Application and results of automatic validation of sewer monitoring data

M. van Bijnen¹,³ and H. Korving²,³

¹ Gemeente Utrecht, P.O. Box 8375, 3503 RJ, Utrecht, The Netherlands
² Witteveen+Bos Consulting Engineers, P.O. Box 233, 7400 AE, Deventer, The Netherlands
³ Department of Sanitary Engineering, Delft University of Technology, P.O. Box 5048, 2600 GA, Delft, The Netherlands

*Corresponding author, e-mail m.van.bijnen@utrecht.nl

ABSTRACT
In Utrecht, a monitoring network has been installed in order to check the reliability of the theoretical hydrodynamic sewer model and study the hydraulic performance of the sewer system. Flows, water levels, rainfall and turbidity are monitored at several locations in the system. The total monitoring network consists of 138 sensors and will be extended in the near future. Managing, analysing and presenting measured data, however, is a very extensive job due to the enormous amount of measurements that are daily registered. Therefore, an automatic validation tool has been developed for validation of large data sets of sewer measurements. The tool is based on the correlation between sensors and deviations in system behaviour. Depending on the type of instrument (water level, flow, rainfall or turbidity), one or more statistical models are used to automatically diagnose the quality of measurements (‘correct’, ‘uncertain’ and ‘incorrect’). In order to prevent erroneous quality labels, e.g. label ‘incorrect’ due to construction works in the sewer system, frequent consultation of the management authority is needed. After evaluation, the data are suitable for planning and design purposes, such as decisions on investments and model calibration.

KEYWORDS
Data quality, data validation, sewer measurements, monitoring network, validation tool

INTRODUCTION
In order to comply with emission standards and reduce combined sewer overflows (CSOs), numerous and expensive investments are required, including building of settling tanks and sewers, enlarging pipes and reducing (paved) areas connected to the sewer system. In Utrecht (the Netherlands), an investment of approximately 40 million Euro is mainly based on the results of theoretical hydrodynamic models. In general, however, the accuracy of the theoretical model in comparison with reality is unknown. Consequently, uncertainty of model results affects the decision-making on the investments.

Reliable model calculations require a high quality sewer database, estimates of dry weather flow and surface run off. The theoretical model of Utrecht comprises 19,404 nodes and 21,091 conduits, with 227 overflows and 232 pumps. The total dry weather flow consists of 2,814 m³/h domestic and 1,419 m³/h industrial wastewater. The run off area estimates 1,482 ha contributing to the combined sewer system en 124 ha to the improved separate system. The combined sewer system in Utrecht is divided in 22 sub systems and wastewater is transported...
between the districts by pumps. The system consists of approximately 650 km of sewers and 184 CSOs.

In order to check the reliability of the theoretical model and understand the hydrodynamic behaviour of the system, a monitoring network has been installed in the combined sewer system. Flows, water levels, rainfall and turbidity are monitored at several locations. The total monitoring network consists of 138 sensors and will be extended in the near future. The extension includes approximately 140 water level sensors, 6 rain gauges and several flow sensors.

Every day at least 55,000 measurements are stored in a database. Without data processing and validation, this results in a large, inaccessible data set with unknown quality. Therefore, validation of measured data is necessary. This not only provides information on the functioning of the measuring equipment, but also limits the large amount of measurement data in order to provide accessibility. Finally, validated data increase the reliability of model results and investments. There are several examples of automatic data validation in the field of sewer systems and wastewater treatment plants (e.g. Mourad and Bertrand-Krajewski 2002, Yoo et al. 2006). However, these applications mainly concern small research projects.

This paper discusses the practical use of an automatic validation tool for validation of large data sets of sewer measurements. The application is illustrated with several examples.

DATA VALIDATION

Measurements in sewer systems need validation before they can be used for planning and maintenance. It is widely acknowledged that this partly results from the extreme conditions in sewers in terms of fouling and corrosion of instruments (see e.g. Mourad and Bertrand-Krajewski 2002, Rosen et al. 2003). Therefore, data validation and processing are vital elements in measuring campaigns. Usually, the amount of data collected in permanent measuring campaigns is very large. As a result, manual validation is very time consuming and possibly inaccurate.

