

A REAL TIME SYSTEM FOR DETECTING EVENTS IN WATER NETWORKS

Shana Meeus¹, Dene Marshallsay²

^{1,2}Artesia Consulting Ltd, Unit 2, Badminton Court, Yate, Bristol, BS37 5HZ

¹shana@artesia-consulting.co.uk

ABSTRACT

Water distribution networks will need to evolve to meet the future challenges of population growth, climate change, aging infrastructure and ever increasing customer expectations. Part of this evolution will be the use of smarter systems to manage the network based on advanced data analytics. This paper presents a new system for automatically detecting abnormal events that occur in water networks and automatically reporting alarms back to the operators. The system makes use of existing flow data, without the need for new hardware or changes to the logging frequency of flow data. The paper presents the development of the system and examples from a case study with Dŵr Cymru Welsh Water.

Keywords: Water leaks, pipe bursts, event detection

1 INTRODUCTION

Water companies will need to manage their water systems more efficiently in the future to deal with the demands of population growth, climate change, aging infrastructure and increasing customer expectations. There are existing challenges as well, with the need to meet Outcome Delivery Incentives (ODIs) [1] developed with the regulators for parameters such as leakage, customer minutes lost and supply interruptions.

The detection of abnormal events in water networks presents a challenge. Flow profiles typically follow daily, weekly and seasonal patterns, making the use of observational detection and simple alarm limits very difficult; with these approaches either missing important events or being overwhelmed with alarms. Automatic methods for detecting events need to overcome these shortfalls, and ideally have the capability of being used in different parts of the water network, for example in district metered areas (DMAs) or on trunk mains feeding the DMAs.

The successful application of automatic event detection systems in water networks should allow events to be classified with confidence automatically, confirmation alarms and statistics to be generated for the operator for out of control events, detection and confirmation that the system is back in control after the event, and automatic storage of event information for detailed analysis of network failures. This paper presents the development of automatic event and alarm reporting system, and its evaluation on two years of data from DMAs and trunk mains within Dŵr Cymru Welsh Water's distribution system.

2 METHODS

2.1 Methodology Review

Several methods can be used to detect events in monitored areas such as district metered areas (DMAs) and water treatment works (WTW). Historically approaches to this employed methods such as statistical lifetime models, regression models and stochastic methods. Shamir and Howard (1979) [2], attempted to model expected pipe burst rates using pipe age and material, however they concluded that in order to develop realistic results a number of judgement factors were required in

addition to age and material. Their analysis assumed an exponential function in the prediction of burst rates. There is debate as whether this is accepted, since other models have been built using alternative relationships such as power relationships, (Constantine *et al.*, 1996) [3], and proportionate hazard models (Kaara, 1984) [4]. In all of these methods, a large amount of data is required with many input variables to achieve realistic results. As such the availability and quality of data poses a problem, thus many models are not applicable.

Acknowledging research developments in the pharmaceutical industry, we explore process control methods as an option for detecting events and anomalies in water networks [5], Process control charts are typically used to detect deviations or changes in outputs likely to affect quality. However, these same methods can be adapted to water networks by focusing on reliability as opposed to water quality in the traditional sense. These methods do not require large amounts of data points from difference sources, but rather continuous data from a single source. In water networks, this is simply the flow data which Water Companies have available.

The methods discussed herein can in essence be divided into two groups; (a) an independent model per time point, here denoted as “Single Time Point Models”, and (b) one model per time series, denoted as “Global Time Series Models”. There are two single time point models investigated for this paper, the first is defined as the mean and standard deviations, and the second is the median and interquartile range (IQR).

The mean and median define the normal operating conditions in each model and the upper and lower limits of the distribution are defined using the standard deviations and interquartile distances respectively. These methods provide a quick and easy approach to assess whether a new flow reading is within the normal distribution of flow values for that time point. The disadvantage of this system is that multiple independent models need to be built and validated, (for 15 minute flow data, there are 96 bins per day x 365 days \approx 35,000 models). The sample size per time point is also dependent on the number of annual flow profiles available for that particular monitored area. The sample size and hence the power will be markedly reduced if the boundary conditions change for that site or if data is missing for that time point. So if a four year dataset is available, the upper and lower limits of the distribution for that time point will be based on the inter-annual variability of those four samples.

The application of the median and Tukey control limits, reduces attenuation in the control limits from the outliers, but issues relating to combining variance from weekend; weekdays and holidays are still present. This method reduces the effect of outliers, however, the control limits are still attenuated by leakage which can be seen in Figure 1.

