
Proceedings Paper:
Gleeson, K. orcid.org/0000-0002-3767-3001, Husband, S. orcid.org/0000-0002-2771-1166, Gaffney, J. et al. (1 more author) (2023) Determining the spatio-temporal relationship between water quality monitors in drinking water distribution systems. In: IOP Conference Series: Earth and Environmental Science. 14th International Conference on Hydroinformatics, 04-08 Jul 2022, Bucharest, Romania. IOP Publishing , 012046.

https://doi.org/10.1088/1755-1315/1136/1/012046

Determining the spatio-temporal relationship between water quality monitors in drinking water distribution systems

Determining the spatio-temporal relationship between water quality monitors in drinking water distribution systems

Killian Gleeson[1], Stewart Husband[1], John Gaffney[2] and Joby Boxall[1]

[1] The Department of Civil and Structural Engineering, The University of Sheffield, Sheffield, S10 2TN, United Kingdom
[2] Siemens UK, Manchester, M20 2UR, United Kingdom

kgleeson1@sheffield.ac.uk

Abstract. A novel method to both assess the strength of connectivity and determine hydraulic transit times between water quality monitors from time series data is reported. It was developed using a network of over 50 mobile multi-parameter sensors deployed for 18 months across a UK drinking water distribution system, and then validated using a network of 18 sensors from a different UK utility. Correlation coefficients are calculated at different time shifts for each possible sensor pair, with strength of connectivity represented by the highest correlation coefficient, and with the temporal lag of this highest correlation also designates the transit time. The results demonstrate the potential to derive valuable spatio-temporal information, with potential to increase understanding of system performance and connectivity. This information can be used to assist with further analytics such as tracking water quality events and improving hydraulic and disinfection residual decay modelling.

1. Introduction

Drinking water quality is typically monitored by periodic discrete sampling that fulfils regulatory purposes, but this can only provide limited information in understanding daily system performance or water quality deterioration processes that are known to occur between treatment and tap, such as discolouration [1] and residual disinfection decay. Technological advances now make it possible to deploy high-frequency (between 1 and 15 samples/minute) water quality monitors along drinking water distribution networks (DWDS), to monitor variables such as turbidity and residual free chlorine. The UK is currently pioneering the deployment of such sensor networks, largely driven by performance targets that include reducing discolouration customer contacts. The resulting data has focussed on identification and recording of events with little analysis to extract greater network understanding and hence inform network management.

There are several challenges and barriers that must be overcome to enable these datasets to be transformed into actionable information, many of which stem from the difficulties in obtaining good quality data. Water quality sensors measuring parameters like free chlorine and turbidity are sensitive scientific instruments that can produce erroneous data when deployed remotely within DWDS. This has led to poor quality data, limiting what can be done with analysis. Determining optimal sensor deployment strategy [2] and how best to analyse the subsequent datasets are also major challenges. To date, water quality analytics has mainly focused on event detection [3–5] and little work has been done

to understand how the spatio-temporal combination of water quality sensor data can be used to enhance DWDS water quality data analysis.

The topic of how simultaneously recorded time series are related to each other spatio-temporally has been studied in areas such as seismology [6], astronomy [7], ultrasound imagine [8], and psychology [9]. Cross-correlation is the most commonly used method for determining the strength of relationship and time lag between two time series signals [10]. This involves shifting one time series relative to another and calculating a correlation coefficient at each step, with the step giving the highest correlation taken as the time lag. Pearson's correlation coefficient (PCC) is the most used coefficient as it measures the linear relationship between two variables. Many variants on cross correlation, such as detrended cross correlation analysis have been developed to deal with non-stationarity and the presence of unwanted periodicity [11].

The aim of this work is to demonstrate the suitability of applying cross-correlation analysis to DWDS water quality time series data, to determine strength of relationship and transit times between sensors. To achieve this, a method is developed to calculate the cross-correlation coefficient between water quality time series and is tested on multiple real-world UK DWDS datasets. As cross-correlation analysis is particularly susceptible to the presence of outliers and erroneous data such as flatline periods, suitable data quality assessment is needed for the correlations to be meaningful.

