This is a repository copy of *Water quality event detection and customer complaint clustering analysis in distribution systems*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/83989/

Version: Submitted Version

Article:
Mounce, S.R., Machel, J.M. and Boxall, J.B. (2012) Water quality event detection and customer complaint clustering analysis in distribution systems. Water Science and Technology: Water Supply, 12 (5). 580 - 587. ISSN 1606-9749

https://doi.org/10.2166/ws.2012.030

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Water quality event detection and customer complaint clustering analysis in distribution systems

Stephen Mounce[1], John Machell[1] and Joby Boxall[1]

[1]Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, S1 3JD, UK. (Email: *s.r.mounce@sheffield.ac.uk, j.machell@sheffield.ac.uk, j.b.boxall@sheffield.ac.uk*)

Abstract
Safe, clean drinking water is a foundation of society and water quality monitoring can contribute to ensuring this. A case study application of CANARY to historic data from a UK drinking water distribution system is described. Sensitivity studies explored appropriate choice of algorithmic parameter settings for a baseline site, performance was evaluated with artificial events and the system then transferred to all sites. Results are presented for analysis of 9 water quality sensors measuring six parameters and deployed in three connected District Meter Areas, fed from a single water source (service reservoir), for a one year period and evaluated using comprehensive water utility records with 86% of event clusters successfully correlated to causes (spatially limited to DMA level). False negatives, defined by temporal clusters of water quality complaints in the pilot area not corresponding to detections, were only approximately 25%. It was demonstrated that the software could be configured and applied retrospectively (with potential for future near real time application) to detect various water quality event types (with a wider remit than contamination alone) for further interpretation.

Keywords
water distribution networks; water quality; online monitoring; event detection; data analysis

INTRODUCTION

Customers regard a reliable supply of water as one of the most important aspects of the water supply service. Although treatment works are closely monitored and controlled in the developed world, this is not the case for Water Distribution Systems (WDS) which carry a much higher degree of uncertainty. High quality water leaving treatment facilities generally deteriorates as it travels through extensive, often convoluted, distribution networks, via a number of mechanisms associated with distribution network materials, hydraulic conditions, chemical and biological reactions, or ingress of polluting materials. Detection of water quality events before customers are affected is paramount to prevent possible public health impacts and even regulatory action, ranging from fines, to loss of operating license if the breach is of sufficient gravity.

There are three types of contamination event which can threaten water quality in WDS: natural, accidental and deliberate. Security concerns regarding intentional contamination of water supplies have been heightened following the 9/11 terrorist attack on the World Trade Center, however natural disasters and accidental contamination can be just as damaging. Minor local incidents are also of (possibly greater) concern, and are most definitely more uncertain - for example pipeline deterioration leading to intrusion of pathogens and contaminants. Figure 1 illustrates the trade-off between impact and likelihood when considering events. Minor events are more likely, but mainly go undocumented. The most serious events have very significant impact but occur rarely. It is events in between these which are fairly likely but also have a significant level of impact (denoted as ‘outbreaks’) that are of particular interest.

A variety of water quality sampling is conducted both to meet regulatory requirements and to inform decisions about operations. The traditional sampling method has been the collection of discrete spot samples followed by laboratory based analysis. Various products are now available for

application in WDS to continually measure parameters such as temperature, turbidity, colour, conductivity, pH, DO, free and total chlorine (Aisopou et al. 2012).

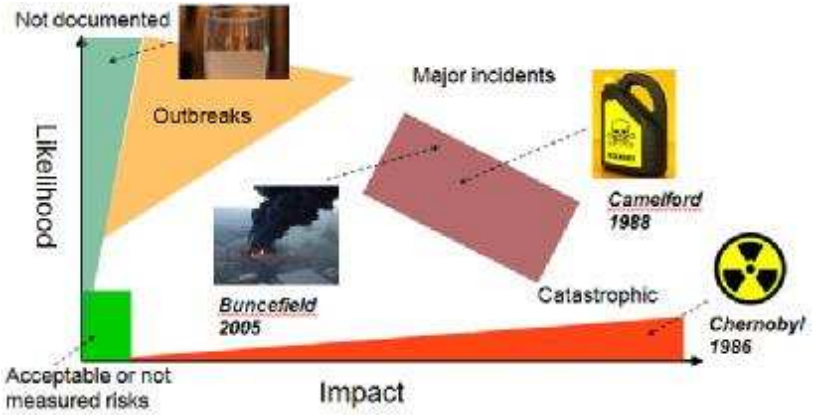

Figure 1. Impact vs. likelihood for WDS incidents (adapted from Block 2010)

