Urban water demand prediction using human mobility data

Project team: Kamil Smolak, Barbara Kasieczka, Katarzyna Siła-Nowicka, Katarzyna Kopańczyk, Witold Rohm, Wiesław Fiałkiewicz

Geo-located data

Human mobility data is collected through **smartphones**

Geo-located data consist of **ID, timestamp and coordinates**

Every time an user runs a smartphone application his/her location is recorded

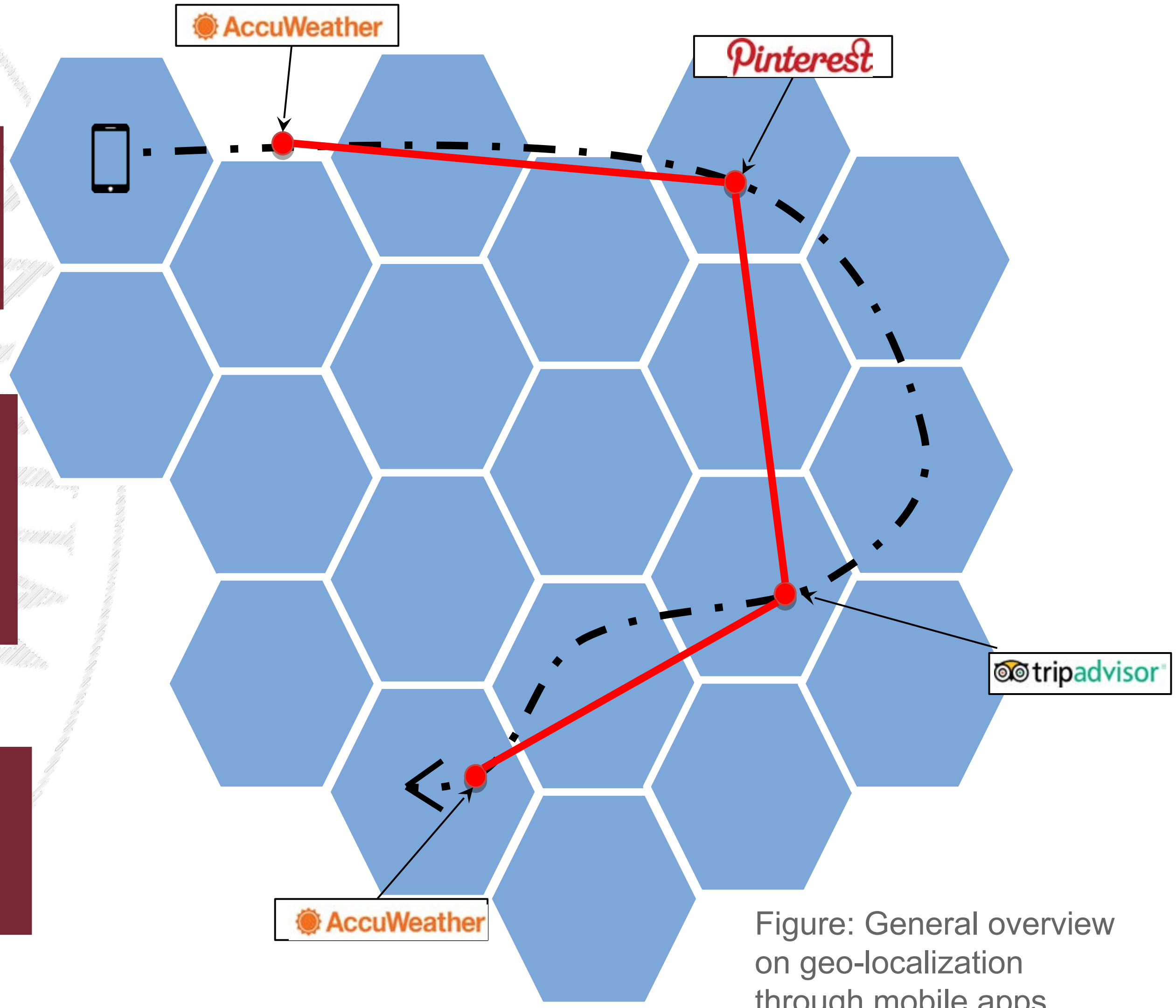

Figure: General overview on geo-localization through mobile apps

Water usage data

Water usage data is **constantly** collected by **measuring devices** installed on the network

