Geostatistical techniques for approximate location of pipe burst events in water distribution systems

Michele Romano, Zoran Kapelan and Dragan A. Savić

ABSTRACT

This paper focusses on the customisation and further enhancement of the recently developed data-driven methodology for the automated near real-time detection of pipe bursts and other (e.g. sensor faults) events at the district metered area (DMA) level. Assuming the availability of pressure/flow data from an increased number of sensors deployed in a DMA, the aim is to: (i) overcome the limitations of the probabilistic inference engine when dealing with the increased data availability; and (ii) exploit the event information resulting from the analysis of the larger number of DMA signals for determining the approximate location of the pipe burst events within the DMA. This is achieved by making use of a multivariate Gaussian mixtures-based graphical model and geostatistical techniques. The novel detection and location methodology is demonstrated and tested on a series of simulated pipe burst events that were performed by opening hydrants in a real-life DMA in the UK. The results obtained illustrate that the new methodology can successfully determine the approximate location of pipe bursts within a DMA (in addition to detecting them in a fast and reliable manner). The performance comparison of several geostatistical techniques shows that the Ordinary Cokriging technique outperforms all other techniques tested.

Key words | artificial intelligence techniques, burst detection and location, geostatistical techniques

Michele Romano (corresponding author)
Zoran Kapelan
Dragan A. Savić
Centre for Water Systems,
College of Engineering,
Mathematics and Physical Sciences,
University of Exeter,
Exeter,
Devon,
UK
E-mail: *mr277@exeter.ac.uk*

INTRODUCTION

The problem of pipe burst events in water distribution systems (WDSs) is a compelling issue for the water companies worldwide. Pipe bursts not only cause economic losses to the water companies (Colombo & Karney 2002) but also represent an environmental issue (i.e. waste of water and energy) and a potential risk to public health (Sadiq *et al.* 2006). Furthermore, they have a negative impact on the water companies' operational performance, customer service and reputation.

In the UK, despite the significant effort put into the ongoing rehabilitation and maintenance of the water supply infrastructure, the number of incidents caused by pipe burst events is still significant. This situation is mainly because the majority of water supply pipes were installed in the first part of the twentieth century and hence today many are in relatively poor condition. Additionally, because of the stochastic nature of these events, it is impossible to predict and completely eliminate them. As a consequence, in their day-to-day operations, the water companies are not only tasked with operating their WDSs optimally to minimise the costs and to meet the required standards of service (e.g. in terms of water quality and providing adequate pressure at the customers' taps), but also with managing contingency situations when pipe burst events occur. This additional burden is especially heavy because the water companies are mainly judged by the public and the regulatory agencies alike based on how well (or otherwise) they perform this task. At present, the resulting potential unplanned interruptions to water supply and the damaging consequences of pipe bursts are tolerated to a lesser extent. Consequently, the timely and reliable detection and location of pipe burst events provides opportunities for improving the water companies' operational efficiency and customer service.

doi: 10.2166/hydro.2013.094

Currently, many pipe burst event detection and location techniques that are based on various principles exist (Puust *et al.* 2010). However, none is ideal and the number of techniques currently practised by the water companies is limited. In many cases, pipe bursts are brought to the attention of a water company only when someone calls in to report a visible event. Water companies that embrace modern leakage management technologies devote considerable manpower and resources to proactively detect and locate the pipe bursts by utilising techniques that make use of highly specialised hardware equipment (e.g. leak noise correlators, pig-mounted acoustic sensors, ground penetrating radar devices, etc.). Some of these techniques are the most accurate ones in use today (Puust *et al.* 2010), but they are also expensive, labour-intensive, slow to run and may require the cessation of pipeline operations for long periods of time. Furthermore, these techniques are generally used after the pipe burst events have occurred and not in near real time. Consequently, much research has been focussed on finding equally effective, but faster and non-invasive, techniques that cost less money to run.

Several techniques exist that promise low-cost solutions by endeavouring to solve the pipe burst detection and location problem by numerical analyses only (Puust *et al.* 2010). Among these techniques, those that make use of statistical and artificial intelligence (AI) data analysis tools for automatically processing the operational variables (e.g. pressure and flow) in an online fashion are of particular interest for providing a rapid response to pipe burst events. Primary examples are the techniques presented by Fenner & Ye (2011), Mounce *et al.* (2010a, 2011), and Palau *et al.* (2012). These techniques are promising because they automate the mundane tasks involved in the data analysis process. They can efficiently deal with the vast amount of, and often imperfect, sensor data collected by modern supervisory control and data acquisition (SCADA) systems and extract information useful in making reliable operational decisions. Statistical/AI-based techniques also present several advantages over other methods such as steady-state analysis-based (e.g. Pudar & Liggett 1992; Wu *et al.* 2010), transient analysis-based (e.g. Liggett & Chen 1994; Kapelan *et al.* 2003), and negative pressure wave techniques (e.g. Misiunas *et al.* 2005; Srirangarajan *et al.* 2012). Firstly, they have a requirement for pressure and/or flow measurements that are sampled much less frequently (e.g. every 15 min) than those required for transient analysis. Furthermore, they rely on the empirical observation of the behaviour of the pipeline network, thus precise knowledge of the pipeline and instrumentation parameters is not required.

It has to be stressed that the above statistical/AI-based techniques have recently started to appear mainly because of the latest developments in hydraulic sensor technology and online data acquisition systems. These developments have enabled the water companies to deploy a larger number of more accurate and cheaper pressure and flow devices and allow data collected by these devices to be received in near real time. Nowadays, the UK district metered areas (DMAs) are usually observed by using pressure and flow sensors located at the DMA entry/import/export points and a pressure sensor located at the DMA critical point (i.e. the one located either at the point of highest elevation or alternatively at a location farthest away from the inlet). The data streams (i.e. signals) from these sensors provide a potentially useful source of information for detecting and locating pipe burst events both quickly and economically. As water companies recognise this fact more and more, and that several other important benefits are yielded by the monitoring of their WDSs in near real time (e.g. improved network visibility and management, higher compliance with regulatory targets, etc.), an increase in the density of coverage of pressure/flow monitoring locations is expected in the near future.

Despite their initial success, the aforementioned statistical/AI-based techniques can be further improved in terms of both event detection reliability and detection time. Furthermore, at present, the available statistical/AI-based techniques allow the discovery of a pipe burst event in a particular area within a WDS (e.g. at the DMA level) without giving any information about its more precise location. Thus, they can be also improved in terms of pipe burst event location accuracy (i.e. to indicate more precisely the likely location of the pipe burst – to restrict the pipe burst search area).

Romano *et al.* (2012) described the development of a methodology for the automated near real-time detection of pipe bursts and other events (which induce similar abnormal pressure/flow variations) at the DMA level. It works by analysing simultaneously all the signals coming online

from the small number of pressure and/or flow sensors usually deployed in a UK DMA. This methodology offers noticeable improvements over the existing techniques. The main improvements involved: (i) using advanced techniques for more efficient and effective processing of the hydraulic data gathered (e.g. wavelets for removing noise from the measured flow and especially pressure signals); (ii) taking advantage of a number of different ensembles of statistical/AI techniques (i.e. statistical process control (SPC) and artificial neural networks (ANNs)) for recognising the various types of event-occurrence evidence; and (iii) using a probabilistic inference engine based on Bayesian networks (BNs) for combining the above evidence and inferring the probability of an event occurrence. Subsequently, Romano *et al.* (2013) further enhanced the above methodology by using: (1) an evolutionary algorithm (EA) optimisation strategy for automatically selecting the best ANN input structures and parameters; and (2) an expectation maximisation (EM) strategy for semi-automatically (re)calibrating the values in the conditional probability tables (CPTs) of the BN (whose output is used for raising the detection alarms). It was shown that the use of this methodology resulted in more effective, reliable, and faster detection of pipe bursts and other events in a DMA.

This paper focusses on the customisation and further development of the event recognition system (ERS), which implements the methodology recently developed by the authors (Romano *et al.* 2012, 2013). Given the trends in availability of pressure/flow data from an increased number of sensors deployed in a DMA, the main aim of the work presented here is to enhance and extend the capabilities of the ERS by enabling it to: (i) more efficiently deal with the increased number of event-occurrence evidence when inferring the probability that an event has occurred; and (ii) exploit the event information resulting from the analysis of the larger number of DMA signals to determine the approximate location of the pipe bursts within the DMA (in addition to detect them in a fast and reliable manner). In view of this objective, the ERS's customisation involves replacing the BN used in the ERS's probabilistic inference engine for inferring the event-occurrence probability at the DMA level with a multivariate Gaussian mixtures-based graphical model (Duda & Hart 1973). This customisation is done to overcome the BN limitations when dealing with increased data availability. The ERS's further development involves using geostatistical techniques for building a model to predict the probability value of a burst associated with each DMA pipe. The intention is to provide a means by which to identify the group of DMA pipes that most likely include the failed pipe.