Problem definition

The need to efficiently manage WDS has highlighted the need to develop asset management tools to assist operators to evaluate the condition of the water distribution area, potential risk of failure, visualize those areas of high risk and propose repair strategies and prioritize work based on impact and cost. Automated software can provide an intelligent communication interface between the monitoring stations and central control. An Event Detection System (EDS) is an automated system for analysing data collected from online monitors and alerting operators to unusual conditions based on anomalous readings taken by sensors relative to background normal data. EDSs analysing hydraulic data have been explored by water utilities, for example for burst detection (e.g. Mounce et al. 2010). A project called NEPTUNE developed a Decision Support System (DSS) pilot for hydraulic network management (Morley et al. 2009) incorporating online data analysis of sensor signals to generate alerts (Mounce and Boxall 2010). The next generation of DSS will benefit from water quality and other sensor types, as well as decentralised intelligence within sensors.

Research has shown that many contaminants of possible concern will cause detectable changes in measurable water quality parameters and hence they can act as indicator (surrogate) measurements. Hall et al. (2007) tested the response of several commercially available water quality sensors in the presence of nine different contaminants introduced to a pipe loop in an experimental facility at different concentrations and found that at least one of the surrogate parameters changed in response to the presence of every contaminant. Although security concerns have been a key motivation, water utilities are also seeking additional benefits through improved operational information for other event types (including treatment failures, service reservoir issues, bursts, and sensor malfunctions).

This paper describes a case study application of the CANARY software to historic data from a UK distribution system. The emphasis has been on using the EDS in a wider context than contamination detection alone and on full event evaluation using water company information sources. The research presented here is part of the UK multidisciplinary "Pipe Dreams" project (http://www.sheffield.ac.uk/pipedreams).

CANARY

Contamination warning systems (CWSs) have been proposed as a promising approach for reducing the risks associated with contamination of drinking water by early detection and management

(AWWA 2005). A CWS can include various approaches to monitoring including water quality sensor deployment in the distribution system, spot sampling and laboratory analysis and customer complaint (contact) monitoring programmes (it could be argued that customers are in one sense the best sensors but this is no longer acceptable). The U.S. Environmental Protection Agency (EPA) has been deploying and evaluating CWSs at a series of drinking water utilities since 2006 (Skadsen et al. 2008). Outputs of this have been tools and strategies for applying EDS as part of a CWS and one of these is CANARY (Hart et al. 2009).

CANARY is an open source software platforms for EDSs which can be used online to monitor and analyse data from available water quality sensors. CANARY can read in Supervisory Control and Data Acquisition (SCADA) data (water quality signals and possibly operations data), perform an analysis in near real-time and then return the evaluated probability of a water quality event occurring at the current time step. CANARY uses statistical and mathematical algorithms to identify the onset of periods of anomalous water quality data, while at the same time limiting the number of false alarms that occur. The software does not seek to provide an indication of the cause of the anomaly. A two step process is adopted: state estimation for future water quality value prediction (using established traditional time series and multivariate statistical processes by default but which could alternatively be an Artificial Neural Network) and a second stage of residual classification for determination of expected or anomalous value (an outlier). It implements several change detection algorithms, one of these being a multivariate nearest-neighbour (MVNN) algorithm (Klise and McKenna, 2006). A binomial event discriminator (BED) examines multiple outliers within a prescribed time window to determine the onset of either an anomalous event or a water quality baseline change as determined by an Event Time Out (ETO) configuration parameter (McKenna et al., 2007). An alternative approach would be to consider using, for example, a Fuzzy Inference System for classification.

CANARY can operate in both on-line (operations) and offline (historic analysis, but simulated online by processing in time series order one step at a time) modes, and it is the latter which is utilised for the study presented in this paper. Each monitoring station is analysed independently using CANARY. The values of the configuration parameters for each station might vary from one utility to the next and could vary across monitoring stations within a utility (EPA 2010).