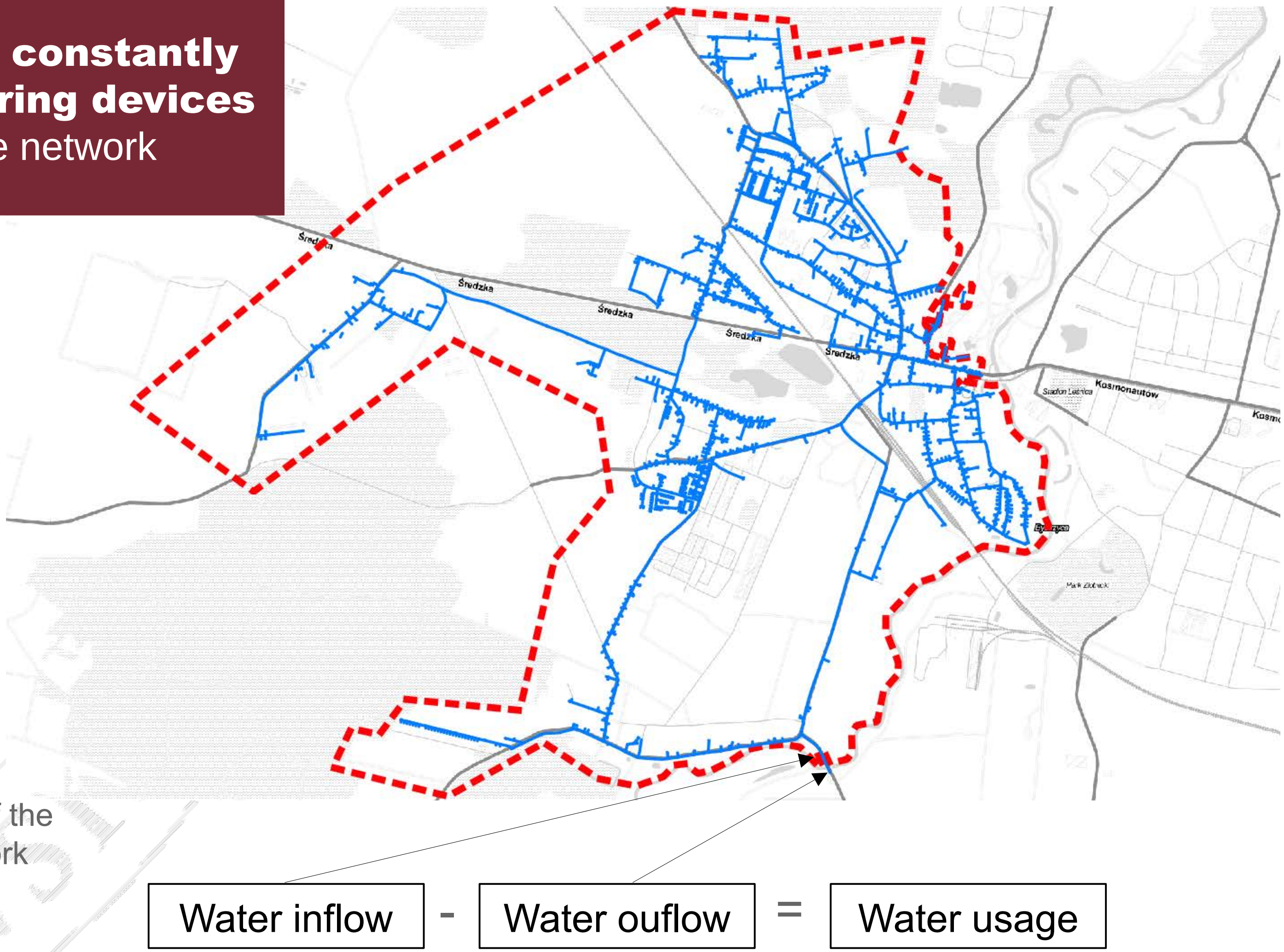

Figure: Part of the pipeline network

Water usage data

Water usage data are **aggregated** over a one-hour period

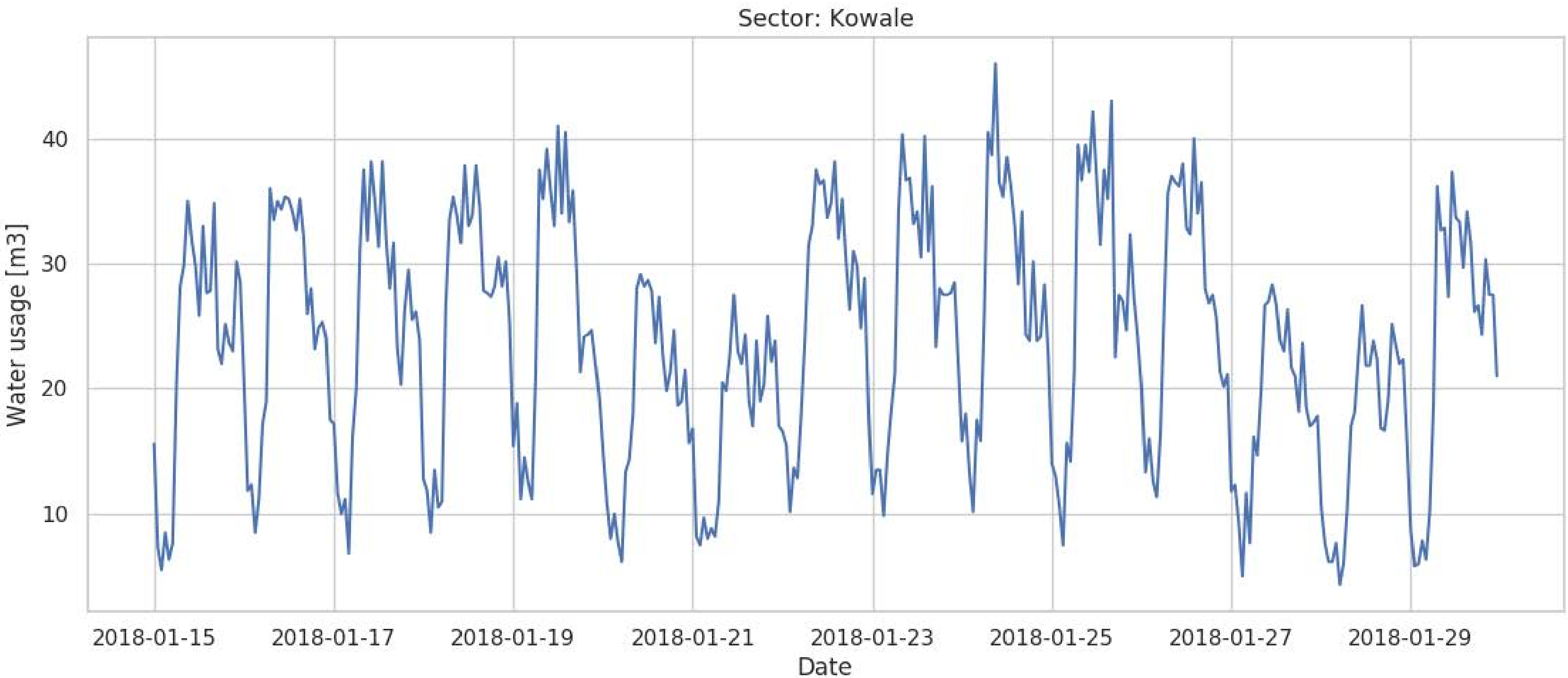

Figure: Sample of water demand data

