# Enhanced Water Demand Analysis via Symbolic Approximation within an Epidemiology-Based Forecasting Framework

Claudia Navarrete-López [1] , **Manuel Herrera** [2] , **Bruno M. Brentan** [3] , **Edevar Luvizotto Jr** [3] **and Joaquín Izquierdo** [4,*]

[1] Faculty of Environmental Engineering, Universidad Santo Tomás, Bogotá 110231, Colombia; claudianavarrete@usantotomas.edu.co
[2] Institute for Manufacturing—Department of Engineering, University of Cambridge, Cambridge CB3 0FS, UK; amh226@cam.ac.uk
[3] LHC—Faculty of Civil Engineering, Universidade Estadual de Campinas, Campinas 13083-970, Brazil; brunocivil08@gmail.com (B.M.B.); edevar@fec.unicamp.br (E.L.J.)
[4] Fluing—Institute for Multidisciplinary Mathematics, Universitat Politècnica de València, 46022 Valencia, Spain
* Correspondence: jizquier@upv.es; Tel.: +34-628-028-804

Abstract: Epidemiology-based models have shown to have successful adaptations to deal with challenges coming from various areas of Engineering, such as those related to energy use or asset management. This paper deals with urban water demand, and data analysis is based on an Epidemiology tool-set herein developed. This combination represents a novel framework in urban hydraulics. Specifically, various reduction tools for time series analyses based on a symbolic approximate (SAX) coding technique able to deal with simple versions of data sets are presented. Then, a neural-network-based model that uses SAX-based knowledge-generation from various time series is shown to improve forecasting abilities. This knowledge is produced by identifying water distribution district metered areas of high similarity to a given target area and sharing demand patterns with the latter. The proposal has been tested with databases from a Brazilian water utility, providing key knowledge for improving water management and hydraulic operation of the distribution system. This novel analysis framework shows several benefits in terms of accuracy and performance of neural network models for water demand.

Keywords: water distribution systems; epidemiology; time series analysis; pattern recognition; dimension reduction

1. Introduction

Water demand forecasting is of paramount importance for water supply planning and operation, to ensure continuous service at suitable levels of quantity and pressure. Accurate water demand models are thereby key for approaching intelligent asset management of such physical assets as pumps, valves and water-meters [1–3]. Also, thanks to deeper understanding of water demand, it is possible to better propose near real-time operation of urban water supply. As other forecasting problems, water demand forecasting is essentially a regression problem and has been approache by both single and multivariate regression models [4] and by classical time-series models [5]. These last models often consider time and calendar variables since users demand water usually at the same hours. However, they behave differently in weekends or holidays than in weekdays. The application of innovative techniques for short-term forecasting has gained force in the recent years to beat the

accuracy of classical short-term forecasting models [6]. This has been mainly due to the growth of the computer capacity and the development of efficient machine learning tools, within which artificial neural networks (ANNs) can be highlighted as an important approach to forecast water demand [7]. The use of extreme learning machines [8], adaptive neuro-fuzzy inference systems [9], and hybrid machine learning and optimisation techniques [10] can be presently considered state-of-the-art research in water demand. However, as the new techniques for demand forecasting evolve, the available data also grows exponentially. New challenges to address are, then, how to deal with big-size time series databases and how to get near real-time response from the predictive models, keeping the highest accuracy levels. Hybrid algorithms can improve the accuracy of forecasting tools, by better matching the climatic and social effects with the time-series trends [11,12]. In this direction, several authors have proposed hybrid methods highlighting the use of ANN and ARIMA models [13–15] or the use of machine learning methods linked to various transformation models, such as wavelets [16] or Fourier series [17].

This paper proposes the development of a novel approach on top of the previously mentioned plethora of possible methodologies to analyse water demand. The building blocks are tool-sets and methods adapted from epidemiological data analysis approaches (EDAAs). Our hypothesis is that Epidemiology-based forecasting methods, which consider spatio-temporal data to analyse a disease spread in a population, are suitable to provide useful knowledge in water supply. Working at different temporal and spatial scales for analysing the water demand of a water distribution system (WDS) makes it possible better control of the system [18,19]. This naturally considers the behaviour of the different types of consumers within each district metered area (DMA) [20]. In addition, taking into account the wide variability and spatial distribution of WDS areas, the analysis can be expanded further towards an integral analysis taking into account other urban utilities such as gas or electricity networks. Still, considering spatial relationships in water demand analysis benefits the provision of more accurate water balances, as the analysis can focus now on each of the DMA encompassing the WDS [21]. At the same time, it eases asset management and control of the system components and leak location and quantification, as well [22]. Water supply asset performance, which must meet users' demand behaviours, may take advantage of water demand patterns of customers with similar features or those who are under similar environmental conditions [23]. This provides a clear advantage regarding other methodologies for analysing water demand, as EDAAs provide a multidisciplinary framework to investigate customer water demand behaviour. EDAAs are not only an alternative to ARIMA approaches for water demand forecasting, but also to clustering and probability-based models for water demand modelling [24]. In addition, EDAAs naturally have the potential to host new capabilities to analyse various risks and vulnerabilities related to water distribution [25]. Reasonably, epidemiology-based models, given its all-encompassing nature, are susceptible to be extended to other various complex processes in urban hydraulics affecting water supply and drainage either individually or jointly [26].

Traditional approaches on Epidemiology are connected with health-related issues over populations and control of the various problems that arise in such a context [27]. Most of the key developments in the area of Epidemiology have been based on using different models and methods for data analysis [28]. Within these tools, network-based models provide a proper way to study a population and how the individuals are interrelated [29,30]. The nodes of the network represent individuals, and the links represent the interactions between individuals or the disease contagion, within this framework. It is interesting to note that similar network representations can be used in several contexts, such as transport and utility networks [31,32], communication networks [33], and social networks [34].

Data analysis methodologies and tools adapted from Epidemiology have shown to be useful in various Engineering applications. This is the case of energy and buildings [35], hydraulics [36,37], and communication systems [38], among others. In energy related issues [39], Epidemiology provides a holistic approach involving professionals in various areas: physicists, engineers, sociologists and

economists, in an interdisciplinary work that leads to consider trends and patterns of energy demand. This takes to strengthening the evidence base for policy makers and to assessing the impact of regulatory actions regarding consumption. The application of methodologies and tools inherited from data analysis in Epidemiology into hydraulics is a hardly investigated approach. However, Bardet and Little [36] pioneered this idea by establishing a parallelism between outbreaks of clustered pipe breaks and human mortality observed during extreme climatic events, such as heat waves [40].

The aim of this work is to provide a suitable framework for urban water distribution system (WDS) management by considering simultaneously multiple network areas within a feature-sharing framework. One of the main advantages of this proposal is to support further methodologies to analyse water demand at several supply areas, and to provide a way of comparison of water use and costumer behaviour depending on the characteristics of each zone [41].

This paper proposes to address the necessary computations through Symbolic-Aggregate-approXimation (SAX) [42] of the individual values of each of the target time series using sequences of symbols. SAX was initially developed to reduce the dense, lengthy information in a time series into chains of characters built to be shorter than the original time series. SAX has also been found useful for various data mining tasks, in particular, indexing [43], clustering [44,45], and classification [46]. The main vocation of SAX-based methods is to provide a suitable avenue for pattern recognition models associated with time series [47].

