Water Resources Management
https://doi.org/10.1007/s11269-021-02768-9

Burst Detection by Water Demand Nowcasting Based on Exogenous Sensors

Caspar V. C. Geelen[1] · Doekle R. Yntema[2] · Jaap Molenaar[1] · Karel J. Keesman[1,2]

Received: 9 June 2020 / Accepted: 6 January 2021/Published online: 02 March 2021

Abstract
Bursts of drinking water pipes not only cause loss of drinking water, but also damage below and above ground infrastructure. Short-term water demand forecasting is a valuable tool in burst detection, as deviations between the forecast and actual water demand may indicate a new burst. Many of burst detection methods struggle with false positives due to non-seasonal water consumption as a result of e.g. environmental, economic or demographic exogenous influences, such as weather, holidays, festivities or pandemics. Finding a robust alternative that reduces the false positive rate of burst detection and does not rely on data from exogenous processes is essential. We present such a burst detection method, based on Bayesian ridge regression and Random Sample Consensus. Our exogenous nowcasting method relies on signals of all nearby flow and pressure sensors in the distribution net with the aim to reduce the false positive rate. The method requires neither data of exogenous processes, nor extensive historical data, but only requires one week of historical data per flow/pressure sensor. The exogenous nowcasting method is compared with a common water demand forecasting method for burst detection and shows sufficiently higher Nash-Sutcliffe model efficiencies of 82.7% - 90.6% compared to 57.9% - 77.7%, respectively. These efficiency ranges indicate a more accurate water demand prediction, resulting in more precise burst detection.

Keywords Water distribution system · Burst detection · Exogenous disturbances

✉ Karel J. Keesman
karel.keesman@wur.nl

[1] Mathematical and Statistical Methods – Biometris, Wageningen University, PO Box 16, 6700 AA Wageningen, The Netherlands

[2] Wetsus, European Centre of Excellence for Sustainable Water Technology, Oostergoweg 9, P.O. Box 1113, 8900 CC Leeuwarden, The Netherlands

1 Introduction

Water distribution networks form an extensive and complex underground infrastructure, coping with a water demand that changes over time and per location. Due to this complexity, optimal management and operation of the distribution network is a challenging task. Suboptimal management has wide ramifications, such as faster deterioration of pipes, insufficient water pressure, increased burst frequency and higher operational costs (Billings and Jones 2008; Kozłowski et al. 2018). Forecasting water demand will help optimize network management and facilitates fault detection (Brentan et al. 2017). Water demand forecasting is challenging, since water consumption depends on many environmental, economic, and demographic factors with temporal and spatial variation (Hutton and Kapelan 2015b). One high priority use of short-term water demand forecasting is burst detection. Burst detection methods are typically based on detecting significant deviations between measured and predicted water demand. Conventionally, the measured water consumption in a District Metering Area (DMA) is compared to a forecast based on historical measurements of water consumption on e.g. the same day in the week and the same time on that day. Significant deviations between the forecasted and the current water consumption indicate a burst, if a suitable and accurate forecasting method is used (Hutton and Kapelan 2015a). The most frequently used methods for water demand forecasting are based on univariate time series models, such as autoregressive moving average (ARMA) models (Hutton and Kapelan 2015b). ARMA models are suitable for short-term forecasts of water demand, as these models are strong in capturing the specific periodic patterns of water consumption. However, water demand is not only a function of periodic water consumption, but is also influenced by exogenous processes, such as holidays, festivals, the weather, pandemics, or other non-periodic deviations in water consumption. Ordinary ARMA models do not take into account these exogenous processes, resulting in an increased false positive rate of burst detection (Billings and Jones 2008). In order to take into account exogenous processes, (multiple) (non-)linear regression or exogenous ARMA models were used, under the condition that extensive data on each of these exogenous processes are available (Adamowski et al. 2012; Papageorgiou et al. 2015; Froelich 2016; Candelieri 2017).

Recent methods make use of neural networks (NN) or other supervised machine learning methods, or hybrid methods that combine NN with univariate/regression forecasting models (Babel and Shinde 2011; Bai et al. 2014; Xu et al. 2018; Pacchin et al. 2019). Similar to exogenous ARMA models, these methods are capable of incorporating exogenous data and boast reliable forecasts, but require extensive historical data for training and are accompanied by large forecast uncertainties, which cannot always be quantified (Hutton and Kapelan 2015b; Anele et al. 2017). Although powerful, even these NN and hybrid models still require identification of all relevant exogenous processes with corresponding data. Identifying the many environmental, economic and demographic exogenous processes that influence drinking water demand as well as collecting all the corresponding data, might not be realistic or feasible for most water distribution companies. A method that does not depend on data of exogenous processes would be invaluable to water demand forecasting and would greatly improve burst detection precision.

Up to now, data of exogenous processes for water demand forecasting was obtained from external sources (such as weather institutes or statistical agencies). However, the multitude of installed pressure and flow sensors in the network present a new, internal data source. These sensors can all be considered as real-time exogenous factors, as they reflect all of the exogenous processes, without having to identify what is the underlying cause of these processes. Hence, instead of using a short-term forecast of water demand based on forecasted

exogenous processes, a water demand nowcast per sensor or DMA water balance based on exogenous flow and pressure sensors within the distribution network can be used. Where forecasting water consumption typically relies on the seasonal nature of collective human water consumption, water demand nowcasting not solely accounts for seasonal water consumption, but also various other water demand patterns caused by exogenous processes, such as weather, holidays, or valve position changes.