Case study

The study site is located in **Wrocław.**

The site is divided into **District Metering Areas (DMAs).**

All the calculations are **referenced to the DMAs.**

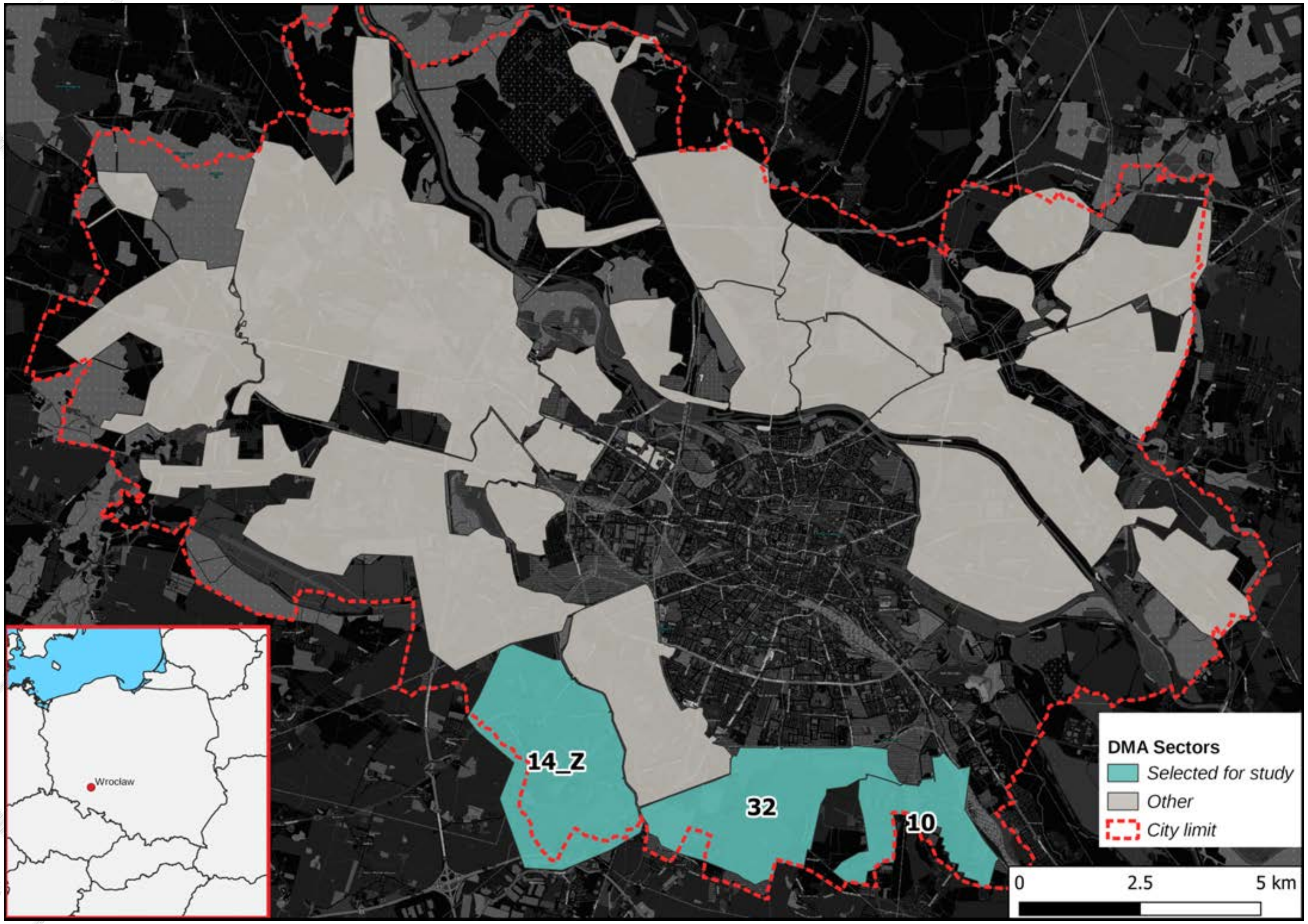

Figure: Map of DMA sectors.

