Urban hourly water demand prediction using human mobility data

Kamil Smolak*, Barbara Kasieczka*, Katarzyna Siła-Nowicka†, Katarzyna Kopańczyk*, Witold Rohm*
and Wiesław Fiałkiewicz‡
*Institute of Geodesy and Geoinformatics
Wrocław University of Environmental and Life Sciences, Wrocław, Poland
Email: kamil.smolak@upwr.edu.pl
†Urban Big Data Centre
University of Glasgow, Glasgow, UK
‡Institute of Environmental Engineering
Wrocław University of Environmental and Life Sciences, Wrocław, Poland

***Abstract*—The efficient management of a water supply system requires precise water demand forecasts as inputs. This paper compares existing prediction methods and improves their performance by integrating human-related factors with water consumption in an urban area. Furthermore, a framework for processing and transforming mobility data into time-series is presented. Results show that using human mobility data improves forecasting accuracy reaching 87.6%.**

***Index Terms*—urban water demand, human mobility, time-series, machine learning**

I. Introduction

Accurate water demand prediction ensures a reliable water distribution system and provides users with water in adequate volumes at an acceptable pressure. Furthermore, it allows for the detection of leakages in a pipeline when observed consumption significantly differs from the forecasted water demand but also improves energy efficiency through lower pumping energy consumption.

Human mobility is a significant factor for water demand distribution in an urban area. One of the possible data sources providing dynamic information on human behaviour in real-time is geolocated mobile phone data. It consists of at least geographic coordinates and a timestamp, which may be used to reconstruct human movements and mobility patterns.

This paper studies how geolocated data can help to improve water demand forecasting and compares the performance of classical and machine learning algorithms. The selected methods are the auto-regressive integrated moving average model (ARIMA) extended by a seasonal component, support vector regression (SVR), random forests and extremely randomized trees. To the best of our knowledge, this research is the first attempt at utilising geolocated data for public utilities demand forecasting.

II. Application

The study area is located in Wrocław. Historical water consumption readings and mobile phone data records cover the time range of 111 days for the period from the 1st of September 2017 to the 20th of December 2017 and the spatial extent of three residential district metering areas (DMA) - no. 10, 14_Z and 32 - which corresponds to 15,3% of the total city hydraulic sectors area. Water consumption readings are collected for each DMA sector and aggregated with one-hour resolution.

In order to use geolocated data as an exogenous predictor, transformation into time-series is required. For each DMA, data are aggregated into one-hour periods where a number of records is counted. This operation creates a time-series.

To maximise mobile and water data time-series correlation, the mobility data are processed in three steps. First, both data series are standardised and transformed into a typical week for each of the studied DMAs. In the last step, the geolocated series are controlled by two parameters named decay and offset. The former informs how long a single record is accounted for, that is if a mobile phone logs at a specific time, for non-zero decay values it will be still considered to be in an area for the determined period. The offset parameter shifts the geolocated series by a given value, so if the applications on people's mobile phones are logging just before they arrive home, that would align mobile phone records with the water demand series (Fig. 1). Those parameters are selected individually for each DMA during the model training phase and allow to raise the correlation level to 48% - 55%.

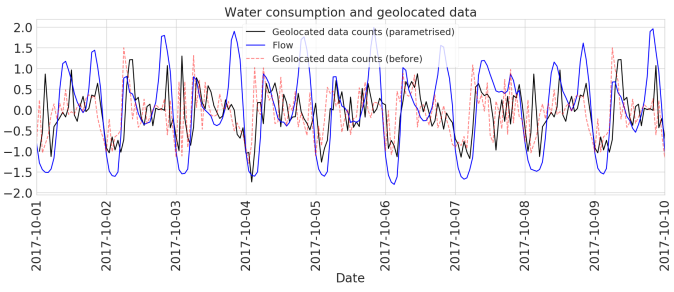

Fig. 1. Standardised geolocated and water consumption time-series after applying offset and decay parameters.

1) Test and validation: The water consumption and processed geolocated data are split into three datasets: 88 days of learning and testing sets and 23 days of validation data (80% and 20% respectively). Learning and testing data are taken together and shuffled during the 10-fold cross-validation process, whereas the validation dataset is used only once at the end of the study.