The remainder of this paper is organised as follows. After this introduction, an overview of the new (i.e. detection + location) methodology is given. This overview is then followed by two sections presenting the theoretical background and implementation details of the techniques used for the ERS's customisation and further development, respectively. The latter sections constitute the core of the new contribution presented in this paper. Next, the results of tests in a UK DMA with simulated pipe burst events (i.e. engineered events (EEs)) are presented in the case study section. Finally, the main conclusions are drawn and acknowledgements given. Note that several abbreviations are used in this paper. A list of these abbreviations can be found in Table 1.

EVENT DETECTION AND LOCATION METHODOLOGY

An automated ERS has been developed recently by Romano *et al.* (2012, 2013). This section presents a brief overview of this system necessary to describe the improvements associated with: (i) its *Inference* subsystem (i.e. ERS's probabilistic inference engine), which is enhanced in this paper; and (ii) its *Location* subsystem, which is newly introduced here as part of the ERS's further development. A more detailed description of the ERS is available in the above references.

The data processing in the ERS starts by receiving the data communicated by the DMA sensors. For each DMA signal and at each communication interval u readings are obtained. For example, assuming 15-min sampled data that are communicated every 30 min to improve the sensors' battery life, the value of u is as equal to 2. This said, it is evident that the value of this parameter depends on the frequency at which the data are communicated and the sampling frequency of the measurements. Therefore, this value can be different for different water companies/SCADA setups. In view of this situation, it has to be stressed that the ERS's data processing

Table 1 | List of abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
BBA	Boundary based analysis
BN	Bayesian network
BIS	Bayesian inference system
CPT	Conditional probability table
DBA	Discrepancy based analysis
DMA	District metered area
EA	Evolutionary algorithm
EE	Engineered event
EM	Expectation maximisation
ERS	Event recognition system
IDW	Inverse distance weighted
LP	Local polynomial
NOP	Normal operating pattern
OC	Ordinary cokriging
OK	Ordinary kriging
RMSE	Root mean square error
SCADA	Supervisory control and data acquisition
SPC	Statistical process control
TBA	Trend based analysis
WDS	Water distribution system

framework is capable of dealing with different communication/sampling frequency scenarios. The u readings obtained then update a time series record, which is stored in the Time Series database. Once all the DMA signals are fully processed as described below, the resulting u probability values that an event has occurred in the DMA and the additional output information useful for enabling the approximate event location (i.e. event-occurrence probability values estimated at the sensor locations) are stored in the Alarms database. If any of the u probability values exceeds a fixed threshold, an alarm is generated. Following the generation of an alarm, the additional output information provided by the ERS is processed further to determine the approximate location of the event occurring.

Figure 1 shows a diagrammatic representation of the ERS. As can be observed from this figure, processing of the pressure/flow data is performed in the five ERS components (i.e. dashed/dotted rectangles) as follows:

1. Capturing of the normal operating patterns (NOPs) of the DMA pressure/flow signals.
2. Identification and estimation of the event-induced deviations between observed (i.e. measured) and captured DMA signal patterns.
3. Inference about the probability that an event has actually occurred, based on above deviations.
4. (Re)calibration of the probabilistic inference engine based on the past events information.
5. Determination of the approximate location of an event within the DMA.

Note that the first three ERS components are used for the actual event detection. The fourth ERS component is used for the initial calibration and the follow-on periodic recalibrations of the ERS's probabilistic inference engine. The fifth ERS component is used for the approximate event location.

Figure 1 also shows that the aforementioned five ERS components are further organised into seven subsystems (i.e. solid snipped corner rectangles) each containing a number of different modules (i.e. solid rectangles). The seven ERS subsystems are as follows: (1) the *Setup* subsystem; (2) the *Discrepancy Based Analysis (DBA)* subsystem; (3) the *Boundary Based Analysis (BBA)* subsystem; (4) the *Trend Based Analysis (TBA)* subsystem; (5) the *Inference* subsystem; (6) the *Bayesian Inference System (BIS) parameters learning* subsystem; and (7) the *Location* subsystem.

The first ERS subsystem (equivalent to first ERS component) is used to perform the pressure/flow signal pattern capturing. The first two modules (i.e. data retrieval and data pre-processing) are used for retrieving the historical data from the Time Series database and assembling a set of data that best represents the most recent NOP of the DMA signal being analysed (i.e. NOP data set). The latter is achieved by automatically discarding the pressure/flow measurements that can be considered as outliers and/or that are not consistent with the expected pressure/flow variations, assuming that no event occurred in the DMA. Next, the third module (i.e. statistics estimation) is used to estimate (from the NOP data set) several vectors of descriptive statistics (i.e. averages and standard deviations). These vectors provide basic statistical information about the

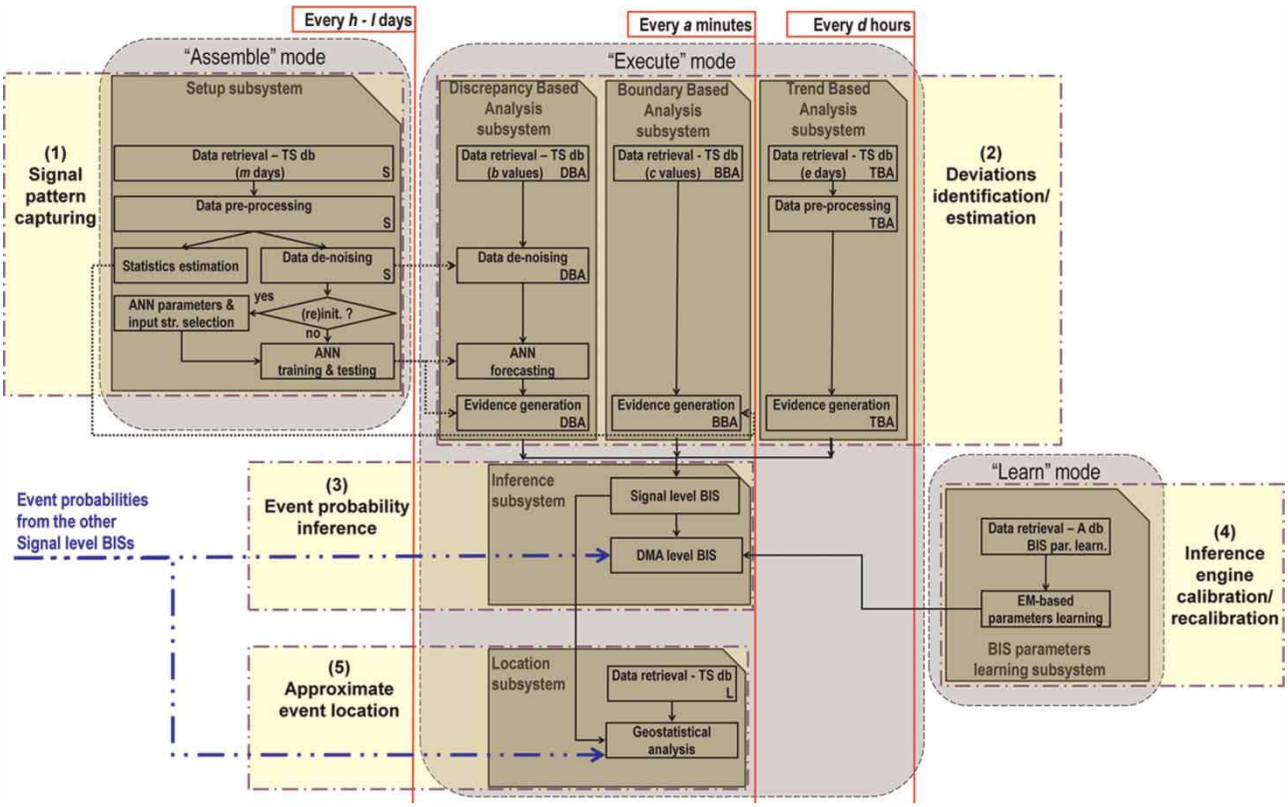

Figure 1 | Diagrammatic representation of the event recognition system components, subsystems and modules.

DMA signal NOP. The remaining modules (i.e. data de-noising, ANN parameters and input structure selection, and ANN training and testing), on the other hand, are used firstly to remove noise from the NOP data set and then for: (i) training and testing an ANN model for the short-term prediction of future DMA signal values; and (ii) to estimate the ANN model prediction error's variability. Note that as the resulting ANN model assumes that no event occurred in the DMA, it provides a model-based type of information about the DMA signal NOP.