The SAX output is a string of characters or "word" approximating the time series [48]. It is thereby essential to work with appropriate methods to capture patterns formed by words characterising and representing the time series data. Suffix trees [42,49] have been widely used for this purpose, and, thus, provide a way to organise sub-strings of SAX words, which eases pattern extraction processes. Suffix trees have various applications [42,49], because of their ability to provide linear-time solutions to complex string configurations [50]. One of the most important problems approached by suffix trees are those related to DNA contamination issues [51] supporting the finding of unexpected patterns in large DNA chains. The compression capability of suffix trees is so efficient that enables to effectively map even the human genome, which is composed by an order of 20,000 protein-coding genes [52,53] and supports searching procedures within millions of nucleotides present in human genome studies [54].

The proposed EDAA working framework, which uses a joint SAX-suffix tree methodology, has been tested using the WDS utility of Franca, in Brazil, to improve demand forecasting. The database under study comes from four sectors of Franca's water network. By adapting EDAAs to analyse these four time series of water demand, better insight of the underlying relationships between their corresponding supply areas is provided. This is useful to develop enhanced forecasting analysis methods, and to speed up operation and control of the water system through better detection of anomalies. In addition, this eases the study on how the different parts of a system respond in front of undesirable events, something obtained by analysing the relationships between exposure and outcome, within a holistic urban water management context. The new framework has shown great suitability for improving water demand predictive models. This is achieved by reinforcing the predictive ability for a WDS area using the feedback from another areas that have shown similarities in terms of their respective demands.

The structure of the paper is the following: Section 2, a fully methodological section, introduces epidemiology-based forecast methods, which are consequently adapted to analyse urban water demand; SAX is also proposed as a useful tool for this aim; the resulting SAX word strings are analysed through a pattern discovery process based on suffix trees. Section 3 presents the case-study of a utility network from which water demand of various sectors is analysed. After presenting the main results for this case-study, Section 4 provides a discussion on the main findings of the current research. Finally, Section 5 closes the paper putting forward conclusions and further work.

2. Epidemiology-Based Data Analysis for Water Demand

Predictive models based on Epidemiology data analysis often include the evaluation of multiple relationships between two or more time series. This is the case of understanding how data changes over time and how to suitable apply forecasting methods. When working with multiple time series models it is usual to establish associations between exposure to a threat and health outcome. For example, from the study of associations between exposure to such elements as air pollution, weather variables or pollen, important results regarding ways to show disease symptoms and consequences may be derived [55,56]. In water supply, similar associations show, for instance, how high temperatures impact on progressively producing higher levels of water demand for a certain population [57]. As a consequence, it is useful to investigate cause-effect relationships between exogenous variables and levels of water demand.

The databases for carrying out these studies are naturally available at regular time intervals. This explains why time series analysis plays such an important role in Epidemiology. The main aim is to explain changes in the main time-dependent variable of the process through changes in the levels of the other variables under analysis. This makes that Poisson regression and ARIMA models have been widely used within this research area [58]. However, as the aim of this work is to find relationships between time series concerning exposure and outcome, other more complex models such as multivariate ARIMA and dynamic models have also been commonly used [59]. It is also usual to work with transfer functions and intervention models, whose use and performance is frequently limited by the complexity of such models [60].

EDAAs ease the approach of multiple water demand analysis when considering the various district metered areas (DMAs) in which a WDS is often partitioned [21]. As Epidemiology methods model cause-effect relationships, EDAAs are easily extended to investigate the impact on water consumption related to the effects of events ranging from valve operations to policies related to extreme weather conditions. EDAA forecasting models make it possible to disaggregate those previously mentioned effects for the individual DMAs [61]. This is done by a patient-level data analysis that allows to develop a different predictive model per DMA. The approach is naturally completed taking into account the existing correlations between DMAs of the same WDS [62,63]. EDAAs provide, in this way, a spatial sense of analysis to the common temporal approach when approaching water demand modelling at urban level. This is the main idea developed in this paper.

2.1. Epidemiology-Based Forecast and SAX

This paper proposes a symbolic aggregate approximation of time series [42] to represent the time series as a sequence of symbols. SAX is a suitable methodology to work under an epidemiology-based framework. The reason is that SAX can be used as a tool to compute distances between time series, to study how various time series are related among themselves (similar patterns), and to ease the analysis of large time series.

2.2. Introduction to SAX

SAX was primarily developed to reduce the dense, lengthy information of a time series to a shorter sequence of characters. SAX has also been found suitable for various data mining tasks, including indexing [43], clustering [44,45], and classification [46]. SAX follows three main steps:

1. Divide a time series with l data into w segments of equal length (number of data), l/w.
2. Compute the arithmetic average of the time series data for each segment.
3. Build proxies of these average values using symbols from an alphabet of size m.

In the first step, a Piecewise Aggregate Approximation (PAA) process is performed. Then, the average for each segment in which the time series splits is obtained - step 2. Finally, SAX associates these average values with symbols, where each symbol represents a certain grade of the average value computed per segment - this is step 3. SAX is based on the assumption that time series values

follow a Normal distribution for each of the segments into which the PAA has divided the series. On computing the final average values, $w - 1$ breakpoints, $\gamma_1, \ldots, \gamma_{w-1}$, that divide the area under the Normal distribution into w equi-probable areas are considered. For each area the arithmetic mean value is, then, obtained [42]. In a more formal way, let $Q = (q_1, \ldots, q_l)$ be a vector of real numbers that normalizes the time series of size l. The PAA representation of Q is a new vector $\bar{Q} = (\bar{q}_1, \ldots, \bar{q}_w)$ of size w whose ith element $\bar{q}_i$ can be computed as in Equation (1):

$$\bar{q}_i = \frac{w}{l} \sum_{j=1+(i-1)l/w}^{il/w} q_j. \tag{1}$$

After the third step, a sequence of symbols (sequence of letters from an alphabet, Σ, of size $m > 2$) or "word" is composed. Each symbol or letter represents a grade for the average at each segment of the original data, i.e. the values obtained in Equation (1). So, using an alphabet, a word has as many letters as segments in the PAA. One cand note by $\hat{Q} = \hat{q}_1, \ldots, \hat{q}_w$ the word obtained this way.

Figure 1 presents a simple example of the output of the SAX process representing a 128-long time series into a word of just 8 characters: `baabccbc`, which uses and alphabet of three symbols, $\Sigma = \{\texttt{a},\texttt{b},\texttt{c}\}$.
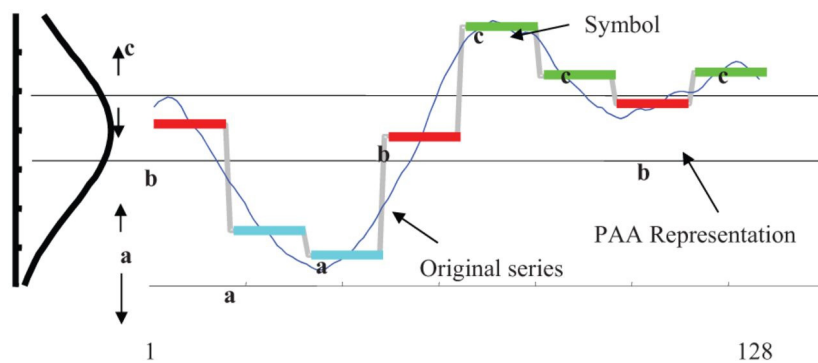

Figure 1. SAX representation of a series adapted from [42].