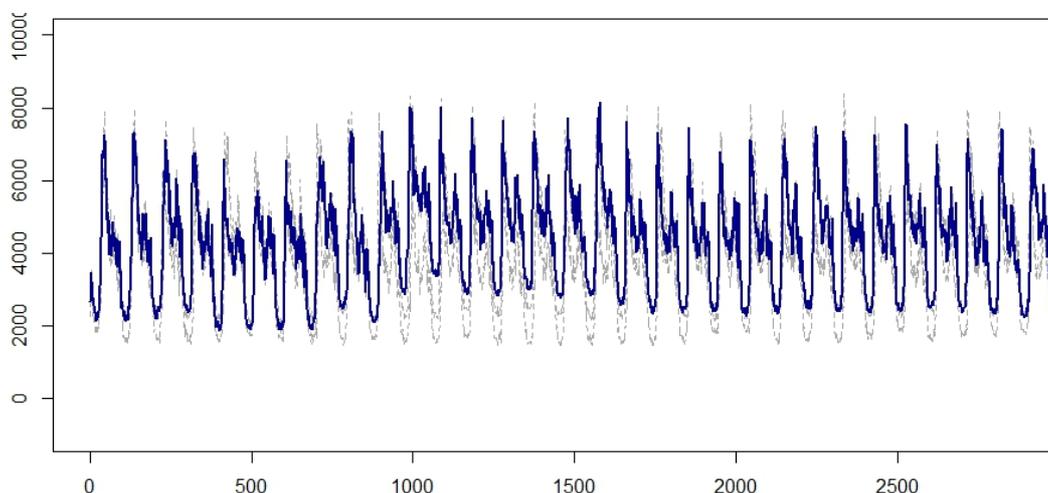


Figure 1 Median and Tukey control limits

The second set of models are based on global time series models. The first are process control (Shewhart) charts such as X bar, R and S charts. The second are EWMA (exponentially weighted moving average) and cumulative sum (CUSUM) charts.

Similarly to single time point models, the mean and standard deviation are used to provide the normal conditions and the upper and lower control limits. The advantage of these models is that rather than developing and validating ~35,000 models, only one model is required. Using global time series methods, the distributions would be less attenuated by missing samples, since ~35000 samples are used to define the upper and lower limits in comparison with single time point models.

Both of these global methods are types of process control charts. Control charts, are tools used to determine if a process is in a state of statistical control. By applying this to the data, it is possible to find which points are 'out of control' indicating an event is occurring.

The assumption of process control charts is that the variables in the sample are independent and that there is no correlation and seasonality between two time points. Since flow data will contain autocorrelation, this will lead to the growth in the number of false alarms on the application of the process control chart. Alwan and Roberts (1988) [6], show that many false alarms may occur in the presence of moderate levels of autocorrelation, as shown in Figure 2. The data used in producing the plot is the raw flow data from a large water treatment works site. The red points indicate that the system is not in control, with the orange points indicating that the system is starting to show signs of abnormal behaviour. The number of orange and red points is excessive, and this particular flow data is considered to be 'normal'. Therefore, the data will need to be adjusted before any process control charts can be used.

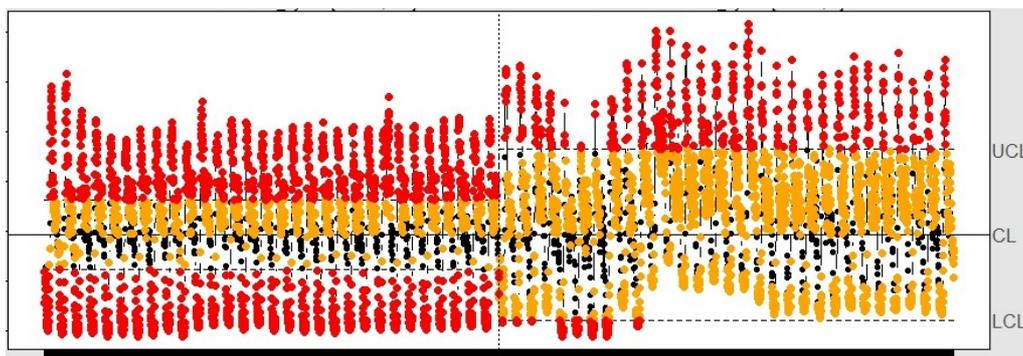


Figure 2 Alarm rate using a system with moderate levels of autocorrelation

A EWMA chart is a control chart for variable data and plots weighted moving average values. The chart averages the data in a way that gives less and less weight to data as they are further removed in time. Because the EWMA chart uses information from all samples, it detects much smaller process shifts than a process control chart would.

