Improved real-time data anomaly detection using context classification

Nemanja Branisavljević, Zoran Kapelan and Dušan Prodanović

ABSTRACT

The number of automated measuring and reporting systems used in water distribution and sewer systems is dramatically increasing and, as a consequence, so is the volume of data acquired. Since real-time data is likely to contain a certain amount of anomalous values and data acquisition equipment is not perfect, it is essential to equip the SCADA (Supervisory Control and Data Acquisition) system with automatic procedures that can detect the related problems and assist the user in monitoring and managing the incoming data. A number of different anomaly detection techniques and methods exist and can be used with varying success. To improve the performance, these methods must be fine tuned according to crucial aspects of the process monitored and the contexts in which the data are classified. The aim of this paper is to explore if the data context classification and pre-processing techniques can be used to improve the anomaly detection methods, especially in fully automated systems. The methodology developed is tested on sets of real-life data, using different standard and experimental anomaly detection procedures including statistical, model-based and data-mining approaches. The results obtained clearly demonstrate the effectiveness of the suggested anomaly detection methodology.

Key words | anomaly detection, context-classification-based detection, data pre-processing, sewer data

Nemanja Branisavljević (corresponding author)
Dušan Prodanović
University of Belgrade,
Faculty of Civil Engineering,
Institute for Hydraulic and Environmental Engineering,
Kralja Aleksandra bld. 73,
11000 Belgrade,
Serbia
E-mail: *nemanja@hikom.grf.bg.ac.rs*

Zoran Kapelan
Centre for Water Systems,
University of Exeter,
Harrison Building,
North Park Road,
Exeter EX4 4QF,
UK

INTRODUCTION

The measured data time series are a fundamental part of water distribution and sewer information systems and play a key role in forecasting, decision making or analysis of the present state of these systems. One of the crucial issues about data usage is its quality (Mourad & Bertrand-Krajewski 2002; Branisavljević *et al.* 2009*c*, *d*). If data quality is unknown or unspecified, the result obtained from any kind of data analysis cannot be fully trusted. It should be noted that anomalous historical data is hard to repair and improve as the additional information about how the data are collected, decreases with time. Some questions are often left unanswered, for example, whether the sensors were calibrated or not, whether the conditions at the measurement location were adequate or not, etc.

This paper deals with the first step in data quality assessment – anomaly detection. Anomaly detection is a type of data validation that classifies the data into two classes: one of regular data values and the other of anomalous data values. There are many tools and methods used in practice and science that can answer the question of whether the particular data value considered is anomalous or not (Venkat *et al.* 2003*a*, *b*, *c*). Some of validation methods are even proposed for sewer-data time-series (Bertrand-Krajewski *et al.* 2000; Mourad & Bertrand-Krajewski 2002). Having said this, there is no perfect or universal tool for anomaly detection and the success of the tool's application depends on the number of factors (e.g. the type of variable monitored, the overall measurement conditions, the sensor/monitoring

doi: 10.2166/hydro.2011.042

equipment used, the characteristics of the phenomenon being captured, etc.).

The correct data validation and checking procedure requires exhaustive use of all available information and mathematical tools. Also, if more than one anomaly detection method is used on the same dataset, its sequence in the detection procedure has to be properly defined (Branisavljević *et al.* 2009*b*). This is an important factor which can significantly improve the detection results. For example, some of the data anomaly tools should be used before other tools as they are best used for data pre-processing. These methods normally detect data values with major errors and, if necessary, exclude these values from the analysed data series. Once this is done, more sophisticated anomaly detection methods may be used to identify further, more subtle errors. In addition to the above, out of all the anomaly detection methods available (Patcha & Park 2007), only a limited number are suitable for online application (i.e. they do not require (frequent) expert involvement).

To ensure satisfactory results from the anomaly detection methods some rules have to be followed.

- One should not rely on just one method for anomaly detection (Rosen *et al.* 2003).
- Data have to be examined by the expert first and the most suitable group of methods, that will give the best results, has to be selected (Rosen *et al.* 2003).
- Some of the selected methods have to be applied successively in predefined order. It is suggested that data should be prepared using specified data pre-processing.
- Anomaly detection methods are most effective when they are tuned according to the most characteristic features of monitored process.
- The results of anomaly detection methods (usually more than one) can then be summarized to unique grade for every data value or data interval, considering data usage (Branisavljević *et al.* 2009*d*).

To further enhance the effectiveness of data anomaly methods, data values may be classified into contexts and some carefully designed pre-processing methods may be applied. This paper addresses the issue of how (and if) the context classification-based detection and the data pre-processing influence the performance of anomaly detection methods. Using a real-life test case, first the anomaly detection without context data was performed. Then automatic context classification is applied and anomaly detection repeated, with detected context classes the data values are in, and with retuned methods. In the Conclusion, it is shown that context data can improve anomaly detection up to the level where it becomes feasible for usage in real-time automatic systems.

METHODOLOGY FOR THE ANOMALY DETECTION

The design of an anomaly detection system in general has two main stages.

1. Selection, development and tuning (i.e., calibration) of the anomaly detection methods ensemble.
2. Application of the above to the observed data stream(s) in real-time.

The first stage has two alternative approaches. The first is to develop the methods that model the regular data characteristics, together with the corresponding threshold values. The threshold values then represent the boundary between regular and anomalous data values (and can be designed to change temporally, if necessary). The second approach requires the modelling of expected anomaly signatures in a regular data series, together with the development of a classification algorithm that will determine the boundary between classes of regular and anomalous data. Since it is hard to predict all the possible anomalies that may occur in sewer system data, the first approach is used here.

Methodology background

Anomaly detection methods have to be selected in a way that ensures that all the technical, expert based and relational characteristics of the examined data are checked (Branisavljević *et al.* 2009*a*). First, to ensure the detection of all major anomalies, multiple detection methods have to be used and designed according to the technical characteristics of the analysed system. Second, since a combined sewer system is influenced by two types of inflows, dry (used water) and wet weather (rainfall, snow), the appropriate anomaly detection methods have to be developed to take this into account. The evidential pattern of sewer flows during dry weather

conditions (easily determined by the expert) also has to be modelled. Third, relational characteristics (linear, non-linear and physical) must also be considered. Considering all requirements mentioned in this section, to ensure proper anomaly detection, some methods may run in parallel but some have to be ordered in sequence to ensure the pre-processing effect for following ones, as shown in Figure 1.

If the anomaly detection system is to be automated (with very rare re-tuning), its tuning parameters have to be adjusted to all possible conditions found in the data (e.g. wet/dry weather conditions, etc.). To enhance online effectiveness of the anomaly detection system, the data analysed should be classified in different contexts. Even though frequently challenging to develop, the effective context detection is important as it should ensure the increased effectiveness of the anomaly detection techniques used. The examples of contexts that the sewer monitoring data can be classified in are as follows:

- weather conditions: wet/dry weather;
- hydraulic conditions: submerged (backwater effect)/non submerged flow;
- sewer management: Pumping ON/OFF;
- social rules: working day/weekend, regular day/holiday/big public event;
- season: Spring/Summer/Autumn/Winter;
- hour of the day: 0/1/…/23
- Other.

To further improve the generic tools for anomaly detection, the data analysed can be pre-processed. Sometimes the data pre-processing is crucial for a measuring method and it is included in a measurement procedure in the measuring equipment (it cannot be accessed by user). For example, ultrasonic measuring devices have internal data pre-processing (data averaging) with the objective to get more sensible and accurate results. A number of different tools and methods for data pre-processing exist. Some examples include:

- re-sampling – selection of representative subset from a large population of data suitable for some anomaly detection tool;
- aggregation and fusion – combining data in clusters that demonstrate some characteristics that were not dominant if the original (i.e. raw) data is used;
- interpolation – method of constructing new data points within the range of known data points;
- noise removal – removing high frequency data;
- normalization – organizing data for more efficient use;
- scaling – adjusting the data to fit in predefined boundaries;
- feature extraction – transforming the input data into the set of features.

