Categorisation of Urban Water Consumptions

Joaquim Leitão*
*jpleitao@dei.uc.pt
CISUC, Department of Informatics Engineering, Univesity of Coimbra, Portugal

Abstract—Lol, here goes the abstract

Index Terms—Water Consumption Patterns, Clustering, Time series, Pattern Recognition.

1 Introduction

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.

The current work aims at categorising water consumptions in a given region of a median size city, contributing towards an improved understanding and knowledge of water consumption patterns. Therefore, in this work, historic water consumptions are intended to be processed and organised in different groups according to their similarity.

In the field of pattern recognition, tasks of this nature are referred to as *clustering* approaches. Since a time-ordered series of data (water consumptions, in this case) is intended to be processed and clustered, the proposed problem can be considered as *time series clustering*. In this sense, it is important to define this term. We present a definition proposed by Aghabozorgi *et al.* [1]:

Definition 1 (Time Series Clustering)**.** *"Given a dataset of n time series data* $D = \{F_1, \cdots, F_n\}$*, time series clustering can be defined as the process of unsupervised partitioning of* D *into* $C = C_1, \cdots, C_k$ *in such a way that homogeneous time series are grouped together based on a certain similarity measure."*

The remainder of this document is organised as follows: Section 2 presents the main objectives of this work. In section 3 water consumption data used in this work is briefly presented. The methodology adopted throughout this work is discussed in section 4 and an overview of the experimental setup is provided in section 5. Sections 6 to 9 cover the main steps of the work, presented in the methodology section. Finally, section 10 concludes this document.

FIXME: Review outline

2 Objectives

The main objective of the proposed work is related with the 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) pattern recognition techniques - in this case clustering techniques; (iii) analysis of the experimental results.

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*.

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

2. http://www.aguasdecoimbra.pt/

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

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 mentioned objectives the following steps were considered in the developed work:

1) **Time Series pre-processing**, comprising two major tasks: computing the unit of analysis for the time series; and imputing missing values
2) **Time Series Representation**, consisting in the application of dimensionality reduction techniques. As will be explained later in this section, both raw time series data and dimensionality reduced time series data is intended to be clustered, in order to compare the results obtained with both methods.
3) **Segmentation of water consumptions**, consisting in the application of clustering techniques to group similar samples together.
4) **Results analysis and assessment**

Extensive research has been made in the field of time series clustering, resulting in the adoption and proposal of a wide variety of clustering approaches and techniques. In a 2015 survey on time series clustering, Aghabozorgi *et al.* [1] highlights techniques such as *hierarchical clustering*, *k-means clustering* and *k-medoids clustering*. Other techniques have also been applied, such as neural networks and variations of the mentioned algorithms, namely of *k-means* ans *k-medoids*.

Determining the most appropriate technique to be applied in our problem is far from being an easy task. The same methods applied on data of different natures can produce completely different results, and the same can even be considered with data of the same nature, but collected from distinct sources and (most probably) with a different methodology.

As a result, in the current work, different techniques are intended to be explored in an attempt to determine the ones that appear to be more suitable to the collected data. These techniques fall in four distinct categories: *Missing Data Imputation*, *Time Series Representation*, *Time Series Distance Metrics* and *Clustering Algorithms*.

Regarding missing data imputation, three distinct techniques are considered to be applied to provide sounding estimations for missing water consumptions data: fitting a polynomial expression, fitting an ARIMA model or the application of Kalman Filters.

With respect to time series representation methods, two alternatives are intended to be considered: on one side, the popular *raw time series* representation - where time series are represented as an ordered list of values collected over time; as an alternative, the applications of dimensionality reduction methods via *Principal Components Analysis* and *Stacked Autoencoders* is also explored.

Several clustering algorithms have been proposed and studied in the literature; however, based on the studied and referenced works, more traditional algorithms such as *Hierarchical* and *K-Means* clustering still remain popular and valid choices among researchers in this field. In this sense, their performance when applied to our data will be assessed. Another popular algorithm applied in time series clustering problems will be implemented and evaluated: this is the case of the recently proposed *TADPole* [2] density-based clustering algorithm, featuring a pruning strategy that softens the computational burden of the application of traditional clustering algorithms to time series.

In the next section, Table **??** presents the different combinations of clustering algorithms and parameters ranges used. Subsequent sections provide a more detailed overview of the different steps followed in this work.