As with other process control charts, an important assumption for the EWMA chart is that the samples obtained over time be independent. If the assumption is violated, then the control limits will either be too narrow, giving rise to an increase number of alarms (positive autocorrelation), or too wide, which will mean alarms are missed (negative autocorrelation). Another chart which is used to detect smaller shifts in the mean is a CUSUM chart, which plots the cumulative sums of the deviations of each sample value from a target value. This chart also has the independence assumption, which is a theme among control charts.

2.2 Removing Seasonality and Dependence

Since the global time series models have many benefits over the single time point models, the assumption of stationarity needs to be adhered to. A time series is considered stationary if its statistical characteristics do not change with time. Essentially this means a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations [7].

Flow data could not be considered stationary since a typical flow profile contains daily, weekly and even annual seasonality. The points also contain autocorrelation; the likelihood of a high flow reading is more likely if the preceding point is also high. This is due to the time needed for the water to completely travel through the pipe, as it is not instantaneous that the water levels change. This is why Figure 2 does not perform well. Therefore pre-processing is required.

Seasonal ARIMA, or SARIMA, is a method which can be used to decompose a time series and remove seasonality using seasonal differencing. The method was developed by Box and Jenkins (1976) [8], and has been successfully applied in various water and environmental management applications, as well as in conjunction with process control [9]. After testing the method in conjunction with statistical process control, it was evident that a lot of seasonality remained within the signal, specifically in the weekend to weekday variation. Therefore, alternative methods such as decomposition were explored.

Dokumentov and Hyndman (2015) [10], explored different methods for decomposing time series data, and based their procedure on regression. This allowed for multiple seasonality types and its provision of confidence intervals. However, they comment that the disadvantage to the process is the extended time it takes to run with large amounts of data. More traditional decomposition methods, such as the one developed by Cleveland et al. (1990) [11], can also be used to remove seasonality and/or correlation. These methods can handle most types of seasonality, allowing for a more tailored decomposition of the original time series unlike some ARIMA methods. More control is available for development as the seasonal component is allowed to change over time. The method is robust to outliers, so occasional data points will not affect the trend estimates.

2.3 Selected Methodology

From the analysis of available methods, global time point models were preferred as they were found to be less sensitive to outliers and less attenuated by leakage. Only one model is built per time series, increasing the quantity of training data available per site. Due to the requirement that the data is stationary, decomposition methods are used to remove the seasonality and dependencies from the time series. This can be done iteratively, removing as much seasonality as possible within the data.

In addition to decomposing the data, whilst developing the system some residual seasonal behaviour still remains within the data. Therefore, a dynamic training window has been developed and used to limit the effects of the changing behavioural patterns throughout the year. The combination of techniques, creates a dynamic system which automatically updates through time. This permits seasonality and dependencies to be removed, without the need to define the factors driving the variation, leaving behind accurate and meaningful alarms. Each set of data is treated independently, creating unique results and maximises the confidence of the alarms.

2.4 Developing the System

Once the methodology had been selected, optimisation steps were undertaken to enhance the system performance. DCWW provided almost 5000 sites including DMAs, trunk mains and pressure sites to enable the process to be sufficiently tested using different types of data.

Until this point, tests were only conducted using DMA data, which we know to be seasonal and auto correlated. However, the same can be said for trunk mains and pressure sites, as the behavioural aspect of water use is still prevalent in both. The large amount of available data enabled separate systems to be developed for these alternative data types, with optimisation steps carried out individually to account for any differences in the data.

Decomposition methods require the type of seasonality to be defined in order to remove their effects from the data. So, tests were conducted to determine the frequency and type of the seasonality and it was found that seasonality could be effectively defined for DMA flow, trunk main and pressure data, which when combined with the moving training window was sufficient for use within the statistical process control charts.

During the optimisation period, it was found that although seasonality had been removed, it was possible for unexpected weather changes to cause inflated water usage in an otherwise traditionally stable time period. This level of uncertainty in these few months were factored into the algorithm, defining the upper and lower control limits for certain portions of the year. This reduced the number of demand events which may be inadvertently detected during unexpected hot weather.

In addition, alarm warning limits were added using defined rules to create a further warning to the lead up to an event. This is to try and minimise the time it takes to notify of an event. The developed algorithm was designed to detect abnormal changes, constituting an event.

2.5 Evaluation Trial

The trial allowed the performance of the model to be evaluated in real time, and an opportunity to observe the impacts. Initially, four years of historic data was provided, two years of which were used to train the model. All 176 sites (see Table 1) were retrospectively analysed using 2015 as the year of interest, in which events were detected. Then, the algorithm was run on a weekly basis using 2016 data, with reports of any bursts or events sent directly to DCWW. This permitted any detected events to be audit checked against company records, allowing a detection rate to be determined.