An important characteristic of SAX is to have associated a metric, a distance between each two SAX-coded time series, which is a lower bound of the distance between time series when computed in the original space. The traditionally used distance between two SAX words is the so-called $MINDIST$, and entails looking up the distances between each pair of symbols or letters. Given two time series Q and C of the same length l, Equation (2) shows how $MINDIST$ is defined:

$$MINDIST(\hat{Q}, \hat{C}) = \sqrt{\frac{l}{w}} \sqrt{\sum_{i=1}^{w} (dist(\hat{q}_i, \hat{c}_i))^2}, \tag{2}$$

where w is the number of segments into which both time series are divided, and $\hat{Q} = \hat{q}_1, \ldots, \hat{q}_w$ and $\hat{C} = \hat{c}_1, \ldots, \hat{c}_w$ are the words into which both time series, Q and C, are respectively encoded using Equation (1). The function $dist()$ can be implemented using a table look-up containing pre-computed distances between letters of the SAX alphabet in use $\Sigma = \{\alpha_1, \ldots, \alpha_m\}$. Those are minimum distances between letters and so $MINDIST$ also returns the minimum distance between the time series of the original values through their corresponding words. Specifically, the distance between two symbols $\alpha_i, \alpha_j \in \Sigma$ is calculated by Equation (3),

$$dist(\alpha_i, \alpha_j) = \begin{cases} 0, & \text{if } |i - j| \leq 1. \\ \beta_{\max(i,j)-1} - \beta_{\min(i,j)}, & \text{otherwise.} \end{cases} \tag{3}$$

where β is the value of the standardised Normal at the corresponding break-point of the PAA configuration.

Working with SAX opens multiple research avenues on methods such as hashing [64], Markov models [65,66], and suffix tree methods [49,67], as considered next. In addition, SAX naturally has an associated sliding window architecture in which every time-frame is encoded by a letter. This will aid to face further challenges for time series analysis.

2.3. Pattern Recognition in SAX Words

A table containing an entry for each SAX word can be a useful tool to count the number of occurrences of certain sub-strings. After creating and saving a suitable table, elements of the same size of the SAX string for which patterns are found out can be easily identified. Observe that any query in the table takes constant time [50]. A convenient way to organise a directory of sub-strings from a word is the technique called suffix tree [68]. The suffix tree is a type of digital search that represents a set of strings over a finite alphabet. Suffix trees are ideal for mining periodic patterns [49] as those commonly found in urban water demand. In addition, the suffix tree technique supports finding anomalies or surprising patterns, as well, which can be understood as consequences of operation manoeuvres in water supply (open/close valves, e.g.,) or the presence of disruptive events (pipe bursts, e.g.,).

A formal definition of the suffix tree, T, associated with a string x of n characters drawn from an ordered alphabet Σ is the following. The string x is first noted as $x[1:n]$, and any sub-string of the form $x[i:n]$ will be called a suffix of x. Let $ be a special character, matching no character in Σ. The character $ is a right end-marker, and its purpose is to separate in T suffix $x[i:n]$ from suffix $x[j:n]$, for $i > j$, whenever the former is a prefix of the latter. This ensures that no suffix is a prefix of another suffix, and that there will be n leaf nodes, one for each of the suffixes of T. Consequently, each leaf of T is labelled with a distinct integer j such that the path from the root to the leaf j corresponds to the suffix $x[j:n]$ [69]. Table 1 provides a summary of the main steps for the suffix tree creation process.

Table 1. Suffix tree construction. An overview.

Let T be a Tree Corresponding to a String x of Length n:
1. T has exactly n leafs, numbered from 1 to n
2. Every edge has a label, which is a sub-string of x
3. Every internal node has at least 2 children
4. Labels of two edges starting at an internal node do not start with the same character
5. The label of the path from the root to a leaf numbered i is the suffix of x starting at position i, i.e., $x[i, \dots, n]$

There are compact representations for the suffix trees that allow encoding a large number of different sub-strings, handling at most $2n$ nodes [70]. This is important as this fact enables indexing large words through several operations of order $\mathcal{O}(n)$.

The decision to take the suffix tree method for pattern extraction form SAX comes from its computational efficiency, the representativeness of the patterns it manages to find, and its huge applicability. This have led this technique to play a central role also in a most complex practice of pattern extraction from strings, such as the case of the human genome [52].

As an example of how the suffix tree process works, the word `cabdcabb` built using the alphabet $\Sigma = \{$`a,b,c,d`$\}$ is proposed. Figure 2 shows the process of pattern extraction for this example.

In the example of Figure 2, the first action is to take the smallest suffix `b$`, starting at the right of the word. After creating a root, a first branch related to `b$` is added to the root and indexed by 8 which is the original position of the suffix in the word. Moving to position 7 of the word one founds the suffix `bb$`; as a branch `b` already exists, another branch starting from the first is added (see step 2 of Figure 2) and the previous one is split giving rise to another one with the end-marker `$`. The next suffix is `abb$`. There is not branch representing `a`. A new branch starting from the root is created to represent the whole suffix, carrying out the index 6 (see step 3 of Figure 2). The process is iterated following the

theoretical approach explained above until getting each suffix in *T* conveniently indexed (step 8 of Figure 2).
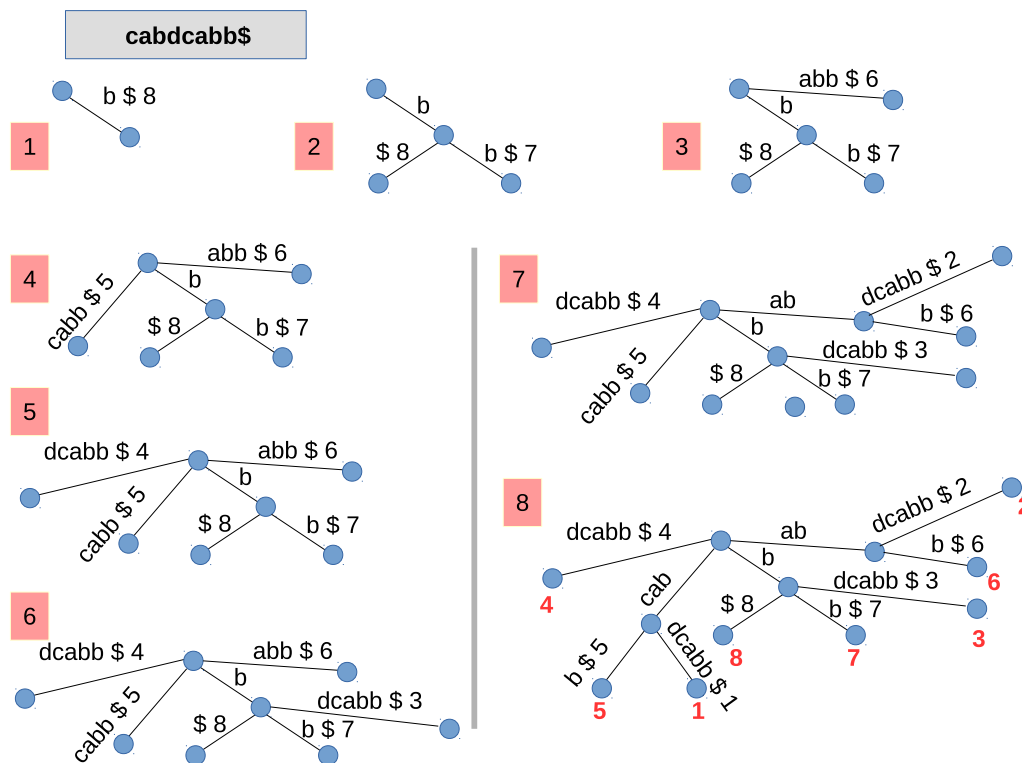

Figure 2. Example of a suffix tree.

