# *Urban Water Consumption Prediction and Categorisation*

Joaquim Pedro Bento Gonçalves Pratas Leitão - 2011150072

University of Coimbra
Doctoral Program in Information Science and Technology
Connectivity and Pattern Recognition

April 5, 2017

## 1 Background

In most cities and urban areas investment in equipment related to water supply and drainage is substantial, being responsible for a considerable share of the budget of many municipalities.

Substantial attention has been given to the construction and installation of new pipe and sewage infrastructures, as well as to maintenance operations on existing infrastructures; however, relatively small efforts have been made to study water consumption patterns in these environments.

Indeed, understanding and accurately predicting such behaviours is extremely important to water companies, as it allows them to improve the management of their infrastructures, namely water storage facilities (such as water tanks and towers): Many water utilities repeatedly overestimate the volume of water needed to supply the population of a given region. Therefore, water companies tend to operate existing water storage tanks close to their full capacity, resulting in higher energy-related costs (since water has to be pumped into the tanks) and in larger volumes of retained water.

By being capable of characterising water consumption patterns in a given region water companies could adjust water volumes in existing storage tanks[1], improving the management of their finances and reducing excess water volumes retained in storage tanks.

## 2 Objective

The current project proposal comprises the following two major objectives:

- Prediction of urban domestic water consumptions in a given region of a median size city.

  That is, the current project aims at developing a system capable of processing input historical water consumption data (for a given region) and predict future water consumptions (in that region) in a given time horizon.

- Categorisation of urban domestic water consumptions in a given region of a median size city.

  That is, taking into account information about an entire civil year (from January 1st to December 31st) the system must be capable of identifying recurrent consumption patterns and determine the moments in time when such patterns were repeated: an intuitive (and somewhat expected) result is that water consumptions during summer months will have a similar shape, which substantially differs from those recorded during winter months. This could be justified by the fact that many people in that region usually spend summer months away, on vacation.

In this sense, the appropriate steps of a pattern recognition system will be considered, namely: (i) data pre-processing; (ii) dimensionality reduction (feature extraction and selection); (iii) pattern recognition techniques - in this case clustering and regression/prediction techniques; (iv) analysis of the experimental results.

[1] For obvious reasons this volume must always be overestimated; however, such overestimation could be controlled.

# 3 Dataset Description

The dataset to be used in the proposed project contains domestic water consumptions in different regions of a median size city. The available data corresponds to the entire 2016 year and was obtained as a result of a collaboration with *Águas de Coimbra*[2], the public water company in the city of Coimbra, Portgual.

Water consumption's measurement is carried out in strategic locations of the city, through measurement and control zones - *ZMC*[3]. These infrastructures provide a more efficient management of the water distribution systems, enabling the network's management in logical zones of analysis and allowing the adoption of measures to control water losses.

Collected data corresponds to a time series of total water volume distributed to the region in question - recorded as a floating point number whose units are $m^3/h$. Therefore, pairs of values $(time, volume)$ are provided. The collection time is provided in the format *dd/mm/yyyy HH:MM:SS*.

During the data acquisition equipment errors and malfunctions can occur, resulting in the failure to record the total consumptions. In such moments, a value of *"n/a"* is presented for the total distributed water volume, signalling the missing value. The developed pattern recognition system must adopt proper techniques to deal with these missing values.

# 4 Methodology

To accomplish the two mentioned objectives the following steps will be considered in the work to be performed:

1. Missing values imputation
2. Dimensionality Reduction
3. Prediction of future water consumptions
4. Segmentation of water consumptions
5. Results analysis and assessment

Regarding the first task, the goal is to apply missing values imputation techniques suitable for univariate time series. According to Moritz et al. (2015), three main categories of methods can be identified: (i) *univariate algorithms* - comprising simpler methods such as mean and median imputation; (ii) *univariate time series algorithms* - containing methods such as linear interpolation, ARIMA and SARIMA models, or the application of Kalman filters; and (iii) *multivariate algorithms applied on lagged data*. Additional methods can also be considered, such as *K-NN*, *Expectation-Maximisation*, *Artificial Neural Networks*, *ARMA* and *ARMAX* models.