This observation becomes especially relevant regarding burst detection. When solely forecasting the seasonal water consumption, significant deviations between the forecasted and the measured water demand will contain many false positive burst alarm caused by non-seasonal water demand due to exogenous processes. Nowcasting water demand at a specific location and based on exogenous sensors will result in a significantly reduction of the false positive rate, as not only seasonal water demand, but also diverging water demand due to exogenous processes can be taken into account. Consequently, when the nowcast deviates from the measured water demand at a specific location, a burst alarm is triggered.

When investigating the measurements of a flow sensor situated close to a burst, the sensor will record a corresponding water demand pattern. However, since most exogenous flow and pressure sensors used in the nowcasting of this sensor will not detect this local burst, the nowcast will reflect the normal diurnal water demand pattern. The resulting difference between measured and nowcasted water demand will thus signal that a bursts has occurred. However, if a more widespread event, such as a holiday, causes a non-diurnal water demand pattern in the investigated sensor, most exogenous sensors will also show a similar pattern. Therefore the nowcasted and measured water demand will not deviate, and this event will thus not trigger a burst warning. The nowcasted water demand based on sensors in proximity as exogenous regressors will more accurately reflect actual water demand compared to water consumption forecasts based on exogenous methods, and thus allows for robust, high certainty and high precision burst detection, without needing vast historical data sets.

The objective of this study was to investigate and evaluate a water demand nowcasting method based on exogenous data from sensors in proximity to the nowcasted sensor or water balance. Our exogenous water demand nowcasting method is compared with a univariate water demand forecasting method that does not depend on data of exogenous processes and is based on RANdom SAmple Consensus (RANSAC) weighted linear regression using up to 20 weeks of past flow measurements (Fischler and Bolles 1981). The water demand nowcast is constructed from the signals from multiple flow and pressure sensors in the distribution network as exogenous factors in a RANSAC Bayesian ridge regression model (MacKay 1992). A 95% prediction uncertainty interval is determined for both methods, to evaluate the uncertainty of the forecasted and nowcasted water consumption. Where other methods reduce exogenous false positives by finding sensor signals with a relatively high distance compared to the signal of other exogenous sensors (Wu et al. 2018), exogenous nowcasting uses the nowcast's uncertainty interval to determine burst occurrence. Three data sets were subjected to the forecasting and nowcasting methods, after which the model efficiency scores were calculated in order to evaluate their performance.

2 Materials and Methods

The exogenous nowcast method and the univariate forecast method were applied to a DMA water balance (data set DMA1), a city-wide sub-DMA water balance (data set DMA1.1) and a

single flow sensor (data set Q1) (Fig. 1), sampled each five minutes from 01/06/2017 up to 01/11/2019, except for DMA1.1, which was sampled from 22/02/2018 up to 01/11/2019, since this DMA was not operational before this date. All data sets were provided by the Dutch drinking water company Vitens. DMA1 is situated in a mainly rural area with a population of more than 100,000 inhabitants spread over 800 km^2. DMA1.1 covers the largest city within DMA1 of more than 30,000 inhabitants and sensor Q1 is located near one of the water production facility within DMA1. For the exogenous nowcast method, data from up to 42 sensors from within DMA1 were used as the exogenous regressors (17 pressure sensors, 25 flow sensors of which 12 industrial water demand flow sensors). The data sets of these sensors were also sampled each five minutes from 01/06/2017 up to 01/11/2019.

2.1 Univariate Water Demand Forecast

In order to forecast up to one week of water demand for a district metering area (DMA) or at a flow sensor in the net using a univariate water demand forecasting method, past measurements from that DMA or sensor are required. For each time t up to one week in the future, a forecast $\widehat{y}_t$ can be made based on past measurements. For that we took those measurements from the preceding 20 weeks that correspond with the same day in the week and the same time on that day. The corresponding linear regression problem is formulated as:

$$Y = X\beta + \epsilon \tag{1}$$

$$\widehat{\beta} = \left((WX)^T X\right)^{-1} (WX)^T Y \tag{2}$$

Here, $N = 20$ are the number of prior measurements considered, $Y = [y_1 y_2, \ldots, y_N]^T$ is an N-dimensional vector with measured water demands for 1, 2, …, N weeks prior to time t, ϵ is the