2. Method

A process involving sensor data quality checks [12] and subsequent cross-correlation analysis was written in Python, primarily using open-source library Pandas [13]. The flow chart in Figure 1 sets out the functionality of the code. This method was initially written to analyse a network of over 50 turbidity and chlorine multi-parameter sensors deployed for 18 months across a UK network. Data quality checks were developed for turbidity and chlorine data to overcome issues experienced by scientific monitors and remote communications from field deployed instruments. Cross-correlation is then used to determine the strength of relationship and transit time between two water quality sensors. These sensors must have sufficient good-quality data in common for the correlations to be meaningful, set as 50% of total window length for this work. This 50% commonality limit was selected to ensure that correlations were meaningful while also allowing for the long periods of missing or low-quality data experienced. PCC's are calculated at different time shifts for each possible sensor pair, with strength of connectivity represented by the highest correlation coefficient, and with the temporal shift of this highest correlation also designating the transit time. The transit time is only meaningful if the maximum PCC is sufficiently high. For this work, a threshold of 0.7 was used as any values above this are widely accepted to indicate a strong correlation [14].

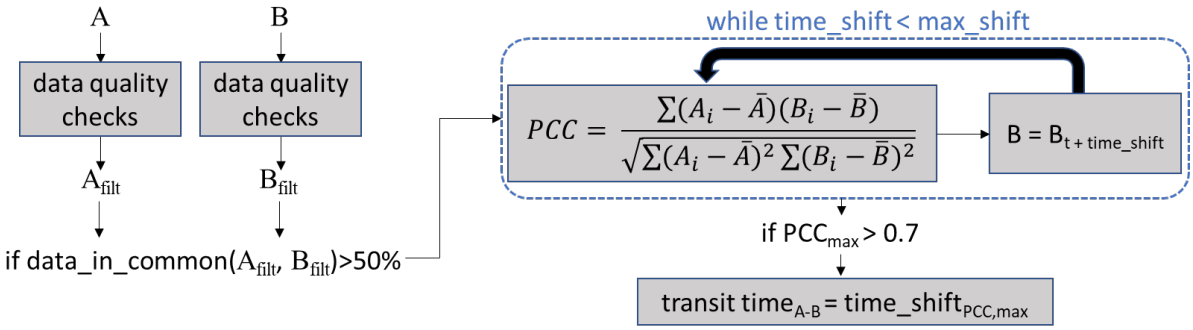

Figure 1. Process flowchart for calculating strength of connectivity and transit time between time series A and time series B.

2.1. Data quality assessment

A rules-based data quality assessment process was developed for turbidity and chlorine to detect the presence of specific anomalies and trends [12]. These are listed in Table 1, and result in the data being categorised into 'removed', 'flagged' or 'remaining' classes. Time stamp errors refer to datapoints that have an unintended sampling interval, compared to the previous datapoints, which can be problematic for analysis and may indicate malfunctioning instrumentation. For mobile sensors as used in this study, missing data is likely due to battery or communication issues, automated checks applying these rules can provide notification to minimise ongoing errors. Single point outliers refer to values that are unrepresentative relative to data before and after. These can occur in turbidity sensor data due to the presence of air bubbles or single highly reflective particles and as potentially unrepresentative are removed before further analysis. Flatlining data occurs when sensors return the same value repeatedly and would not be expected for sensitive instruments in a dynamic environment. Extended periods above a threshold can indicate a sensor error or external interference but could also be a real event so are flagged but not removed. For chlorine sensors, extended periods below a threshold can equally indicate a sensor error and so are flagged. Excessive noise was an issue identified as specific to the turbidity sensors during this trial when the sensor fluctuated between two distinct data points (considered unlikely for turbidity; as a result of data assessment analysis this was later identified as a power-cycle related fault). Drift can occur in turbidity sensors due to optical lens fouling from material accumulation or deterioration of membranes, usually manifesting itself in a slow gradual baseline increase over several weeks.

Table 1. Rules-based data quality assessment.