CASE STUDY

In this section, one sensor is selected for use in determining appropriate algorithmic parameter values through sensitivity studies. The performance is then assessed via the addition of artificial events to the baseline file. Finally, the system is transferred to all available sensors to perform a full evaluation using additional information sources. This is based on the assumption that for the same make of sensor, measuring the same parameters in a relatively close geographical proximity, with a single water source, that the use of one set of parameters derived by sensitivity analysis should be reasonable. MVNN was selected for this study as EPA (2010) concluded from extensive testing that the difference between it and other statistical algorithms is minimal.

Overview

Solomat water quality Sondes (a self contained, submersible, multi-parameter measuring instrument) incorporating a Censar chip were deployed by the water utility for a pilot study. These were situated in an urban distribution system (in total comprising approximately 110 km of pipework) in three connected urban District Meter Areas (6200 properties approximately), fed from a single water source (service reservoir). The water mains vary in size from 50 mm to 400 mm and span a range of ages from circa 1900 to new mains. The majority of the mains infrastructure is

comprised of iron pipes (approximately 70%). The rest is a mixture of plastic including PVC and MDPE and a small number of steel pipes. Parameters measured were water temperature, pH, dissolved oxygen, conductivity, turbidity and pressure (the latest instrumentation now available has more comprehensive measurement possibilities). Historic data was collated from multiple files and formatted into CANARY format. Analysis was restricted to those instruments without any major gaps in data (discarding those with less than 20% data availability). Review of viable data sets resulted in 9 monitoring stations and an approximate one calendar year of logging at a five minute resolution (80% average data availability for any measurement at that time step). Information from the work management system (e.g. from mains repair records), from relevant customer contacts and DMA inlet flows was also obtained for cross correlation / detection confirmation.

Parameter sensitivity studies for configuration

A sensitivity study was conducted exploring the MVNN prediction algorithm (with BED used to provide event probability over a time window) investigating the key parameters (EPA 2010) of window length used for prediction (MVNN) and outlier threshold (measured in in units of standard deviation - sigma). Ranges selected were based on EPA findings and a complete enumeration was conducted for pairs of values. The parameters were otherwise:
a) ETO=288 (5min) time steps (1 day) b) BED window=24 (5min) time steps (2 hrs) c) Probability of outlier 0.9 (CANARY default) d) BED event probability threshold =0.5 (CANARY default).

The ETO parameter determines the point at which a baseline change can be concluded to have occurred due to continuous alarming over this period (because the early stages of a baseline change are the same as an event) and one day was judged a reasonable value. The size of the BED window was defined at 2 hours since the study was not focussed on short duration anomalies and instead significant events. Integrating results over greater numbers of time steps prior to increasing the probability of event detection generally results in fewer false positive detections, but at the expense of faster detection time, so for 5 minute sample data this value could be lower for an online system albeit with an expectation of a higher number of false positives. Figure 2a shows results for station 71003 for 380 days for a varying MVNN window size (from 0.5 to 2 days) and the threshold (0.5 to 2 sigma). Event clusters are defined by CANARY to be distinct contiguous sequences of event classifications bounded on either end by periods of normal background water quality data. It is evident that the lower the threshold value the greater the number of outliers that will be identified and hence the greater the number of event clusters. The majority of event clusters in Figure 2a are considered as false positives for the purposes of this benchmarking (several known events are actually present as verified in later evaluation). Figure 2b shows how average residual changes with MVNN window size indicating that the residual is minimised with a one day window.

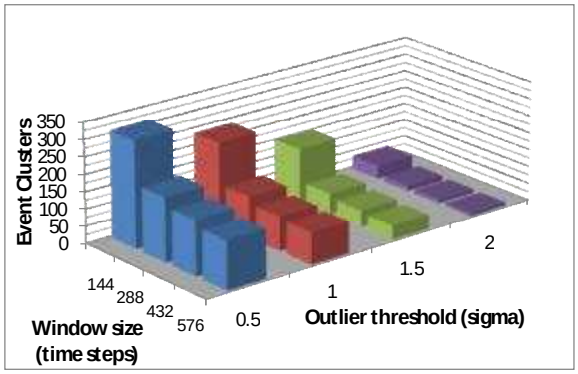

Figure 2a. Station 71003 sensitivity analysis of window size and threshold for MVNN (event clusters)