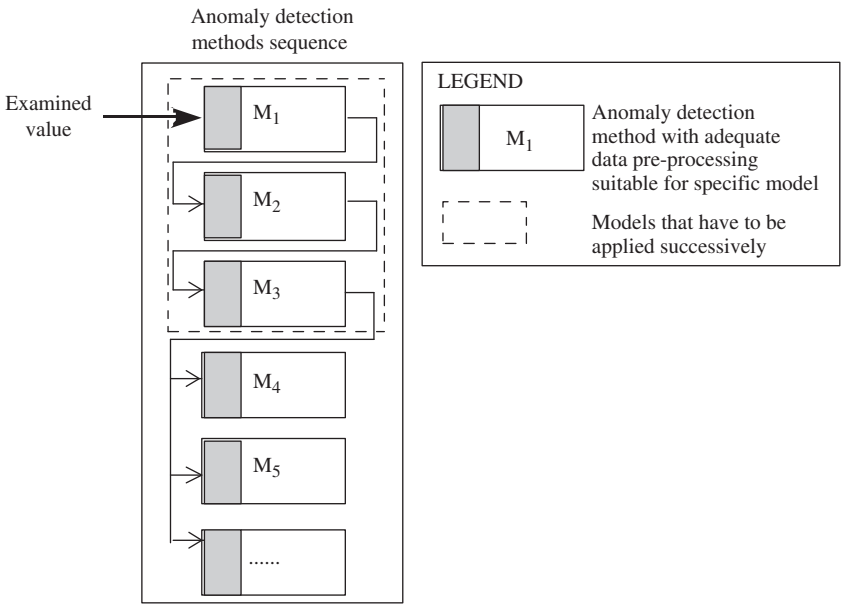

Figure 1 | Sequence of several anomaly detection methods.

In the context of data pre-processing for anomaly detection, it is necessary to apply the adequate method, tool or idea that would give the expected results and improve anomaly detection methods. However, no general, ready-made solutions exist for this. In order to select the most suitable pre-processing method (which is useful for the specific anomaly detection tool) it is necessary to provide an expert study on the impacts of the pre-processing on the data. The main idea is to provide such a data pre-processing procedure that will increase the possibility for anomalous data to be detected without an increase in false alarms. Sometimes if the pre-processing tool is wrongly selected the data anomaly detection results can be disappointing, and the performance of the applied anomaly detection tool can even be reduced. Also, data pre-processing can be applied in the anomaly detection methods design and tuning procedures as well as in its application. This means that even during the model selection and tuning, data pre-processing can be useful in getting more suitable methods for further application. Some methods, like the Artificial Neural Network (ANN) or the support vector machine (SVM) are hard to design if the data is not pre-processed adequately (e.g. scaled).

Context-based detection and data pre-processing

The position of context-based detection is specified in Figure 2.

Data pre-processing can be implemented at several locations in the anomaly detection procedure (Figure 2). As shown in the Figure 2, the process starts with the data being acquired from different sources. Note that the internal sensor data pre-processing is not presented here as it is assumed that the measurement equipment is installed properly. The general data pre-processing is the first pre-processing module that is applied. It can be designed to be applied automatically to the data before the data enters the database and it should be a part of a database ETL (extract, transform and load) integration system. The examples of general data pre-processing include unit conversion, checks for the data format or data normalization.

After the data enters the database it has to pass the anomaly detection module. For a successful application of specified tools in this module it is necessary to provide data adjustment for the easier and more flexible data manipulation. Examples include: finding gaps in data, marking double entries and forming multiple time series by data aggregation (data with multiple timelines), data interpolation, data selection, etc. All this is part of the general data pre-processing for anomaly detection. In this part of data transformation it is convenient to extract some redundant data series that would provide easier data transformation (data normalization).

Finally, some anomaly detection tools require specific pre-processing for better performance. The selection of specific data pre-processing is based on the characteristics of the data and characteristics of the anomaly detection method used. After the anomaly detection process the validation grades are transferred back to the database as a metadata.

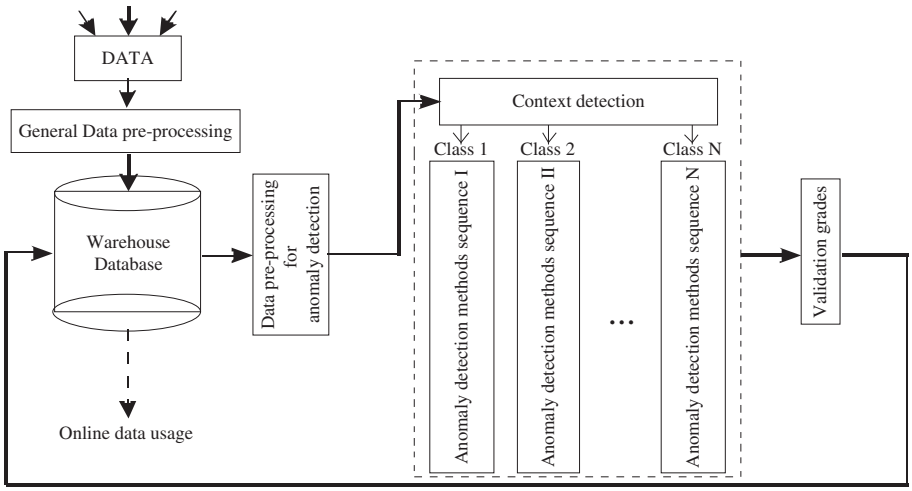

Figure 2 | Position of context detection and data pre-processing in the anomaly detection flowchart.

Performance indicators for anomaly detection methods

It is hard to provide good grading framework that will give insight into the suitability of an anomaly detection method (and further, in turn, on the pre-processing method). The reason for this comes from the fact that the anomaly detection has to provide results that cannot be checked by other methods or measurements and are usually not visually noticeable (except major errors that can be identified using visual inspection). There are two possible ways of dealing with this problem. One is to rely on common sense, experience and good knowledge about the methods and procedures used in anomaly detection and data pre-processing. The key role in this approach is an expert's role in providing the successful anomaly detection module. The other approach is based on adjusting the selected anomaly detection methods to the time series analysed with already detected anomalous data values.

An exhaustive process of manual anomaly detection done by an expert has to be considered with caution. Even if the expert has relevant experience, and does their job carefully, it is still most likely that only major anomalies will be detected (see Figure 3). As it can be seen from this figure, the expert managed to mark only the obvious errors comprised of: 1) zero values, 2) dubious spikes or 3) sequences of data that don't fit the expert's experience about the expected data pattern.

Artificial anomalies may be introduced into the regular data time-series to check the performance of the anomaly detection methods. These anomalies can be divided into two groups: additive and multiplicative anomalies (Branisavljević 2009*a*, *b*, *c*, *d*). Examples for additive anomalies include spikes, constant offset or linear offset. Multiplicative anomalies are induced with, for example, sensor calibration curve span change.