5 SUMMARY OF THE EXPERIMENTS

Colocar a tabela com label "experiments_summary_table"

6 TIME SERIES PRE-PROCESSING

As stated in section 4, time series pre-processing comprises two main iterations: computing the unit of analysis of the time series and imputing missing values in time series' readings. These steps will be covered in more detail in their own subsections.

6.1 Unit of Analysis

When identifying patterns of recurrent behaviour in time series data it is important to determine the periods of time with the higher contributions to the overall signal. In other words, the most intense and important periods of time are sought.

Similarly to the works of Abreu *et al.* [3] and Calabrese *et al.* [4], a *Fourier* analysis can be conducted on the collected time series data to determine the frequencies with higher contribution. In this sense, a *Fast Fourier Transform* (FFT) was performed on the time series data.

Analysing the results of such transformation, it was determined that the frequency with the largest contribution to the resulting signal was the frequency $0\ Hz$. Even though this may seem an unexpected result, a valid explanation can be found: by performing the FFT of a given signal, the frequency with higher contribution is usually an approximation of the fundamental frequency. As the signal being analysed (the time series data) is not periodic, its fundamental period is infinite, rendering its fundamental frequency to be $0\ Hz$. In such scenarios, the second frequency with higher contribution is usually considered as the unit of analysis. In this case such a frequency was around $24h$, suggesting this to be the unit of analysis.

With this finding, an aggregation of the time series values was conducted: the values of the original time series were recorded with a one-minute interval; since the unit of analysis is $24h$, each sample to be clustered contains readings of one day, with an interval of $1h$[4].

6.2 Missing Data Imputation

In almost all real-life applications, errors in the data collection procedures are prompted to occur, forcing data analysis and processing methods to properly deal with missing and invalid values. Substantial research has been performed on the topic of missing data imputation, resulting in different methods and techniques being proposed and tested in a variety of problems.

Nevertheless, choosing a valid missing data imputation technique to be applied in a given problem is far from being an easy task. The same method can have completely distinct performances when applied to different data, of diverse natures and/or collected in different ways. In the particular case of this work, no previous studies of the performance of missing data

4. Therefore, each sample consists in a list of 24 values, one per hour. These individual hourly consumptions were obtained simply by adding all partial consumptions in the same hour.

imputation methods had been conducted on the collected data. As a result, a choice was made to browse the literature on this topic and test the application of the most cited and used ones in the collected data.

Steffen *et al.* [5] studied the application of different missing data imputation methods on univariate time series. In their work, the authors obtained interesting results for approaches exploring linear interpolation, ARIMA and SARIMA models and Kalman Filters.

Alternative methods can also be considered. Usually more active in time series prediction and classification, techniques featuring artificial neural networks, decision tress or k-nearest neighbours [6], [7] can also be applied to compute estimations for missing values. In addition, traditional time series analysis methods such as autocorrelation and trend analysis have also been applied [3], as well as expectation-maximization and multivariate algorithms on lagged data [5].

Inspired by the results reported on the literature, and after an initial study of the different methods, a choice was made to select the following three missing data imputation methods for further analysis: (i) linear interpolation; (ii) ARIMA; (iii) Kalman Filter.

At this point, the goal of this analysis was to studying the performance of these methods when applied to the collected data. This study was conducted as follows: Initially individual samples were divided into samples that contained missing values and samples without missing value[5]. For each complete sample missing values were artificially introduced. Each method was then applied to the individual samples and the correctness of their imputations was assessed using the *root mean squared error* (rmse) metric.

Overall the Kalman Filter produced better estimates than the competing alternatives: An average rmse of about 263.83 was registered (against a rmse of about 571.37 for the ARIMA model and 12403.17 for the linear interpolation), and the Kalman filter produced better daily estimates (smaller rmse on individual samples) in 189 days (against 72 for the ARIMA model and only 4 for the linear interpolation).

Based on such results the Kalman Filter was chosen as the imputation approach for our data. Therefore, for each day where missing data were registered, a Kalman Filter model was computed and estimations for the missing data in that day were obtained.

7 Time Series Representation

When working with high dimensional data, computational requirements grow both in terms of memory and execution time: if each time series contains more values then it will require more space in memory and, therefore, if a larger number of time series is intended to be clustered or processed in some way substantially more memory may be required to load all the time series data. In addition, specially in the case of clustering, distance computation between data samples is often performed; if a larger number of samples is considered, each comprising several dimensions, distance computation between pairs of samples will take longer to compute, as will the overall process.