3. Analysis of Urban Water Demand

This paper computes SAX for several water consumption time series that correspond to various DMAs composing a utility WDS. The aim of the proposal is to use SAX to extract patterns from those time series. As already seen, SAX also supports computing distances between time series. This makes it possible to study different water uses and pattern demands depending on the DMA and the customers' characteristics. The process brings a practical application for the SAX epidemiology-based forecast model introduced in Section 2.

3.1. An Epidemiological Framework for Water Demand Analysis

Epidemiology forecasting often works with a spatial analysis to supplement the time series analysis. In this paper, we claim that based on similarities that can be found among various DMAs, it is possible to improve water demand background knowledge for a specific target DMA. This is achieved by considering the information found on those DMAs exhibiting similar demand patterns to the targeted DMA. These similarities are based on the above described methods regarding SAX distance and suffix trees. There are several alternative ways to approach distances or (dis)similarities between time series [71]. In many cases, these alternatives are based on algorithms developed for whole time series clustering by taking the main characteristics of the static clustering algorithms. Then the process continues by modifying the similarity definition to a new appropriate one, or by applying a transformation to the time series so that static features are obtained [72]. The method in which this paper lies has the advantage of working well with large time series databases since the dimension of the problem is naturally reduced by the SAX related methods. The overall process is shown in the flowchart of Figure 3, where DMA$_i$ is targeted for improving the predictive accuracy on its demand.
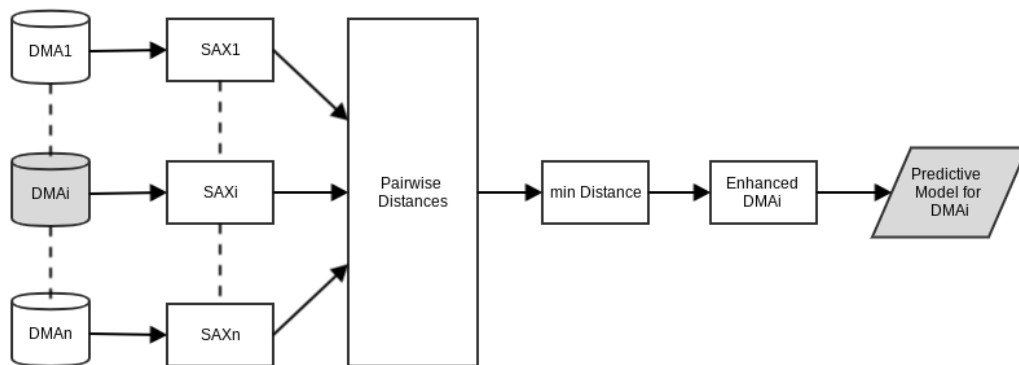
Figure 3. Flowchart of the overall process for enhancing DMA-demand predictive models.

The enhancement prediction process shown in Figure 3 starts by selecting a DMA to be targeted. Then, each water demand time series is transformed into SAX for all the DMAs. All the pairwise distances between the target and the other DMAs are evaluated. The closest DMA to the target is selected to enhance the information of the predictive model on the targeted DMA. This process has a clear advantage over just one single predictive model and, in addition, shows one way to apply the knowledge produced by SAX.

An artificial neural network [73] is the model chosen herein to show the water demand prediction improvement. An ANN is an interconnected group of neurons [74] that execute non-linear computations based on the input values, and the resulting output is used to fed other neurons. Neurons are usually arranged as a series of interconnected layers, as in the so-called multi-layer perceptron neural network used here. Hidden nodes with appropriate nonlinear transfer or activation functions are used to process the information received by the input nodes, each one associated with one of the predictors. Usually, a three-layer feed-forward model is used for forecasting purposes. Based on the data presented to the network a backpropagation error algorithm is used to iteratively adjust the neuron connection weights in such a way that the predictive performance of the network is improved. In this case-study, PAA values of the historic water demand time series at various time lags is the input presented to the ANN.

3.2. Case-Study Description

In this Section, the presented methodology is applied to four DMAs of the Franca WDS. Franca is Brazilian city located to the West of the country (see left map in Figure 4, and a detail on the right inset). Franca is a city approximately 315,000-strong.

The study on these four DMAs takes into account spatial distribution and hydraulic properties of water demand on the whole WDS. Figure 5 shows how Franca WDS is partitioned into several DMAs. Such a partition was performed considering pressure conditions for the main areas. The grey blocks in Figure 5 correspond to the DMAs involved in the case-study of this work. In this WDS, water is supplied from a single water treatment plant to all DMAs by a main pipe line. The most important tanks for water distribution are highlighted in the Figure, and are responsible to control and regularise the supply for the DMAs that integrate the entire WDS.

Three out of the four DMAs chosen for analysis are of similar size regarding the number of demand nodes. Specifically, the number of household connections for SA-3 is 2168, for SL is 2506, and for SA-ZA is 2728. These three DMAs are urban districts with small businesses, a common configuration for Brazil's residential areas. The fourth DMA involved in the case-study is bigger than the other three. It is called ETA, has 10,439 household connections and also supplies a prison; thus, it is a sector of particular importance for the city water management. The time series for each DMA

under study collects 4000 h water demand consumption data metered in litres per second. So, the data of the case-study covers a period of approximately five months.

Figure 4. Location map of the city of Franca, Brazil.
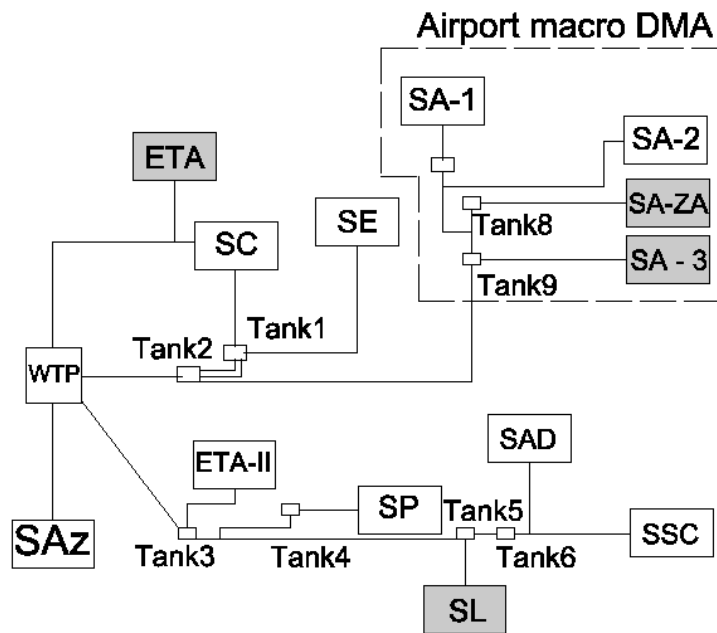

Figure 5. DMA layout of Franca's WDS.

