGM 101 (2008) Evaluation of measurement data—supplement 1 to guide to the "expression of uncertainty in measurement"—propagation of distributions using Monte Carlo method. BIPM report. http://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf

Kondragunta C.R., (2001) An outlier detection technique to quality control rain gauge measurements. In: AGU, Spring Meeting, Boston, Massachusetts.

Kramer M.A. (1991) Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal. 37(2): 233-243.

Lee J.M., Yoo C.K., and Choi S.W. (2004) Nonlinear process monitoring using kernel principal components analysis. Chemical Engineering Science. 59: 223-234.

Lempio G., Podlasly C., Einfalt T. (2010) NIKLAS-automatical quality control of time series data. The six European conference on radar in meteorology and hydrology.

Martinez, M. A., Trivino, B.F., Gomez, S.A.B., Gutierrez, G. A.A. and Rubiano, C.M. (2001). Uncertainty and redundancy in flow metrology. Actes de la Conférence A&E2001 Automatique et Environnement", Saint-Etienne (France), 4–6 Juillet 2001, 8 p.

McKenna, S. A., M. Wilson, AND K. A. Klise (2008). Detecting Changes in Water Quality Data. Jounal of the American Water Works Association. 100(1):76-85

Mourad M., Bertran-Krajewski J.-L.(2002) A method for automatic validation of long time series of data in urban hydrology. Water Science and Technology. 45(4-5):263-270.

NDBC (2009) Handbook of Automated Data Quality Control Checks and Procedures. National Data Buoy Center Stennis Space Center, Mississippi 39529-6000

Olisson G., Nielsen M., Yuan Z., Lynggaard-Jensen A., Steyer J.-P. (2005) Instrumentation, Control and Automation in Wastewater Systems. Scientific and technical report, No 15. IWA publishing.

Pedersen, L., Jensen,N. E., Madsen. H., (2010). Calibration of Local Area Weather Radar—Identifying significant factors affecting the calibration. Atmospheric Research. 97(1–2): 129-143.

Pelczer I., Cisneros-Iturbe H.L. (2007) Automated validation of data from a rainfall network. Novatech. Lyon, France.

Pranatyasto T.N., Qin S.J. (2001) Sensor validation and process fault diagnosis for FCC units under MPC feedback. Control Engineering Practice 9: 877–888.

Qin S.J., Li W. (1999) Detection, identification, and reconstruction of faulty sensors with maximized sensitivity. AIChE journal. 45(9): 1963-1976

Ren L., Xu Z.Y. and Yan X.Q.(2011) Single-sensor incipient fault detection. IEEE Sensors Journal. 11(9):2102-2107.

Sun R., Tsung F., and Qu L., Evolving kernel principal component analysis for fault diagnosis. Computers and Industrial Engineering. 53:361-371.

Ustoorikar K., Deo M.C. (2007) Filling up gaps in wave data with genetic programming. Marine Structures. 21(2-3,):177-195

Venkatasubramanian V., Rengaswamy R., Yin K., Kavuri S.N. 2003. A review of process fault detection and diagnosis Part I: Auantitative model-based methods. Computers and Chemical Engineering 27: 293-311

Winkler, S., Bertrand-Krajewski, J.-L., Torres, A., Saracevic, E. (2008). Benefits, limitations and uncertainty of in-situ spectrometry. Water Science and Technology, 57(10), 1651-1658.

Yoo C.K., Villez K., Lee I.B., Van Hulle S., & Vanrolleghem P.A. 2006 Sensor validation and reconciliation for a partial nitrification process. Water Science and Technology. 53(4–5):513–521