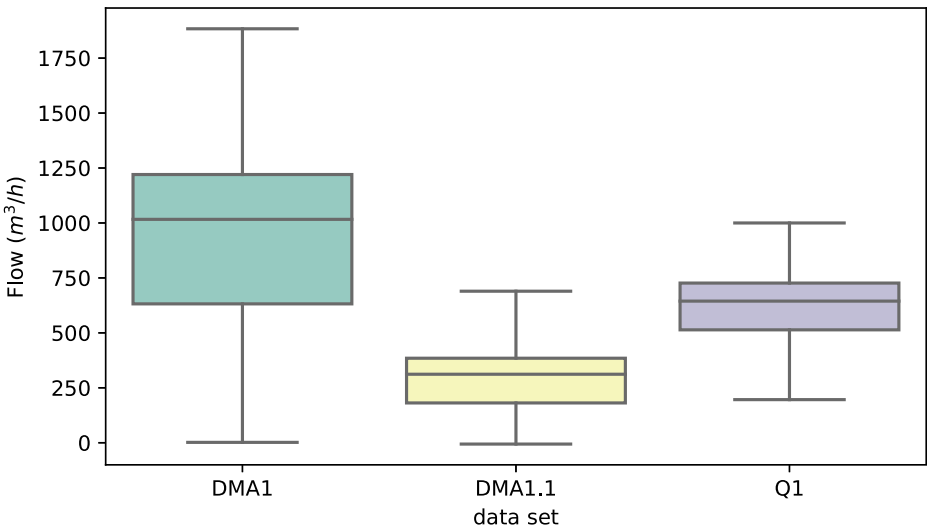

Fig. 1 Boxplots of the data sets of DMA1, DMA1.1 and Q1

corresponding residual vector, $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N \end{bmatrix}$ is the regressor matrix, and $\widehat{\beta}$ is the parameter vector with weighted least squares estimates of the regression coefficients, which are the intercept and slope of the line fitting the 20 data points. Since more recent water demand has more predictive value compared less recent water demand, exponentially weighted least squares is applied to rely relatively more on the most recent measurements, instead of ordinary or generalized least squares. This weighting is achieved by using the diagonal weighting matrix $\mathrm{W} = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_{N,N} \end{bmatrix}$, where $w_{i,i} = (1-p)\text{^(N-i+1)}^{N-i+1}$ with $i = 1, 2, \ldots, N$ and weighting factor $p = 0.2$.

To ensure robust regression, RANdom SAmple Consensus (RANSAC) is applied in order to eliminate the outliers from the training data measurements (Fischler and Bolles 1981). This is achieved by selecting all sets s consisting of every possible combination of N_s rows from X (keeping them in the original order), the corresponding N_s values from Y, and N_s rows and columns from W with $2 \le N_s \le N$. Each resulting combination $\widetilde{X}$, $\widetilde{Y}$, and $\widetilde{W}$ is used in Eq. (2), resulting in a corresponding estimate of the parameter vector $\widehat{\beta}_s$. The "optimal" parameter vector $\widehat{\beta}_{opt}$, that maximizes the RANSAC cost function, and accompanying inlier combinations of regressors X_{opt}, responses Y_{opt} and weights W_{opt} can be found from all sets s using the RANSAC cost function for each set s:

$$\widehat{\beta}_{opt}(s) = \underset{\widehat{B}_s}{\operatorname{argmax}} \sum_{i=1}^{N_s} \begin{cases} \widetilde{W}_{i,i} & if \left(\widetilde{y}_{s\,i} - \widetilde{X}_{i\,s}\widehat{\beta}_s\right)^2 \le \delta_d \\ 0 & otherwise. \end{cases} \tag{3}$$

Outliers are excluded from the regression, if the squared residuals $\left(\widetilde{y}_{s\,i} - \widetilde{X}_{i\,s}\widehat{\beta}_s\right)^2$ are larger than a residual threshold δ_d chosen as the Median Absolute Deviation (MAD) of the responses $\delta_d = \mathrm{MAD}(Y) = \mathrm{median}(|Y - \mathrm{median}(Y)|)$. If RANSAC or missing data results in less than $N_{min} = 12$ inliers, a residual threshold $\delta_d = 2 * \mathrm{MAD}(Y)$ is used instead. If this more tolerant threshold still results in less than $N_{min} = 12$ inliers, RANSAC is not used, as RANSAC apparently does not help to improve the linear fit. Consequently, in that case all 20 measurements are used. Regarding the application of water demand forecasting, using multiples of MAD of responses Y were chosen as the RANSAC residual threshold δ_d, since this is assumed to result in robust results when responses Y have low noise. Thus, the cost function in Eq. (3) makes use of the exponential weights $\widetilde{W}_{i,i}$ in order to penalize past measurements, since recent measurements are assumed to strongly resemble future water demand. The lower threshold of 12 inliers was used to ensure sufficient data is retained to fit the model. Combined with the exponential weights, this ensures sufficient measurements from the recent past are still taken into account.

Estimates of the predicted value $\widehat{y}_t$ and variance $\Sigma_{\widehat{y}_t}$ as well as the 95% prediction uncertainty interval $\left[\widehat{y}_{t*},\ \widehat{y}_t^*\right]$ can be calculated based on the squared residuals Σ_ϵ, significance level 0.05, and $\widetilde{N}_{opt}-2$ degrees of freedom (Chatfield 1993):

$$\Sigma_{\epsilon} = \left\| Y_{opt} - X_{opt}\widehat{\beta}_{opt} \right\|_2^2 \tag{4}$$

$$\Sigma_{\widehat{y}_t} = \left(1 + X_t\left(X_{opt}^T X_{opt}\right)^{-1} X_t^T\right)\Sigma_{\epsilon} \tag{5}$$

$$\widehat{y}_t = X_t\widehat{\beta}_{opt} \tag{6}$$

$$\left[\widehat{y}_{t*},\ \widehat{y}_t^*\right] = \left[\widehat{y}_t \pm t_{0.975,\widetilde{N}_{opt}-2}\sqrt{\Sigma_{\widehat{y}_t}}\right] \tag{7}$$