3.3. Results

The dimension of the time series corresponding to the four DMAs of the case-study is reduced using SAX. Figure 6 shows the performance of the PAA tuning process. The reduction selection considers: (a) the accuracy of the PAA, in terms of the sum of squares—SS—computed as the distance between the arithmetic average for a PAA and the original values of the time series; and (b) the number reduction, represented by the number of segments. Under both criteria, it is taken 400 as the number of segments forming the PAA partition, with a dictionary or alphabet of four letters {a,b,c,d}. The letters are related to the various levels of water demand values in the time series; where a is the lowest level and d is the highest.
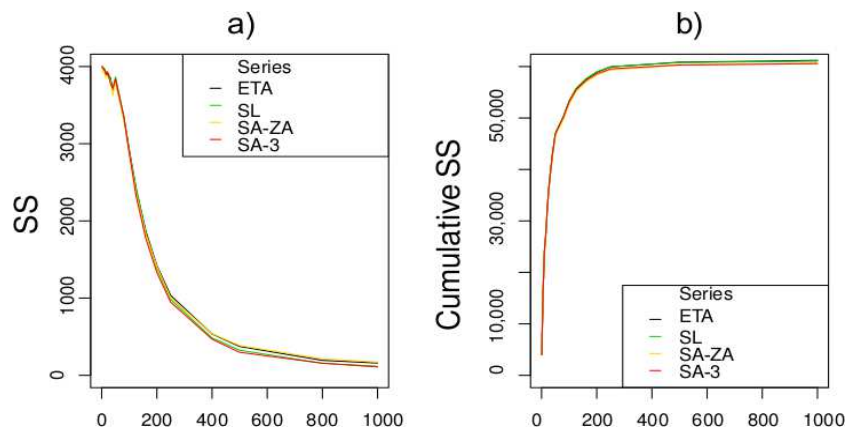
Figure 6. Number of segments for the PAA configuration; (**a**) SS vs. number of segments; (**b**) cumulative SS vs. number of segments.

The accuracy of the various PAAs is better explained by Figure 7. This Figure shows various levels of PAA reconstruction of water demand for the case of the district SL: Figure 7a–d show PAAs for the time series reconstructions into 50, 100, 200 and 400 segments, respectively. This is a good representative example of how the accuracy level of a PAA depends on the number of segments in which the time series is divided, ranging from low fidelity time series reconstruction for a PAA with 50 segments (Figure 7a) to the higher PAA fidelity when using 400 segments (Figure 7d).

Table 2 shows the *MINDIST* distances between the four DMAs when using 400 segments for computing the PAA, as this is several segments suitable to explain most of the variability found in the time series under study, according to both Figures 6 and 7. SL, SA-ZA and SA-3 are the DMAs which are closer according to the *MINDIST* distance. Furthermore, it is worth noting that the SAX word for the water demand time series corresponding to SA-ZA can be directly obtained by combining the other two DMAs.

The sub-string `aaddcdbaaddddbaacdddcaa` is the longest SAX pattern common to all the series found through suffix trees applied to the SAX words. Considering that each letter represents values for 10 water consumption registers, this pattern shows a similar tendency for 230 h of the original time series (it means over 9 consecutive days with water demands of similar level at all the DMAs). Other shorter patterns have also been found in the SAX words of the case-study and each one is found out at least twice within each series. Similar patterns to those found in SL are in both SA-ZA and SA-3.

One of the main hypotheses of this paper is to combine these similarity patterns between DMAs with those obtained from *MINDIST*. This would help support further predictive methods based on available historical data by using information on similar water consumption behaviours in other areas. Our claim is that this would improve the accuracy of predictive models and, in addition, managing conditions of similar demand response in different DMAs would ease the detection of abnormal readings at flow-meters, or even the presence of unexpected water uses or leaks. This idea (as described by Figure 3) is proposed to improve the predictive model for the DMA SL using some *MINDIST*-close DMA. In this case, SA-3 is chosen, instead of SA-ZA (which is at null distance), for diversity reasons. Please note that using SA-ZA would have led to the use of too similar (almost identical) features, and this would not have enriched the process. In contrast, SA-3 introduces increased variability to the process.

The ANN herein considered to work along with SAX has one hidden layer, and neurons use log-linear activation functions. Several configurations were tested through combination of the typical ANN parameters: 3, 6 and 9 nodes for the hidden layer; and 0.01 and 0.1 for the learning rate parameter. The 2250 values per DMA of the data series have been chopped down to 500, 400 and 200 values per time series in order to retrain the ANN at different samples. This allows to work with the accuracy average values for the models instead of using just one single value. The input variables for the ANN

are in all cases measures of water demand 20 min, 40 min, and 1 week ahead of the current time *t* (output). The databases have been split into partitions of 70% for training and 30% for validation.
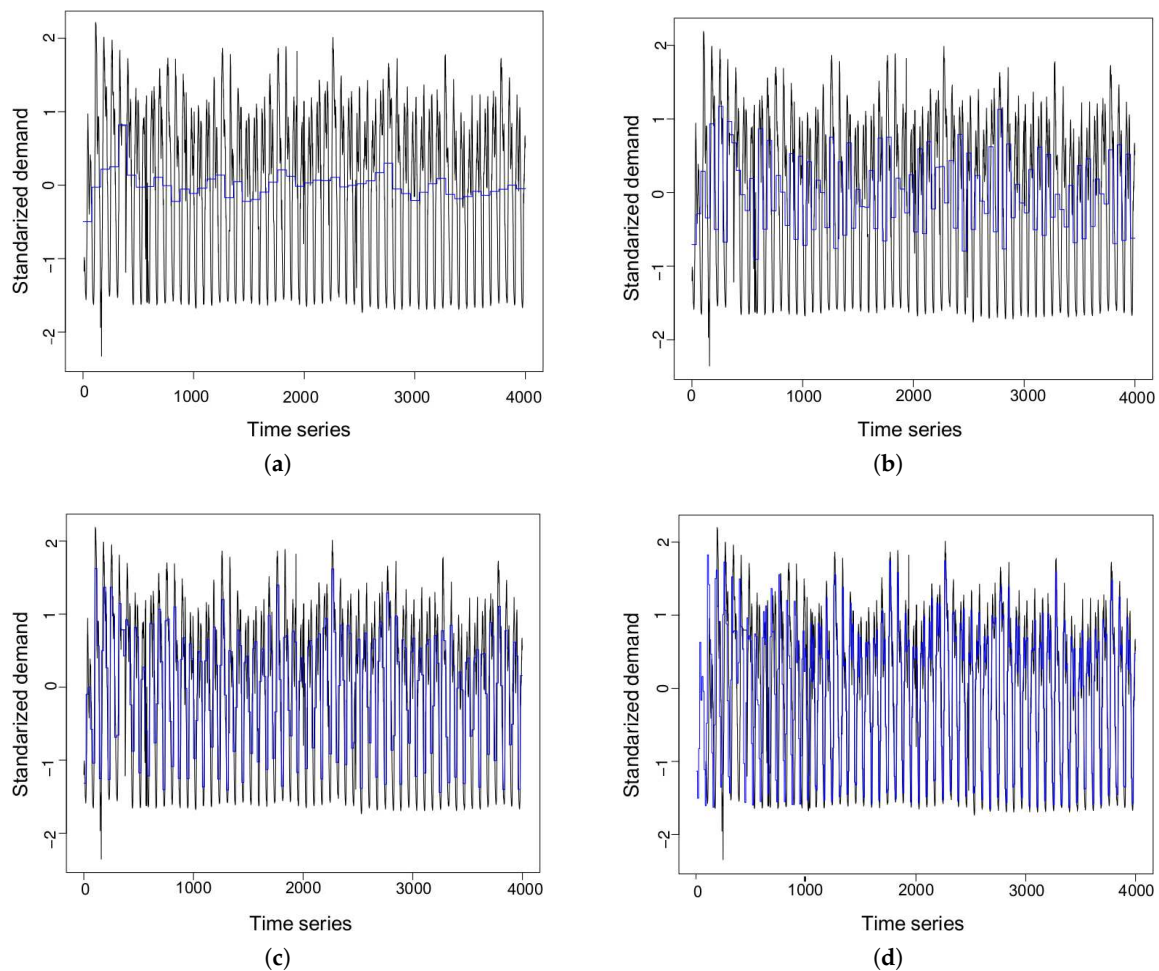
Figure 7. Sensitivity of PAA reductions. Case of SL's DMA. (**a**) SL DMA—Original water demand values and PAA computed through 50 segments; (**b**) SL DMA—Original water demand values and PAA computed through 100 segments; (**c**) SL DMA—Original water demand values and PAA computed through 200 segments; (**d**) SL DMA—Original water demand values and PAA computed through 400 segments.