Assuming that the anomalous data values are marked, the anomaly detection performance indicator can be defined as follows:

$$p = \frac{N_{registered}}{N_{anomalies} + N_{missed} + N_{registered\ \ nonanomalies}} \tag{1}$$

where $N_{registered}$ is the number of registered (detected) anomalies, $N_{anomalies}$ is the number of the anomalies in the data, N_{missed} is the number of the missed anomalies ($N_{anomalies} - N_{registered}$), $N_{registered\ \ nonanomalies}$ is the number of registered values with the models used that are not anomalies. It can be

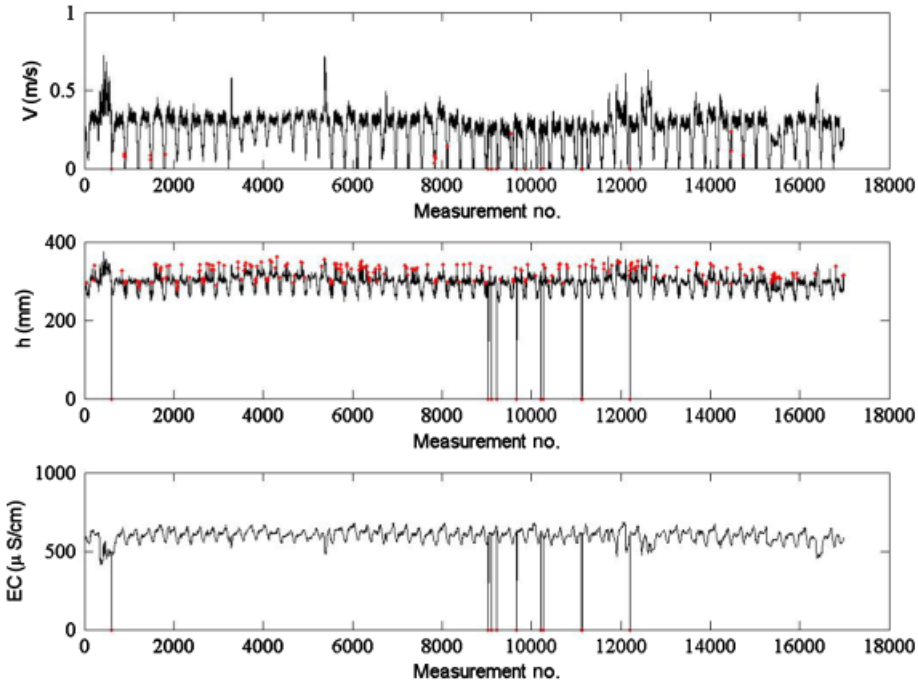

Figure 3 | The original time data series with manually marked anomalous data by the expert.

seen from Equation (1) that p is equal to 1 only if all anomalies are detected, that is, if $N_{registered\ \ nonanomalies} = 0$ and $N_{missed} = 0$.

Alternatively, if the data is over-sampled, and the detection of regular data as anomalous can be assumed as reasonable loss, but the data values with anomalies have to be detected with more reliability (false alarm tolerant system) Equation (1) may be rearranged as follows:

$$p_{false\ \ tolerant} = \frac{N_{registered}}{N_{anomalies} + N_{missed}} \tag{2}$$

In the same manner, a scenario when the data is scarce and the data with mild anomalies can be accepted as regular, a grading may be defined as follows:

$$p_{false\ \ sensitive} = \frac{N_{registered}}{N_{anomalies} + N_{registered\ \ nonanomalies}} \tag{3}$$

Anomaly detection methods

The following anomaly detection methods are designed and tuned: M1 – Detection of zero values; M2 – Detection of flat lines (if the same value is recorded for longer than 30 minutes); M3 – Min./Max. according to the physical limits (e.g. the diameter of the outlet); M4 – Min./Max. values according historical data; M5 – Statistical univariate test (Grubb's test) that labels the spikes and marks the data that is not consistent with daily pattern during dry weather; M6 – Statistical multivariate test based on principal component analysis (PCA) that labels the spikes and the data that are not correlated with the majority of the data exploring liner relationships; M7 – ANN non-linear regression model; M8 – One-class SVM – classification tool for extracting the outliers; M9 – Model based on physical relations that uses Manning's equation.

The first three methods (M_1, M_2 and M_3) are used for detection of major errors. All methods are tuned firstly according to the best practice and experience. For the real-time systems, the period between two tunings has to be specified according to the characteristics of the monitored system (measurement micro location characteristics, monitoring equipment characteristics, etc.).

M1: Zero value detection

This method performs a simple check to find out whether a particular data value is equal to zero or not. The method has no parameters and no thresholds so its application is trivial.

M2: Flat line detection

The flat line method parameter (length of the flat line that is considered as anomalous) is determined so that the measured variable cannot have the same value for more than 30 minutes. This type of error may be induced by many causes. For example, the sensitivity of the ultrasonic measurement device will decrease for low values of the measured variable and, as a consequence low velocities and water depths may induce flat line type anomalies. Another example is the electro-conductivity measurements where flat line type anomalies may occur due to conductivity probes being plugged by for example, floating impurities which in turn, may interrupt the water flow between the electrodes.

M3: Min./Max. detection

The parameters for Min./Max. determined are formed according to geometric, hydraulic and data quality constraints. The sediment deposition and possible bio-film layer on the outlet wall were not considered. No pre-processing was used for major error detections because methods like filtering, for example, may sometimes reduce the erroneousness of anomalous data and even transfer some of it to the neighbouring data values.

M4: Min./Max. thresholds based on historical values

The threshold values provided by the limitations of the system are quite broad. More specific threshold values may be obtained if the historical data is considered. However without classifying data in contexts, it is possible just to provide some general historical minimums and maximums. One should be aware of seasonal effect on data, especially in lower limits of the variable since for example, the base flow rate may increase during the spring season. Since the example in this paper is sampled during winter only, the seasonal effect is not considered.

M5: Grubb's test

It was noticed that the observed velocity and water depth follow the Student's t distribution if observed on the time, and not on date, basis. This enables the application of some univariate statistical test. Method M5 represents Grubb's test based on statistics of Student's t distribution (Kottegoda & Rosso 1998). The G statistic is calculated as follows:

$$G = \frac{Y - \overline{Y}}{\sigma}$$

where Y is the data that is subject to checking, $\overline{Y}$ and σ are mean value and standard deviation of the sample. The threshold value is be calculated as follows:

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

where N is number of data samples and $\alpha = 95\%$ is the confidence level.

M6: PCA – multivariate statistical test

One of the most popular outlier detection methods is PCA (Yoo *et al.* 2006). PCA is based on the linear relationship between data and will transform the data according to its variability. Data is transformed according to the correlation matrix to the new coordinate system that is oriented to the direction of greatest data variability. Hotteling's t^2 is the parameter that represents the Mahalanobis distance of transformed data value from the origin of new coordinate system. This parameter can be calculated as follows:

$$t^2 = n(x-\mu)^T W (x-\mu)$$

where n is data dimension; x is the multidimensional data value; μ is the mean value of x; and W is the covariance matrix. The latter matrix is a measure of how far the data value is from the majority of data. 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 Hotteling's t^2 can be compared with a threshold value. The threshold value is determined as the largest Hotteling's t^2, calculated using the time series for model development.

M7: ANN non-linear regression model

ANN (Cherkassky & Mulier 2007) is a popular tool for non-linear data regression. The feed forward type ANN with six hidden layers, trained by the back propagation algorithm, is used here to model the water depth h_i (ANN output) as a function of time t_i, (or more precisely, hour of the day when the modelled variable occurs), velocity V_i and electro-conductivity EC_i, representing the input layer of ANN:

$$h_i = f(t_i, V_i, EC_i)$$

In the application stage, the absolute difference between modelled and measured data is compared to the threshold value. The threshold value is determined as the largest difference between modelled and measured values.

M8: One-class SVM

Data classification is a data mining method used to classify the data based on predefined set of classes. The one-class SVM classification (Schölkopf *et al.* 1999; Chang & Lin 2001; Cherkassky & Mulier 2007) is a special case where the training data is classified into just one class forming a minimum radius sphere around the data. Since the minimum radius sphere is not always the optimal answer (data are not evenly spread within the sphere), the analysed data is usually transformed into the higher dimensional space using kernel functions. The radial basis kernel function is used here.