The second, third and fourth ERS subsystems are used together (i.e. synergistically) to perform the deviations identification and estimation data analysis in the second ERS component as follows: (i) the *DBA* subsystem checks that the discrepancies between the incoming observed DMA signal values and their ANN predicted counterparts do not exceed pre-defined limits based on the estimated ANN model prediction error's variability; (ii) the *BBA* subsystem checks that the incoming observed DMA signal values lie inside a 'data envelope' whose boundaries are defined by using the vectors of descriptive statistics estimated from the NOP data set; and (iii) the *TBA* subsystem monitors, on a Control Chart (Shewhart 1931), how the mean of the historical DMA signal values recorded during a particular time window during the day (e.g. from midnight to 4, 4 to 8 a.m., etc.) varies over time. The reason for using three analysis subsystems is that, by using an ensemble of different statistical and AI techniques, each of them focusses on recognising a specific type of evidence that an event has occurred. Furthermore, as they perform tasks in parallel they allow simultaneous assessment of how an event affects the pressure/flow measurements from different perspectives (e.g. short-term and long-term effects).

The fifth ERS subsystem (equivalent to the third ERS component) is used to perform the event probability inference analysis. Starting from the event-occurrence evidence generated as described above, various (i.e. one for each

DMA signal) discrete input BN-based Signal level BISs are used to infer the event-occurrence probability values at the sensor locations. These probability values are then further processed in the multivariate Gaussian mixtures-based DMA level BIS to infer the 'global' (i.e. for the DMA being studied) event-occurrence probability value.

The sixth ERS subsystem (equivalent to the fourth ERS component) is used to perform the inference engine (re) calibration data analysis. An EM strategy is used here for semi-automatically (re)calibrating (i.e. learning) the DMA level BIS parameters based on information about the past events that have occurred in the DMA being studied.

Finally, the seventh subsystem (equivalent to the fifth ERS component) is used to perform the approximate event location data analysis. In this subsystem, the output information resulting from the various Signal level BISs is processed further by means of geostatistical techniques. Based on the locations of the deployed sensors, these techniques enable performing spatial interpolation of the event-occurrence probability variable. As a result, a probability value of a burst is associated with each DMA pipe.

As shown in Figure 1, the ERS has three main modes of operation: (1) the 'Assemble' mode; (2) the 'Execute' mode; and (3) the 'Learn' mode. These modes of operation define the time schedule according to which data analyses are performed in each subsystem. The 'Assemble' mode is used to 'tune' the data-driven ERS when it is initialised (i.e. used for the first time in a DMA). Later on, it is used: (i) regularly (e.g. weekly) when the ERS is updated (to capture the latest normal operating conditions of a DMA), thereby providing a continuously adaptive ERS; and (ii) periodically (e.g. every 3 months) when the ERS is reinitialised (to account for the seasonal variations in the DMA's pressure/flow regime, growing demand over time, etc., or following occasional operational/other DMA changes, e.g. re-valving). The 'Execute' mode is the normal operating mode used at every communication interval to detect and approximately locate the events. Finally, the 'Learn' mode may be used for the initial calibration and for the follow-on periodic recalibration of the ERS's probabilistic inference engine. As the data analyses performed in this mode of operation have a requirement for the past events information, its actual utilisation depends on whether or not this information is available/considered.

MULTIVARIATE GAUSSIAN MIXTURES-BASED DMA LEVEL BIS

The objective of the DMA level BIS is to infer, at each time step during a data communication interval, the probability that an event has occurred in the DMA being studied. This calculation then enables (by means of a user-defined detection threshold) the raising of the detection alarms if and when necessary.

In the ERS presented in Romano *et al.* (2012, 2013), the DMA level BIS consists of a BN (Edwards 2000; Jensen 2001), which combines all the evidence of an event occurrence resulting from the different ERS analysis subsystems (i.e. *DBA*, *BBA* and *TBA*) and coming simultaneously from all the DMA signals. Each node of that BN represents a variable (e.g. event-occurrence evidence from a particular analysis subsystem, coming from one of the DMA signals at a specific time) and is discretised into states (e.g. high, moderate or none). A CPT is associated with each node and contains the parameters (i.e. probability values) that are used to perform inference. It has to be stressed that, if an increase in the number of DMA signals that are simultaneously analysed in the ERS is assumed, the number of BN inputs and consequently the overall number of BN nodes and parameters in the CPTs will increase drastically. This situation may affect the computational efficiency of the inference process. Furthermore, manual specification of all the required CPT parameters is tedious, and necessitates not only domain knowledge, but also an understanding of the probabilistic calculus and of the probabilistic graphical models. Even when the learning of the CPT parameters from the past events is considered (Romano *et al.* 2013), the increased number of 'missing' parameters (i.e. those in the hidden BN nodes) together with the large multidimensional data structure pose serious challenges to the efficiency of the algorithm for learning from incomplete data that has to be used.

To avoid the potential computational inefficiency during the inference process and to circumvent the other difficulties/limitations outlined above, in the customised ERS presented here, the DMA level BIS consists of a two-class (i.e. alarm on, alarm off), two-component multivariate Gaussian mixtures-based graphical model (Duda & Hart 1973). This DMA level BIS works in an n-dimensional feature space by inferring, at any time step, the probability that an event has occurred in the DMA – based on the continuous

output (i.e. event-occurrence probability value) of the n Signal level BISs. This situation implies that, unlike the BN-based DMA level BIS used previously (Romano *et al.* 2012, 2013), the various types of event-occurrence evidence resulting from the different data analyses performed on each DMA signal are not used as input directly. Here, they are first processed in the Signal level BISs. The outputs of these BISs are then fed into the multivariate Gaussian mixtures-based graphical model. This said, note that by processing the event-occurrence evidence in such a way the new multivariate Gaussian mixtures-based DMA level BIS still enables synergistic combination of the event information from all the analysed DMA signals, thereby aiming at increasing the reliability of the detection alarms. It is also worth stressing that the above results in an ERS with a 'hierarchical architecture' (i.e. Signal level BISs–DMA level BIS). This 'hierarchical architecture' has the potential to enable (if desired and/or beneficial) embedding the local detection intelligence (i.e. data analyses performed on the single DMA signal) into the sensor device itself ('smart device' in AI jargon). The central detection intelligence (i.e. the data analysis performed by the DMA level BIS), on the other hand, can become part of the overall decision support type system used in the water company's control room.

Figure 2 shows the structure of the new. It can be observed from this figure that this graphical model has three nodes, two of which have discrete parameters for each of the allowed states of the relevant variables. In the third node (i.e. input node) the probability distribution is described using the multivariate Gaussian mixture function formula. The parameters of the multivariate Gaussian mixtures (i.e. means and covariances) together with the other

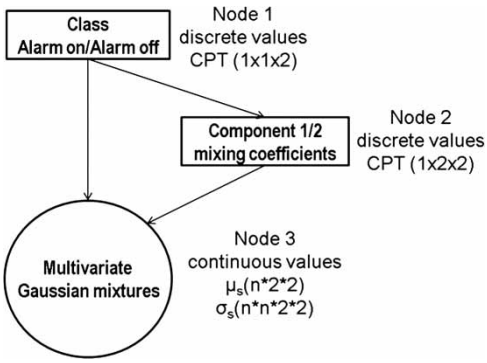

Figure 2 | Structure of the district metered area level Bayesian inference system.

model's parameters are determined using the EM algorithm (Dempster *et al.* 1977). The EM algorithm represents a general method for estimating likelihood functions. It is useful in situations in which simpler optimisation methods fail and it is the most commonly employed algorithm for learning from incomplete data (Jensen 2009). This algorithm was developed in the statistics community by Dempster *et al.* (1977) and adapted for the use with the graphical models by Lauritzen (1995). For a given database of 'cases' the EM algorithm determines estimates of the model's parameters that are optimal within a neighbouring set of solutions. It starts with initial values (e.g. chosen at random) for all the model's parameters, and then iteratively refines them. Each iteration ensures that the likelihood function increases and eventually converges to a local maximum. The iteration process consists of two steps, namely the Expectation (E) and Maximisation (M) steps, which are performed in alternating manner until convergence. Note that a detailed description of the EM algorithm, as can be applied to estimating the parameters of a mixture of multivariate Gaussian densities, can be found in Redner & Walker (1984).