Case study

Geo-located data consist of **564,069 records**

After filtration **38,148** were used for further studies

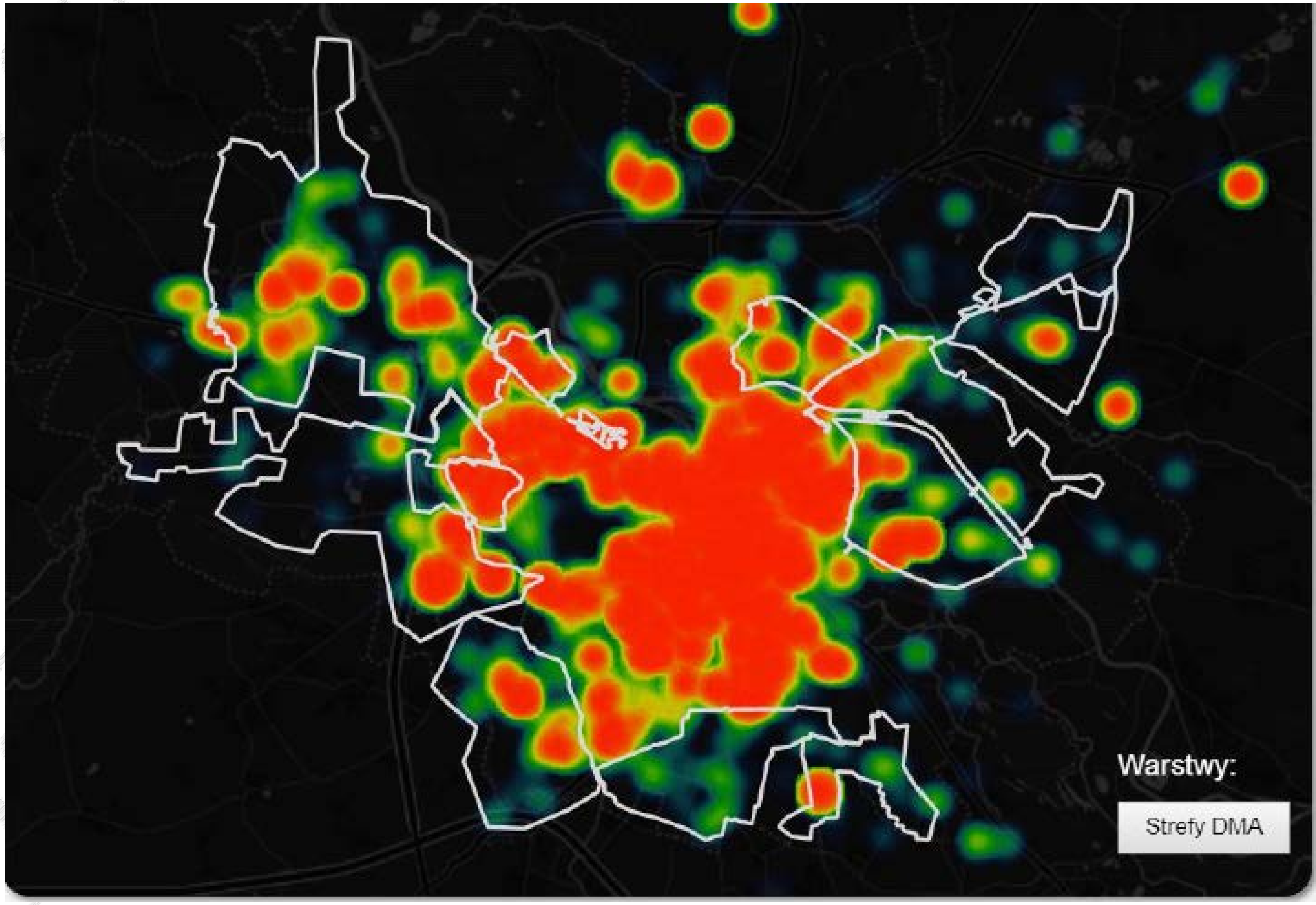

Figure: Heatmap of geo-located data and DMAs (author: Barbara Kasieczka)

Forecasting framework

The forecasting process has been fully automated within a consistent framework

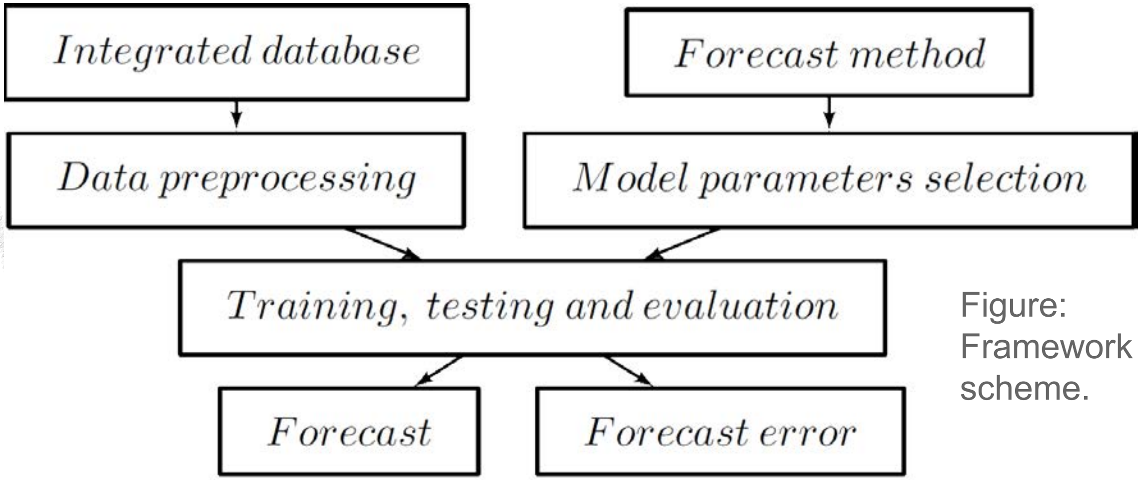

Figure: Framework scheme.

Model splits data into **88 days of learning and testing** sets and **23 days of validation** set.

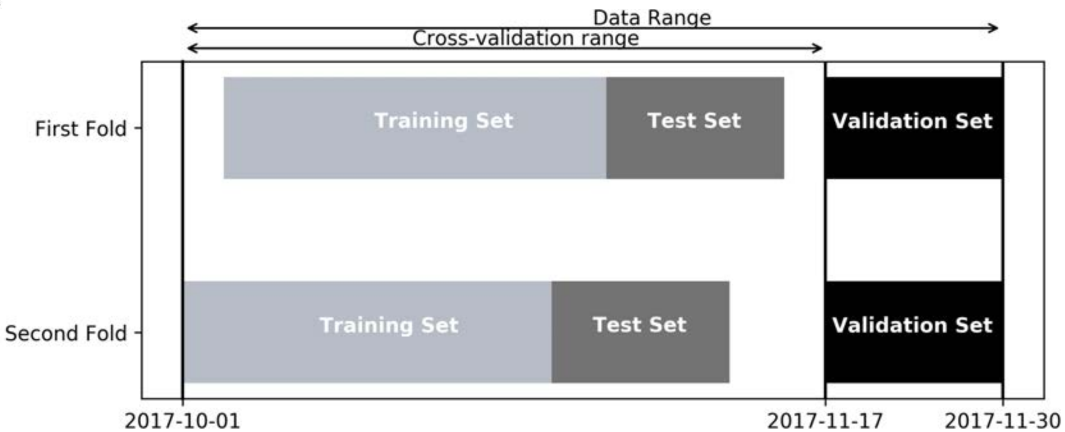

Figure: Cross-validation scheme. Training and testing sets are moved insinde of 88 days range.