Turbidity	Chlorine	Removed/Flagged
Time stamp errors	Time stamp errors	Removed
Missing Data	Missing Data	Removed
Single Point Outliers		Removed
Flatlining Data	Flatlining Data	Removed
Extended periods above a threshold	Extended periods above a threshold	Flagged
Excessive noise	Extended periods below a threshold	Flagged
Drift	Drift	Flagged

An example of these rules applied to a turbidity and chlorine sensor is shown in Figure 2. As the legend indicates, any periods greater than 1 hour where the turbidity data was above 1 NTU were highlighted. Likewise, any periods greater than 1 hour where the chlorine data was below 0.2 or above 1 mg/l were highlighted. Any periods where the value did not change for at least 6 hours were highlighted as flatlining data. In this example, this sensor produced very little useful data despite monitoring for over a year, highlighting the multiple issues that can occur. From these results later investigation identified problems with the monitor fitment and sample line flows.

2.2. Cross-correlation analysis

Figure 3 demonstrates how cross-correlation can be used to determine the transit time between two sensors. The top plot shows two synthetic time series, over the course of four days. Time Series B is a copy of Time Series A, with a lag of 4 hours and an offset applied. The bottom plot displays the cross-correlation curve, the peak of which is the time shift which results in the strongest correlation. The maximum correlation coefficient was found to occur for a time shift of 4 hours (indicated by the dotted red vertical line). This method was tried out on multiple real water quality parameters, to determine which are most suited, and the calculated transit times were inspected visually on the time series data. In some cases, high quantities of missing data meant that there was insufficient data in common between two sensors for the cross-correlation to be meaningful. To prevent this, a 50% total

window length commonality was added as a process step following data quality assessment as shown in Figure 1. If the two time series are over a year in length, cross-correlation analysis was done over shorter monthly periods and reporting an average result for the entire length. This is done to avoid the correlations being dominated by seasonal trends shared by many unrelated locations, with shorter time frames more likely to produce correlations that are meaningful in terms of hydraulic connectivity.

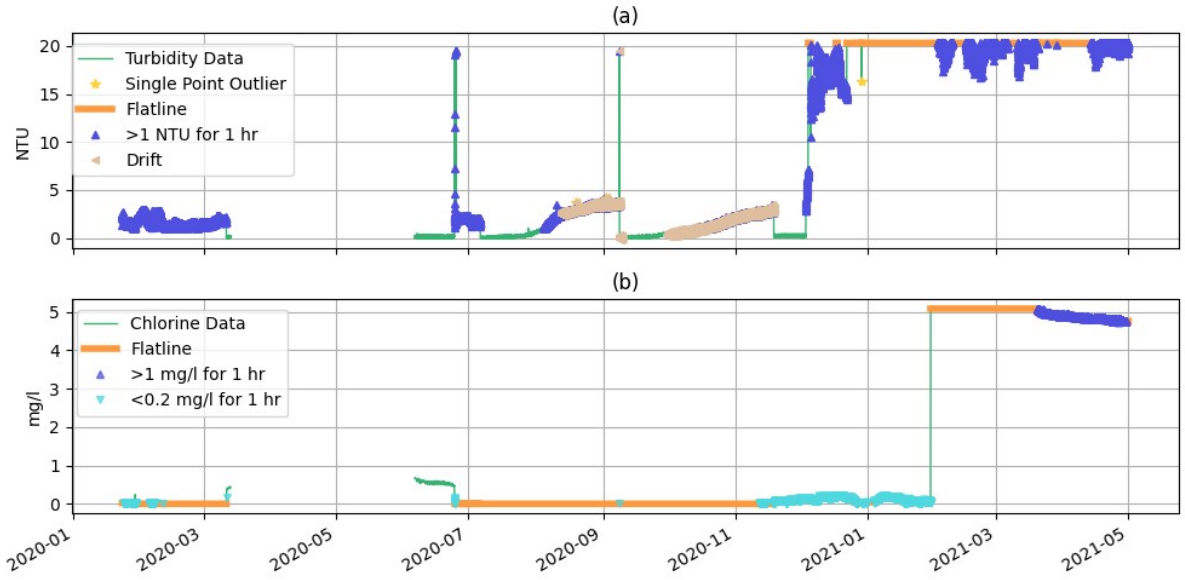

Figure 2. Data-quality rules applied to turbidity (a) and chlorine (b) time series.