For validation of large data sets of sewer measurements an automatic validation tool has been developed (Ottenhoff et al. 2007). This tool has been used in several monitoring projects of sewer systems in the Netherlands. The tool is based on the correlation between sensors and deviations in system behaviour. Depending on the type of instrument (water level, flow, rainfall or turbidity), one or more models are used to determine whether a measurement is correct. The validation tool comprises several standard checks (such as double measurements, out of range measurements and missing values) and a more site specific control model (Figure 1).

The control model consists of relatively simple regression models that are calibrated using stepwise regression. The main advantage of this approach is that malfunctioning sensors are automatically left out. The data are also checked for monotone and sudden trends. The tool automatically diagnoses the quality of measurements (‘correct’, ‘uncertain’ and ‘incorrect’), if possible, by individually validating each measurement of a sensor. In order to prevent erroneous quality labels, e.g. due to construction works in the sewer system, frequent consultation of the sewer management authority is needed. After evaluation, the data can be applied for planning and design purposes.
The validation tool automatically diagnoses the quality of measurements (‘correct’, ‘uncertain’ and ‘incorrect’), if possible, by separately validating all measurements of one sensor. Data validation answers the following questions:

- is the sensor working?
- are the measurements reliable?
- can the measurements be used for planning purposes?
- can the measurements be used for calibration of the hydrodynamic model?
- which measurements have to be investigated in more detail?

In the next part of this paper the validation procedure as described in Figure 1 will be explained using practical examples.

**Quick scan and pre-treatment**

Prior to validation, pre-treatment can be necessary, e.g. in case of a mismatch between measuring frequency and sampling time. This requires a synchronisation of intervals between subsequent measurements to, for example, 5 minutes. Another example of pre-treatment is filling gaps due to missing data. Both routines are based on interpolation. Level measurements are very suitable for interpolation, because of their nearly continuous character. The behaviour of turbidity in a sewer system, however, is much less continuous. Consequently, important information can be discarded due to interpolation.

Furthermore, aggregation of data can be necessary because of low correlation between values at the original time scale. This often results from fast variations in process dynamics. For example the correlation between precipitation and water level is low at a sampling frequency of 5 minutes due to the time needed for run off. Aggregation to hourly values increases correlation. A disadvantage of aggregation is that measurements cannot be validated...
individually at the original time scale. However, it enables a more reliable assessment at a larger time scale.

**Double measurements**
The next step is a check for double measurements. This means that the same date and time labels occur more than once in the data set but with a different observed value. Incorrect programming of readout software can cause such errors. Figure 2 shows an example of shifted dates between two readouts in 2005, the first one on November 11, the second on December 22. Measured values are shifted 10 minutes between the two readouts. Another cause of double measurements is a sudden switch of month and day (Figure 3).

![Figure 2. Measured values shifted in time](image)

![Figure 3. Switching day and month in date format](image)

**Out of range measurements**
Measurements exceeding the maximum and minimum limits of the sensor or the physical range (e.g. below manhole bottom) are labelled as ‘incorrect’. Figure 4 shows an example of out of range flow measurements at a pumping station. The maximum flow is 720 m$^3$/h (60 m$^3$/5 minutes). The registered values exceed the maximum value several times. Therefore, further research on site is needed.
Figure 4. Out of range measurements at pumping station Neutronweg

Missing values
Missing values represent incorrect data. In order to save storage capacity a variable sampling interval can be applied. However, incorrectly programmed sampling interval causes missing measurements. Loss of data can also result from limited storage in the sensor, problems with telemetry, loss of power supply or malfunctioning of equipment.