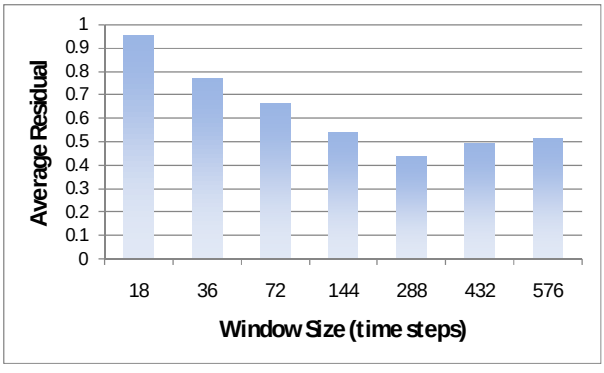

Figure 2b. Station 71003 varying window size with threshold one sigma for MVNN (average residual over parameters)

The EDS results were examined qualitatively to determine the best threshold value that resulted in event detection on obvious significant changes in water quality and which minimised events and outliers throughout the rest of the data set. A MVNN window length of one day, 288 time steps, and threshold value of two standard deviations was used within CANARY (resulting in 15 event clusters, see figure 2a). This is in agreement with the EPA who previously found that, across different types of monitoring station, a window size of between one and two days is enough to provide reasonably accurate and useful predictions of future water quality values.

Performance evaluation with artificial events

Contaminants may affect water quality signals, but their signals are difficult to distinguish from the background noise. Data quality over a one year period in a live system may be patchy and prone to hardware errors. Reliable and accurate reporting cannot be taken for granted, and complete operational information is often not available. To evaluate the performance of an EDS with respect to false negatives, it is necessary to have a water quality data set that contains actual events. Hence, simulated events with known lifetime were added to logged data in order to quantify performance. Synthetic patterns were added to baseline normal data to assess and benchmark performance. The simulated events change the water quality by adding a deviation to the background. A similar technique was adopted to that used in EPA (2010):

$$Z_E(t) = Z_0(t) + E_{ind}(t) \cdot e \cdot E_{\max} \cdot \sigma_z \qquad (1)$$

where $Z_E(t)$ is the event modified water quality value at time *t*, $Z_0(t)$ is the original background water quality at that time step, E_{ind} is an event indicator between zero and one during an event or zero otherwise, *e* defines a decrease/increase in the parameter in response to some event and $E_{\max}$ is a coefficient applied to σ_z (standard deviation of the data stream). Simulated events represented some (theoretical) changes in parameters over variable time periods in response to possible events. Two synthetic patterns were used: square wave and sinusoidal wave (which determines E_{ind}). The simulated events were added to the baseline data, into periods of stable measurements, and CANARY applied as previously. The overall performance was evaluated as summarized in Table 1. The penultimate column describes the percentage of the actual known event time periods (real or simulated) that are classified as an event. The final column provides the average delay from the start of the event to the alarm state due to the BED classification window.

Table 1. CANARY results on training and testing data (before/after addition of events)

Data set	Event clusters (ETO)	Average event cluster length	Proportion of true events with an event cluster	Proportion of total time of overlap (true vs. estimated)	Average delay in detection (2h BED)
Station A base data (1 known real event)	15 (6)	799.7 mins	100% (1/1)	76%	2h 40 mins
Station A base data with 9 added artificial events	24 (6)	662.7 mins	100% (10/10)	72%	1h 34 mins

Results of transferability to full evaluation

The parameters derived in the sensitivity studies for the MVNN with BED algorithm were applied to analysis of all available sensors. It was decided to increase the MVNN window size to two days

(576) since this further reduced the event clusters for the base data from 15 to 9 (removing 6 detections with (drifting) temperature as the principle contributing signal) – see figure 2a. Consequently, parameters were selected based on the intention to analyse a relatively long period of historical data and detect reasonably dominant events which could be correlated with network information, while reducing false positives. In practical operational use, it is likely that these parameters would be set more sensitively to bias the trade-off towards faster detection times. A pilot conducted in the U.S. city of Cincinatti with 15 online sensors determined that 50 alarms per month were manageable for the water utility at this scale (Allgeier 2010).

