CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Journal Paper

Multivariate time series clustering and forecasting for building energy analysis: Application to weather data quality

In Press, Journal Pre-proof

Ênio Filho*
Luís Sanhudo
João Manuel Coelho Rodrigues

*CISTER Research Centre

CISTER-TR-201105

2020/11/09

Multivariate time series clustering and forecasting for building energy analysis: Application to weather data quality

Ênio Filho*, Luís Sanhudo, João Manuel Coelho Rodrigues

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: enpvf@isep.ipp.pt, lpnsanhudo@fe.up.pt, jmcr@fe.up.pt

https://www.cister-labs.pt

Abstract

In recent years, several tools for building energy performance simulation and analysis have been developed to assist in increasing building energy performance, harvesting its computing capabilities for a reliable and accurate energy performance prediction. To perform this analysis, energy tools typically require crucial data regarding the building 19s surrounding environment, which is acquired from neighbouring weather stations. However, these stations often experience hardware malfunctions, resulting in either erroneous or missing data. Traditionally, these values are rectified through empirical and geostatistical methods, which, while reflecting several decades of practice, may prove to be inadequate when considering a purely data-driven approach. To this end, the present study introduces a machine learning methodology proposing the application of regression algorithms to rectify the erroneous values in datasets, and the clustering of weather stations, based on their recorded climatic conditions, to enhance the regression models. A shape-based approach for clustering time series of different climatic parameters and weather stations is pursued, using the k-medoids algorithm alongside dynamic time warping as the similarity measure. Both Artificial Neural Networks (ANN) and Support Vector Regression (SVR) models are evaluated as exemplary regression algorithms, with different sets of predictors. Mean Squared Error is used as the performance metric. A data set of different climatic parameters from southeastern Brazil was used, with air temperature being chosen as the response variable, given its importance in energy consumption. Results indicate that a machine learning approach to the problem is indeed viable. ANN slightly outperforms SVR in the prediction of the studied weather variable.

Journal of Building Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: http://www.elsevier.com/locate/jobe

Multivariate time series clustering and forecasting for building energy analysis: Application to weather data quality control

Luís Sanhudo [a,*], João Rodrigues [a], Ênio Vasconcelos Filho [b]

[a] CONSTRUCT, Faculty of Engineering (FEUP), University of Porto, Porto, Portugal
[b] CISTER Research Centre, ISEP, Polytechnic Institute of Porto, Porto, Portugal

ARTICLE INFO

Keywords:
Building energy analysis
Weather data quality control
Time series clustering
Artificial neural networks
Support vector regression

ABSTRACT

In recent years, several tools for building energy analysis and simulation have been developed to assist in increasing building energy performance, harvesting its computing capabilities for a reliable and accurate energy performance prediction. To perform this analysis, energy tools typically require crucial data regarding the building's surrounding environment, which is acquired from neighbouring weather stations. However, these stations often experience hardware malfunctions, resulting in either erroneous or missing data. Traditionally, these values are rectified through empirical and geostatistical methods, which, while reflecting several decades of practice, may prove to be inadequate when considering a purely data-driven approach. To this end, the present study introduces a machine learning methodology proposing the application of regression algorithms to rectify the erroneous values in datasets, and the clustering of weather stations, based on their recorded climatic conditions, to enhance the regression models. A shape-based approach for clustering time series of different climatic parameters and weather stations is pursued, using the k-medoids algorithm alongside dynamic time warping as the similarity measure. Both Artificial Neural Networks (ANN) and Support Vector Regression (SVR) models are evaluated as exemplary regression algorithms, with different sets of predictors. Mean Squared Error is used as the performance metric. A data set of different climatic parameters from southeastern Brazil was used, with air temperature being chosen as the response variable, given its importance in energy consumption. Results indicate that a machine learning approach to the problem is indeed viable. ANN slightly outperforms SVR in the prediction of the studied weather variable.

1. Introduction

Increasing energy efficiency and reducing energy consumption are some of the leading research objectives in the Architecture, Engineering and Construction (AEC) industry [1–3]. In recent years, these goals have even been backed by international strategies, such as the European Union's (EU) 2030 Climate & Energy Framework. This framework aims for a 32.5% improvement in energy efficiency, as well as a 40% reduction in greenhouse gas emissions (based on 1990 levels), with values expected to be revised upwards in 2023 [4]. Considering that buildings are responsible for approximately 40% of EU's energy consumption, as well as 36% of the Carbon Dioxide (CO2) emissions, the successful achievement of these strategies is tightly linked with the existing building stock [5], namely, its energy retrofitting. In fact, previous studies indicate that the energy retrofitting of the current building stock has not only to be considered if the above-mentioned goals are to be accomplished, but also presents itself has one of the most efficient solutions to do so [3,6,7].

However, there are multiple challenges faced by AEC professionals when tackling energy retrofitting [8–11]. To this end, in the last few years, several tools for building energy analysis and simulation have been developed to assist in examining a building and swiftly exploring, comparing, and optimizing retrofitting solutions. In fact, in the "Building Energy Software Tools Directory" [12], provided by the United States Department of Energy, over 200 tools are listed, being proposed as suitable solutions to manage and assess a building's energy performance [13–16]. These tools incorporate features such as whole-building energy analysis, solar radiation study, artificial lighting and daylight examination, thermal performance, HVAC system comparison, water usage or acoustic examination [3,12,17–19].