Forecasting framework

The framework is independent of the forecasting method. Therefore, **machine learning and classical** approaches are compared.

Forecasting methods used in the study

Machine Learning	Classical
Random Forest	SARIMA
Extremely Randomized Trees	SARIMAX
Support Vector Regression	

Data preprocessing

Water demand data

Data were preprocessed to remove outliers and fill data gaps when possible

Geo-located data

Data were preprocessed to remove outliers

Furthermore, geo-located data are preprocessed to maximise its impact on predictions.

Geo-located data preprocessing

First, geo-located data are transformed into time-series

Initially, correlation with water usage time-series is very low

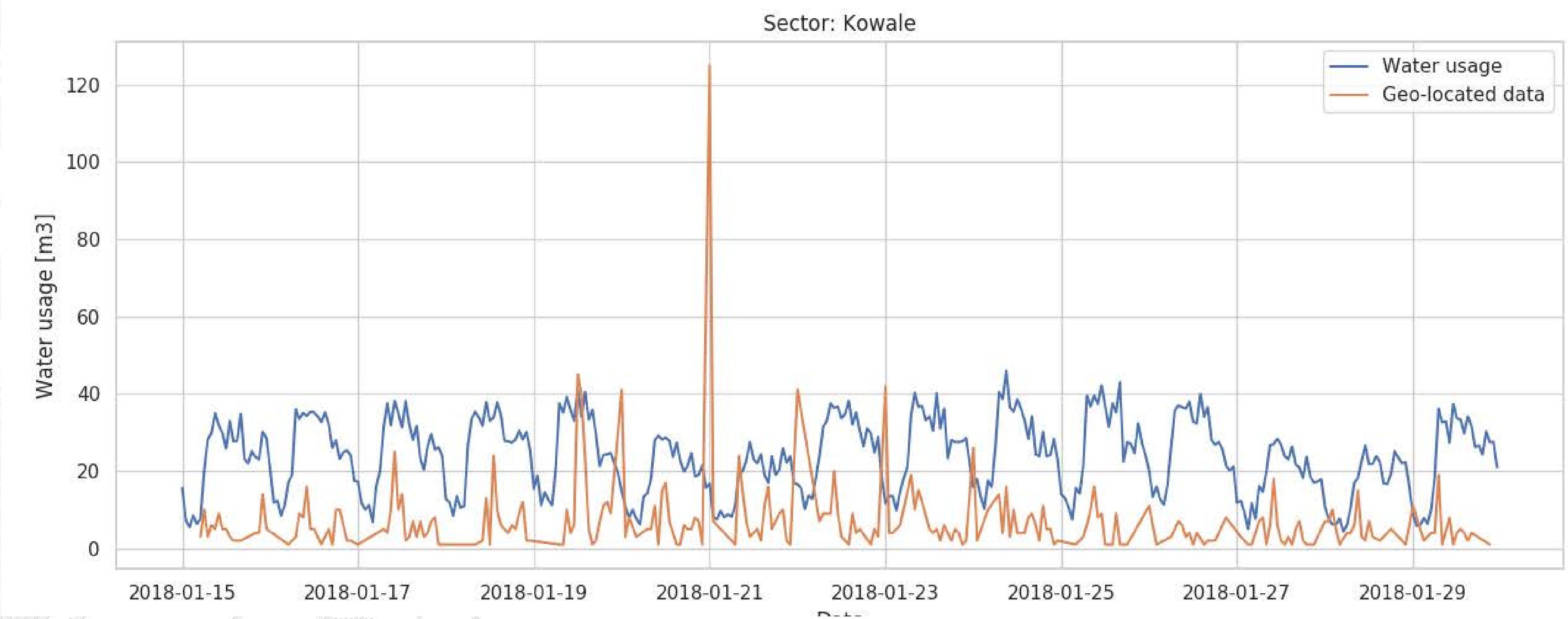

Figure: Water demand and unprocessed geo-located data series.

Geo-located data preprocessing

Geo-located data are transformed into a „typical week”