M9: Manning's equation

Physical models can also be utilised for anomaly detection. The major advantage of a physical model is that, unlike data driven models, it can provide reliable estimates even when extrapolating, that is, when used to make a prediction which

Table 1 | Classes, the data values are classified in according to defined contexts

Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
All Data	MOR.	DAY	EVN.	NIGHT	MOR	DAY	EVE.	NIGHT
All Data	WET	WET	WET	WET	DRY	DRY	DRY	DRY

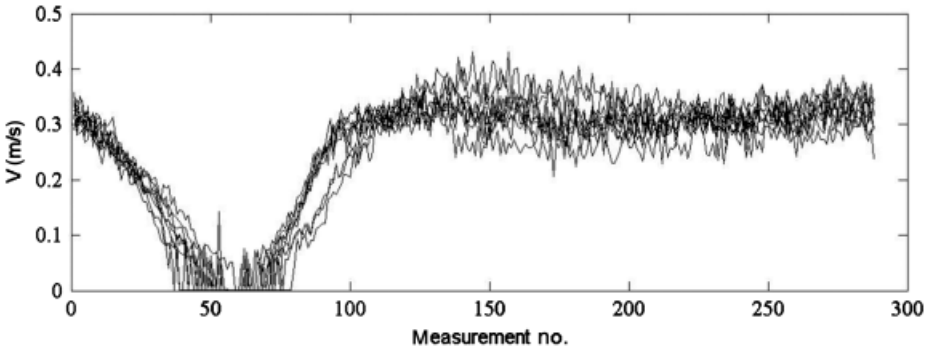

Figure 4 | Time series of measured water velocities (10 days) during dry weather.

is out of the limits of the data used for its calibration. In this study, a Manning's equation is used to link sewer flow velocity and its water depth:

$$V = \frac{1}{n} R^{\frac{2}{3}} \sqrt{I_d}$$

where n is the Manning's resistant factor; R is the hydraulic radius and I_d is the sewer bed slope. Since Manning's equation explicitly represents dependency of velocity on water depth, the inverse form is usually not available in explicit form. That is why the inverse form of Manning's equation has to be determined implicitly, using some optimisation method or expanding the equation in the Taylor's series. Nevertheless the regular form is used in this paper, as it didn't provide any significant changes in the result, when applied in inverse form (with calibrated parameters, the equation is close to linear). The calibration parameter is n.

Context classification methods

In this paper the monitored data samples were classified into nine classes using the following two contexts: time period of the day and weather conditions.

The first context defined (time period of the day) has four classes: 1) morning (6:40–8:10), 2) day (8:15–0:40), 3) evening (0:45–2:45) and 4) night (2:50–6:35). The second context is based on the fact that during the wet weather episodes data patterns and relationships are different from those during the dry weather episodes. Therefore, the second context is formed using the following two classes: 1) wet

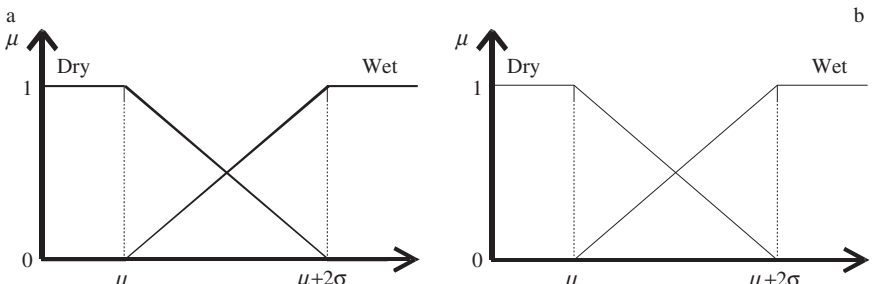

Figure 5 | Fuzzy sets that represent the belonging of the data value to a wet/dry weather context class for: (a) velocity and water depth and (b) electro-conductivity.

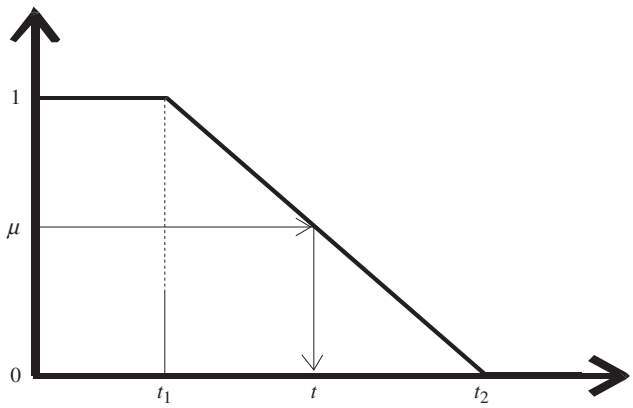

Figure 6 | Threshold value determination according to the data value's class membership.

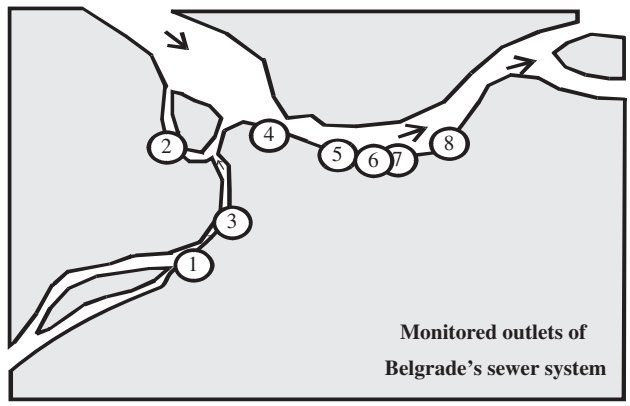

Figure 7 | Locations of eight monitored outlets in the Belgrade sewer system.

weather episodes and 2) dry weather flow data. Based on the above, all the data analysed can be classified into one of the nine classes presented in Table 1.

To define classes for the first context the velocity data values collected during the dry weather conditions (10 days) are selected and re-sampled. It can be noticed (Figure 4) that the daily pattern of the measured velocities varies from day to day, both in quantity and time.

A simple statistical algorithm empowered with fuzzy set theory was developed and used here to identify the start and end of wet weather events. The following (simple) logic was used: during the wet weather episodes, the velocity and water depth will be greater from the average values obtained during the dry weather episodes, but, at the same time, the electro-conductivity will decrease since the fresh rain water has lower ion concentration than the sanitary (i.e., dry weather) water. This logic can be formulated using fuzzy sets, as shown in Figure 5.

Since each data value has to be classified into one of two classes (wet/dry weather), the following fuzzy rule is determined to classify the data value into the wet weather class:

$$\mu_{rain}(V_i) \text{ AND } \mu_{rain}(h_i) \text{ AND } \mu_{rain}(EC_i) > 0.5 \text{ or}$$

$$\min(\mu_{rain}(V_i),\ \mu_{rain}(h_i),\ \mu_{rain}(EC_i)) > 0.5$$

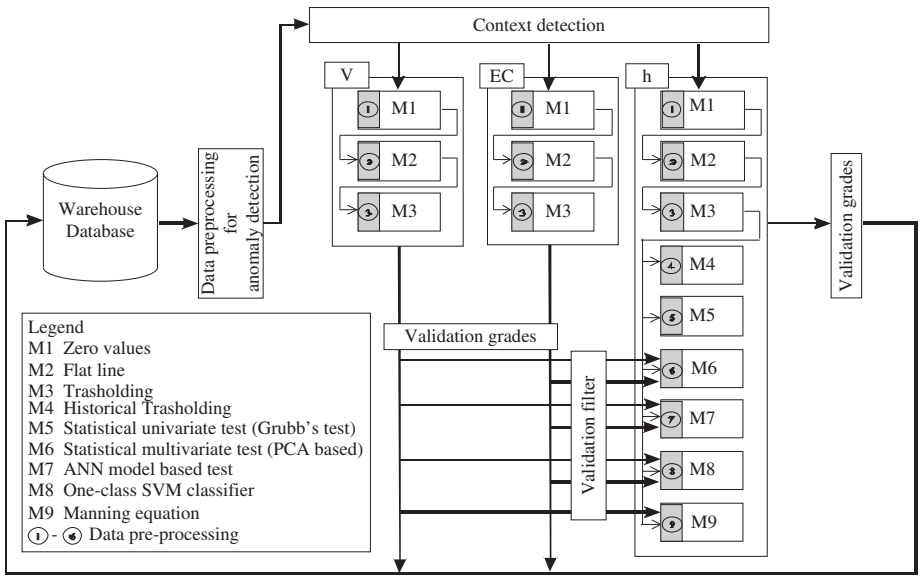

Figure 8 | The sequences of anomaly detection methods for velocity, conductivity and water depth.