2.2 Exogenous Water Demand Nowcast

The exogenous nowcasted water demand $\widehat{y}_t$ was constructed using windows with $N = 2016$ measurements, corresponding to one week of sensor data sampled each five minutes, using P flow and pressure sensors in the same DMA. In the case of DMA-wide water demand nowcasting, data from inflow, outflow, and water production location sensors were excluded as regressors, as these are constituents of the DMA water mass balance (Hutton and Kapelan 2015a). For every window of 2016 measurements, sensor signals with a standard deviation smaller than 5 *kPa* or 5 m^3h^{-1} were excluded from the analysis, as signals with a small standard deviation contain no or hardly any information. Consequently, these non-persistently exciting signals were omitted from the analysis as these hardly contain useful information and lead to increased multicollinearity.

The exogenous nowcasted water demand is constructed using a RANSAC Bayesian ridge regression model. For each week of N measurements, a real-time estimate $\widehat{y}_t$ for a specific sensor or water balance can be calculated, based on past measurements with a sampling interval of five minutes. Thus, in this case: $Y = [y_1, y_2, \ldots, y_N]^T$, and $X = \begin{bmatrix} 1 & x_{1,2} & \cdots & x_{1,P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,2} & \cdots & x_{N,P} \end{bmatrix}$ for P exogenous regressors. The resulting Bayesian ridge prediction is formulated as:

$$Y = X\beta + \epsilon, \qquad \epsilon \tilde{} \mathcal{N}\left(0, \widehat{\Sigma}_{\epsilon}\right) \tag{8}$$

$$Y \tilde{} \mathcal{N}\left(X\widehat{\beta}, \widehat{\alpha}\right) \tag{9}$$

$$\beta \tilde{} \mathcal{N}\left(\widehat{\beta}, \widehat{\lambda}^{-1} I_P\right) \tag{10}$$

$$\widehat{\beta} = \left(X^T X + \widehat{\lambda} I_P\right)^{-1} X^T Y \tag{11}$$

Here, ϵ is the residual vector, α the variance of the noise, and λ the Tikhonov regularization penalty. Weakly informative Gaussian priors were used for the uncertainty in the regression coefficients, i.e. $\lambda \sim \Gamma(10^{-6}, 10^{-6})$ and noise variance $\alpha \sim \Gamma(10^{-6}, 10^{-6})$ with initial guesses $\widehat{\lambda}_0 = 1$ and $\widehat{\alpha}_0 = \frac{1}{\Sigma_Y}$, in order to ensure fast and accurate optimization of the regression parameters, before solving for $\widehat{\beta}$, $\widehat{\lambda}$, and $\widehat{\alpha}$ (MacKay 1992; Tipping 2001). By not setting the regularization penalty λ beforehand, but treating it as a random variable, it can be automatically tuned to the data, concurrently with α, λ, and β.

Ridge regression was chosen for its capacity to reduce multicollinearity caused by sensors displaying similar diurnal water demand patterns. Without regularization, thus with $\lambda = 0$, this would lead to a nearly singular matrix $X^T X$. Regularization improves efficiency of the nowcasting and reduces variance, by introducing a small amount of bias (Pacchin et al. 2019). For the practical application of water demand nowcasting, the small amount of bias introduced is deemed acceptable in the bias-variance tradeoff, as it prevents the prediction from being over-dependent on the signal of a single exogenous sensor. Bayesian LASSO (least absolute shrinkage and selection operator) regression was also considered, but was ultimately rejected, as prioritizing a low number of regressors resulted in overfitting on just a few exogenous sensors, which makes the prediction highly sensitive to local anomalies in a few exogenous sensors. By relying on multiple exogenous sensors under ridge regularization, a more robust prediction was created.

Similar to using the cost function (Eq. 3), RANSAC was used to remove outliers from the training data, retaining at least 90% inliers of all measurements ($N_{min} = 1814$). Every possible inlier combination s of N_s rows from X and the corresponding N_s values from Y with $2 \leq N_s \leq N$ is subjected to the Bayesian ridge regression model (Eq. 8–11) in order to find the corresponding regression coefficient estimates in $\widehat{\beta}_s$ and their precision $\widehat{\lambda}_s$. The "optimal" parameter vector $\widehat{\beta}_{opt}$, accompanying precision $\widehat{\lambda}_{opt}$, and set s of inlier combination of regressors X_{opt} and responses Y_{opt} can be found using the RANSAC cost function (Eq. 3), disabling the weighting of more recent data by using $w_{i,\,i} = 1$ for $i = 1, 2, \ldots N_s$. A residual threshold $\delta_d = 0.2MAD(Y)$ was used, unless this results in less than 90% inliers ($N_{min} = 1814$), in which case $\delta_d = MAD(Y)$ was used. If neither resulted in more than 90% inliers or, due to missing data points, less than 90% of the total window size is available, RANSAC was not used. Removing a small fraction of data may indicate some outliers were present in the data. However, when RANSAC disregards a large fraction of the data ($\geq 10\%$), this most likely indicates an anomalous signal that cannot be appropriately fitted by the chosen model. In this case, prioritizing fitting the model to the data instead of editing the data to fit the model will most likely explain more of the phenomena present in the data.