APPROXIMATE PIPE BURST EVENT LOCATION

Geostatistical techniques

Geostatistics is a branch of statistics that focusses on spatial datasets. Applications of geostatistical techniques can be found in mining engineering, hydrology, geology and many other fields (see Isaaks & Srivastava 1989; Cressie 1993). Geostatistical techniques focus on the relationship between the value of a variable at a given geographical location and the values of the same and (possibly) other variables at locations some distance from it (i.e. at a set of measured locations). Their basic goal is to interpolate the value of a variable at locations that have not been measured, using data from the surrounding measured locations. Ultimately, they allow the creation of a model (i.e. interpolation surface) of how the variable's values are distributed across the entire domain of interest.

In this study the following geostatistical techniques have been considered: (i) the inverse distance weighted (IDW) interpolation technique (Shepard 1968); (ii) the local

polynomial (LP) interpolation technique (Gandin 1963; Cleveland & Devlin 1988); (iii) the ordinary kriging (OK) technique (Krige 1951; Isaaks & Srivastava 1989); and (iv) the ordinary cokriging (OC) technique (Myers 1982). Note that in-depth discussions about these and other geostatistical techniques for spatial interpolation can be found in the above references and in Cressie (1993), Deutsch & Journel (1998) and Banerjee *et al.* (2004). Here, an outline of the main characteristics of the above techniques is given only.

Inverse distance weighted interpolation

IDW interpolation is a deterministic geostatistical technique that estimates the variable's value at a prediction (i.e. unmeasured) location using a linear combination of the measured variable's values surrounding the prediction location. Those values are weighted by an inverse function of the distance from the prediction location to the measured locations. IDW assumes that the measured variable's values closest to the prediction location have the greatest influence on the predicted variable's value. The variable's value at a prediction location s_0, is computed as follows:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{1}$$

where n is the number of measured variable's values surrounding the prediction location, λ_i are the weights assigned to each measured variable's value, and $Z(s_i)$ is the measured variable's value at the location s_i. The weights are determined as follows:

$$\lambda_i = \frac{d_{i0}^{-p}}{\sum_{i=1}^{n} d_{i0}^{-p}} \quad \text{where} \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{2}$$

where d_{i0} is the distance between the prediction location and each of the measured locations. As d_{i0} increases, a weight approaches zero. The factor p is a power parameter that influences how fast the weighs decrease as d_{i0} increases. The choice of the p value is arbitrary (Webster & Oliver 2001). The most popular choice of p is 2 (Li & Heap 2008).

IDW interpolation is one of the simplest geostatistical techniques. However, it does not take advantage of the spatial correlation structure of the data explicitly. Furthermore, IDW is an exact interpolation technique. That is, it predicts a variable's value identical to the measured variable's value at a measured location. This situation implies that this technique may generate surfaces with sharp peaks or valleys. Finally, the interpolated values at any location within the domain of interest are bounded by the maximum and minimum of the measured variable's values. This factor is considered to be an important shortcoming because, in order to be useful, an interpolation surface should predict accurately certain important features of the 'true' surface. For example, the locations and magnitudes of maxima and minima, even when they are not included in the set of measured variable's values (Lam 1983).

Local polynomial interpolation

LP interpolation is a deterministic geostatistical technique that finds its roots in the trend surface analysis theory. Trend surface analysis assumes that the variable's values in the domain of interest are a function of the geographic coordinates. Each measured variable's value $Z(s)$, is considered to be the sum of a deterministic polynomial function of the geographic coordinates $f(x, y)$ (i.e. trend surface), plus a random error term ε (Webster & Oliver 2001):

$$Z(s) = f(x, y) + \varepsilon \tag{3}$$

The polynomial function can be expanded to any desired degree. The coefficients of the polynomial function are found by the method of least squares, which makes sure that the sum of the squared deviations from the trend surface is minimised. The variable's values at the prediction locations are then estimated by substituting the coordinates of the prediction locations into the polynomial function (i.e. the predicted variable's values are approximated by the fitted trend surface).

In the framework outlined above, LP interpolation uses the variable's values at the measured locations within localised (and overlapping) windows rather than using the variable's values at all the measured locations. This situation implies that a set of 'local' trend surfaces are fitted to the measured variable's values and then patched together to

construct the final interpolation surface. The window can be moved around and the surface value at the centre of the window is estimated at each point.

Similarly to the IDW interpolation, LP interpolation does not take advantage of the spatial correlation structure of the data explicitly. Indeed, it only uses the geographical coordinates to predict the variable's values. However, LP is an inexact interpolation technique. When the input dataset exhibits short-range variation, LP interpolation can be a good method to capture the finer details (Akima 1970).

Ordinary kriging

OK is the most widely used version of Kriging. It is similar to the IDW interpolation technique, in that it uses a weighted linear combination of the measured variable's values surrounding the prediction location to estimate the variable's value at the prediction location. Because OK is a stochastic technique, however, two statistical data analysis steps have to be followed for generating an interpolation surface. These steps are: (1) quantifying the spatial correlation structure of the measured variable's values (also known as variography); and (2) estimating the variable's values at the prediction locations. During the first step, a spatial dependence model is fit to the variable's values at the measured locations. During the second step, the fitted model from variography, the spatial data configuration of the prediction locations and the measured variable's values surrounding the prediction locations are used to perform prediction. By following this procedure, OK provides a solution to the problem of estimation of the interpolation surface that takes into account the spatial correlation structure of the data.

In the OK technique, the predictions are based on the following model:

$$Z(s) = \mu + \varepsilon(s) \tag{4}$$

where $Z(s)$ is the variable's value at location s, μ is an unknown constant mean, and $\varepsilon(s)$ is the spatially correlated part of variation. The predictions are made according to Equation (1). Here, however, the weights are based not only on the distance between measured and prediction locations, but also on the overall spatial arrangement among the measured locations and their values.

OK presents several advantages over the IDW and LP interpolation techniques. Firstly, OK provides the best linear unbiased estimate as it attempts to optimise the weights assigned to the variable's values at the neighbouring measured locations. Secondly, the methodology also provides the Standard Error at the prediction locations and gives an indication of the reliability of the estimate. Furthermore, it does not produce edge effects resulting from trying to force a polynomial to fit the data.

Ordinary cokriging

OC is a stochastic technique that can be seen as an extension of OK to the case of more than one variable (Journel & Huijbregts 1978). Similarly to OK, the method computes the interpolated values of the variable of interest by optimising the weights assigned to the variable's values at the neighbouring measured locations based on the spatial correlation between the measured variable's values. However, OC also relies on the relationship between the variable of interest and other variables (secondary/auxiliary variables) and uses the information from other variables in an attempt to create a better prediction model.

In the case of two variables U and V that are spatially correlated, the OC predictions are made according to the following formula:

$$\hat{Z}_U(s_0) = \sum_{i=1}^{n_U} \lambda_{Ui} Z_U(s_i) + \sum_{j=1}^{n_V} \lambda_{Vj} Z_V(s_j) \tag{5}$$

where $\hat{Z}_U(s_0)$ is the estimate for U at the prediction location s_0, n_U and n_V are the number of measured values (used for the prediction) of the primary variable U, and of the secondary variable V, respectively, $Z_U(s_i)$ is the measured value of the primary variable at the location s_i, $Z_V(s_j)$ is the measured value of the secondary variable at the location s_j, and λ_{Ui} and λ_{Vj} are the associated weights. In OC, besides the experimental semivariograms for both U and V, information on the joint spatial co-variation (i.e. interaction) of both variables is taken into consideration as well.

OC presents the same advantages of OK over the IDW and LP interpolation techniques. OC, however, may improve the predictions and reduce the variance of the estimation error by drawing on the additional information from the other spatially correlated variables to help with interpolation.

Pipe burst event location procedure

Considering the event-occurrence probability value as the variable of interest and, in the case of OC only, the measured pressure as the secondary variable, the geostatistical techniques described in the previous sections are used in the *Location* subsystem of the further developed ERS for estimating the probability values of a burst event associated with the DMA pipes. This outcome is achieved according to the procedure outlined below.

Once an alarm is raised by using the customised ERS, the available values of the variable of interest estimated at the sensor locations (i.e. output of the various Signal level BISs) together with, in the case of OC only, the values of the secondary variable measured at the sensor locations are used to build a geostatistical interpolation surface. This interpolation surface, in turn, is used to estimate the probability values of a burst event at every network junction (node, i.e. prediction locations). Note that, to be consistent, each of these probability values has to be used only as a first estimate of the probability of a burst event occurring in the pipes connected at the particular network junction. Once this estimate is done, a probability value of a burst event occurring at a pipe is calculated for each DMA pipe by taking the average of the burst probabilities estimated at the two pipe end nodes. Note that, in the case of the two stochastic geostatistical techniques (i.e. OK and OC), an average standard error of interpolated junction values is also calculated for each pipe in order to assess the uncertainty of the predictions. Finally, the DMA pipes are grouped based on their burst probability values. Here, this grouping is carried out using the Jenks natural breaks classification method (Jenks 1967). This method allows determination of the best arrangement of the burst probability values into a user-defined number of classes by seeking to minimise each class's average deviation from the class mean, while maximising each class's deviation from the means of the other classes. In other words, the method seeks to reduce the variance within classes and maximise the variance between classes. As a result of the application of this procedure, the group of DMA pipes with the highest predicted burst probability values indicates the DMA area where the pipe burst event has most likely occurred.