Sources of information for the evaluation when correlating with events included the following:

- Customer Contact (CC) database (customer reports of discolouration, milky water or taste/odour problems)
- Work Management System (WMS) record of main repairs database
- DMA inlet flows where available for confirmation of bursts

Figure 3 shows an example detection of a known burst with principal contributing signals being a pressure decrease and turbidity increase. CANARY provides details of the major contributing signals (parameters) associated with each event cluster. It was anticipated that turbidity and pressure would be strongly contributing in the case of abnormal flow based events (such as bursts or repairs), and turbidity and possibly pH in the case of discolouration or taste and odour issues. Actual responses are configuration dependent as it is known that system hydraulics are significant in determining the magnitude of potential discolouration (Husband and Boxall 2011).

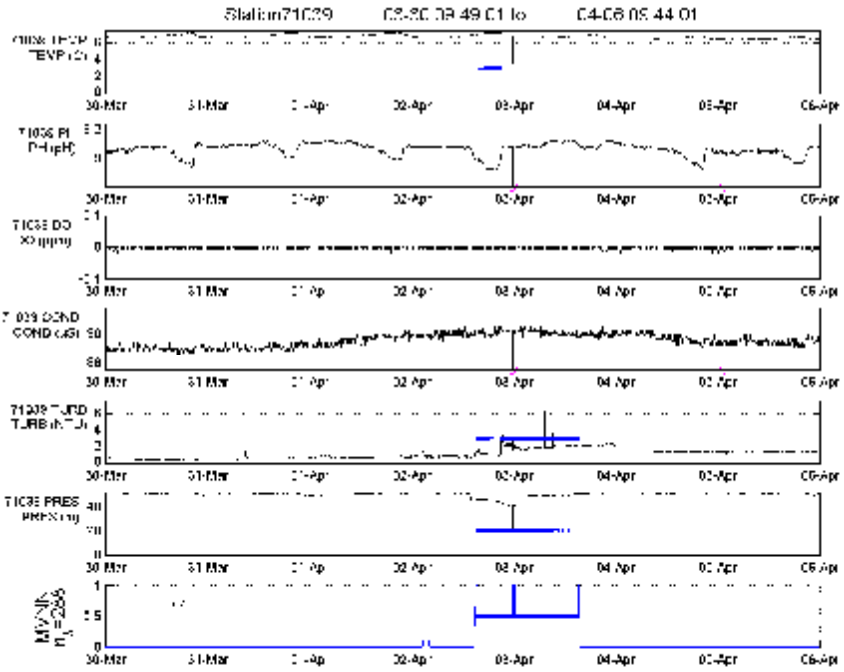

Figure 3: One week sensor 71039 data with known burst

Event clusters for all nine sensors were evaluated using the additional information sources. Correlating these events to causes was achieved spatially (at DMA level only) and temporally (with at most one day proximity) with the DMA flow and pressure data providing further corroboration of cause timings. The system successfully detected real world incidents over the one year period including correlation with pipe bursts and repairs, abnormal flows, customer reports of

discolouration and sensor failure. A total of 139 event clusters were evaluated and classified using the following scheme:

- Burst/repair or flow – correlates to burst main repair or large flow, sometimes with CCs
- CCs (multiple) – correlates to more than one customer report of discolouration, milky water or taste/odour
- Unknown event – a significant change across multiple parameters determined by visual inspection of time series but with no correlation by available information sources
- Sensor failure – sensor problems and data corruption
- Ghosts – no obvious defined change determined by visual inspection of time series, but can include drift and minor deviation. A WDS is a complex non-linear reactor subject to a large degree of uncertainty and additionally data quality can be limited for operational records that are generally available for this type of evaluation.

Figure 4a provides a summary of classifications. Figure 4b shows which parameters were contributing factors to detections for particular categories; for example pressure and turbidity were predominant for burst/flow hydraulic events. Although figure 4a shows a good explanation of causes of event clusters, with only 14% unexplained, it does not provide any information on false negatives i.e. missed significant events. For evaluating this, customer contacts of discoloured water, milky water and taste/odour were judged to be the best indication of significant incidents. One call or email from a customer regarding an issue is not usually indicative of a major problem and quite possibly associated with domestic plumbing; instead (temporal) clustering of close contacts are more likely to be representative of noteworthy issues. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) a density-based data clustering algorithm was utilised to cluster contacts temporally (Ester et al., 1996). The algorithm requires two parameters ε (neighbourhood distance) and the minimum number of points to form a cluster *minPts*.