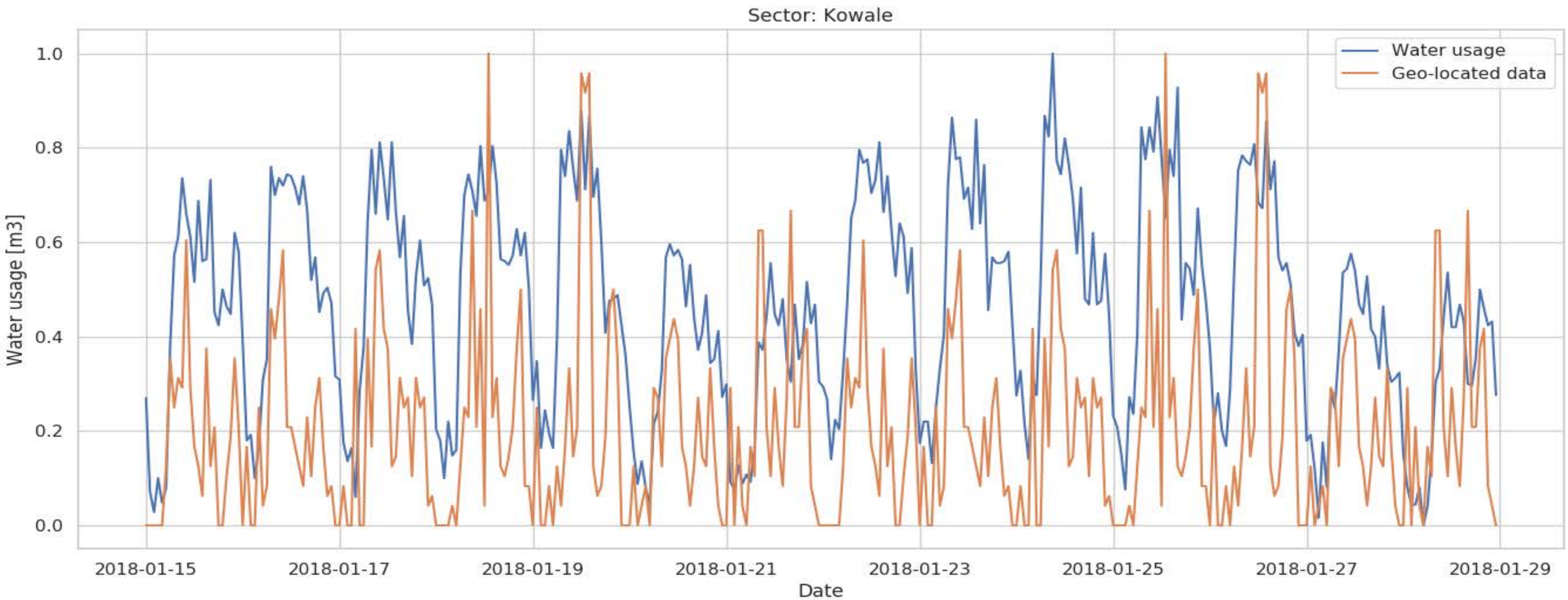

Figure: Water demand and regularised and normalised geo-located data series.

Geo-located data preprocessing

Next, geo-located series are transformed using **decay** parameter

It informs how long a single record is accounted to stay in the DMA

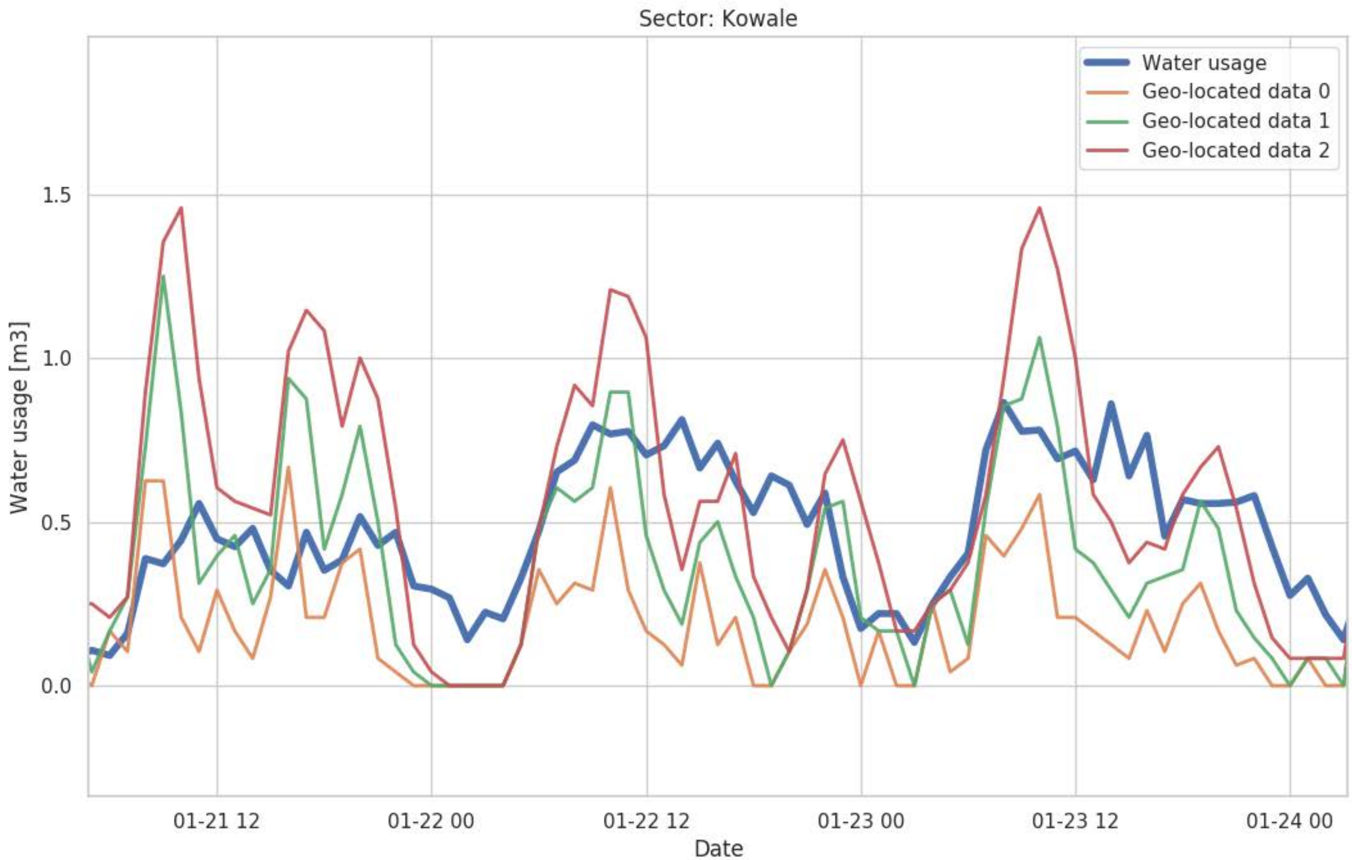

Figure: Results of geo-located series tranformations using various decay parameters.