Figure 5. Correctly programmed variable measurement interval

In Utrecht, the sampling frequency of water level sensors at CSOs changes from 5 minutes to 1 minute when the water level exceeds the weir level. Figure 5 shows a correctly programmed sensor. The weir level of this CSO is 0.88 m+NAP. Switching of the interval causes the deviations in the figure. Figure 6 shows an example of an incorrectly programmed sampling interval. The interval again depends on the water level. However, a clear relationship between
water level and sampling interval is missing. Due to the enormous variation in sampling interval the measurements are less suitable for further analyses.

**Figure 6.** Incorrectly programmed measurement interval

**Trends**
The time series are also checked for monotonous and sudden trends. Trends can be indicative of drift of the sensor as well of gradual changes in the system itself. Most classical trend tests, however cannot properly deal with signals with a large auto correlation and fast fluctuations of sewer processes. As a result, the applied trend test has to account for both aspects. A seasonal Kendall test with correction for covariance of the signal is most appropriate for detecting a linear trend (Hirsch and Slack 1984, Dietz and Killeen 2006). A step trend can be detected by comparing local variance with the variance of the complete time series. When a cause for a detected trend is missing, measurements are labelled as incorrect.

**Figure 8.** Detected step trend at two locations
Figure 8 shows the impact of reconstruction works on system performance. The sudden decrease of the water level on these two locations resulted from reconstruction works in the sewer system. Since the cause of the trend is known, the measurements are labelled as correct. However, jumps in the measurements due to auto-calibration of the sensor results in the label ‘incorrect’ because the measurements are unsuitable for further application. This may result from maintenance of the sensor without a good protocol for reinstallation.

**Statistical model**

In order to check the measurements with the statistical model, all sensors are clustered on the basis of an extensive correlation and regression analysis. The statistical model is site specific. Clusters consist of sensors with a large correlation. If necessary, the clustering is adjusted based on system characteristics. For Utrecht 15 clusters are defined.

The statistical model consists of a combination of models depending on the type of sensor. They are trained for each cluster individually. They are re-calibrated during each validation round using stepwise regression (Draper and Smith 1981). The main advantage of stepwise regression is that malfunctioning sensors are automatically left out. It results in a regression model in which the most significant parameters are included. If an already included location appears to be not significant enough, it is removed again.

An example is shown in Figure 9. The measurements at the internal weir of the storage tank do not correspond with the measurements at the external weir. The storage tank fills up although the water level in the sewer system remains below weir level.

**Quality labels**

The results of the control steps are combined in an overall assessment of the measurement values. A quality label (‘correct’, ‘uncertain’ and ‘incorrect’) is attached to each individual measurement. If possible, the labels are made more specific (e.g. ‘trend’ or ‘out of range’) in order to save time when solving the problem.
The validation results are graphically presented to the user by means of a web-based plug-in. An example of the results is shown in Figure 10. In order to prevent erroneous labelling of measurements, e.g. due to construction works in the sewer system, frequent consultation of the sewer manager is needed.

![Pie chart showing the quality of water level sensors for January 2007.](image)

**Figure 10.** Reported quality of all water level sensors for January 2007

**RESULTS AND CONCLUSION**

The objective of this paper is to describe the application of automatic validation for large data sets of sewer measurements. The tool consists of a combination of tests to determine the quality of a measured value, including logical tests (missing values, out of range values, etc.), regression models and trend detection. It is provided to the user as a web-based plug-in. The results show that data validation provides indispensable information on the performance of measurement equipment. Only measurements labelled ‘uncertain’ require further investigation. In addition, data validation provides an excellent basis for contracts on reliability and availability of measurements between sewer manager and data provider.

In Utrecht, data validation is used for several purposes, such as the correction of erroneous records in the database of the sewer system. Furthermore, cleaning of sewers can be based on the interpretation of validated measurements, e.g. an increasing amount of CSOs as an indicator for blockage of a siphon. In the near future, the validated measurements will also be applied for calibration of the hydrodynamic model of the sewer system. Based on the results of the monitoring network and the validation procedure, approximately 8 million Euro will be invested in Utrecht in 2008. Other measures costing over 12 million Euro will be investigated in more detail because doubts have risen regarding their effectiveness.

**REFERENCES**