Table 2. $MINDIST$ for the 4 DMAs of the case-study.

|	ETA	SL	SA-ZA	SA-3
ETA	0	–	–	–
SL	4.77	0	–	–
SA-ZA	5.22	0.00	0	–
SA-3	4.27	2.13	2.13	0

The results obtained from the errors for the DMA SL and the DMA combination SL + SA-3 show a stable behaviour, with prediction errors as expected, namely, the values obtained in the prediction of the networks combining the two DMAs show consistently lower values in all the cases. A *t*-test was performed to compare the error means for each prediction: SL and SL + SA-3 obtaining a *p*-value = 0.0033, which shows significant differences between the means of the two groups. The global average error using the single DMA SL is 1.0106 while for the combination of both DMAs it

is 0.8779. Table 3 summarises these results; *t* is the test statistic and *df* stands for degrees of freedom. The confidence interval indicates 95% of the sample lying between its bounds.

Table 3. Mean Error of predictive models using DMA SL and DMA combination SL + SA-3. Comparison by Student's *t*-test for equality of means.

Mean Error SL	Mean Error SL + SA-3	*t*	*df*	*p*-Value	Confidence Interval
1.011	0.878	2.955	476	0.003	0.044–0.221

Figure 8 shows the error distribution per ANN configuration. Configuration 1 is formed by 4 blocks of 500 values of water demand; Configuration 2 is formed by 8 blocks of 500 values; Configuration 3 is formed by 20 blocks of 200 values; and Configuration 4 is formed by 8 blocks of 400 values. The rest of parameters are the same for the four configurations. The minimum average error is found for Configuration 3.
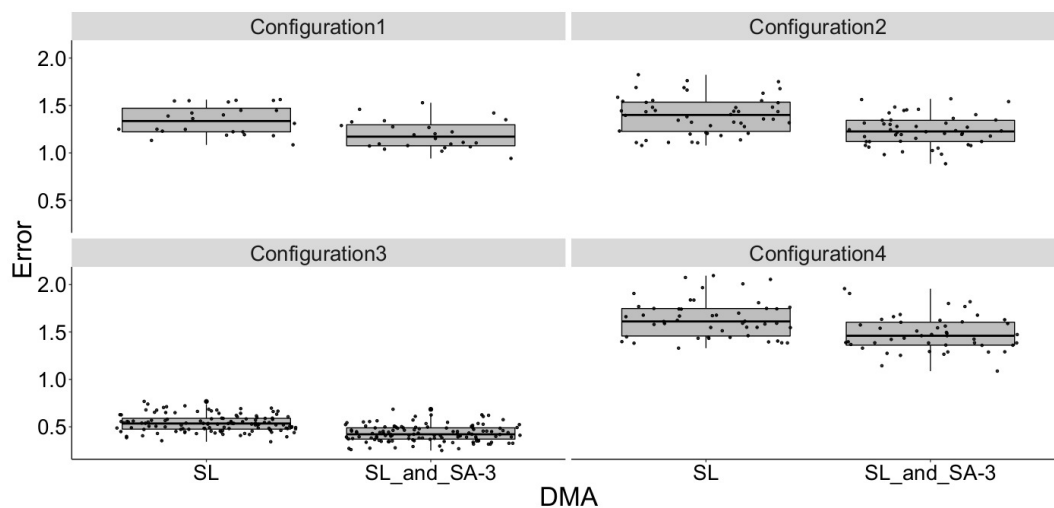
Figure 8. Box-plots of the associated errors by ANN configuration.

Errors computed in Figure 8 come by iteratively run the ANN with new fresh data for training and validation purposes at each time. By repeating several times the same model with different data, it is possible to reach more robust computations. The final predictive model is selected on average and it is not biased by the effect of its dependency from a unique sample. Table 4 shows the results for the top ANN configuration (configuration 3).

Table 4. Prediction Mean Error (top ANN configuration) for DMA SL and for DMA combination SL + SA-3.

Learning Rate	Nodes in Hidden Layer	Mean Error SL	Mean Error SL + SA-3
0.01	3	0.549	0.455
0.01	6	0.533	0.410
0.01	9	0.526	0.425
0.1	3	0.567	0.488
0.1	6	0.529	0.401
0.1	9	0.517	0.399

Two main thoughts come from analysing the error distributions shown in Figure 8. Firstly, the enhanced model provides more accurate results in all the cases. This means that the strategy of using several DMAs information for the predictive model at a single DMA is adequate. Secondly,

there is an upper bound for the optimal length of the information used to make the predictive model of the water demand. This happens as the minimum error is found by using predictive models of 200 values.

4. Discussion

The paper proposes a novel framework based on adapting Epidemiology data analysis methodologies to deal with challenges from water utility management. To show the usefulness of the proposal, an innovative methodology to enhance the predictive models of water demand is developed. This methodology uses the SAX distance as a basis to find out similar DMAs to the target DMA, in which the predictive model is developed.

The second proposal of this work is to use SAX for epidemiology-based forecast modelling. In particular, SAX performs well as a support for finding motifs and surprising patterns in time series of water demand and for approaching visualisation analytics. All of these features are suitable to work well for epidemiology issues. However, this is also important in the water demand case, as SAX is able to:

- extract regular consumption patterns, useful for planning various operations in drinking water management and further rehabilitation plans;
- support anomaly detection, an issue able to directly point out causes of pressure transients in pipelines, or to locate contamination sources and their impact on treatment of effluents;
- work with long time series and be an accurate framework for computing similarities. In water demand applications, this enables to approach relationships between consumption patterns in different sector areas of the water distribution system. As a consequence, also effluents can be traced back with enhanced accuracy.

The SAX approach for water demand time series is combined with suffix trees, which are also used by human genome researchers in algorithms to find out the most important structures hidden in biological sequences. This is possible given the high capability of suffix trees to store and compress a large amount of symbolic data in linear time. This feature is susceptible to be used to near real-time applications in shorter time series such as it may be the case of water demand. In the current proposal, suffix trees support the extraction of consumption patterns and the comparison of water demand at various DMAs.

With information from various DMAs, the analysis of this paper shows how predictive models can be enhanced by adding information from time series databases with similar features to the target DMA. For the sake of clarity, it has been proposed the use of a single ANN as a predictor. However, any predictive model might be suitable to be embedded in the process.

SAX can be further expanded to also use quantiles or local regressions for each of its partitions, instead of straightforwardly using the average. This would enhance the current outcomes, providing better modelling capabilities for peaks of water consumption and for finding even more accurate demand patterns. Since SAX is also a method to reduce the database dimension, working with SAX allows to extend the methodology proposed to big-size time series models.

5. Conclusions

The proposal of this work is to tailor an epidemiological data analysis approach as a novel approach for urban hydraulics. Among the wide scope of applications that EDAAs open to research further, this paper successfully deals with an epidemiology-based forecasting methodology for water demand prediction. This suggests EDAAs' suitability to approach key water supply management issues. Still, EDAAs need further investigation in urban hydraulics as a promising framework for tasks related to drinking water network operation and management.