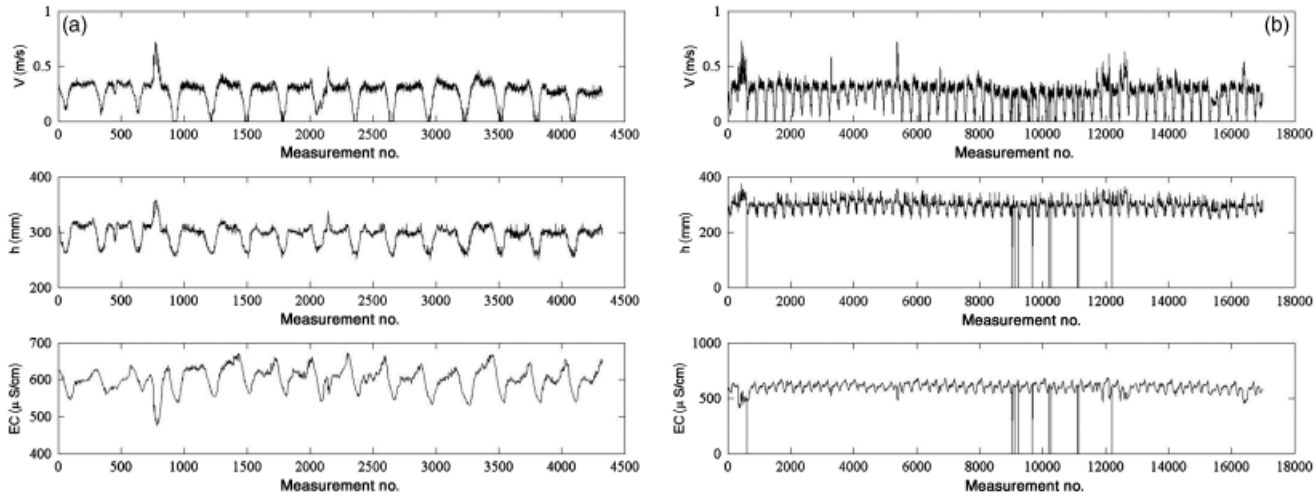

Figure 9 | Time series of variables measured at the Belgrade combined sewer system outlet: (a) time series for anomaly detection methods and context detection development; (b) time series for testing the suggested methodology.

The variables V_i, h_i and EC_i are the velocity, water depth and conductivity on a specific time point during the day (i). $\mu_{rain}(V_i)$, $\mu_{rain}(h_i)$ and $\mu_{rain}(EC_i)$ are mean values of velocity, water depths and conductivities on a specific time point during the day, obtained from the selected dry weather measured episodes. $\sigma(V_i)$, $\sigma(h_i)$ and $\sigma(EC_i)$ are standard deviations of velocity, water depths and conductivities on a specific time point during the day, obtained from the selected dry weather measured episodes.

After data classification, the anomaly detection methods are tuned for each class separately, using the same rules as in the anomaly detection analysis without contexts. Since the membership to the specific context class may be different from one, the threshold values are determined to compensate for potential misclassification. The threshold values (for methods M4, M6, M7 and M9) are estimated according to the membership of the data value to the specific context class. This procedure is graphically represented in Figure 6.

The value of membership to a specific class is mapped on the fuzzy set (Figure 5), where t_1 and t_2 are the threshold fuzzy set parameters determined according to the best practice and experience.

Table 2 | Historical minimum and maximum values of monitored variables

	V[m/s]	*h*[mm]	EC[μS/cm]
Min.	0.002	247	476
Max.	0.721	358	673

CASE STUDY

System description

The monitoring system of the Belgrade sewage system was established in 2006 with the aim to monitor water quantity and quality at eight principal outlets that cover more than 80 per cent of Belgrade's waste water collected from both households and industry (Figure 7). Time series of measured velocity, water depth and conductivity (see Figure 3) have been chosen here to demonstrate how adequate context detection and pre-processing tools can improve performance of the designed anomaly detection module.

System set-up

Velocity (V), water depth (h) and conductivity (EC) of sewer water in one outlet of the Belgrade's combined sewer system were continuously measured at a time step of 5 minutes (see Figures 9(a) and (b)). It was noticed that the time series of water depth have suspicious spikes not related to any known processor technical characteristic of the system. Unlike water depth, the velocity and

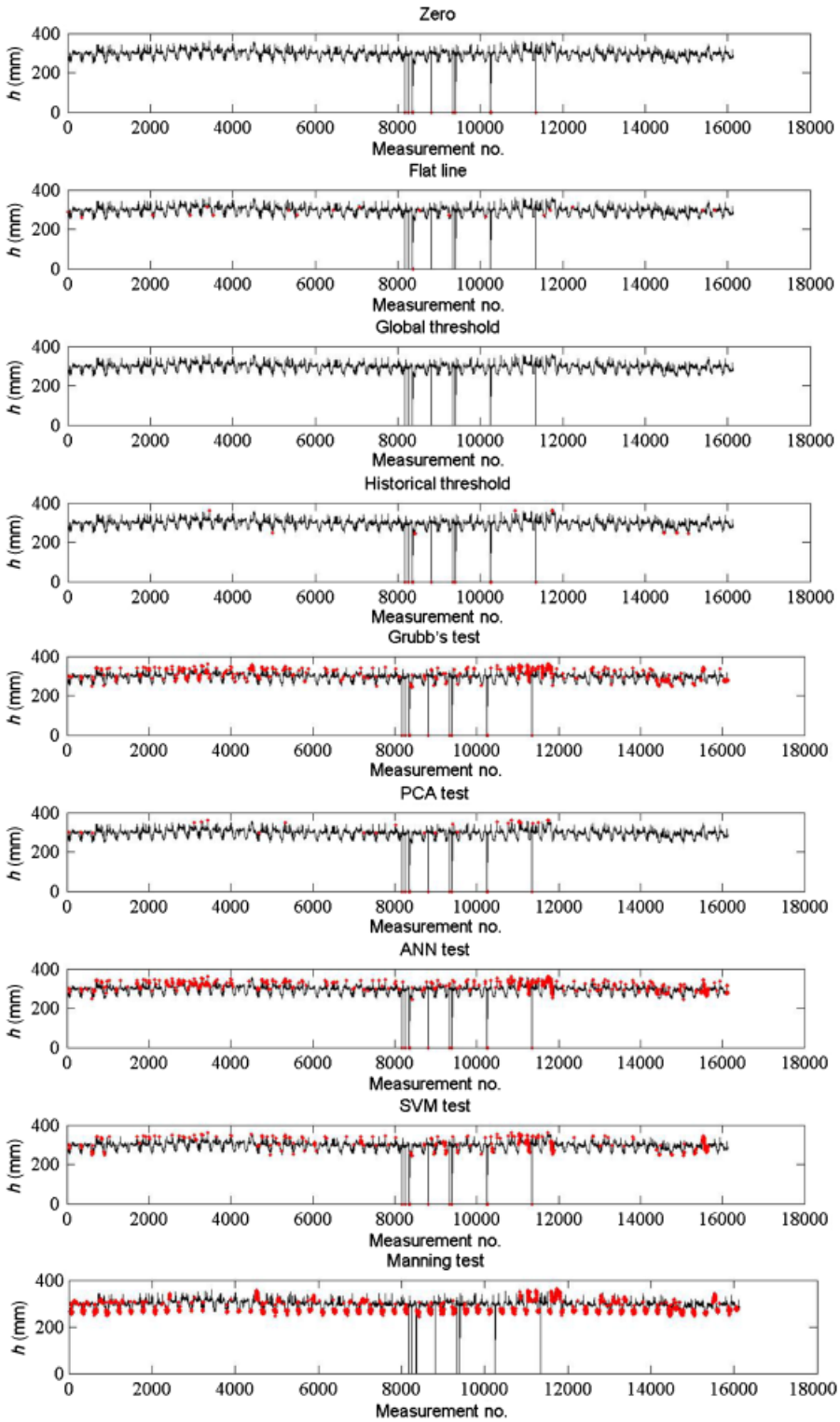

Figure 10 | Time series of measured data with marked anomalies without data pre-processing or context classifications.

