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. Sections **??** to **??** cover the main steps of the work, presented in the methodology section. Finally, section 9 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*.

During the data acquisition equipment errors and malfunctions can occur, resulting in the failure to record the total

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*.

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.

Distance computation between raw time series data tends to be an heavy task in terms of computational resources. The main reason for this has to do with the high dimensionality of time series data. In this sense, in some publications the application of dimensionality reduction techniques has been explored. In the scope of this work both raw time series data and reduced dimensionality time series data is intended to be applied. Regarding dimensionality reduction techniques, *Principal Components Analysis* and *Stacked Autoencoders* are intended to be explored.

When dealing with raw time series, the use of more traditional distance metrics, such as the euclidean distance, has been shown to fail in capturing similarities in the series' shape. Alternative distance metrics, among which *Dynamic Time Warping* (DTW) [2] is highlighted, have been extensively adopted when dealing with raw time series data [3], [4] and is considered a more robust distance metric for time series [5]. Nevertheless, euclidean distance is still applied in time series clustering works, specially if dimensionality reduction techniques are employed. An example of such application can be found in the work of Abreu *et al.* [6]. In this sense, DTW is intended to be used when dealing with raw time series data, whereas the euclidean distance will be the choice for the dimensionality reduced data.

Clustering Algorithm: **FIXME: Para já estou apenas a considerar HC + K-Means e apenas vario a metrica de distancia e a forma como calculo os centroides. No entanto posso vir a considerar um outro algoritmo, o TADPole, que aparentemente existe em R e é batante utilizado - Para já deixar este ponto em standby**

FIXME: Adicionar uma tabela com os grupos de testes que fizémos

5 Time Series Pre-Processing

Unit of analysis and missing data imputation

6 Time Series Representation

Começar a falar de raw time series data, dizer que é apenas o resultado de agrupar os dados originais na unidade de análise anteriormente calculada, aplicar método de missing values imputation e pronto.

Falar de PCA e Stack Autoencoders - Comparar performances, dizer que SAEs nao conseguiram aprender –¿ e deixar já nota para future work?

7 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 [7] 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 **?? FIXME: Referência tabela a adicionar na secção 4** 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 silouette coefficient value was kept.

8 Results Assessment

FIXME: Texto introdutório

8.1 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.

8.2 Further Results Analysis

Apresentar novas métricas, resultados e reflexão

9 Conclusion

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2009.