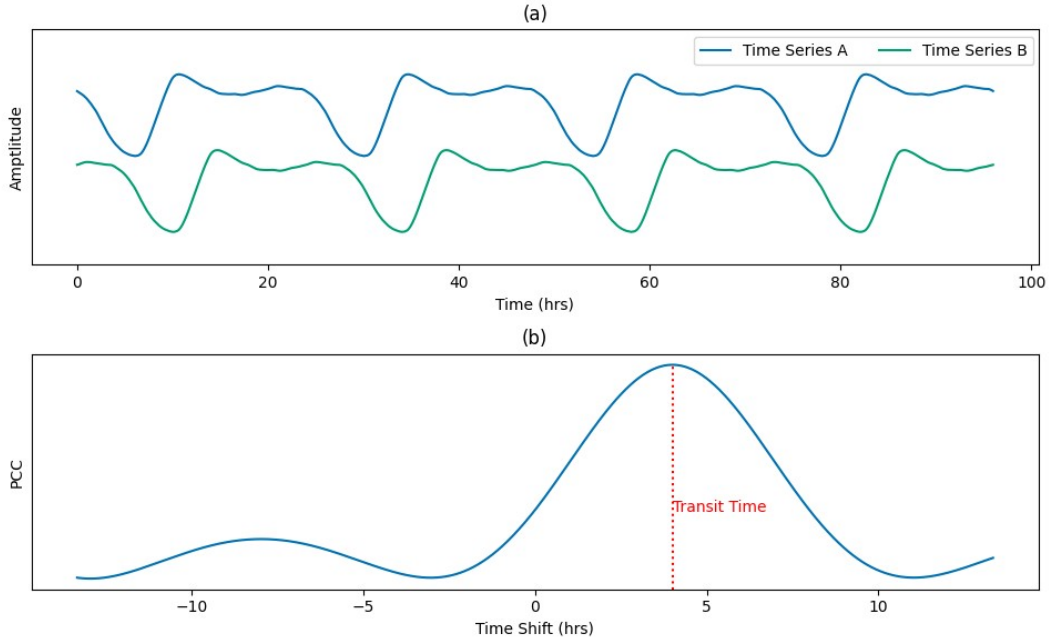

Figure 3. Two example time series (a) and their corresponding cross-correlation curve (b).

3. Results

Cross-correlation analysis using PCC was done for over 60 sensors, across 2 DWDS from different UK water utilities. Selected results are presented in figures 4 and 5.

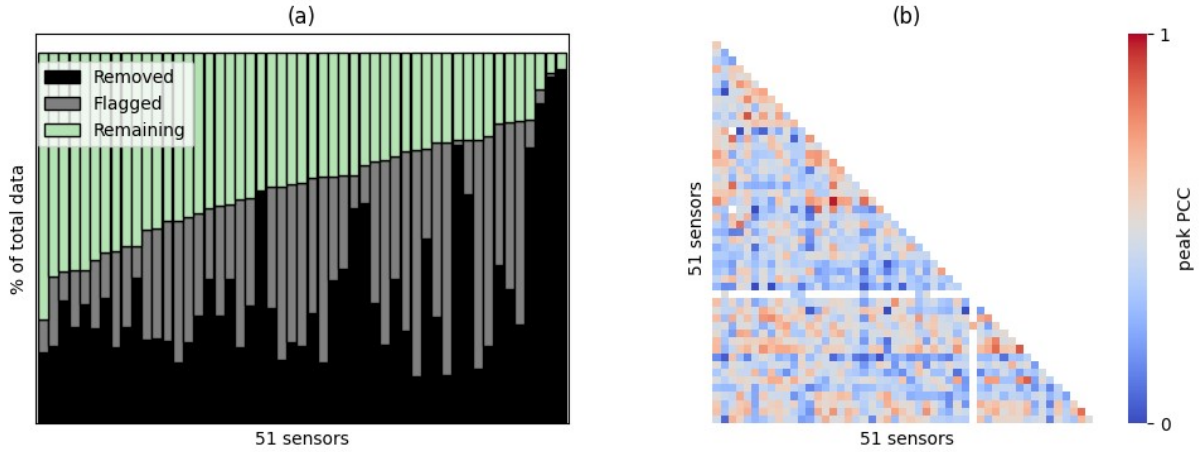

Figure 4. (a) Bar chart showing data quality rules applied to 51 sensors; (b) Heatmap showing peak PCC for same set of 51 sensors.