2) SARIMA and SARIMAX model tuning: SARIMA and SARIMAX model parameters selection is based on the autocorrelation (ACF) and partial autocorrelation functions (PACF) analysis. First, time-series are decomposed to remove seasonality effects. Then, ACF and PACF are used to determine a set of potential model parameters. Each combination is tested for its performance using 3-fold cross-validated approach. Finally, the selected parameters are $p = 3$, $d = 1$, $q = 2$, $P = 3$, $D = 1$, $Q = 4$ and $s = 24$.

3) Machine learning (ML) methods tuning: Tested ML methods have their initial parameters (called hyperparameters), which have to be set up before running the algorithm. For each of the algorithms, the best performing combination is selected using a random search and a learning set with 3-fold cross validation applied.

Tree-based models are tested for a number of trained trees (from 10 to 2000) and a type of split evaluation criterion (mean squared error and mean absolute error). These parameters are selected to 590 trees and the mean squared error criterion for random forests and to 680 trees and the mean squared error criterion for extremely randomized trees. SVR model is tested for various types of kernels (linear or radial basis function), various epsilon values (from 0.001 to 5) and a penalty parameter C (from 0.001 to 5). Epsilon value determines the threshold of acceptable error where no penalty is given during the training process. Penalty parameter is used to control the trade-off between bias and overfit. The best performing combination is a linear kernel, $\epsilon = 2.831$ and $C = 1.661$.

The machine learning methods were adapted for time-series forecasting. This requires selecting of internal and external lags parameters, which determine the number of previous time-series records considered during the prediction task. For water consumption data (Fig. 2), root-mean-square error (RMSE) drops for 168 lags, which is the length of the season (a week) and this value is set. The same test is run for a number of external lags. However, for this parameter, the solution does not vary significantly, since the geolocated data are provided in real-time. Hence, a low value of 5 lags is used.

III. Results

Three types of prediction models are tested. First, denoted as $G(D, O)$ uses all the accessible data along with the modified geolocated time-series. Second solution ($G(0, 0)$) neglects decay and offset parameters. The third model (W) is used for comparison and is based only on historical water consumption. Models performance is expressed by RMSE calculated as an average root mean square error from 10-fold cross-validation in each of the three DMA sectors. The results are included in the Tab. I.

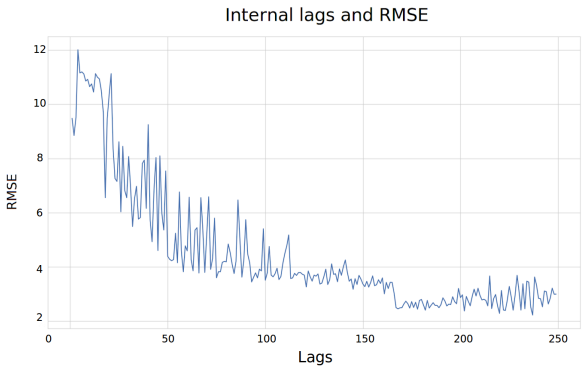

Fig. 2. Number of considered lags and model RMSE computed in 5-fold cross-validation.

TABLE I
RMSE FOR FORECASTING MODELS

Method	W	G(D,O)	G(0,0)
Random forests	0.138	0.130	0.149
ExtraTrees	**0.132**	**0.129**	**0.124**
SVR	0.207	0.175	0.166
SARIMA / SARIMAX	0.199	0.167	-

The results indicate the superiority of ExtraTrees and random forest methods above SVR and SARIMA in terms of prediction accuracy. It can also be noted that the implementation of geolocated data with decay and offset parameters increased the performance of the prediction models.

IV. Conclusions

The aim of this study was to use mobile phone data to 1) indicate dependencies between water usage in a highly populated urban area and its citizens' mobility patterns; 2) investigate the potential of using mobility-related exogenous variables to forecast water demand. A complete forecasting framework was presented.

The best performing algorithm was extremely randomized trees, reaching 87,6% prediction accuracy at an average for a two-weeks ahead forecast. It has also been shown that a moderate (over 50%) correlation of the geolocated time-series and water demand data can be achieved through introducing decay and offset parameters, used for the human mobility data modification.

Acknowledgment

This work was supported by the Climate-KICs Pathfinder Programme under Citizens behaviour patterns for smart utilities and service management (CHASE) project. We wish to thank the Selectivv Mobile House company and local water infrastructure manager MPWiK S.A. for sharing the data for this study.