Table 1 Results from 3 month evaluation trial

Outcome	DMA flow	Trunk Main flow	Pressure
Number of sites used	145	19	12
No. of events detected	507	43	24
Sites with at least one event detected	73%	65%	67%
Sites with no detected events	27%	35%	33%
Average no. of events per site per year	5.2	3.8	3
Event detection rate	> 95%	> 92%	Sample too small

Table 1 displays results from the evaluation trial, in which events were detected in the eight months between January and September. In total, there were 574 events detected, across the different data types. This is a significant improvement on the number of alarms which were being generated using the situational awareness system in DCWW (previously ~3000 per day across all sites). There were an average of 5.2 events per site per year for DMAs, in a total of 507 events. The detection rate of 95% was computed using a combination of company records for the same time period for all sites, as well as visually inspecting the flow profiles for each site to identify other events not detected by the system. This figure assumes that all events which were flagged were real, so that there are no false positives. This was corroborated using company records, however we recognise that some smaller scale events may not be logged by DCWW due to no customer reports.

Similarly, the trunk mains detection rate was found to be over 92%, when testing 19 sites. The detection rate using the pressure sites was also determined as being very high, however as only twelve sites were examined, the figure has little confidence and is not reported. The trial found that the events were detected earlier than currently accomplished, and in some cases they were identified before the perturbation is detectable by the human eye. To illustrate the performance of the algorithm, examples are provided using a DMAs and trunk mains.

2.6 Case Studies

Figure 3 shows an example of a typical output from the network detection algorithm when used on a large trunk main. This image shows that an alarm was raised five hours before the burst event occurred, meaning the impact of the event could have been minimised considerably, if this were used in real time. The algorithm initially raised an alarm at 06:45am on 3rd November 2015, the burst then occurred at 11:45am that day. This particular incident caused detrimental effects to many customers; no water for an extended period.

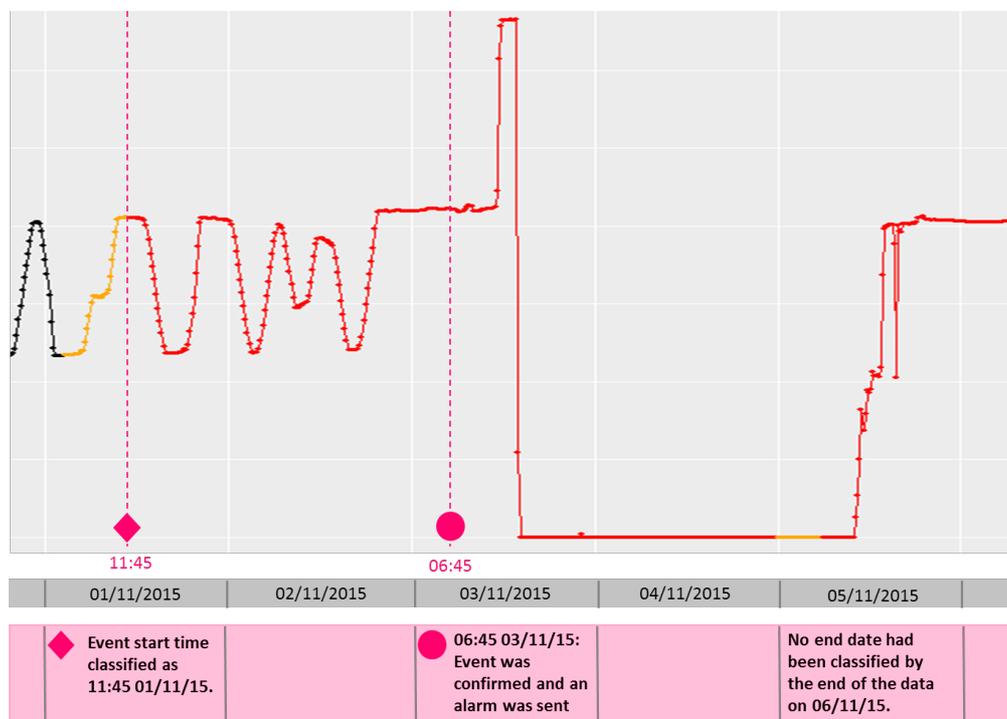


Figure 3 Case study of results found using a large trunk main in Wales.