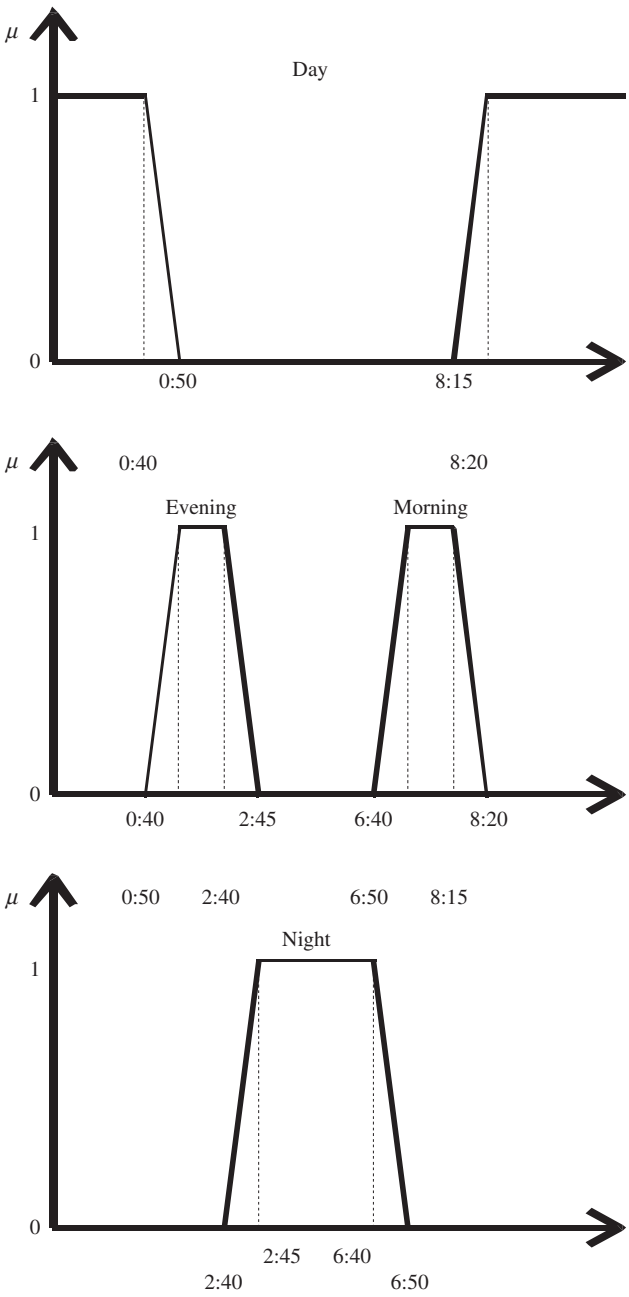

Figure 11 | Results of the fuzzy c-means clustering algorithm applied on time series of measured water velocities (10 days) during dry weather.

conductivity measurements presented more stable performance. Figure 8 depicts the anomaly detection procedure developed for the three aforementioned data time series.

The results of all methods were binary values, where 0 represents the regular and 1 represents the anomalous value. Methods M1–M3 were positioned at the start of the detection procedure, in the sequential order. These methods ware applied to all data.

The anomaly detection module is designed according to the time series inspection and authors' experience in the anomaly detection field. It is designed to provide labelling followed by exclusion of the following types of anomalies: a) zero values; b) flat line values; c) too large and too small values; d) spike type anomalies; e) data that are not in the correlation with most of the data, and f) data that are not consistent with the typical daily patterns.

The measured data time series are divided into two parts. The first part (15 days of measurements) is used for the development of context detection procedures and the design and tuning of anomaly detection methods (Figures 9(a)). The second part of data (59 days of measurement) is used for testing the proposed methodology and its improvements (Figures 9(b)). The first part of data is carefully chosen to represent all the features examined in the detection process. The wet weather (i.e., rainfall) episodes and daily data variations are present to ensure proper data anomaly detection model calibration and tuning. The time series used for testing is carefully checked for anomalies using all specified tools online in order to statistically test the performance of automatic approach of anomaly detection.

The diameter of the outlet ($R = 0.8$ m) is used as the constraint for water depth. The corresponding velocity is determined to be 1.5 m/s assuming the fully filled pipe and the Manning's coefficient for concrete of $n = 0.014$ $m^{-1/3}s$. The minima and maxima electrical conductivities are defined as follows: $EC_{min} = 50$ μS/cm for the clean water and $EC_{max} = 53000$ μS/cm for the seawater.

The historical limits on the time series used for method tuning (Figures 9(a)) are presented in Table 2.

The three-dimensional training dataset consisting of water depth, velocity and electro-conductivity is used to form a one-class SVM model with LIBSVM (A LIBrary for Support Vector Machines) (Chang & Lin 2001). At the application stage, each analysed water depth data value is checked to see if it is within the class boundaries or not. If not, the data is marked as anomalous.

It is assumed that the value of $n = [0.001, 0.014]$ $m^{-1/3}s$. The model based on Manning's equation is then calibrated by minimizing the mean square error between the