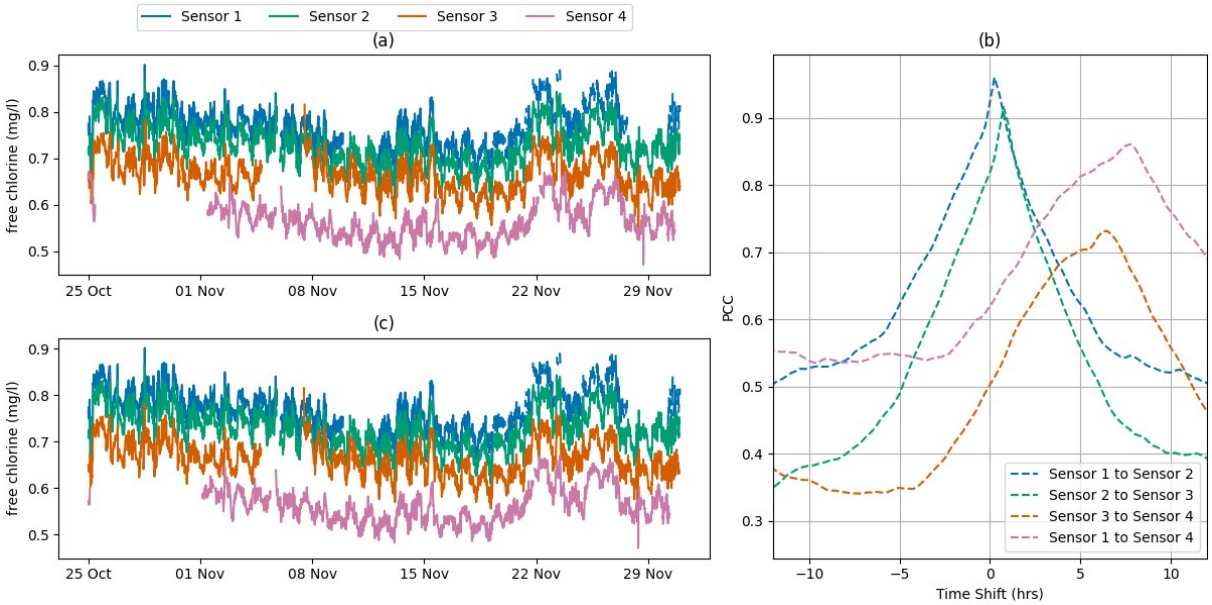

Figure 5. (a) Four interconnected chlorine time series; (b) Cross-correlation curves between the four chlorine sensors; (c) Sensors 2, 3 and 4 synced relative to sensor 1 using the calculating transit times.

Figure 4(a) shows data quality from 51 sensors deployed over 18 months, with 4(b) the corresponding heatmap showing the highest calculated PCC coefficient between each possible sensor pair, with red indicating sensors that were highly correlated and blue weakly correlated. The deployment of instrumentation was not bespoke for this research, but usefully involved deploying multiple sensors across a network resulting in a dataset of variable data quality. After performing the data quality assessment stage for the entire set of time series data, cross-correlation analysis was done

for each month of data, with median PCC values used to obtain the results presented here. Chlorine was the parameter used for this analysis and Figure 4b indicates which sensors are hydraulically connected to each other and is a good starting point before more detailed cross-correlation analysis is done to determine transit times in specific network sections. The empty cells indicate sensor pairs that did not have at least 50% of data in common, after erroneous data was removed, with one sensor having very little data. As the flow chart in Figure 1 explains, transit times are only taken to indicate connectivity and useful information for PCC above 0.7, but the heatmap allows for all PCC's to be visualised. Despite the data quality issues with this dataset, cross-correlation analysis was still able to identify sections of strong connectivity and transit time information.

Figure 5 presents results from four sensors that were found to be closely interlinked through the initial cross-correlation analysis presented in figure 4. The upper left plot shows the four chlorine time-series profiles for just over a month, with the bottom left plot showing the effects of time-syncing the time-series relative to Sensor 1. The right-hand side plot shows the cross-correlation curves for each section, with the transit times calculated to be 15 minutes (Sensor 1 to Sensor 2), 45 minutes (Sensor 2 to Sensor 3) and 6 hours, 45 minutes (Sensor 3 to Sensor 4). As these sensors were sampled every 15 minutes, the transit times cannot be calculated to a higher degree of precision. Though there was some missing data in these sensors, there was enough data commonality (i.e. > 50%) to accurately calculate the cross-correlation.