The procedure outlined above is repeated at every time step after an alarm is raised (e.g. every 15 min). However, due to the dynamics of a pipe network, a burst event might affect the pressure/flow measurements from different sensors at different times. Thus, in order to obtain a more precise indication of the likely burst event location as time progresses, the ERS uses the Signal level BISs' output and, in the case of OC only, the measured secondary variable's values at the time steps following detection, in a cumulative fashion. Further details about this calculation can be found in Romano *et al.* (2011).

Test of prediction performance

The evaluation of the performance of the different geostatistical techniques being tested is based on the cross-validation technique (Devijver & Kittler 1982). Cross-validation removes each of the m sensor locations s_i, one at a time, and estimates the associated burst probability value $\hat{Z}(s_i)$ using the remaining sensor locations. The estimated and actual burst probability values $Z(s_i)$ are then compared and a summary statistics is computed. The summary statistics used in this study is the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[\hat{Z}(s_i) - Z(s_i)]^2} \tag{6}$$

The RMSE is the square root of the sum of the squared residuals. This statistics not only allows seeing how closely a resulting interpolation surface estimates the actual burst probability values, but may also be used to compare the performance of different geostatistical techniques. The smaller the RMSE value, the better is the performance.

CASE STUDY

The data analyses reported here aimed at testing and illustrating the capabilities of the new event detection and

location methodology. Furthermore, they aimed to compare the performances of the different geostatistical techniques considered for the approximate location of an event within a DMA.

The DMA being studied is predominantly rural. It has 17.8 km of pipes and a total of 925 customer connections. Its configuration is a combination of loops and branches. The data used in this study were recorded on 6 and 7 August 2008. They consisted of 15-min readings from 13 pressure sensors, which were deployed in the DMA being studied for the purpose of carrying out a series of EEs. The EEs considered here were carried out on 7 August 2008. Specifically, five hydrants, in different locations and at different times, were opened to create additional network flows to waste, thereby simulating the pipe burst events.

Although only pressure data were analysed in this study, the new detection and location methodology can be used for the analysis of flow data as well as or for the analysis of pressure and flow data simultaneously. However, as the presented methodology is based on the assumption that data from an increased number of sensors deployed in a DMA are available, the rationale for analysing pressure data only is as follows. It is envisaged that, in the near future, the deployment of an increased number of pressure sensors would be the water companies' preferred choice. This prediction is motivated by the fact that pressure sensors can be installed at lower costs than flow sensors and their calibration and maintenance requirements are also far less onerous. Additionally, by analysing the pressure data only, this study also aims to prove that, supported by a suitable data analysis methodology, pressure sensors can play an important role in the context of near real-time event detection and location – despite pressure data being considered less reliable than flow data (Mounce *et al.* 2010b).

In the case study presented here, in order to circumvent the lack of historical data from the 13 sensors considered (as they were *temporarily* installed for the purpose of carrying out the EEs only), several changes to the way the ERS presented in the methodology section normally works (see also Romano *et al.* 2012, 2013) had to be implemented. For example, the *BBA* and *TBA* subsystems had to be omitted as the data analyses in these subsystems have a requirement for several weeks of historical data. This omission was easily achieved due to the fact that ERS is fully modular and did not affect the validity of the results obtained. The use of the historical data from sensors *permanently* installed in the DMA could have only further improved the ERS performance.

In the light of the aforementioned limited data availability, the changes that had to be made to the way the ERS normally works resulted in the data analysis procedure described below. For each signal, the raw pressure data recorded during 6 August 2008 (day of data during which no EEs were carried out) were first checked for erroneous time stamps/missing values, repaired accordingly and then taken as representative of the signal's daily variations, assuming that no burst occurred in the DMA (i.e. NOP data set). Subsequently, noise from the NOP data set was removed, and the de-noised NOP data set was used for training and testing an ANN model for the one-step ahead prediction of future signal values.

The ANN model used here is based on a feed-forward multilayer perceptron ANN (Bishop 1995) with a hyperbolic tangent transfer function used for the neurons in the single hidden layer and a linear transfer function used for the neuron in the output layer. The ANN is trained using the back-propagation method (Rumelhart *et al.* 1986). With regard to the training and testing of this ANN model, note that the ANN parameters & input structure selection module (see Romano *et al.* 2013) was not used. Here, the ANN parameters and input structure employed for all the ANN models (i.e. relative to the 13 pressure signals) were the same and chosen in such a way that the resulting ANN prediction models were able to closely approximate the training sets whilst allowing good generalisation performance.

Once the ANN prediction model for the signal being analysed was available, the *DBA* subsystem was used for: (i) comparing the DMA signal values recorded during 7 August 2008 to their ANN predicted counterparts; (ii) identifying/estimating, at each time step (i.e. 15 min), significant (i.e. indicative of an event occurrence) discrepancies between those values; and (iii) further processing the identified discrepancies by using Control Rules (Shewhart 1931) to provide reliable evidence of an event occurrence. The resulting evidence was analysed next in the relevant Signal level BIS. At each 15-min time step, this Signal level BIS inferred the probability of an event occurring. Finally, once all the pressure signals were fully processed according to the procedure outlined here, the event probabilities from the

13 Signal level BISs were: (i) simultaneously analysed by the customised DMA level BIS (i.e. multivariate Gaussian mixtures-based) in order to raise the detection alarms; and (ii) simultaneously analysed in the *Location* subsystem in order determine the approximate location of the simulated burst event within the DMA. Note that the parameters of the DMA level BIS were learned using the EM algorithm and information (i.e. start time and duration) about the first two EEs carried out on 7 August 2008.

Table 2 reports the obtained ERS detection times for all the simulated pipe burst events and the corresponding hydrant opening and closing times. The actual hydrant flow rates and their corresponding values expressed as percentage of the average DMA inflow are also given. As can be seen from this table, all the simulated pipe burst events were identified at the best possible detection times (bearing in mind the 15-min sampling interval used). Furthermore, no false alarms were raised. Note, however, that because of the way in which the DMA level BIS parameters were learned, it was not unexpected that the first two EEs were correctly and timely detected.

In addition to detecting the simulated pipe burst events in a fast and reliable manner, the further developed ERS also allowed determination of the approximate location of the opened hydrants (i.e. location of the simulated pipe burst events).

Figure 3 shows the location results obtained using the four tested geostatistical techniques when the first pipe burst event simulated on 7 August 2008 was considered. The cumulative procedure mentioned in the approximate pipe burst event location section was applied. In this regard, note that the location results shown in Figure 3 refer to the third time step after the event was detected (i.e. 9:15 a.m.) and that, as an example, Figure 4 shows how the interpolation surface obtained using the OC technique changes with time following the event detection. Figure 3 is divided into four quadrants that refer to the results obtained using: (1) the OC (Figure 3(a)); (2) the OK (Figure 3(b)); (3) the LP (Figure 3(c)); and (4) the IDW (Figure 3(d)) geostatistical techniques. Each of the four quadrants shows the DMA pipes grouped using the Jenks natural breaks classification method and coloured according to their burst event probability value. The higher the value of the burst event probability for a pipe the more likely it is that this pipe is the 'failed' pipe. In each of the quadrants, the locations of the deployed pressure sensors are indicated using square symbols and the real location of the opened hydrant is indicated by a star symbol.