Geo-located data preprocessing

Finally, water and geo-located series offset is determined using Fast Fourier transform

The solution that maximises series correlation is taken for further calculations

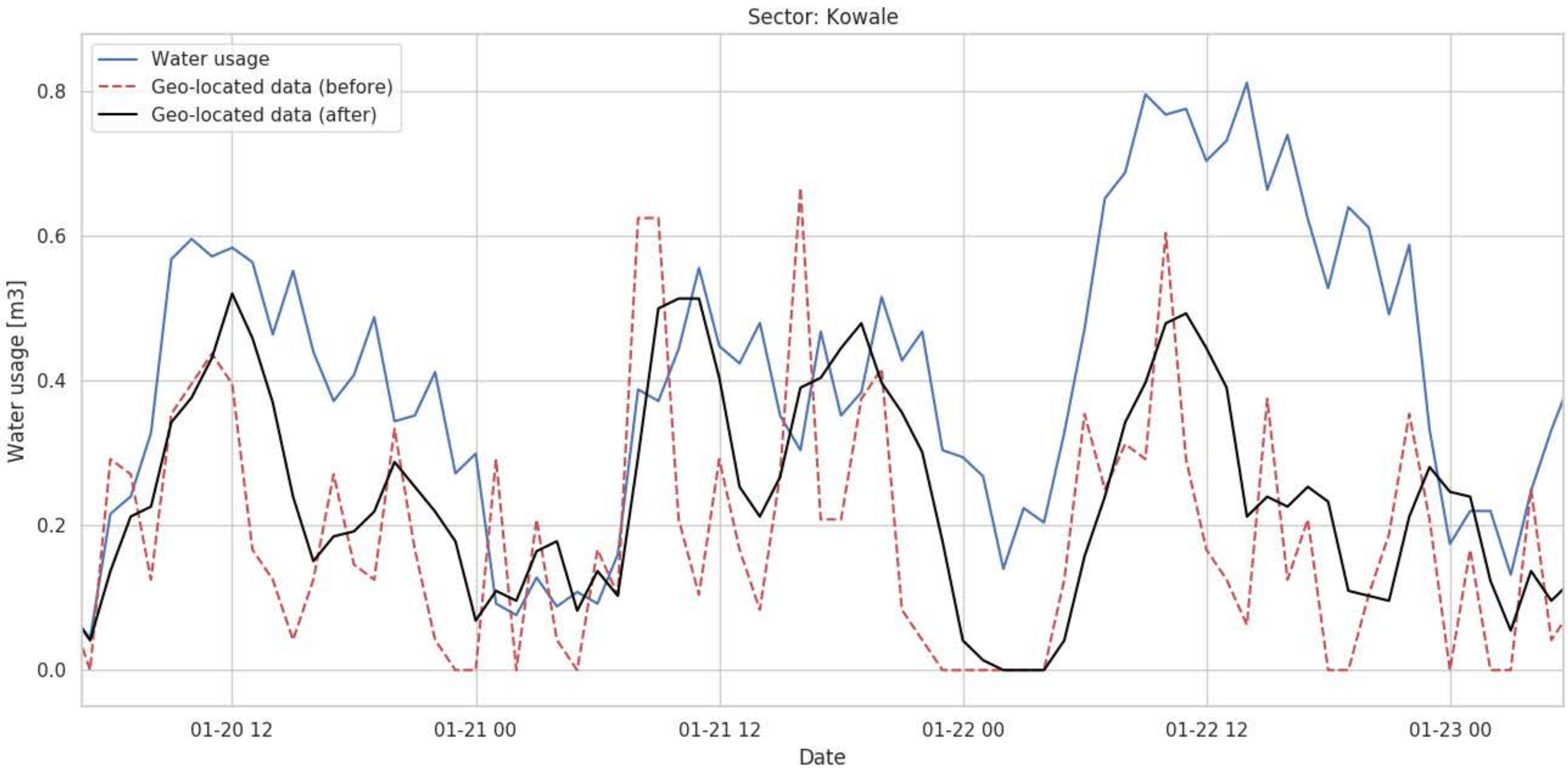

Figure: Water demand and geo-located data before and after processing.

Model parameters selection

Lags determine the number of previous time-series records considered during prediction task.

For water consumption data best solution is obtained for 168 lags (a length of the week)