In order to construct a reliable nowcast of the water demand, the model should be fitted on the basis of representative data with minimal outliers. If anomalous events occur in the training data, masking these as outliers will benefit the model more than the commonly used weighting or replacement (Eliades and Polycarpou 2012; Ye and Fenner 2014). An additional advantage compared to similar methods is that allowing masking of a small percentage of data also ensures that the method does not struggle from a small percentage of missing data points, as these will be masked (Wu et al. 2018).

The resulting inlier regressor matrix X_{opt}, response vector Y_{opt}, and model parameter vector $\widehat{\beta}_{opt}$ and $\widehat{\lambda}_{opt}$ are used to calculate the sum of squared residuals Σ_ϵ and response estimate $\widehat{y}_t$

(Eq. 4 and 6, respectively), as well as the response estimate's variance $\Sigma_{\widehat{y}_t}$ and 95% prediction uncertainty interval $\left[\widehat{y}_{t*},\ \widehat{y}_t^*\right]$:

$$\Sigma_{\widehat{y}_t} = \left(1 + X_t\left(\widetilde{X}_{opt}^T\widetilde{X}_{opt} + \widehat{\lambda}_{opt}I_P\right)^{-1}X_t^T\right)\Sigma_\epsilon \tag{12}$$

$$\left[\widehat{y}_{t*},\ \widehat{y}_t^*\right] = \left[\widehat{y}_t \pm z_{0.975}\sqrt{\Sigma_{\widehat{y}_t}}\right] \tag{13}$$

As the regression coefficients change slowly, it suffices to fit the water demand model only once per day. However, real-time predictions can still be constructed at any time t based on the last fitted regression coefficients $\widehat{\beta}_{opt}$ and regularization penalty $\widehat{\lambda}_{opt}$. To illustrate this approach, water demand nowcasting was performed every five minutes from the latest fitted model, which was updated every day at midnight.

3 Results and Discussion

The univariate forecasting and exogenous nowcasting method were applied to the three data sets, DMA1, DMA1.1 and Q1. For data set DMA1, the results of both methods were compared with a so-called Dynamic Bandwidth Monitor (DBM), a univariate forecasting method developed and currently in use by drinking water company Vitens (Fitié 2014) (Fig. 2). To evaluate and compare the model efficiencies between the methods applied to the same data set, the Normalized Root Mean Squared Error (NRMSE, Eq. 14, where values closer to 0% indicate better performance) was calculated. To compare the model efficiencies using different data sets, the Nash-Sutcliffe model efficiency (NS, Eq. 16, where a value closer to 100% indicates better performance) was calculated (Table 1).

$$NRMSE = \frac{1}{\mu_Y}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2} * 100\% \tag{14}$$

$$NS_p = \left(1 - \frac{\sum_{i=1}^{N}\left|y_i - \widehat{y}_i\right|^p}{\sum_{i=1}^{N}\left|y_i - \mu_Y\right|^p}\right) * 100\% \tag{15}$$

Burst detection can be done by evaluating the actual flow measurements with respect to the prediction uncertainty intervals of the predicted water demand. Where some studies rely on manually selected or validation data based burst detection thresholds (Huang et al. 2018; Wu et al. 2018), exogenous water demand nowcasting relies on the calculated prediction uncertainty intervals.

The performance of the nowcasting method as compared to the univariate methods is illustrated in Fig. 2. Where the univariate forecasts deviates from the measured flow due to non-seasonal exogenous processes, and thus trigger significantly more false positive burst

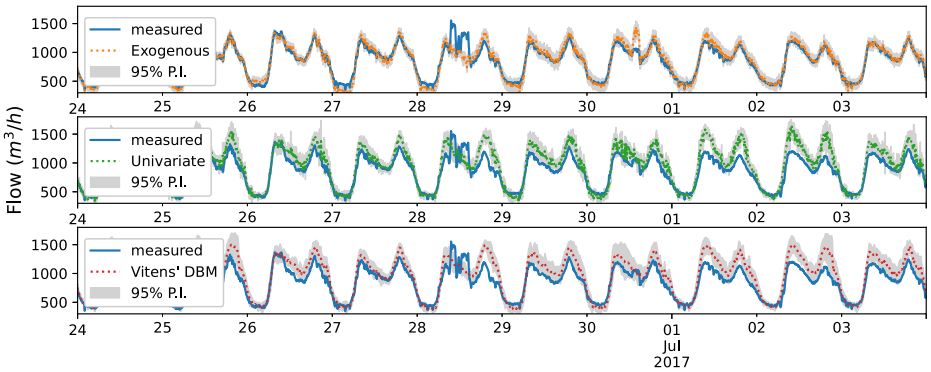

Fig. 2 Comparison between measured and exogenous water demand nowcast (top), univariate water demand forecast (middle) and Vitens DBM forecast (bottom) for data set DMA1, along with their respective 95% prediction certainty intervals from 24th of June 2017 up to the 3rd of July 2017

alarms, the exogenous nowcast is able to account for these phenomena and thus prevent these false alarms. Within DMA1, at 16:20 on the 28th of June a lengthwise tear burst occurred in a 630 mm PVC pipe dating out of 1976 and at 14:20 on the 30th of June a 'simulated burst' occurred in a T-junction between 400 mm PVS pipes dating from 1989 and 1994 when water was lost during placement of new pipes. Both bursts are detected correctly by the exogenous nowcasting method. Although the burst on the 28th is also detected by the univariate methods, the difference is less pronounced and the burst on the 30th is not detected by the univariate methods. The reduced burst alarm rate of the exogenous method and the combination of these water demand prediction methods are detailed in Appendix.