As it can be seen in Figure 3, when the OC or the OK techniques were used, the group of DMA pipes in the proximity of the 'failed' pipe was successfully identified. This situation was not the case when the two deterministic

Table 2 | Event recognition system alarm start/end times, hydrant opening/closing times, actual hydrant flow rates, and hydrant flow rates as percentage of the average district metered area inflow

	Alarm start time	Alarm end time	Hydrant opening/ closing time	Actual hydrant flow rate (l/s)	Hydrant flow rate as % of the average DMA inflow
First event	08:30	09:15	08:25	0.00	–
			08:26	5.00	51
			08:34	6.67	68
			09:29	0.00	–
Second event	09:45	10:30	09:33	0.00	–
			09:34	5.17	53
			09:35	5.75	58
			09:36	6.25	63
			10:24	18.83	191
			10:37	0.00	–
Third event	11:00	11:45	10:56	0.00	–
			10:57	5.00	51
			10:58	6.67	68
			11:56	18.33	186
			11:58	0.00	–
Fourth event	12:15	13:00	12:06	0.00	–
			12:07	6.17	63
			12:08	7.33	74
			13:05	13.00	132
			13:09	0.00	–
Fifth event	13:30	14:15	13:17	0.00	–
			13:18	5.00	51
			13:19	7.00	71
			13:30	7.50	76
			13:32	5.00	51
			14:18	5.00	51
			14:20	0.00	–

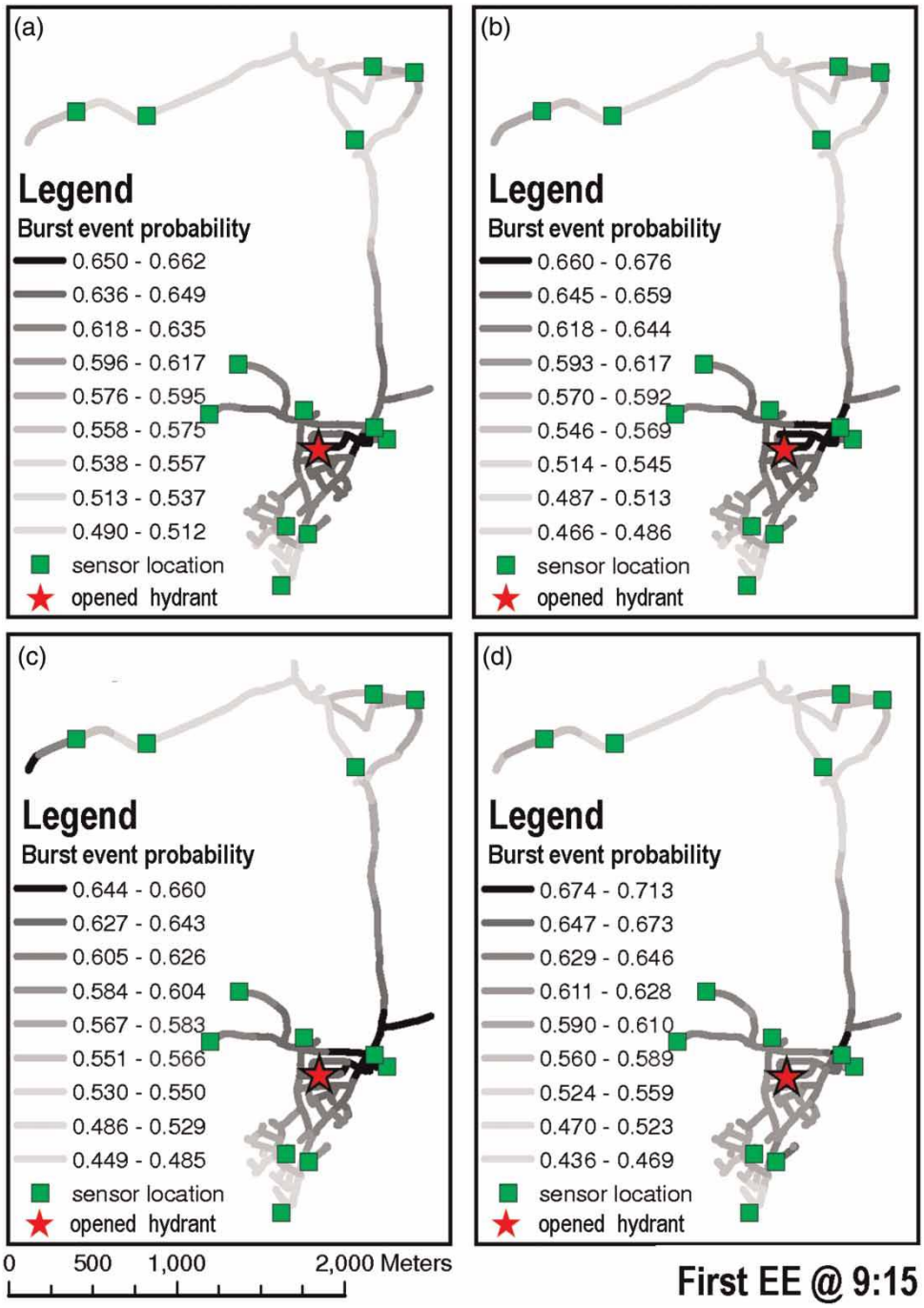

Figure 3 | The first engineered event location results using the ordinary cokriging (a), the ordinary kriging (b), the local polynomial (c), and the inverse distance weighted (d) geostatistical techniques.

geostatistical techniques (i.e. LP and IDW) were used. Furthermore, by comparing the results shown in the two top quadrants, it is possible to observe that the OC technique allowed a more precise indication of the likely location of the 'failed' pipe to be obtained. Indeed, the number of DMA pipes in the group of DMA pipes with the highest burst event probability values is reduced. Therefore the burst event search area is reduced as well. Note that similar

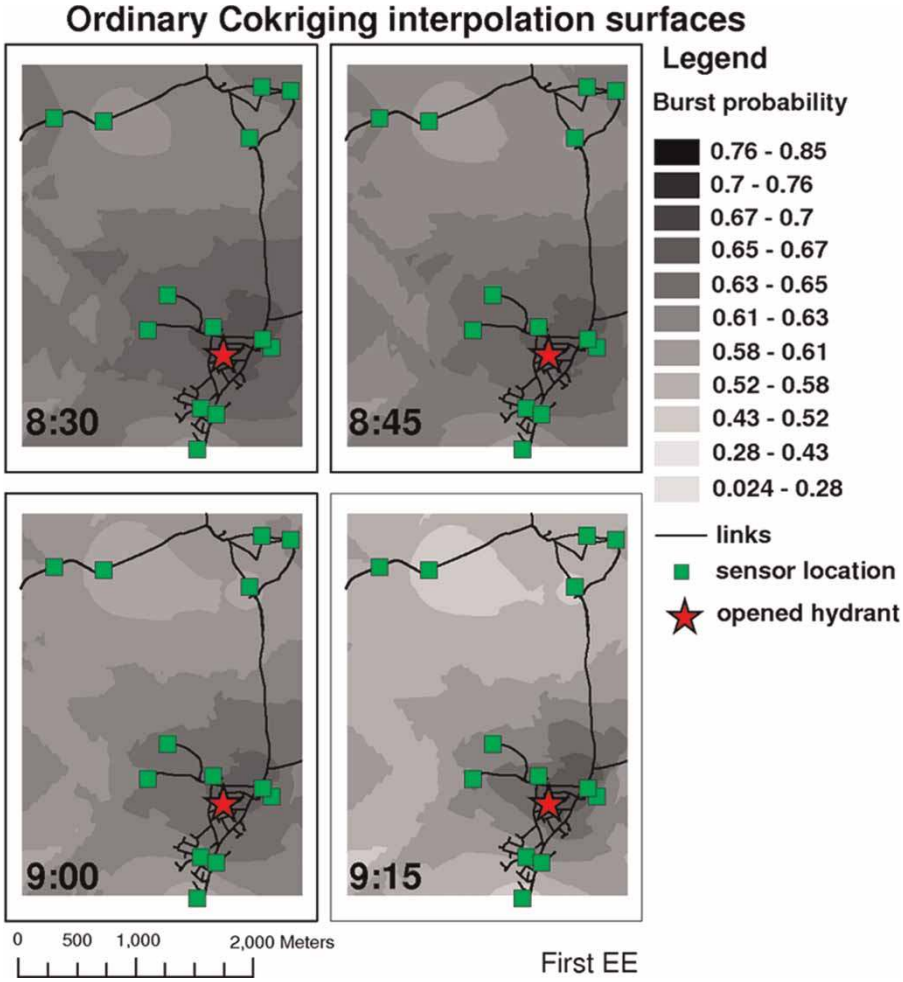

Figure 4 | Ordinary cokriging interpolation surface over four consecutive time steps starting from the time the first engineered event was detected.

results were obtained for the other four simulated burst events. However these results are not shown here due to space restrictions.

To support the above findings, Table 3 reports the values of the RMSE (i.e. performance indicator) calculated for each of the four tested geostatistical interpolation techniques and for all the simulated pipe burst events. This table clearly shows that the interpolation surfaces obtained using the OC technique outperform the interpolation surfaces obtained using the other geostatistical techniques. Thus, it confirms that better prediction models can be obtained by simultaneously considering the information from the correlated pipe burst probability and pressure variables.