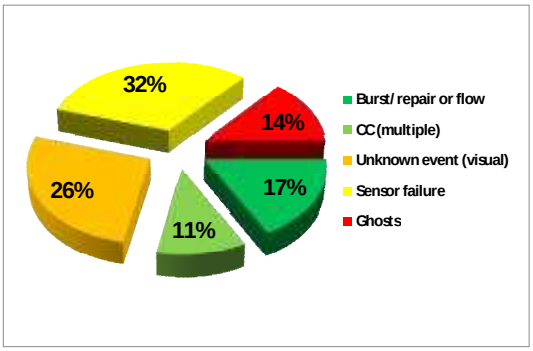

Figure 4a. Event cluster evaluation summary for one year period

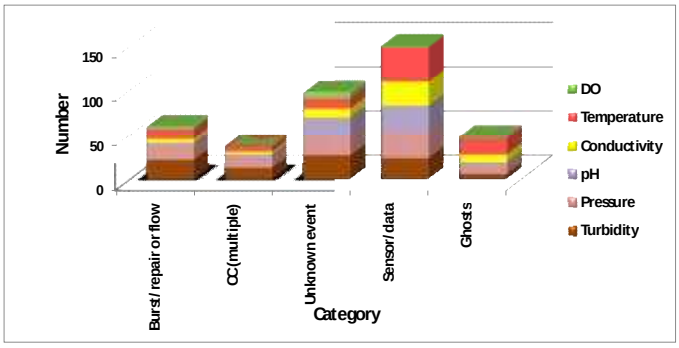

Figure 4b. Category of event with corresponding contributing factors to CANARY detection

When clustering the contacts temporally, it was opted to do this as one group across the three DMAs. The DMAs were interconnected and served by one water source and studies have shown that over 40% of CCs are explained by clusters affecting multiple DMAs (Husband et al., 2010). The total individual customer contacts in the three DMAs for the one year period was 223. DBSCAN was applied in MATLAB on normalised date stamps with varying algorithmic parameter values (ε = 1-7 days and *minPts*= 2-7 contacts) and detections then correlated, where possible, to the CC clusters (detection at one sensor at some time in period of cluster) as summarised in Figure 5. Over the range of DBSCAN parameters, approximately 75% of CC clusters could be linked to an event cluster produced by CANARY analysis (that is, a detection within the duration defined by the first and last contact, allowing one day prior to this).

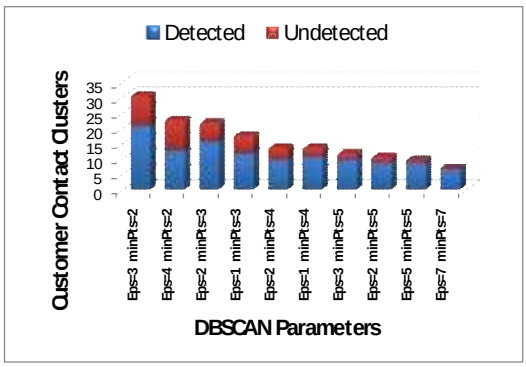

Figure 5: Temporal correlation of event clusters and contact clusters using DBSCAN

Further work could explore both temporal and spatial relationships through improved model and GIS integration. Furness et al. (2011) describe an integrated network model constructed from CC datasets regarding water quality, maintenance records and network hydraulics.

CANARY detects anomalies from baseline data and does not attempt to provide an indication of the cause of the anomaly; however, relative contributing parameters for a particular detection could be formalised into fuzzy system inputs to provide indicative event type classification. Future work will explore online application of CANARY for state of the art sensors. In particular, the combination of a wider range of water quality parameters as well as hydraulic parameters on a single sensor will facilitate incorporation of pattern matching and development of a fuzzy logic 'event fingerprint' interpretation system over multiple parameters.