The exogenous nowcasting method also outperforms the univariate methods regarding reduced alarm rate (Table 1). A possible reason for the very good performance of the exogenous method is its real-time nature in contrast with the one-week-ahead forecasts of the univariate methods. However, regarding real-time burst detection, the forecast window size is not relevant. The univariate method more often significantly deviates from the measured flow (Appendix Table 2), as these deviations could be caused by any of many exogenous processes, ranging from holidays, festivities, extreme weather to unexpected peak water consumption (Fig. 2).

The better performance of the nowcasting method is less pronounced when looking at DMA1.1. DMA1.1 reflects the water demand in a large city, while the majority of the

Table 1 Model performance scores

Model	Data set	NS_1 (%)	NRMSE (%)
Exogenous	DMA1	90.6	6.5
Univariate	DMA1	75.1	11.4
Vitens' DBM	DMA1	77.7	10.0
Exogenous	DMA1.1	82.7	10.4
Univariate	DMA1.1	74.6	14.1
Exogenous	Q1	83.5	6.0
Univariate	Q1	57.9	11.0

exogenous sensors in DMA1 that are used to nowcast the DMA1.1 water balance are situated in more rural areas. The resulting difference in demographics influences water demand, making the rural exogenous sensors used suboptimal for predicting the DMA1.1 city water demand. For DMA1 and Q1, the exogenous sensors reflect water demand from both rural and city areas, which may explain the better performance.

The investigated DMA's from Vitens contained enough sensors with signals that could serve as exogenous regressors. However, sensor density differs between DMA and water company. The sensitivity of the exogenous water demand nowcasting method with respect to the number of exogenous regressors considered was also investigated by applying our method to the Q1 data set for different number of exogenous regressors. Of the 40 exogenous sensors data sets available for sensor Q1, 30 were used in fitting the model, as the remaining 10 either did not significantly contribute to the Bayesian ridge analysis or had a too low variance to be included in the analysis. In order to determine how much each sensor contributed to the exogenous prediction, the mean of the absolute regression coefficient estimates ($\mu(\beta_i) = \mathrm{mean}_t\left(\left|\beta_{opt,i}(t)\right|\right)$ for $i = 1, 2, \ldots, P$) was determined for each of these sensor signals. Data set Q1 was again subjected to the exogenous method, where in each consecutive iteration the regressor with the smallest μ_i was removed. For less than three regressors, the method was not able to produce a prediction for at least 95% of all measurements, thus the resulting number of sensors investigated was ranging from 30 to 3. The NRMSE, Mean Absolute Percentage Error (MAPE, Eq. 16) and mean of the 95% prediction uncertainty interval bandwidth over the duration of the data set ($\mu(95\,\%\,P.\,I.)$) were calculated for each number of exogenous regressors used by comparing the respective predictions with the actual measurements (Fig. 3).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| * 100\% \tag{16}$$

Using less regressors may result in a lower precision of burst detection, since there may be insufficient data on the local consumption pattern present in a limited number of regressors. In addition, using less regressors may result in a lower recall of burst detection, especially when a burst occurs that is reflected in the data of all regressors. Including more regressors reduces this possibility and increases recall. This result may also explain the few instances of increase in NRMSE and MAPE when including more sensors, instead of the expected decrease. Consequently, from the top panel in Fig. 3, thus for the given period of data set Q1, approximately 13–20 sensors are needed to obtain appropriate water demand predictions. Consequently, this analysis facilitates the choice of sensor density for "optimal" detection. The other two data sets, DMA1 and DMA1.1, were subjected to the same approach and showed similar results (Fig. 3, middle and bottom panel, respectively).

4 Conclusions

Exogenous nowcasting is a more robust and accurate alternative to univariate water demand forecasting based on historical data. An advantage of nowcasting based on exogenous sensors in the distribution net is that no exogenous processes that influence the water demand have to be identified and no data from these processes need to be available. Our novel exogenous