It is important to note that the approximate pipe burst event location methodology effectiveness/accuracy depends on the number of pressure (and/or flow) devices deployed in the DMA – the more the better. Furthermore, it depends on the spatial layout of these devices within the DMA. To support these statements, the four scenarios analysed are shown

Table 3 | Root Mean Square Errors

	OC	OK	LP	IDW
First event	0.175	0.177	0.178	0.204
Second event	0.135	0.135	0.136	0.140
Third event	0.126	0.129	0.129	0.132
Fourth event	0.102	0.147	0.152	0.154
Fifth event	0.094	0.100	0.104	0.154

OC, Ordinary cokriging; OK, Ordinary kriging; LP, Local polynomial; IDW, Inverse distance weighted.

in Figure 5. In each of these scenarios a different number of sensors, which are arranged according to a specific spatial configuration, is considered. In all these scenarios, the OC geostatistical technique was used to determine the approximate location of the first pipe burst event simulated on 7 August 2008.

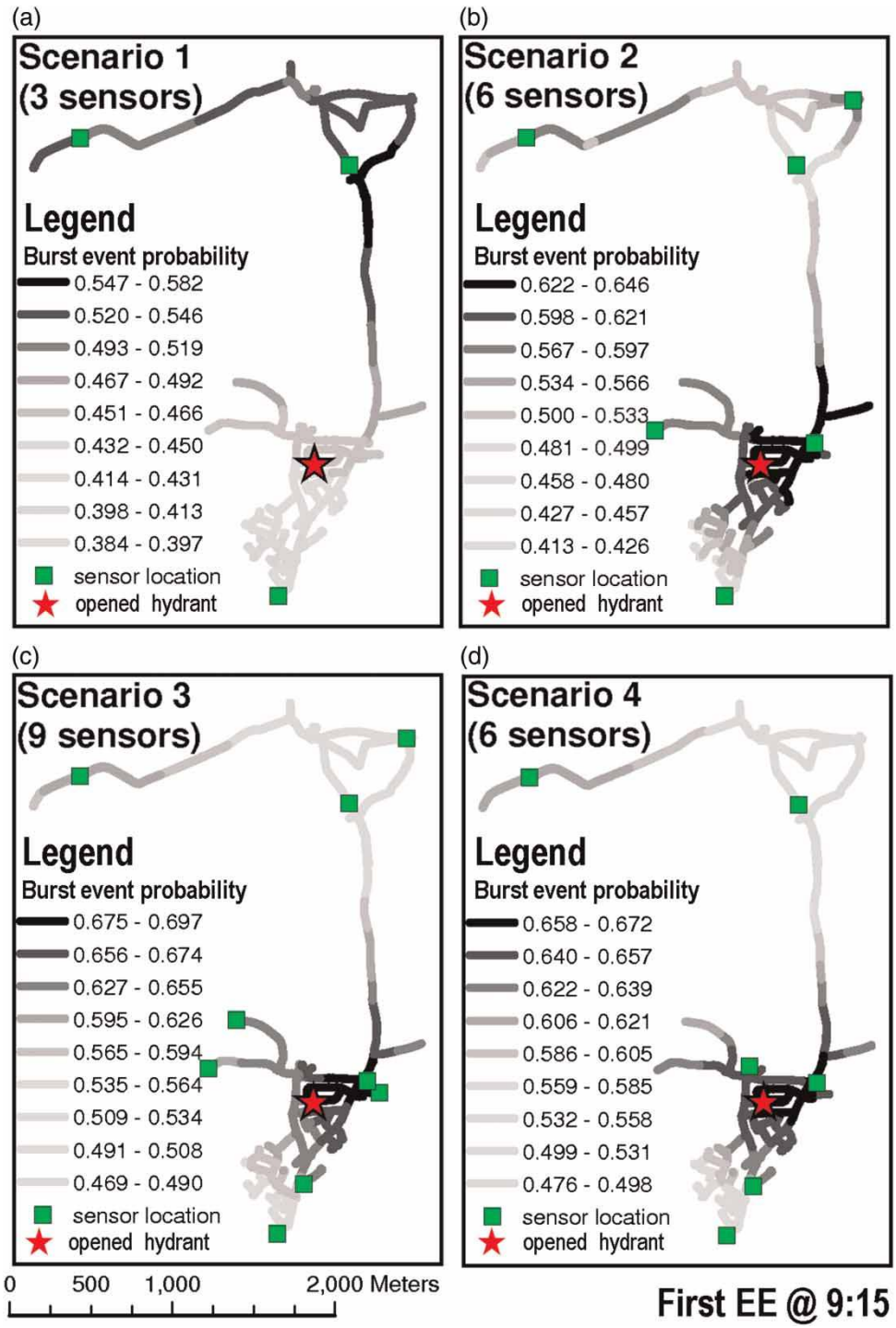

Figure 5 | The first engineered event location results using the ordinary cokriging geostatistical technique for analysing the pressure data from three sensors (a), six sensors located as in scenario 2 (b), nine sensors (c), and six sensors located as in scenario 4 (d).

By observing Figure 5, it can be seen that in scenario 1 the ERS was not able to identify the group of DMA pipes in the proximity of the 'failed' pipe. In the remaining three scenarios, on the other hand, the group of DMA pipes in the proximity of the 'failed' pipe was identified successfully. However, in scenario 3 (i.e. nine sensors arranged as shown in the bottom-left quadrant of Figure 5) the ERS provided a more precise indication of the likely location of the 'failed' pipe than the one provided in either scenario 2 (i.e. six sensors arranged as shown in the top-right quadrant of Figure 5) or scenario 4 (i.e. six sensors arranged as shown in the bottom-right quadrant of Figure 5).

Furthermore, by comparing scenarios 2 and 4, it is possible to observe that, although the same number of sensors was used, the spatial arrangement of the deployed sensors considered in scenario 4 enabled the ERS to indicate the likely location of the 'failed' pipe more precisely. In view of this observation, it is worth highlighting that the use of the approximate pipe burst event location methodology presented in this paper could/should be supported by the development and use of a methodology for optimising the number and spatial arrangement of the sensors to be employed in order to achieve a required degree of location accuracy, no matter where in the DMA the pipe burst event occurs. The development of such a methodology is an active area of research (e.g. Farley *et al.* 2010, 2012). Having said this, note that this topic is beyond the scope of the work presented here and hence will not be discussed in greater detail.

Finally, as mentioned in the approximate pipe burst event location section, a further advantage yielded by the use of the stochastic geostatistical techniques over the deterministic ones is the possibility of assessing the 'quality' of the predictions. Figure 6 shows, as an example, a map of the OC Standard Error for the first EE. Similarly for each of the quadrants in Figure 3, in this figure the locations of the deployed pressure sensors are indicated by using square symbols, the real location of the opened hydrant is indicated by a star symbol, and the Jenks natural breaks classification method is used for grouping the DMA pipes. Here, however, the DMA pipes are grouped and coloured according to their Standard Error value. The lower the value of the Standard Error for a pipe, the higher the confidence that the predicted pipe burst probability value for that pipe is close to its 'real' value (i.e. good estimate). From this figure, it is possible to observe that the predicted pipe burst probability values are more accurate the closer they are to the sensor locations. Note that a similar qualitative behaviour was observed for the other EEs and when the OK technique was used.

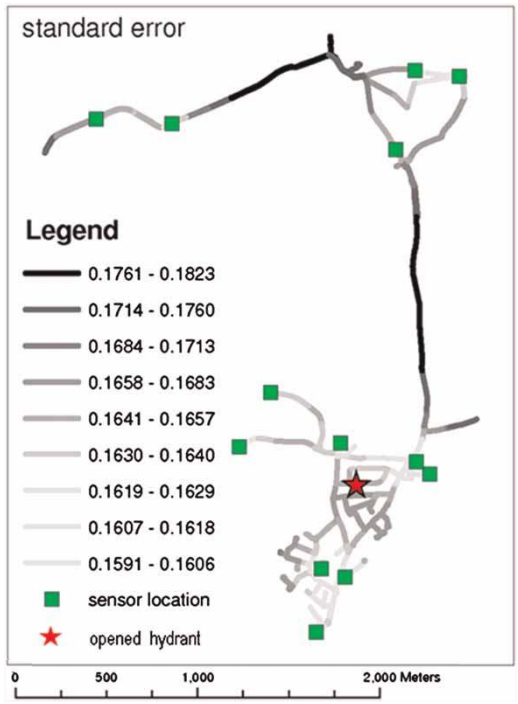

Figure 6 | Ordinary cokriging standard error map for the first engineered event.