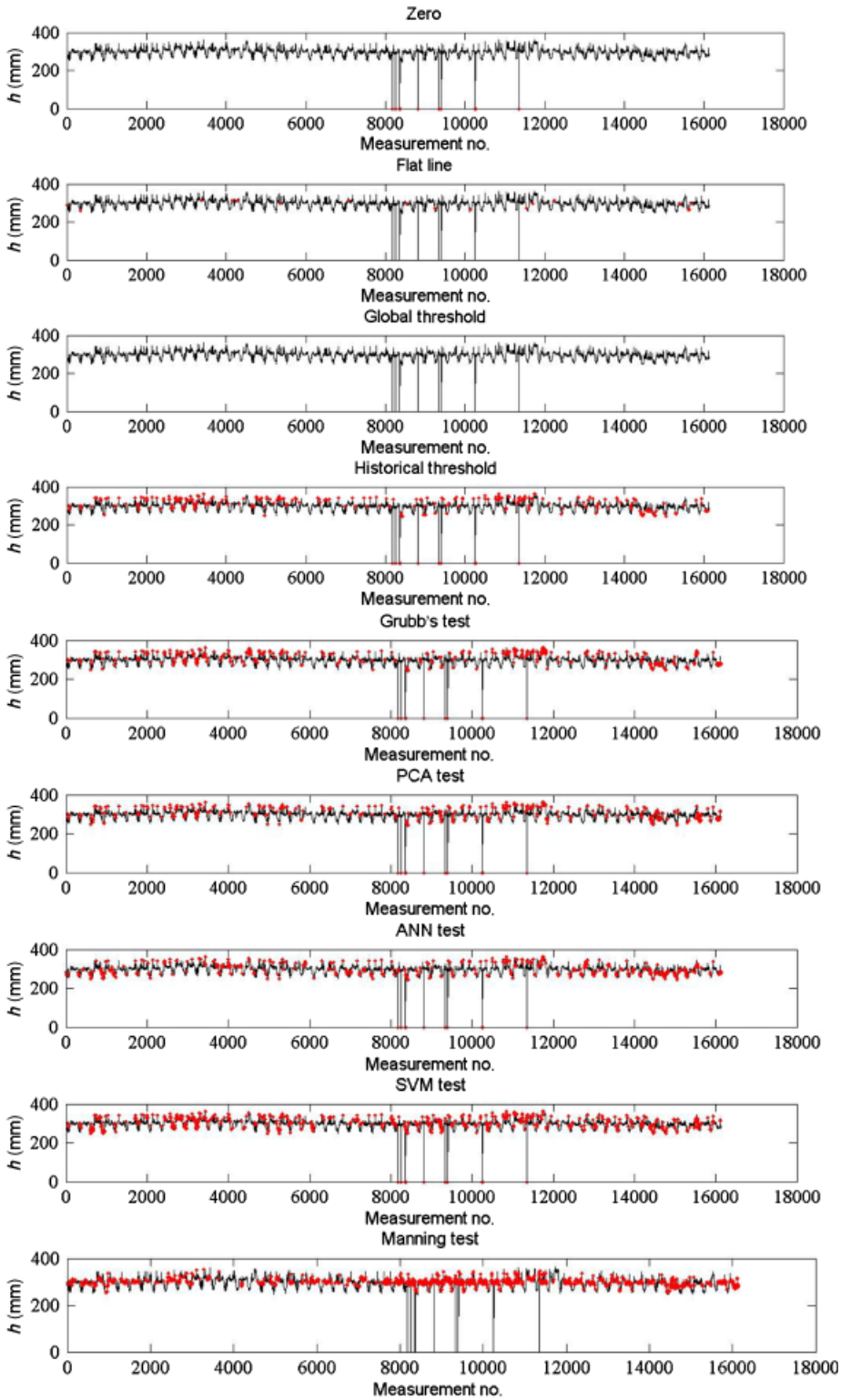

Figure 12 | Results of anomaly detections with data pre-processing and context classification.

Table 3 | Summary of anomaly detection results without data pre-processing or context classifications

	No anomalies	No detected	No missed	No false	P	$P_{\text{false tolerant}}$	$P_{\text{false sensitive}}$
M_1	244	24	220	0	0.517	0.052	0.098
M2	244	2	242	27	0.0039	0.004	0.007
M_3	244	0	244	0	0.0000	0.000	0.000
M_4	244	25	219	24	0.0513	0.054	0.093
M_5	244	191	53	621	0.2081	0.643	0.221
M_6	244	46	198	10	0.1018	0.104	0.181
M_7	244	234	10	559	0.2878	0.921	0.291
M_8	244	136	108	477	0.1641	0.386	0.189
M_9	244	67	177	3587	0.0167	0.159	0.017

measured and calculated velocities. The calibration resulted in $n = 0.014$ $m^{-1/3}s$. The threshold value (used to label a particular piece of data, in this case velocity, as good or bad) is determined as the largest residual value, that is, difference between the measured and the modelled velocities in the training dataset (see Figures 9(a)). This ensures that no anomalous data is present in the training dataset.

RESULTS AND DISCUSSION

Results of anomaly detection without context classification

The above nine data anomaly detection methods were firstly tuned (i.e., calibrated) using the training dataset. Once this was done, the methods were applied to the validation dataset in a sequence outlined above. The results obtained are shown in Figure 10 where all anomalous data points identified are marked with red dots. The same results are also summarized in Table 3 by using performance indicators defined in Equations (1), (2) and (3). As it can be seen from this table, using the M7 method 234 out of 244 anomalies were detected which equals a success rate of 92 percent.

The following can be noted from Figure 10 and Table 3: (1) The 'brute force' anomaly detection methods (M1–M3) provided expected results since their implementation is straightforward due to their narrow scope. The historical minimums and maximums marked all the zero values and some data values during wet weather events; (2) The Grubb's test (M5) provided results that identified all the data during the wet weather episodes as anomalous. That was expected since this data is the most outlying data in the statistical distribution; (3) The PCA method (M6) has poor results which can be explained with the fact that the threshold value for the PCA test was formed as the maximum Hotteling's t^2 value from the data time series used for developing the model (see Figures 9(a)); (4) The ANN-based method provided excellent results, while the SVM-based method had many false alarms, most of which occurred during the wet weather events. This is due to the fact that the SVM-based method is unable to test any data inside the minimum radius sphere. The reason for this is the method's nature to hide the anomalous data values if they are surrounded with the regular data; (5) The Manning equation method (M9) has shown good performance on daily data samples but it generated many false alarms during the night time. The reason for this is that the method used is rather simple with only one calibration parameter used for the whole data series.

Based on the above, it can be concluded that the problems encountered with some methods come from the fact that it is hard to develop and especially fine tune these methods in a way that will provide adequate data checks for all data values in the time series (for some data values threshold is too high or the method is not well calibrated). To further improve the performance of the above anomaly detection methods, the context detection and data pre-processing is used in next section.

Table 4 | Available training parameters for some context classes

	Class 0 All Data All Data	Class 1 Mor. Wet	Class 2 Day Wet	Class 3 Evn. Wet	Class 4 Night Wet	Class 5 Mor. Dry	Class 6 Day Dry	Class 7 Eve. Dry	Class 8 Night Dry
Training Parameters	4320	0	221	8	13	285	2749	367	677

Results of anomaly detection with context classification

Figures 11 depicts the results of the fuzzy c-means clustering algorithm (Cherkassky & Mulier 2007) used to divide the data sampled during the dry weather into four classes of the first context (time period context).

After context classification, all the models in the anomaly detection methods are retuned, and new, fuzzy based, threshold values are determined. The results obtained by running the same set of aforementioned nine anomaly detection methods, this time on the pre-processed and context classified data, and retuned are presented in Figure 12 and Table 5.

In the case of the Belgrade sewer system, there was not enough data in some context classes (marked in grey in Table 4) to develop some data driven methods (M6, M7 and M8). Those methods are omitted from the anomaly detection system for these classes.

In addition to context classifications introduced above, data pre-processing was also applied before running the anomaly detection methods. The data pre-processing involved the following: velocity and conductivity data were pre-processed to reduce the measurement noise and water depth data were re-sampled and rearranged in order to apply Grubb's test. The general data pre-processing tool (located between the data and the anomaly detection modules, see Figure 2) is also applied to all data with the objectives to mark any gaps in the data and check the double data entries.

Comparison of obtained results with and without context classification

As can be seen from presented results, when compared to the case of running the same set of nine anomaly detection methods on raw data, the detection results improved. The total number of detected anomalies is now 233 (out of 244), with method M8, resulting in a success rate of 93 per cent, which is higher than the success rate obtained in the case with raw data. It also can be seen that the number of false alarms is reduced for most of the anomaly detection methods.