References

1. Fecarotta, O.; Carravetta, A.; Morani, M.; Padulano, R. Optimal Pump Scheduling for Urban Drainage under Variable Flow Conditions. *Resources* **2018**, *7*, 73. [CrossRef]
2. Creaco, E.; Pezzinga, G. Comparison of Algorithms for the Optimal Location of Control Valves for Leakage Reduction in WDNs. *Water* **2018**, *10*, 466. [CrossRef]
3. Nguyen, K.A.; Stewart, R.A.; Zhang, H.; Sahin, O.; Siriwardene, N. Re-engineering traditional urban water management practices with smart metering and informatics. *Environ. Model. Softw.* **2018**, *101*, 256–267. [CrossRef]
4. Adamowski, J.; Karapataki, C. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: Evaluation of different ANN learning algorithms. *J. Hydrol. Eng.* **2010**, *15*, 729–743. [CrossRef]
5. Caiado, J. Performance of combined double seasonal univariate time series models for forecasting water demand. *J. Hydrol. Eng.* **2009**, *15*, 215–222. [CrossRef]
6. Herrera, M.; Torgo, L.; Izquierdo, J.; Pérez-García, R. Predictive models for forecasting hourly urban water demand. *J. Hydrol.* **2010**, *387*, 141–150. [CrossRef]
7. Msiza, I.S.; Nelwamondo, F.V.; Marwala, T. Water demand prediction using artificial neural networks and support vector regression. *J. Comput.* **2008**, *3*, 1–8. [CrossRef]
8. Tiwari, M.; Adamowski, J.; Adamowski, K. Water demand forecasting using extreme learning machines. *J. Water Land Dev.* **2016**, *28*, 37–52. [CrossRef]
9. Vijayalaksmi, D.; Babu, K.J. Water Supply System Demand Forecasting Using Adaptive Neuro-fuzzy Inference System. *Aqua. Procedia* **2015**, *4*, 950–956. [CrossRef]
10. Ji, G.; Wang, J.; Ge, Y.; Liu, H. Urban water demand forecasting by LS-SVM with tuning based on elitist teaching-learning-based optimization. In Proceedings of the IEEE 26th Chinese Control and Decision Conference (2014 CCDC), Changsha, China, 31 May–2 June 2014; pp. 3997–4002.
11. Zhou, L.; Xia, J.; Yu, L.; Wang, Y.; Shi, Y.; Cai, S.; Nie, S. Using a Hybrid Model to Forecast the Prevalence of Schistosomiasis in Humans. *Int. J. Environ. Res. Public Health* **2016**, *13*, 355. [CrossRef]
12. Cadenas, E.; Rivera, W.; Campos-Amezcua, R.; Heard, C. Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model. *Energies* **2016**, *9*, 109. [CrossRef]
13. Zhang, G.P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]
14. Herrera, M.; García-Díaz, J.C.; Izquierdo, J.; Pérez-García, R. Municipal water demand forecasting: Tools for intervention time series. *Stoch. Anal. Appl.* **2011**, *29*, 998–1007. [CrossRef]
15. Khashei, M.; Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft Comput.* **2011**, *11*, 2664–2675. [CrossRef]
16. Campisi-Pinto, S.; Adamowski, J.; Oron, G. Forecasting urban water demand via wavelet-denoising and neural network models. Case study: City of Syracuse, Italy. *Water Resour. Manag.* **2012**, *26*, 3539–3558. [CrossRef]
17. Brentan, B.M., Luvizotto, E.Jr.; Herrera, M.; Izquierdo, J.; Pérez-García, R. Hybrid regression model for near real-time urban water demand forecasting. *J. Comput. Appl. Math.* **2017**, *309*, 532–541. [CrossRef]
18. Di Nardo, A.; Di Natale, M.; Musmarra, D.; Santonastaso, G.F.; Tzatchkov, V.; Alcocer-Yamanaka, V.H. Dual-use value of network partitioning for water system management and protection from malicious contamination. *J. Hydroinform.* **2015**, *17*, 361–376. [CrossRef]
19. Scarpa, F.; Lobba, A.; Becciu, G. Elementary DMA design of looped water distribution networks with multiple sources. *J. Water Resour. Plan. Manag.* **2016**, *142*, 04016011. [CrossRef]

20. Panagopoulos, G.P.; Bathrellos, G.D.; Skilodimou, H.D.; Martsouka, F.A. Mapping urban water demands using multi-criteria analysis and GIS. *Water Resour. Manag.* **2012**, *26*, 1347–1363. [CrossRef]

21. Herrera, M. Improving Water Network Management by Efficient Division into Supply Clusters. Ph.D. Thesis, Universitat Politècnica de València, València, Spain, 2011.

22. Buchberger, S.G.; Nadimpalli, G. Leak estimation in water distribution systems by statistical analysis of flow readings. *J. Water Resour. Plan. Manag.* **2004**, *130*, 321–329. [CrossRef]

23. Candelieri, A. Clustering and support vector regression for water demand forecasting and anomaly detection. *Water* **2017**, *9*, 224. [CrossRef]

24. Padulano, R.; Del Giudice, G. Pattern Detection and Scaling Laws of Daily Water Demand by SOM: An Application to the WDN of Naples, Italy. *Water Resour. Manag.* **2018**, 1–17. [CrossRef]

25. Bloetscher, F. Protecting people, infrastructure, economies, and ecosystem assets: Water management in the face of climate change. *Water* **2012**, *4*, 367–388. [CrossRef]

26. Bach, P.M.; Rauch, W.; Mikkelsen, P.S.; Mccarthy, D.T.; Deletic, A. A critical review of integrated urban water modelling–Urban drainage and beyond. *Environ. Model. Softw.* **2014**, *54*, 88–107. [CrossRef]

27. Woodward, M. *Epidemiology: Study Design and Data Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2013.

28. Clayton, D.; Hills, M. *Statistical Models in Epidemiology*; Oxford University Press: Oxford, UK, 2013.

29. Ganesh, A.; Massoulié, L.; Towsley, D. The effect of network topology on the spread of epidemics. In Proceedings of the IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, FL, USA, 13–17 March 2005; Volume 2, pp. 1455–1466.

30. Goltsev, A.V.; Dorogovtsev, S.N.; Oliveira, J.; Mendes, J.F. Localization and spreading of diseases in complex networks. *Phys. Rev. Lett.* **2012**, *109*, 128702. [CrossRef] [PubMed]

31. Danila, B.; Yu, Y.; Marsh, J.A.; Bassler, K.E. Optimal transport on complex networks. *Phys. Rev. E* **2006**, *74*, 046106. [CrossRef] [PubMed]

32. Herrera, M.; Izquierdo, J.; Pérez-García, R.; Montalvo, I. Multi-agent adaptive boosting on semi-supervised water supply clusters. *Adv. Eng. Softw.* **2012**, *50*, 131–136. [CrossRef]

33. Maslov, S.; Sneppen, K.; Zaliznyak, A. Detection of topological patterns in complex networks: Correlation profile of the internet. *Phys. A Stat. Mech. Appl.* **2004**, *333*, 529–540. [CrossRef]

34. Lloyd, A.L.; Valeika, S. Network models in Epidemiology: An overview. *Complex Pop. Dyn.* **2007**, 189–214. [CrossRef]

35. Hamilton, I.; Summerfield, A.; Oreszczyn, T.; Ruyssevelt, P. Using epidemiological methods in energy and buildings research to achieve carbon emission targets. *Energy Build.* **2017**, *154*, 188–197. [CrossRef]

36. Bardet, J.P.; Little, R. Epidemiology of urban water distribution systems. *Water Resour. Res.* **2014**, *50*, 6447–6465. [CrossRef]

37. Navarrete-López, C.; Brentan, B.; Herrera, M.; Luvizotto, E., Jr.; Izquierdo, J.; Pérez-García, R. Epidemiological approach to forecast water demand consumption through SAX. In Proceedings of the Mathematical Modelling in Engineering & Human Behaviour 2016, Valencia, Spain, 18–20 July 2016; pp. 1–6.