CONCLUSIONS

There is increasing interest in the use of online water quality monitoring and quality alert generation as indicators of contamination events (intentional or accidental), as well as for helping to identify other types of network event. This study has investigated the use of the CANARY EDS within a sizeable historic water quality data set for a UK field validation case study. Parameters analysed were water temperature, pH, dissolved oxygen, conductivity, turbidity and pressure. Access to water utility records has allowed a comprehensive evaluation (spatially limited to DMA level) of detections, with a wider event remit than purely contamination, resulting in only 14% of event clusters unexplained (ghosts). False negatives, defined by temporal clusters of water quality CCs in the pilot area not corresponding to detections, were only approximately 25%.

ACKNOWLEDGEMENT

This work is part of the Pipe Dreams project supported by the U.K. Science and Engineering Research Council, grant EP/G029946/1. The authors would like to thank Yorkshire Water Services for data provision and the U.S. EPA for freeware distribution of the CANARY tool.

REFERENCES

Aisopou, A., Stoianov, I. and Graham, N. (2012). In-pipe water quality monitoring in water supply systems under steady and unsteady state flow conditions: A quantitative assessment. *Water Research*, **46**, 235-246.

Allgeier, S. C. (2010). Optimizing the performance of a drinking water contamination warning system. Proc. of Water Contamination Emergencies: monitoring, understanding, acting, Mülheim-an-der-Ruhr, Germany, 11th - 13th October 2010.

AWWA. (2005). Contamination warning systems for water: an approach for providing actionable information to decision-makers, American Water Works Association, Denver, CO.

Block, J. C. (2010). Summary on Acting (tentative). Proc. of Water Contamination Emergencies: monitoring,

understanding, acting, Mülheim-an-der-Ruhr, Germany, 11th - 13th October 2010.
EPA (2010). Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development, Testing and Application of CANARY. EPA/600/R-10/036, Washington, DC. http://www.epa.gov/ord.
Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proc. of KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.
Furness, W., Mounce, S. R. and Boxall, J. B. (2011). A data-driven methodology for determining the causes of discolouration in distribution. Proc. of CCWI 2011, Exeter, UK.
Hall, J., Zaffiro, A. D., Marx, R. B., Kefauver, P. C., Krishnan, E. R., and Herrmann, J. G. (2007). Online water quality parameters as indicators of distribution system contamination. *Journal American Water Works Association*, **99**(1), 66–77.
Hart, D. B., and McKenna, S. A. (2009). CANARY user's manual, version 4.1, EPA/600/R-08/040A, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.
Husband,S., Whitehead, J., Boxall, J.B., (2010). The Role of Trunk Mains in Discolouration. *Journal of Water Management*, **163**(8), 397-406, ICE, DOI: 10.1680/wama.900063.
Husband, P. S. and Boxall, J. B. (2011). Asset deterioration and discolouration in water distribution systems. *Water Research*, **45**, 113-124.
Klise, K. A., and McKenna, S. A. (2006). Multivariate applications for detecting anomalous water quality. Proc., 8th Annual Water Distribution Systems Analysis Symposium, ASCE, Reston, VA.
McKenna, S. A., Hart, D., Klise, K., Cruz, V., and Wilson, M. (2007). Event detection from water quality time series. Proc., World Environmental and Water Resources Congress, ASCE, Reston, VA.
Morley, M. S., Bicik, J., Vamvakeridou-Lyroudia, L. S., Kapelan, Z. & Savic, D. A. (2009). Neptune DSS: A Decision Support System for Near-Real Time Operations Management of Water Distribution Systems. In: Integrating Water Systems. Boxall and Maksimovic (eds), Taylor and Francis, pp. 249-255.
Mounce, S. R., and Boxall, J. (2010). "Implementation of an on-line Artificial Intelligence District Meter Area flow meter data analysis system for abnormality detection: a case study." *Wat. Sci. Tech* , **10**(3), 437-444.
Mounce, S. R., Boxall, J.B. and Machell, J. (2010). Development and Verification of an Online Artificial Intelligence System for Burst Detection in Water Distribution Systems. *Water Resources Planning and Management* **, 136**(3), 309-318.
Skadsen, J., Janke, R. J., Grayman, W., Samuels, W., Tenbroek, M., Steglitz, B. and Bahl, S. (2008). Distribution system on-line monitoring for detecting contamination and water quality changes. *Journal American Water Works Association*, **100**(7), 81-94.