Figure 6 presents data from a second different set of instrument deployment elsewhere in the UK. The utility had 18 monitors deployed and following this analysis it was found many had data quality issues and some incorrectly sited such that they were not hydraulically connected as planned.

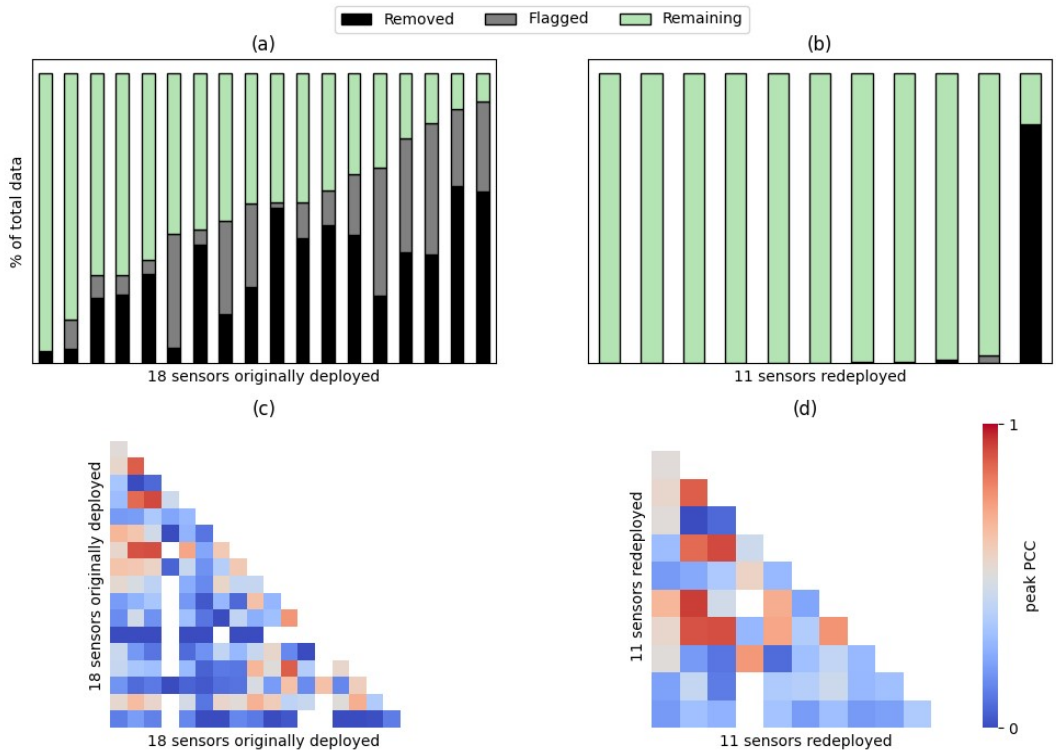

Figure 6. (a) Bar plot showing data quality of 18 sensors originally deployed over course of 12 months; (b) Bar plot showing data quality from 2 months of 11 sensors being redeployed following analysis; (c) Heatmap showing median monthly peak PCC values between 18 sensors originally deployed over 12 months; (d) Heatmap showing peak PCC values in 2 months between 11 sensors redeployed after analysis.

Based on initial analysis and the results in figure 6(a), it was decided to revise the project with a more focused monitor deployment, consisting of 11 monitors in a believed-to-be connected network section. As figure 6 shows, using data from a 12-month window pre and a 2-month period post the focussed sensor deployment, data quality was greatly improved, as fewer sensors now needed to be maintained. For the initial 12-months of data, monthly cross-correlations were done, with median PCC values reported for each pair. Of the connectivity illustrated by the second cross-correlation heatmap (figure 6(d)), there is an overall improvement, and 5 sensors can be observed as highly connected. The remaining 6 sensors were less connected, with 2 redeployed in locations considered of specific water company interest but with this analysis indicating that these sites display distinctly different behaviour. This demonstrates a benefit of cross-correlation analysis to determine if monitors deployed at correct location and if system connectivity is as expected.