38. De Domenico, M.; Granell, C.; Porter, M.A.; Arenas, A. The physics of spreading processes in multilayer networks. *Nat. Phys.* **2016**, *12*, 901. [CrossRef]

39. Hamilton, I.G.; Summerfield, A.J.; Lowe, R.; Ruyssevelt, P.; Elwell, C.A.; Oreszczyn, T. Energy epidemiology: A new approach to end-use energy demand research. *Build. Res. Inf.* **2013**, *41*, 482–497. [CrossRef]

40. Herrera, M.; Ferreira, A.A.; Coley, D.A.; de Aquino, R.R. SAX-quantile based multiresolution approach for finding heatwave events in summer temperature time series. *AI Commun.* **2016**, *29*, 725–732. [CrossRef]

41. Padulano, R.; Del Giudice, G. A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data. *Water Resour. Manag.* **2018**, *32*, 3671–3685.

42. Lin, J.; Keogh, E.; Wei, L.; Lonardi, S. Experiencing SAX: A novel symbolic representation of time series. *Data Min. Knowl. Discov.* **2007**, *15*, 107–144. [CrossRef]

43. Toshniwal, D. Feature extraction from time series data. *J. Comput. Methods Sci. Eng.* **2009**, *9*, 99–110.

44. Aghabozorgi, S.; Wah, T.Y. Clustering of large time series datasets. *Intell. Data Anal.* **2014**, *18*, 793–817. [CrossRef]

45. Martínez-Álvarez, F. Clustering preprocessing to improve time series forecasting. *AI Commun.* **2011**, *24*, 97–98.

46. Yuan, J.; Wang, Z.; Han, M.; Sun, Y. A lazy associative classifier for time series. *Intell. Data Anal.* **2015**, *19*, 983–1002. [CrossRef]

47. Rajaraman, A.; Ullman, J.D.; Ullman, J.D.; Ullman, J.D. *Mining of Massive Datasets*; Cambridge University Press: Cambridge, MA, USA, 2012; Volume 1.

48. McLaurin, E.; McDonald, A.D.; Lee, J.D.; Aksan, N.; Dawson, J.; Tippin, J.; Rizzo, M. Variations on a theme: Topic modeling of naturalistic driving data. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage: Los Angeles, CA, USA, 2014; Volume 58, pp. 2107–2111.

49. Rasheed, F.; Alshalalfa, M.; Alhajj, R. Efficient periodicity mining in time series databases using suffix trees. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 79–94. [CrossRef]

50. Keogh, E.; Lonardi, S.; Chiu, B.c. Finding surprising patterns in a time series database in linear time and space. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–25 July 2002; pp. 550–556.

51. Schmieder, R.; Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **2011**, *6*, e17288. [CrossRef] [PubMed]

52. Välimäki, N.; Gerlach, W.; Dixit, K.; Mäkinen, V. Compressed suffix tree–a basis for genome-scale sequence analysis. *Bioinformatics* **2007**, *23*, 629–630. [CrossRef] [PubMed]

53. Ezkurdia, I.; Juan, D.; Rodriguez, J.M.; Frankish, A.; Diekhans, M.; Harrow, J.; Vazquez, J.; Valencia, A.; Tress, M.L. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum, Mol, Genet.* **2014**, *23*, 5866–5878. [CrossRef]

54. Bermudez-Santana, C.I. Aplicaciones de la bioinformática en la Medicina: El genoma humano. ¿Cómo podemos ver tanto detalle? *Acta Biol. Colomb.* **2016**, *21*, 249. [CrossRef]

55. Cai, L.; Li, X.; Ghosh, M.; Guo, B. Stability analysis of an HIV/AIDS epidemic model with treatment. *J. Comput. Appl. Math.* **2009**, *229*, 313–323. [CrossRef]

56. Jackson, M.; Chen-Charpentier, B.M. Modeling plant virus propagation with delays. *J. Comput. Appl. Math.* **2017**, *309*, 611–621. [CrossRef]

57. Brentan, B.M.; Meirelles, G.; Herrera, M.; Luvizotto, E.; Izquierdo, J. Correlation analysis of water demand and predictive variables for short-term forecasting models. *Math. Probl. Eng.* **2017**, *2017*. [CrossRef]

58. Bhaskaran, K.; Gasparrini, A.; Hajat, S.; Smeeth, L.; Armstrong, B. Time series regression studies in environmental epidemiology. *Int. J. Epidemiol.* **2013**, *42*, 1187–1195.

59. Clayton, D.; Hills, M.; Pickles, A. *Statistical Models in Epidemiology*; International Epidemiological Association (IEA): Chicago, IL, USA, 1993; Volume 161.

60. Helfenstein, U. The use of transfer function models, intervention analysis and related time series methods in epidemiology. *Int. J. Epidemiol.* **1991**, *20*, 808–815. [CrossRef] [PubMed]

61. Herrera, M.; Izquierdo, J.; Pérez-García, R.; Ayala-Cabrera, D. Water supply clusters by multi-agent based approach. In *Water Distribution Systems Analysis 2010*; American Society of Civil Engineers (ASCE): Reston, VA, USA, 2010; pp. 861–869.

62. Herrera, M.; Abraham, E.; Stoianov, I. A graph-theoretic framework for assessing the resilience of sectorised water distribution networks. *Water Resour. Manag.* **2016**, *30*, 1685–1699. [CrossRef]

63. Jung, D.; Choi, Y.H.; Kim, J.H. Optimal Node Grouping for Water Distribution System Demand Estimation. *Water* **2016**, *8*, 160. [CrossRef]

64. Wang, X.; Mueen, A.; Ding, H.; Trajcevski, G.; Scheuermann, P.; Keogh, E. Experimental comparison of representation methods and distance measures for time series data. *Data Min. Knowl. Discov.* **2013**, *26*, 275–309. [CrossRef]

65. Lin, J.; Li, Y. Finding structural similarity in time series data using bag-of-patterns representation. In *Scientific and Statistical Database Management*; Springer: New York, NY, USA, 2009; pp. 461–477.

66. Cassisi, C.; Prestifilippo, M.; Cannata, A.; Montalto, P.; Patanè, D.; Privitera, E. Probabilistic Reasoning Over Seismic Time Series: Volcano Monitoring by Hidden Markov Models at Mt. Etna. *Pure Appl. Geophys.* **2016**, *173*, 2365–2386. [CrossRef]

67. Chen, G.; Puglisi, S.J.; Smyth, W.F. Fast and practical algorithms for computing all the runs in a string. In *Annual Symposium on Combinatorial Pattern Matching*; Springer: New York, NY, USA, 2007; pp. 307–315.

68. Gusfield, D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*; Cambridge University Press: Cambridge, MA, USA, 1997.

69. Grossi, R.; Italiano, G.F. Suffix trees and their applications in string algorithms. In Proceedings of the 1st South American Workshop on String Processing, Federal University of Minas Gerais, Belo Horizonte, Brazil, 13–15 September 1993; pp. 57–76.

70. McCreight, E.M. A space-economical suffix tree construction algorithm. *J. ACM* **1976**, *23*, 262–272. [CrossRef]

71. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]

72. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]

73. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.

74. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [CrossRef]