The only inputs given to the system to achieve this were simply the 15 minute flow trace. Therefore there is no requirement to include other factors which are more difficult to obtain, such as pipe material, density, age or DMA information such as number of households or location. In addition, the upper and lower control limits are algorithm-defined, so no user input is required to adjust thresholds. This removes the manual element to the detection, which is a significant benefits of this type of system. The trunk main in particular seemed to have no obvious repeating shape, making it very difficult to define its ‘normal’ operation. Yet, the algorithm was able to learn this behaviour and react accordingly.

To illustrate this further, the same algorithm was used in a DMA. Figure 4 shows the timeline of outputs. Again, this image shows a burst detected during the 3 month trial, which was not previously known to have occurred prior to running the algorithm.

The system was able to detect changes in the data from 18:30 on 28th January 2016. Statistical confidence was achieved at 00:00 that night and an alarm was issued. Following detection, customer records were scrutinised, and it was found that the first customer contact was not until the morning of the 29th January. Four more calls were received the following day, and the burst was located by DCWW at 08:30 on 30th January. The algorithm was again able to detect an event before the current means of detection, thus potentially saving both time and money in the form of volume of water lost.



Figure 4 Case study of results using a DMA in Wales.

Throughout all of the 547 events detected in 2016 alone, the algorithm was able to identify problems before customer contact, giving the potential of minimising the impact of events on customer minutes lost. In addition, more timely and confident decisions can be made allowing events to be mitigated as they occur. It also contributes towards managing risk and impacts on ODIs, a key driver in a company's targets. Leakage technicians will be given the ability to focus on problem areas quickly and efficiently, without relying on the problem to escalate in order for customers to be motivated to call.

3 CONCLUSIONS AND NEXT STEPS

During the three month trial, the ability of the algorithm was demonstrated through its early detection of events. This was done using a method commonly utilised in the pharmaceutical industry not requiring data rich inputs. The model is based on a global time series model which uses one set of upper and lower control limits per year, ensuring enough data can define limits per period, instead of trying to fit a control window around each 15 minute flow point which may limit and distort the true signal.

The next stage removed seasonality and autocorrelation from the time series to enable use of the statistical process control chart. Thorough analysis showed that decomposition methods were op-

timal due to their unlimited options for seasonal periods. Combined with a moving training window to set alarm limits, the algorithm has shown to be self-learning and requires no intervention to update control limits. The overall provision is fewer and more reliable alarms, in comparison with current Welsh Water processes.

The research can detect events within water networks before they occur, it reduces risk and improves operational responses by advanced warning. Delivery targets such as reliability of supply and leakage can be improved and the technology is adaptable and applicable to other water sector applications, where control of time series data is needed. The authors plan to take this to the next stage, by further collaborating with Dŵr Cymru Welsh Water and testing the system on more networks in a live environment.

4 ACKNOWLEDGEMENTS

The authors wish to thank Dŵr Cymru Welsh Water for their support and collaboration in this project, for data provision and their help in classifying the events.

References

- [1] Setting price controls for 2015-20, Final price control determination notice: policy chapter A2 – outcomes, OFWAT, (2014).
- [2] Shamir, U. and Howard, C. D. D. (1979). “An analytic approach to scheduling pipe replacement.” *AWWA Journal*, 71 (5), 248-258.
- [3] Constantine, A. G., Darroch, J. N., and Miller, R. (1996). “Predicting Underground Pipe Failure.” Australian Water Works Association.
- [4] Kaara, A.F. (1984). A decision support model for the investment planning of the reconstruction and rehabilitation of mature water distribution systems. PhD thesis, MIT, Cambridge, MA.
- [5] Hanna, S. M., (2003), “Application of statistical process control (SPC) in the manufacturing of medical accelerators”. Proceedings of the 2003 particle accelerator conference.
- [6] Alwan, L. C. and Roberts, H. V., (1988), “Time-series modelling for statistical process control.” *Journal of Business & Economic Statistics*, 1988, vol. 6, issue 1, pages 87-95.
- [7] <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm>
- [8] Box, G. E. P. and Jenkins, G. M. (1976). “Time series analysis: Forecasting and control”. Revised Edn., Hodey-Day, San Francisco.
- [9] Naill, P. E. and Momani, M. “Time series analysis model for rainfall data in Jordan: Case study for using time series analysis”. *American Journal of Environmental Sciences* 5.
- [10] Dokumentov, A. and Hyndman, R. J. (2015), “STR: A seasonal trend decomposition procedure based on regression”. Monash University.
- [11] Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning, (1990). “STL: A seasonal trend decomposition procedure based on loess”. *Journal of official statistics*, Vol. 6, No.1, pp 3-73.