4. Discussion

Cross-correlation is shown as an effective method for determining strength of connectivity and hydraulic transit times between water quality sensors. It relies on there being a detectable pattern in each time series. If the time series profile is too flat, the correlation will be difficult to compute due to noise. The method is sensitive to outliers so careful pre-processing and data preparation is required. Data quality checks are an essential prior step to cross-correlation, with outliers and erroneous data such as periods of flatlining removed so they do not negatively impact the correlations. This is particularly essential for water quality monitors which are prone to measurement errors when deployed within DWDS. If a single time series has a strong autocorrelation (i.e., correlates with itself on some time lag, often a diurnal pattern) then the transit time can also not be trusted.

The lessons learned from the initial 51-sensor deployment were taken and applied to the 18-sensor deployment in a different UK utility. The results shown in figures 5 and 6 show how an initial cross-correlation can be applied to large datasets, with the results allowing for a more detailed investigation into connected network sections. Figure 4 demonstrates how a combination of effective data quality assessment and cross-correlation can be used to enhance not only the data quality achieved but the amount of inter-connected high-quality data. This enables a move from single sensor analysis to multi-sensor analysis that can diagnose entire network sections.

Window size is an important consideration, and this work has shown that month long window sizes were effective. This protects against seasonal trends dominating each correlation, while also making the process more robust against periods of erroneous data that have not been removed by the data quality assessment stage. Even short erroneous periods could interfere with a correlation between long time series but would only appear in one of the smaller windows. Shorter windows also allow for changes in transit times over time to be looked at. Though not done in this work, overlapping windows like those used in detrended cross correlation analysis could be employed to increase the temporal resolution of cross-correlation outputs even further.

The methods developed in this work are designed to be agnostic to sensor manufacturer but depends heavily on what parameter is considered. Chlorine is a well-suited parameter for cross-correlation as it retains a similar time series profile, even several hours downstream. Turbidity was not as well-suited due to their being strong local variations even in locations very closely linked. Fortunately, chlorine is a key parameter for DWDS water quality monitoring with companies wanting to optimise amount dosed (cost and concerns with disinfection by-product formation) whilst aiming to retain a residual. The absolute value of chlorine is dictated by the quality and frequency of calibration, but cross-correlation is affected by patterns not absolute values. This gives it a unique data quality requirement that focuses on finding and removal erroneous sub-sequences rather. This work shows that detrending may not be necessary for cross-correlation to be effective on chlorine time series data, but the use of overlapping windows could help understand how transit times are changing, while also reducing the influence of erroneous data. The availability of flow data could be used to explain transit

time variances seen for different time periods. For example, increased transit times between sensors could be explained by a drop in average flow rates.

The connectivity and transit time results can help bridge the current gap between data and actionable information. Examples include:

- Identification of network connectivity, determined through strength of correlation. Useful as utilities can be surprised by which locations are not actually connected and have entirely different water quality profiles.
- Improved sensor data quality assessment, which is vital for the successful deployment of such sensitive traditionally lab-based instruments in remote locations. By comparing connected sensors, erroneous data can be identified and separated from genuine events with greater accuracy.
- Derived transit time information could be used to improve the calibration accuracy of DWDS hydraulic models, typically calibrated using pressure data. This could be particularly useful when adding water quality functionality to hydraulic models, as higher standards of calibration would be required [15].
- Use in characterisation of discolouration events. For example, an event could be described as local to a specific sensor, or global and seen by multiple sensors. Knowing the connectivity and transit times is necessary to be sure about such conclusions. Global events that travel through the network can be assessed with knowledge of hydraulic transit times, which could help in characterising an event.
- Calculation of disinfection residual decay rates for different network sections. As this is difficult to accurately model [16], decay rates for specific sections can be determined, potentially highlighting regions with excessive chlorine decay and improving disinfection decay modelling.

5. Conclusions

This work demonstrates how cross-correlation between water quality sensors can be used to identify DWDS spatio-temporal connectivity and transit times. Data quality assessment is an essential first step to ensure the cross-correlation results are meaningful. The results show that chlorine is a well-suited parameter for such cross-correlation analysis. The derived connectivity and transit time information can be used to enhance data quality assessment and water quality event tracking; and can improve hydraulic and disinfection residual decay modelling.