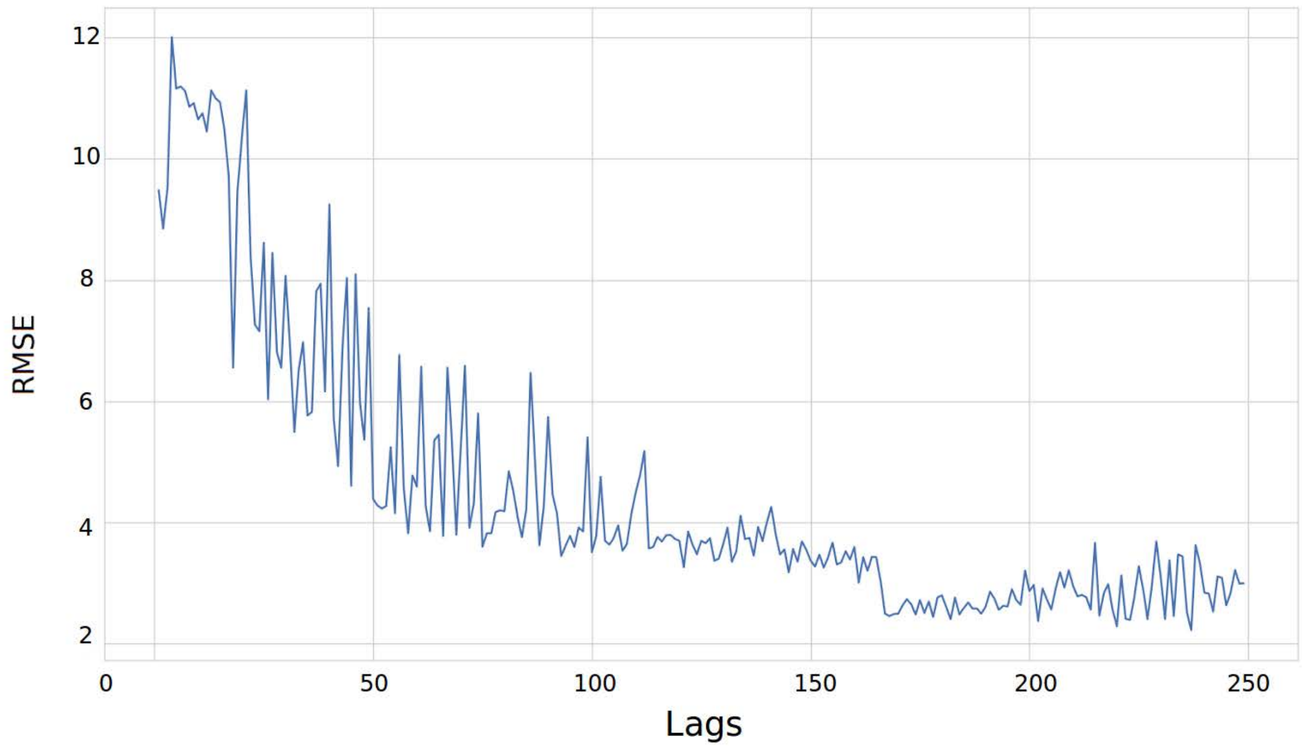

Figure: RMSE depending on numbers of lags used in prediction.

Results

Kowale sector – DMA of industrial type

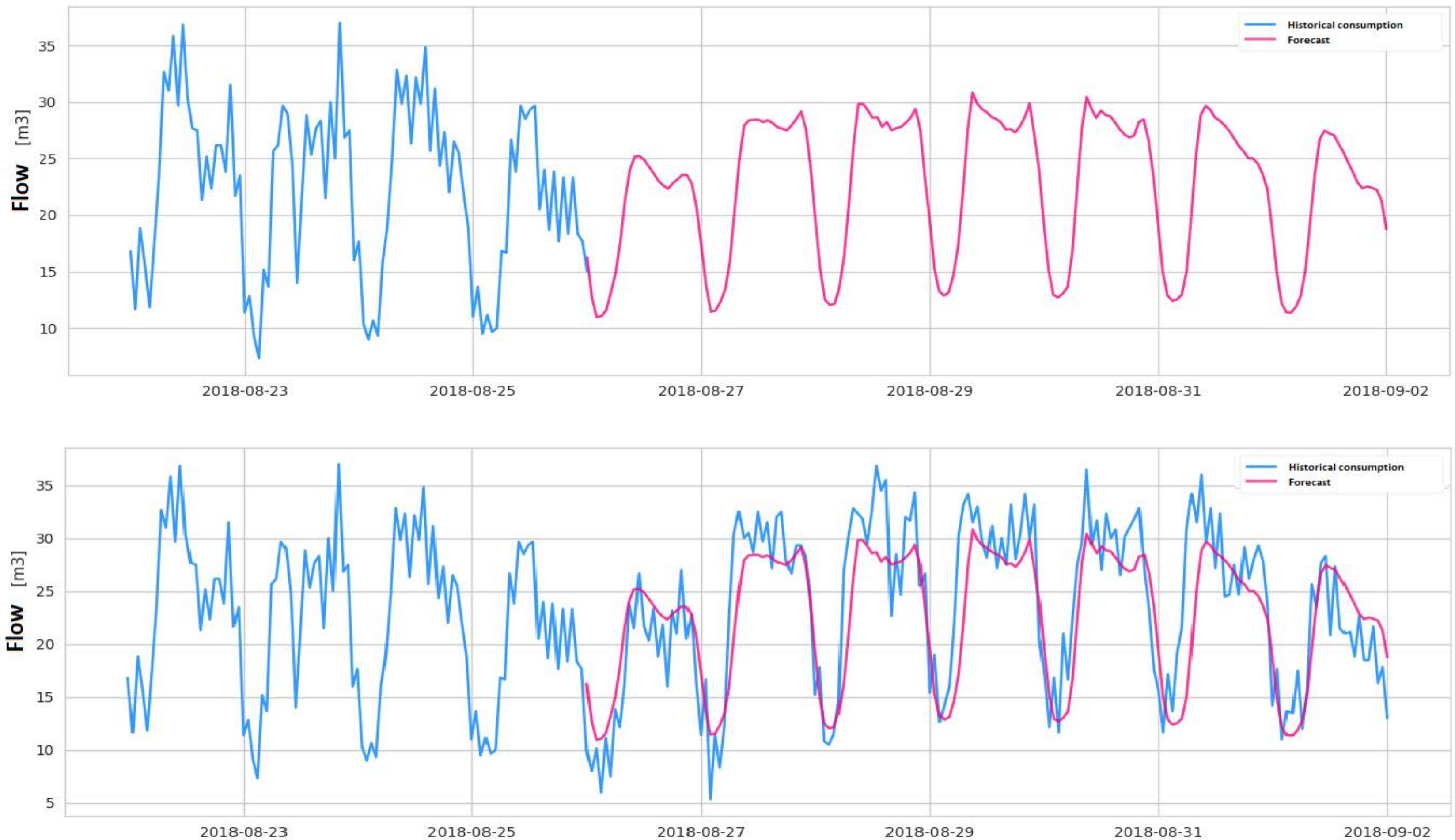

Figure: A comparison of predicted and measured water usage.

Results

Using geo-located data **improves forecasting accuracy**

Average accuracy score was **87.6%**

The best forecasting method was **Extremely Randomized Trees**

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	-

Conclusions and further works

The best performing algorithm was machine learning tree-based method, which outperformed the classical approach

Geo-located data improves predictions accuracy. However, **there is still a room for improvement** of geo-located data processing methods, which will result in further accuracy improvement.

Due to the high series correlation, we plan to investigate if it is possible to base water demand predictions on geo-located data only

Wrocław University of Environmental and Life Sciences

Contact: kamil.smolak@upwr.edu.pl

Urban water demand prediction using human mobility data