CONCLUSIONS

The wider availability of more accurate and cheaper pressure/flow sensors allows us to envisage that an increased number of these devices (particularly pressure sensors) will be deployed in the UK DMAs in the near future. In this paper, a new event detection and location methodology, which effectively exploits the data from a larger (than currently used in the UK practice) number of deployed sensors, has been presented. In particular, the customisation and further development of the ERS recently developed by the authors have been described. The ERS's customisation involved using a multivariate Gaussian mixtures-based DMA level BIS to deal more efficiently with the increased

data availability. The ERS's further development involved using geostatistical techniques for processing in near real time the output information from the Signal level BISs and building a model to predict the probability value of a burst associated with each DMA pipe. As an ensemble, they enable the new ERS to not only detect pipe burst events but also determine their approximate location within a DMA.

The detection and location capabilities of the presented methodology have been tested here by applying the customised and further developed ERS to the analysis of pressure data recorded during a series of EEs in a UK DMA. The use of different geostatistical interpolation techniques has also been investigated. The main findings from these tests are briefly summarised below.

Using the pressure measurements only, the new methodology: (i) detected all the EEs in a reliable (i.e. without false alarms) and timely manner (all the EEs were detected at the best possible time); and (ii) successfully approximately located the simulated pipe burst events at different location within the studied DMA. These results not only demonstrate the capabilities of the new methodology but also provide evidence that the pressure data can play an important role in the context of near real-time event detection and location in WDS.

The OC technique outperformed the other tested geostatistical interpolation techniques. Even though the use of the OK technique already allowed successful identification of the approximate locations of all the EEs, the use of the OC technique allowed more accurate identification (i.e. reducing the burst event search area).

It can be concluded that the customisation and further development of the ERS extends the capabilities of the proactive and fully automated methodology recently developed by the authors. The effective, reliable and timely detection of pipe burst events together with the successful identification of their approximate location, which could be achieved using the proposed methodology, can facilitate prompt interventions and repairs. This situation, in turn, may reduce the potential damage to the infrastructures and to third parties and improve the water company's operational performance and customer service, thereby yielding substantial improvements to the state-of-the-art in near real-time WDS incident management.

ACKNOWLEDGEMENTS

This work is part of the first author's PhD sponsored by the University of Exeter. The DMA data used in the paper have been collected as part of the Neptune project funded by the UK Engineering and Physical Sciences Research Council (EP/E003192/1) and provided by Mr Ridwan Patel from Yorkshire Water, which is gratefully acknowledged. The role of the University of Sheffield in field trials is also acknowledged. The work presented in this paper has been patented (Publication No. WO/2010/131001).

REFERENCES

Akima, H. 1970 A new method of interpolation and smooth curve fitting based on local procedures. *Journal of Association for Computing Machinery* **17**, 589–602.

Banerjee, S., Carlin, C. P. & Gelfand, A. E. 2004 *Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, USA.

Bishop, C. M. 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

Cleveland, W. S. & Devlin, S. J. 1988 Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

Colombo, A. F. & Karney, B. W. 2002 Energy and costs of leaky pipes: toward comprehensive picture. *Journal of Water Resource Planning and Management* **128**, 441–450.

Cressie, N. 1993 *Statistics for Spatial Data*. John Wiley & Sons, New York.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38. (Available from http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z)

Deutsch, C. V. & Journel, A. G. 1998 *GSLIB: Geostatistical Software and User's Guide*, 2nd edn. Oxford University Press, New York.

Devijver, P. A. & Kittler, J. 1982 *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Englewood Cliffs, NJ.

Duda, R. & Hart, P. 1973 *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Edwards, D. 2000 *Introduction to Graphical Modelling*, 2nd edn. Springer-Verlag, New York.

Farley, B., Mounce, S. R. & Boxall, J. B. 2010 Field testing of an optimal sensor placement methodology in an urban water distribution network for event detection. *Urban Water Journal* **7**, 345–356.

Farley, B., Mounce, S. R. & Boxall, J. B. 2012 Development and field validation of a burst localisation methodology. *Journal of Water Resources Planning and Management*.

Fenner, R. A. & Ye, G. 2011 Kalman filtering of hydraulic measurements for burst detection in water distribution systems. *Journal of Pipeline Systems Engineering and Practice* **2**, 14–22.

Gandin, L. S. 1963 *Objective Analysis of Meteorological Fields*. GIMIZ, Leningrad, Russia. English translation from the Russian, Israel Program for Scientific Translation, 1965, Jerusalem, Israel.

Isaaks, E. H. & Srivastava, R. M. 1989 *Introduction to Applied Geostatistics*. Oxford University Press, New York.

Jenks, G. F. 1967 The data model concept in statistical mapping. *International Yearbook of Cartography* **7**, 186–190.

Jensen, F. V. 2001 *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York.

Jensen, F. V. 2009 Bayesian networks. *Wiley Interdisciplinary Reviews: Computational Statistics* **1**, 307–315.

Journel, A. G. & Huijbregts, C. J. 1978 *Mining Geostatistics*. Academic Press Inc., London.

Kapelan, Z., Savić, D. A. & Walters, G. A. 2003 A hybrid inverse transient model for leakage detection and roughness calibration in pipe networks. *Journal of Hydraulic Research* **41**, 481–492.

Krige, D. G. 1951 A statistical approach to some mine valuations problems at the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**, 119–139.

Lam, N. S. 1983 Spatial interpolation method: a review. *The American Cartographer* **10**, 129–135.

Lauritzen, S. L. 1995 The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* **19**, 191–201.

Li, J. & Heap, A. D. 2008 *A Review of Spatial Interpolation Methods for Environmental Scientists*. Record 2008/23, Geoscience Australia, Canberra, Australia.

Liggett, J. A. & Chen, L.-C. 1994 Inverse transient analysis in pipe networks. *Journal of Hydraulic Engineering* **120**, 934–955.

Misiunas, D., Lambert, M. F., Simpson, A. R. & Olsson, G. 2005 Burst detection and location in water distribution networks. *Water Science and Technology: Water Supply* **5** (3–4), 71–80.

Mounce, S. R., Boxall, J. B. & Machell, J. 2010a Development and verification of an online artificial intelligence system for burst detection in water distribution systems. *Journal of Water Resources Planning and Management* **136**, 309–318.

Mounce, S. R., Farley, B., Mounce, R. B. & Boxall, J. B. 2010b Field testing of optimal sensor placement and data analysis methodologies for burst detection and location in an urban water network. *Proceedings of the Ninth International Conference on Hydroinformatics*, Tianjin, China.

Mounce, S. R., Mounce, R. B. & Boxall, J. B. 2011 Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics* **13**, 672–686.

Myers, D. E. 1982 Matrix formulation of co-kriging. *Mathematical Geology* **14**, 249–257.

Palau, C. V., Arregui, F. J. & Carlos, M. 2012 Burst detection in water networks using principal component analysis. *Journal of Water Resources Planning and Management* **138**, 47–54.

Pudar, R. S. & Liggett, J. A. 1992 Leaks in pipe networks. *Journal of Hydraulic Engineering* **118**, 1031–1046.

Puust, R., Kapelan, Z., Savić, D. A. & Koppel, T. 2010 A review of methods for leakage management in pipe networks. *Urban Water Journal* **7**, 25–45.

Redner, R. & Walker, H. 1984 Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Romano, M., Kapelan, Z. & Savić, D. A. 2011 Burst detection and location in water distribution systems. Proceedings of the World Environmental and Water Resources Congress, Palm Springs, USA.

Romano, M., Kapelan, Z. & Savić, D. A. 2012 Automated detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management*.

Romano, M., Kapelan, Z. & Savić, D. A. 2013 Evolutionary algorithm and expectation maximisation strategies for improved detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management*.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. 1986 *Learning Internal Representations by Error Propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, pp. 318–362.

Sadiq, R., Kleiner, Y. & Rajani, B. 2006 Estimating risk of contaminant intrusion in distribution networks using Dempster-Shafer theory of evidence. *Civil Engineering and Environmental Systems* **23**, 129–141.

Shepard, D. 1968 A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the Twenty-third ACM National Conference, New York.

Shewhart, W. 1931 *Economic Control of Quality of Manufactured Product*. Van Nostrand Reinhold, New York.

Srirangarajan, S., Allen, M., Preis, A., Iqbal, M., Lim, H. B. & Whittle, A. J. 2012 Wavelet-based burst event detection and localization in water distribution systems. *Journal of Signal Processing Systems*.

Webster, R. & Oliver, M. 2001 *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.

Wu, Z. Y., Sage, P. & Turtle, D. 2010 Pressure-dependent leak detection model and its application to a district water system. *Journal of Water Resources Planning and Management* **136**, 116–128.

First received 19 May 2012; accepted in revised form 16 November 2012. Available online 2 January 2013