Surveying the literature in the topic of missing data imputation for univariate time series, namely in the works of Abreu et al. (2012) and Moritz et al. (2015), *univariate time series algorithms* produced interesting results, namely linear interpolation and ARIMA techniques. Therefore, such techniques will be subject of attention in this work.

When performing clustering, classification and/or regression tasks the dimensionality of the data used must also be subject of attention. Indeed, distance calculations using raw data can be computationally expensive if a substantial number of dimensions (features) is being considered for each data point. Furthermore, in the case of time series processing (but not only restricted to this problem) when measuring the distance between raw time series highly unintuitive results may be obtained due to the high sensitivity of some distance metrics to distortions and noise in the data.

Among the scientific community, dimensionality reduction is often achieved by the application of feature extraction and selection algorithms. Regarding feature extraction methods, *Principal Component Analysis (PCA)* and *Self-Organising Maps (SOM)* have been applied to univariate time series Calabrese et al. (2010), Abreu et al. (2012), Aghabozorgi et al. (2015). Other popular

[2] `http://www.aguasdecoimbra.pt/`

[3] Or in portuguese, *Zonas de Medição e Controlo*.

methods, namely *Stacked Autoencoders* Vincent et al. (2010) can also be considered for this task. With respect to feature selection methods, *filter* and *wrapper* methods have been proposed.

In the topic of time series prediction several regression techniques have been proposed. Based on the studied literature in this topic, the application of *Artificial Neural Networks*, namely *Recurrent Neural Networks (RNN)* Connor et al. (1994), ElSaid et al. (2016) have been successfully adopted for several years. Other techniques, such as *ARMA* and *ARIMA*Zhang (2003) models can also be found in the literature, as well as the application of *Support Vector Machines* for regression purposes Tay and Cao (2002), Kim (2003).

By its turn, surveyed works on segmentation of univariate time series show a clear tendency towards the application of clustering methods, namely *Hierarchical* and *Partitioning* (HC and PC, respectively). Within HC, agglomerative methods appear to be more popular; among PC methods, K-Means Clustering seems to be the most popular choice. The combination of both HC and PC methods is also popular, namely Hierarchical Clustering and K-Means Clustering Abreu et al. (2012).

Finally, the last task to be performed has to do with the assessment and analysis of the obtained results. With respect to the third task, cross validation evaluation metrics will be adopted; regarding segmentation of water consumptions, typical metrics to assess the quality and tendency of the identified clusters are intended to be applied, such as the *Silhouette* coefficient, *Calinski-Herabaz* index, *Hopkings Statistic* and the *Elbow* method.

## 5 Conclusion

The proposed project aims at predicting and categorising water consumptions in a given region of a median size city. To accomplish this goal the tasks presented in section 4 were defined and are intended to be performed.

In both prediction and categorisation tasks, data pre-processing is an extremely important task since inadequate data can strongly condition both the computed regression (or prediction) model and information categorisation.

Characterisation of water consumptions, by means of both prediction and categorisation, can play a crucial role in the management of a city's water resources: water companies can make use of such knowledge to manage water in storage tanks, minimising excess water volumes retained in such tanks.

## References


Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. Comparison of different methods for univariate time series imputation in r. *arXiv preprint arXiv:1510.03924*, 2015.

Joana M Abreu, Francisco Câmara Pereira, and Paulo Ferrão. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and buildings*, 49:479–487, 2012.

Francesco Calabrese, Jonathan Reades, and Carlo Ratti. Eigenplaces: segmenting space through digital signatures. *IEEE Pervasive Computing*, 9(1):78–84, 2010.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

AbdElRahman ElSaid, Brandon Wild, James Higgins, and Travis Desell. Using lstm recurrent neural networks to predict excess vibration events in aircraft engines. In *e-Science (e-Science), 2016 IEEE 12th International Conference on*, pages 260–269. IEEE, 2016.

G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

Francis EH Tay and LJ Cao. Modified support vector machines in financial time series forecasting. *Neurocomputing*, 48(1):847–861, 2002.

Kyoung-jae Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1):307–319, 2003.