Another motivation for seeking alternative time series representations, as pointed out by Aghabozorgi *et al.* [1], has to do with the fact that when measuring the distance between raw time series samples highly unintuitive results may be gathered resulting in the clustering of series similar in noise, instead of shape. It must be noticed that the adopted distance metric highly influences the clustering results.

In this sense, several time series representation methods have been proposed in the literature. Aghabozorgi *et al.* presents a taxonomy comprising four main categories: i) Data adaptive; (ii) Non-data adaptive; (iii) Model-based; and (iv) Data dictated. The choice of time series representation has a big impact in the outcome of the clustering exercise, in both quality and execution time. Furthermore, time series representation also affects the choice of distance metric (also called dissimilarity function).

In the specific clustering problem being tackled in this work, where time series containing similar behaviours are intended to be grouped together, one of the major challenges to be face has to due with the dissimilarity function (that is, how to determine that two time series are close together). Even though it is computationally expensive, the *Dynamic Time Warping* (DTW) [8] distance metric is quite popular, having been extensively adopted when dealing with raw time series data [1], [9]–[11] and being considered a more robust distance metric for time series [12].

As the DTW targets time sequences, it can be concluded that a substantial number of approaches adopt a raw time series representation. Dimensionality Reduction (DR) techniques such as *Principal Components Analysis* (PCA) have also been explored [3] in an attempt to represent time series samples in fewer dimensions, thus reducing the computational burden of distance computation. As the application of DR techniques loses the temporal sequence of the data, alternative distance metrics need to be considered.

In light of the discussion carried out in this section, in the scope of the current work, it is intended to explore clustering performance obtained using two distinct time series representation approaches: (i) **Raw time series representation**, featuring the use of the popular *Dynamic Time Warping* distance metric; (ii) **"Dimensionality-Reduced" time series representation**, featuring the use of the classical *Euclidean* distance.

7.1 Raw Time Series Representation

With respect to the *raw* time series representation not much needs to be added to what has already been mentioned. Initially, this representation can be seen as considering raw collected time series values.

Taking into account the pre-processing steps mentioned in section 6, imputed time series corresponding to daily lists of hourly water consumption readings should be considered at this point; however, an additional step must be considered: when working with raw time series data, performing a *Z-Normalization* of the data can significantly improve distance computation results with the DTW metric [13].

As a result, when working with raw time series representation, a *Z-Normalization* of the data was conducted. The *Z-Normalization* can be defined as:

$$X = \frac{X - MEAN(X)}{STD(X)}$$

7.2 Dimensionality Reduction for Time Series

Regarding the application of dimensionality reduction techniques to time series data, two popular algorithms of this nature were compared: *PCA* and *Stacked Autoencoders* (SAEs). The comparison was based on the rmse on reconstruction and on visual inspection of the reconstructed series.

5. The reader must keep in mind that individual samples consist in a list with water consumptions for a given day, featuring hourly values.

In PCA the target number of features was selected as a function of the percentage of explained data variance. In this sense, principal components capable of representing 80% of the data variance were selected, yielding a reduced dimensionality of 3 features.

SAEs were then trained for dimensionality reduction, targeting the same number of features as PCA. Different network architectures were considered while training the SAEs (that is, different number of autoencoders with different numbers of neurons per hidden layer, different activation functions, etc); however, in the best scenarios, SAEs were only capable of getting close to PCA in terms of reconstruction error, producing reconstructed time series considerably more distant to the original ones than PCA.

It is also worth mentioning at this point that both PCA and SAEs approaches were tested with normalised data, using the *Min-Max* method. Figure 1 presents a reconstruction of water consumptions for one day using both a PCA decomposition and Stacked Autoencoders.

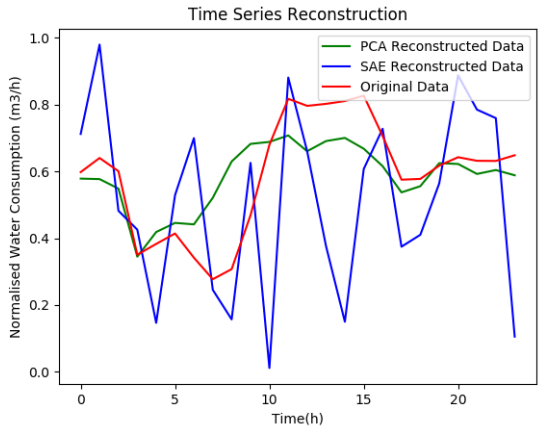

Figure 1. Reconstruction of water consumptions for one day using a PCA decomposition and a Stacked Autoencoder.