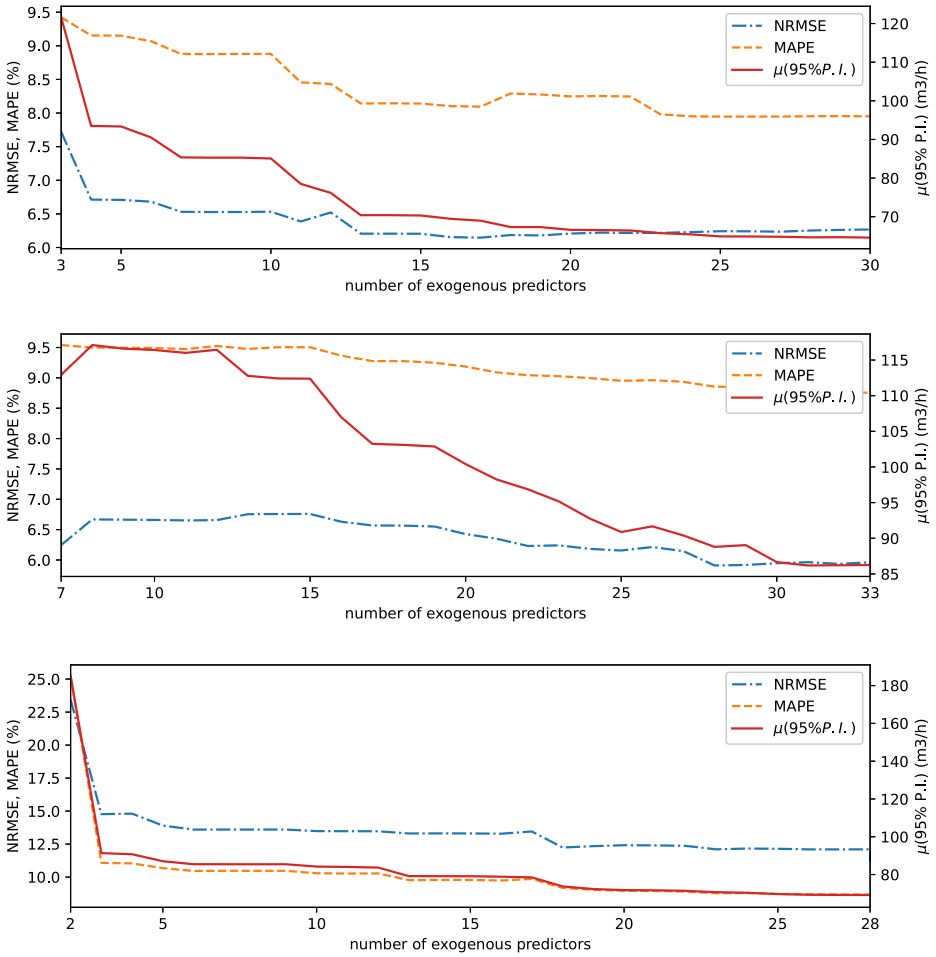

Fig. 3 Model performance (NRMSE, MAPE and average prediction uncertainty interval) of the exogenous method for data set Q1 (top), DMA1 (middle) and DMA1.1 (bottom) as a function of the number of exogenous sensors used as regressors

method performs significantly better than both currently used univariate methods for both data sets, regarding reduced false positive rate (Fig. 2, Table 1).

By combining both the exogenous nowcasting and one of the univariate forecasting methods, a high precision tool is created that only detects a burst when the actual measurements lie outside both the exogenous and univariate 95% prediction uncertainty interval (Appendix, Table 2).

Appendix

The combination of exogenous and univariate methods was also considered. The results are given in Table 2, in which the non-diagonal entries contain the results of a combination of the methods with which the column and row correspond. By using both methods, a high precision

Table 2 Number of identified anomalous measurements with percentages of the total number of measurements that significantly deviate from the nowcasted or forecasted water demand. On the diagonal we present the results for the individual methods and in the non-diagonal entries the results for a combination of methods

	Exogenous	univariate	Vitens' DBM
Exogenous	2140 (2.0%)	824 (0.8%)	851 (0.8%)
univariate	–	11,215 (10.7%)	6603 (6.3%)
Vitens' DBM	–	–	10,191 (9.7%)

tool is created that only detects a burst when the actual measurements lie outside both the exogenous and univariate 95% prediction uncertainty interval (Table 2)

Notation

The following symbols are used in this paper:

N	scalar	number of measurements
P	scalar	number of measurements
Y	vector $N \times 1$	responses
X	matrix $N \times (P+1)$	regressors
β	vector $(P+1) \times 1$	regression coefficients
W	matrix $N \times N$	exponential weights of responses
p	scalar	exponential weighting factor $0 < p < 1$
λ	scalar	regularization penalty
α	scalar	noise variance
t	scalar	time
ϵ	vector $N \times 1$	residuals
Σ_a	scalar	variance of random variable a
s	-	ordered set of inliers with N_s elements, where $2 \leq N_s \leq N$
$\widetilde{a}$ or a_s	-	belonging to the inlier set s
a_{opt}	-	belonging to the "optimal" inlier set s as determined by RANSAC
$\widehat{a}$	-	estimate of variable a
$t_{a,\,b}$	scalar	t-value with significance level a and b degrees of freedom
z_a	scalar	z-value with significance level a
$\left[\widehat{y}_{*t}, \widehat{y}_t^*\right]$	scalar, scalar	95% prediction uncertainty interval for predicted response $\widehat{y}_t$
$\mu(95\,\%\,P.\,I.)$	scalar	mean of the prediction uncertainty interval over time
μ	scalar	mean of vector Y
$RMSE$	scalar	Root Mean Squared Error
APE	scalar	Absolute Percentage Error
$MAPE$	scalar	Mean Absolute Percentage Errorr
NS	scalar	Nash-Sutcliffe score
$\mu(\beta_i)$	scalar	mean of absolute regression coefficient $\lvert\beta_i\rvert$ over time

Acknowledgements This work was performed in the cooperation framework of Wetsus, European Centre of Excellence for Sustainable Water Technology (www.wetsus.eu). Wetsus is co-funded by the Dutch Ministry of Economic Affairs and Ministry of Infrastructure and Environment, the Province of Fryslân, and the Northern Netherlands Provinces. The authors would like to thank the participants of the research theme "Smart Water Grids" for the fruitful discussions and financial support and Vitens in particular for sharing the data sets required to perform this research.