Further comparison of the results obtained using the aforementioned nine anomaly detection methods with and without context classification is presented in Figure 13. As it can be seen from this figure, in majority of cases, the

Table 5 | Summary of anomaly detection results with data pre-processing and context classification

	No anomalies	No detected	No missed	No false	P	$P_{\text{false tolerant}}$	$P_{\text{false sensitive}}$
M_1	244	24	220	0	0.052	0.05	0.10
M2	244	0	244	28	0.000	0.00	0.00
M_3	244	0	244	0	0.000	0.00	0.00
M_4	244	202	42	220	0.399	0.71	0.44
M_5	244	187	57	334	0.294	0.62	0.32
M_6	244	224	20	359	0.360	0.85	0.37
M_7	244	109	135	896	0.085	0.29	0.10
M_8	244	237	11	483	0.321	0.93	0.33
M_9	244	71	173	2725	0.023	0.17	0.02

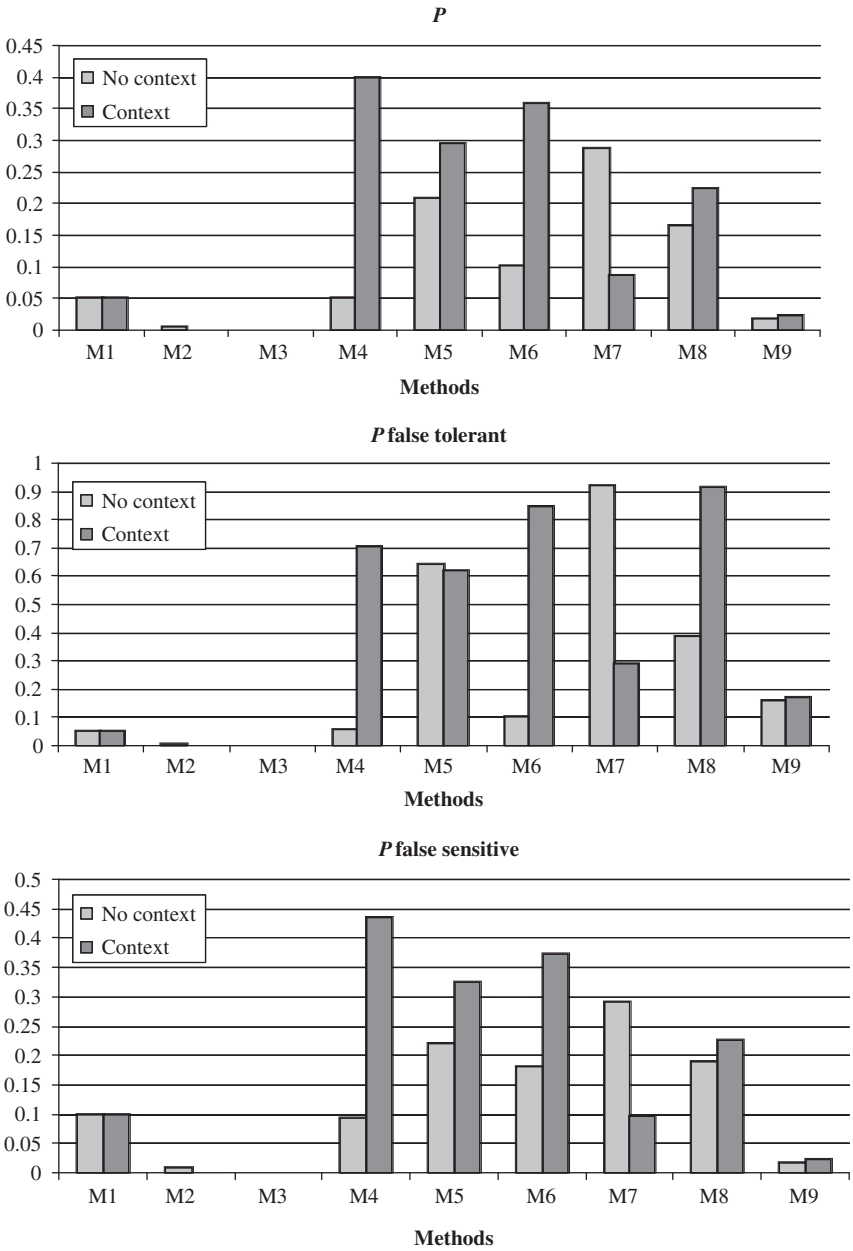

Figure 13 | Comparison of indicators P, $P_{\text{false tolerant}}$ and $P_{\text{false sesnsitive}}$.

effectiveness of the anomaly detection methods increases when context classification is used.

CONCLUSION

Anomaly detection is the first and major step in the quality assessment of any time series data. As shown in this paper, data can and should be classified using context information and pre-processed before it enters the anomaly detection system as this will improve the success rate of the latter system. In general, context-based detection should be considered before the anomaly detection methods are applied. On the other hand, data pre-processing can be done at three principal locations in the data acquisition and anomaly detection system. First, as part of the data integra-

tion system, second, as part of the general pre-processing module (prior to data entering the anomaly detection module) and third, as part of some specialized module for specific anomaly detection.

The above methodology developed and presented here was applied to a real-life case study involving three data streams (water depths, velocities and conductivity) in a sewer system of Belgrade. The results obtained clearly demonstrate the benefits (increased true and reduced false alarm rates) of using smart pre-processing techniques for data anomaly detection. Using context classification in anomaly detection methods increases the effectiveness of these methods in the automated, real-time context. To overcome the potential uncertainty in context classification, it is shown that the proposed fuzzy set methodology for threshold determination may be successfully used.

As mentioned, anomaly detection is just the first step. To implement the whole system and be able to explore and benefit from the results obtained, the integration of different grades to one single, possible continuous grade is needed. Such integration of grade marks has to be problem specific, it can't be universal. And of course, the data end user has to be able to cope with given measure of uncertainty.

REFERENCES

Bertrand-Krajewski, J. L., Laplace, D., Joannis, C. & Chebbo, G. 2000 *Mesures en Hydrologie Urbaine et Assainissement*. Tec&Doc, Lavoisier.

Branisavljević, N., Kapelan, Z. & Prodanović, D. 2009a Online time data series pre-processing for the improved performance of anomaly detection methods, CCWI-2009. In: *Integrating Water Systems*, Boxall, J. & Maksimovic, C. (eds.), Taylor & Francis Group, London, pp. 99–103.

Branisavljević, N., Prodanović, D. & Kapelan, Z. 2009b *Validacija podataka hidrotehničkih procesa*. SDHI, Babe, Srbija.

Branisavljevic, N., Kapelan, Z. & Prodanovic, D. 2009c Bayesian-based detection of measurement anomalies. In: *Environmental Data Series*, 8th International Conference on Hydroinformatics Hydroinformatics, Concepsion-Chile.

Branisavljević, N., Prodanović, D. & Pavlović, D. 2009d Automatic, semi-automatic and manual validation of urban drainage data. In: *Proc of the 8th Conference on Urban Drainage Modelling, 11 September*, Tokyo, Japan.

Chang, C. & Lin, C. 2001 LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cherkassky, V. & Mulier, F. 2007 *Learning from Data: Concepts, Theory, and Methods*. 2nd edition. Wiley-IEEE Press, Hoboken, NJ.

Kottegoda, N. & Rosso, R. 1998 *Statistics, Probability and Reliability for Civil and Environmental Engineers*. McGraw-Hill International Editions, Singapore.

Mourad, M. & Bertrand-Krajewski, J.-L. 2002 A method for automatic validation of long time series of data in urban hydrology. *Wat. Sci. Technol.* **45**(4–5), 263–270.

Patcha, A. & Park, J. M. 2007 An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Networks* **51**, 3448–3470.

Rosen, C., Röttorp, J. & Jeppsson, U. 2003 Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. *Wat. Sci. Technol.* **47**(2), 171–179.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. 1999 Estimating the support of a high-dimensional distribution. Technical report, Microsoft Research, MSR-TR-99–87.

Venkat, V., Raghunathan, R., Kewen, Y. & Surya, N. K. 2003a A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Comput. Chem. Eng.* **27**, 293–311.

Venkat, V., Raghunathan, R., Kewen, Y. & Surya, N. K. 2003b A review of process fault detection and diagnosis Part II: Qualitative models and search strategies. *Comput. Chem. Eng.* **27**, 313–326.

Venkat, V., Raghunathan, R., Kewen, Y. & Surya, N. K. 2003c A review of process fault detection and diagnosis Part III: Process history based methods. *Comput. Chem. Eng.* **27**, 327–346.

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

First received 1 March 2010; accepted in revised form 3 August 2010. Available online 6 January 2011