From the analysis of figure 1 it is clear that, even though SAEs may be able to reconstruct the original data with an error close to PCA, SAEs fail in capturing the shape and tendency of the series. As a proper learning could not be obtained with SAEs, PCA was the dimensionality reduction technique chosen to be applied in the experiments and exercises carried out in the remainder of this work.

8 Clustering

Dizer que escolhemos HC + K-Means **FIXME: TADPole?**

why and how!!!

Mostrar resultados HC

A major characteristic of K-Means clustering is that, depending on how the initial cluster centroids are computed, different results in each run of the algorithm (for the same value of the parameter K) may be produced. This is specially true if a random initialisation is performed.

As a result of this characteristic, K-Means clustering is usually executed several times for the same value of K and the initial cluster selection that produces better results is chosen. The question that immediately raises from such a statement is how to determine if one cluster selection is better than another? To answer this question, Anil Jain [14] proposed the execution of the K-Means algorithm several times, for the same value of k, keeping the partition that produced the smallest squared error.

In K-Means, clusters are sought so that their within-cluster distances are minimised and between-cluster distances maximised. Indeed, this is exactly the same evaluation performed in the computation of the silhouette coefficient. In this sense, Jain's suggestion was slightly adjusted, to keep the partition with the higher silhouette coefficient value, instead of the lowest squared error.

Therefore, for each of the tests presented in table **??** where the K-Means algorithm was used, 10 repetitions for the algorithm were considered for each value of k defined. After the 10 repetitions, only the partition with the highest silhouette coefficient value was kept.

9 Results Assessment

FIXME: Texto introdutório

9.1 Initial Results Analysis

Dizer que os resultados obtidos para cada uma das combinações de métodos apresentados anteriormente (que podem ser consultados na tabela da secção 4) foram avaliados segundo X métricas: **FIXME: Apresentar métricas num itemize e fazer uma breve descrição, colocar fórmulas, dizer que valores são bons, etc**

Fazer uma reflexão sobre os resultados e dizer quais aqueles que merecem ser mais estudados.

Atenção que analisei muita coisa com base no SC mas este pode nao ser o melhor método... Combinar um pouco a análise com o elbow method! (SC e um indicador normalisado, o que tem vantagens porque nao e tao sujeito a subjectividade como e o elbow method, mas por vezes em dados reais pode ser necessario experiencia e analise visual que o SC nao consegue traduzir)

FIXME: Não esquecer de falar do facto de o SC poder nao ser a metrica mais adequada para avaliar os clusters!

9.2 Further Results Analysis

Apresentar novas métricas, resultados e reflexão

Meter gráficos aqui

10 Conclusion

De uma forma geral o numero de clusters parece andar proximo de 4, pelo menos andamos sempre a volta desse valor...

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[2] N. Begum, L. Ulanova, J. Wang, and E. Keogh, "Accelerating dynamic time warping clustering with a novel admissible pruning strategy," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49–58, ACM, 2015.

[3] J. M. Abreu, F. C. Pereira, and P. Ferrão, "Using pattern recognition to identify habitual behavior in residential electricity consumption," *Energy and buildings*, vol. 49, pp. 479–487, 2012.

[4] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: segmenting space through digital signatures," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2010.

[5] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, "Comparison of different methods for univariate time series imputation in r," *arXiv preprint arXiv:1510.03924*, 2015.

[6] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[7] F. Yu and X. Xu, "A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved bp neural network," *Applied Energy*, vol. 134, pp. 102–113, 2014.
[8] S. Chu, E. Keogh, D. Hart, and M. Pazzani, "Iterative deepening dynamic time warping for time series," in *Proceedings of the 2002 SIAM International Conference on Data Mining*, pp. 195–212, SIAM, 2002.
[9] T. W. Liao, "Clustering of time series data—a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.
[10] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, Y. Chen, and E. Keogh, "Dynamic time warping averaging of time series allows faster and more accurate classification," in *Data Mining (ICDM), 2014 IEEE International Conference on*, pp. 470–479, IEEE, 2014.
[11] H. Izakian, W. Pedrycz, and I. Jamal, "Fuzzy clustering of time series data using dynamic time warping distance," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 235–244, 2015.
[12] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, pp. 1–35, 2013.
[13] A. Mueen and E. Keogh, "Extracting optimal performance from dynamic time warping," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2129–2130, ACM, 2016.
[14] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2009.