Acknowledgments

This research has been supported by an Engineering and Physical Sciences Research Council (EPSRC) studentship as part of the Centre for Doctoral Training in Water Infrastructure and Resilience (EP/S023666/1) with support from industrial sponsor Siemens UK, Anna Taliana and Paul Gaskin of Welsh Water, Jez Downs of Southern Water, and Derek Leslie of ATi.

References

[1] Husband PS and Boxall JB 2011 Asset deterioration and discolouration in water distribution systems. *Water Res*. **45**(1) 113–24

[2] Klise KA, Nicholson BL and Laird CD 2017 *Sensor Placement Optimization using Chama* [Internet] (Albuquerque, NM and Livermore, CA; United States) available from: http://www.osti.gov/servlets/purl/1405271/

[3] Hart D, McKenna SA, Klise K, Cruz V and Wilson M 2007 CANARY: A water quality event detection algorithm development tool *Proc. World Environmental and Water Resources Congress 2007* [Internet] (Reston, VA: American Society of Civil Engineers) pp 1–9 available from: http://ascelibrary.org/doi/10.1061/40927%28243%29517

[4] Klise KA and McKenna SA 2008 Multivariate applications for detecting anomalous water

quality *Proc. Water Distribution Systems Analysis Symposium 2006* [Internet] (Reston, VA: American Society of Civil Engineers) pp 1–11 available from: http://ascelibrary.org/doi/abs/10.1061/40941%28247%29130

[5] Mounce S, Machell J and Boxall J 2012 Water quality event detection and customer complaint clustering analysis in distribution systems *Water Sci. Technol. Water Supply* **12**(5) 580–7

[6] Vandecar JC and Crosson RS 1990 Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares *Bull.-Seismol. Soc. Am*. **80**(1) 150–69

[7] Peterson BM, Wanders I, Horne K, Collier S, Alexander T, Kaspi S et al 1998 On uncertainties in cross-correlation lags and the reality of wavelength-dependent continuum lags in active galactic nuclei *Publ. Astron. Soc. Pacific* **110**(748) 660–70 http://iopscience.iop.org/article/10.1086/316177

[8] Bonnefous O 1986 Time domain formulation of pulse-Doppler ultrasound and blood velocity estimation by cross correlation *Ultrason. Imaging* **8**(2) 73–85 https://linkinghub.elsevier.com/retrieve/pii/0161734686900015

[9] Boker SM, Rotondo JL, Xu M and King K 2002 Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series *Psychol. Methods* **7**(3) 338–55 http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.7.3.338

[10] Benesty J, Chen J and Huang Y 2004 Time-delay estimation via linear interpolation and cross correlation *IEEE Trans. Speech Audio Process* **12**(5) 509–19 http://ieeexplore.ieee.org/document/1323087/

[11] Horvatic D, Stanley HE and Podobnik B 2011 Detrended cross-correlation analysis for non-stationary time series with periodic trends *EPL (Europhysics Lett.)* **94**(1) 18007 https://iopscience.iop.org/article/10.1209/0295-5075/94/18007

[12] Gleeson K, Boxall J, Husband S and Gaffney J 2021 Automated data quality assurance for water quality sensors in drinking water distribution systems *Proc. Water Quality Technology Conf.* (Tacoma, WA: AWWA)

[13] McKinney W 2010 Data structures for statistical computing in Python *Proc. Python in Science Conf*. (Austin, Texas: SciPy)

[14] Schober P, Boer C and Schwarte LA 2018 Correlation coefficients *Anesth. Analg.* **126**(5) 1763–1768 http://journals.lww.com/00000539-201805000-00050

[15] Boxall JB, Saul AJ and Skipworth PJ 2004 Modeling for hydraulic capacity *J. Am. Water Works Assoc.* **96**(4) 161–9 https://onlinelibrary.wiley.com/doi/10.1002/j.1551-8833.2004.tb10607.x

[16] Speight V and Boxall J 2015 Current perspectives on disinfectant modelling *Procedia Eng.* **119** 434–41 https://linkinghub.elsevier.com/retrieve/pii/S187770581502576X