Code Availability Custom code written in Python 3.7 was developed for this study.

Funding This work was performed in the cooperation framework of Wetsus, European Centre of Excellence for Sustainable Water Technology (www.wetsus.nl). Wetsus is co-funded by the Dutch Ministry of Economic Affairs and Ministry of Infrastructure and Environment, the European Union Regional Development Fund, the Province of Fryslãn and the Northern Netherlands Provinces. This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No 665874. The authors also like to thank the participants of the research theme "Smart Water Grids" at Wetsus for their financial support.

Data Availability Data sets used in this research were provided by Vitens.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Conflict of Interest None.

References

Adamowski J, Fung Chan H, Prasher SO, Ozga-Zielinski B, Sliusarieva A (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour Res 48:1–14. https://doi.org/10.1029/2010WR009945

Anele AO, Hamam Y, Abu-Mahfouz AM, Todini E (2017) Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction. Water (Switzerland) 9. https://doi.org/10.3390/w9110887

Babel MS, Shinde VR (2011) Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of Bangkok. Water Resour Manag 25:1653–1676. https://doi.org/10.1007/s11269-010-9766-x

Bai Y, Wang P, Li C, Xie J, Wang Y (2014) A multi-scale relevance vector regression approach for daily urban water demand forecasting. J Hydrol 517:236–245. https://doi.org/10.1016/j.jhydrol.2014.05.033

Billings RB, Jones CVTA-TT- (2008) Forecasting Urban Water Demand

Brentan BM, Luvizotto E, Herrera M et al (2017) Hybrid regression model for near real-time urban water demand forecasting. J Comput Appl Math 309:532–541. https://doi.org/10.1016/j.cam.2016.02.009

Candelieri A (2017) Clustering and support vector regression for water demand forecasting and anomaly detection. Water (Switzerland) 9. https://doi.org/10.3390/w9030224

Chatfield C (1993) Calculating interval forecasts. J Bus Econ Stat 11:121–135

Eliades DG, Polycarpou MM (2012) Leakage fault detection in district metered areas of water distribution systems. J Hydroinf 14:992–1005. https://doi.org/10.2166/hydro.2012.109

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24:381–395. https://doi.org/10.1145/358669.358692

Fitié JH (2014) Dynamic bandwidth monitor. Vitens N.V, In https://github.com/Vitens/DBM

Froelich W (2016) Daily urban water demand forecasting - comparative study. Commun Comput Inf Sci 613. https://doi.org/10.1007/978-3-319-34099-9_49

Huang P, Zhu N, Hou D, Chen J, Xiao Y, Yu J, Zhang G, Zhang H (2018) Real-time burst detection in district metering areas in water distribution system based on patterns of water demand with supervised learning. Water (Switzerland) 10:1–16. https://doi.org/10.3390/w10121765

Hutton C, Kapelan Z (2015a) Real-time burst detection in water distribution systems using a Bayesian demand forecasting methodology. Procedia Eng 119:13–18. https://doi.org/10.1016/j.proeng.2015.08.847

Hutton CJ, Kapelan Z (2015b) A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. Environ Model Softw 66:87–97. https://doi.org/10.1016/j.envsoft.2014.12.021

Kozłowski E, Kowalska B, Kowalski D, Mazurkiewicz D (2018) Water demand forecasting by trend and harmonic analysis. Arch Civ Mech Eng 18:140–148. https://doi.org/10.1016/j.acme.2017.05.006

MacKay DJC (1992) Bayesian interpolation. Neural Comput 4:415–447. https://doi.org/10.1162/neco.1992.4.3.415

Pacchin E, Gagliardi F, Alvisi S, Franchini M (2019) A comparison of short-term water demand forecasting models. Water Resour Manag 33:1481–1497. https://doi.org/10.1007/s11269-019-02213-y

Papageorgiou EI, Poczeta K, Laspidou C (2015) Application of fuzzy cognitive maps to water demand prediction. IEEE Int Conf fuzzy Syst 2015-Novem. https://doi.org/10.1109/FUZZ-IEEE.2015.7337973

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1:211–244

Wu Y, Liu S, Smith K, Wang X (2018) Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems. J Water Resour Plan Manag 144:04017084. https://doi.org/10.1061/(asce)wr.1943-5452.0000870

Xu Y, Zhang J, Long Z, Chen Y (2018) A novel dual-scale deep belief network method for daily urban water demand forecasting. Energies 11. https://doi.org/10.3390/en11051068

Ye G, Fenner RA (2014) Weighted least squares with expectation-maximization algorithm for burst detection in U.K. water distribution systems. J Water Resour Plan Manag 140:417–424. https://doi.org/10.1061/(asce)wr.1943-5452.0000344