To properly perform this analysis, energy tools require high-quality, long-term, accurate meteorological data, as it constitutes a critical

* Corresponding author.
E-mail address: lpnsanhudo@fe.up.pt (L. Sanhudo).

https://doi.org/10.1016/j.jobe.2020.101996
Received 2 April 2020; Received in revised form 2 November 2020; Accepted 5 November 2020
Available online 10 November 2020

aspect in characterizing the boundary conditions of the building [20–23]. This weather data is typically acquired from neighbouring weather stations, as measuring weather data at the building location is generally considered cost-prohibitive [24]. However, as seen in Refs. [25,26], these stations often present malfunctions that result in erroneous or missing data. These malfunctions can be originated from simple hardware breaks, to software malfunction and data transmission errors. The resulting data of these malfunctions must then be estimated or corrected, for the dataset to be used in energy analysis. To this end, traditionally, these values are rectified through empirical and geostatistical methods [25–28], which, while reflecting several decades of practice, may prove to be inadequate when considering a purely data-driven approach.

As an alternative to these methods, the present research proposes a novel strategy exploiting machine learning capabilities to rectify the incorrect or null values in datasets. In recent years, many algorithms have been applied in weather-related scenarios, but only in the forecasting of different climatic parameters, as temperature [29–31], precipitation [32–34], or wind speed [35,36]. This research aims at employing regression algorithms to rectify the erroneous values in datasets, and further proposes clustering weather stations based on their recorded climatic conditions, to enhance the regression models. A data set of different climatic parameters from southeastern Brazil was used, with air temperature being chosen as the response variable, given its importance in energy consumption.

The outline of this paper is as follows: section 2 presents the methodology applied in this article, as well as a formulation of the algorithms and a brief literature review on relevant applications. In section 3, the data set considered in this study is described, and some preparation steps are duly justified. The implementation of the algorithms is detailed in section 4, along with the discussion of pertinent results and the comparative evaluation of the performance of two distinct regression algorithms. Section 5 synthetizes main conclusions, pointing out some relevant observations for future work.

2. Methodology

The data set of this study encompasses time series of climatic parameters registered in several weather stations, which are schematically represented by black dots in Fig. 1. In a context of erroneous or missing data at a given weather station, a natural option would be to consider the climate response recorded at other stations with similar behaviour. As such, in the first step of the proposed methodology, clusters of weather stations are obtained, after defining an appropriate similarity measure and implementing the k-medoids algorithm. Then, air temperature estimates in one weather station are calculated through regression, considering as predictors the records from other stations belonging to the same cluster. These results are compared with the ones obtained from simpler models, trained exclusively in the historical data of a single weather station. In all of these problems, both Support Vector Regression (SVR) and Artificial Neural Networks (ANN) models are built up, in order to compare their performances and assess their suitability to provide accurate estimates when the recorded data is incorrect or null.

2.1. K-medoids

Saeed Aghabozorgi et al. [37] defined time series clustering as an unsupervised process for partitioning n time series, $D = \{F_1, F_2, \ldots, F_n\}$, into k clusters, $C = \{C_1, C_2, \ldots, C_k\}$, where $D = \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \varnothing$ for $i \neq j$. This type of clustering is often challenging for several reasons, typically related to the high dimension of the data set, which exponentially decreases the clustering speed [37–39]. This approach can be divided into six different groups of algorithms: partitioning, hierarchical, grid-based, model-based, density-based clustering and multi-step clustering. Saeed Aghabozorgi et al. [37] thoroughly analyses these different approaches, discussing each approach characteristics in detail while also reviewing multiple applications in current research. Additionally, there are three different ways to cluster time series: shape-based, feature-based and model-based. These are dependent on their interaction with the raw-data, respectively: work directly with raw data, indirectly through extracted features, or indirectly with models built from the raw data [37,40–42]. For more information on time series clustering see Refs. [37,40–43], and for its application in weather-related scenarios see Refs. [44,45]. In this study, we pursue a shape-based approach, using dynamic time warping (DTW) as the distance/similarity measure, Lower Bounding Keogh (LB Keogh) as DTW's acceleration method, and k-medoids as the clustering algorithm.

The k-medoids algorithm [46] is a popular clustering technique that clusters the data set D of n objects into k clusters, with k being provided by the user [47–50]. The algorithm operates on the principle of minimizing the sum of dissimilarities between each object and its corresponding medoid, the most centrally located point in a cluster. As such, a medoid can be defined as the object of a cluster whose average dissimilarity to all the remaining objects in the cluster is minimal [51]. To be initialized, the algorithm randomly selects k objects in the dataset D as initial medoids to represent the k clusters [52]. Then, each remaining object is clustered with the nearest medoid. After this initial random selection of k medoids, the algorithm iteratively tries to locate better medoids by minimizing the above-mentioned sum [53]. The iterative process is repeated until no medoids change its placement and the final k clusters are acquired [52]. As such, the k-medoids algorithm works as follows [54]:

1. **Input:** k: number of clusters; D: data set containing n objects.

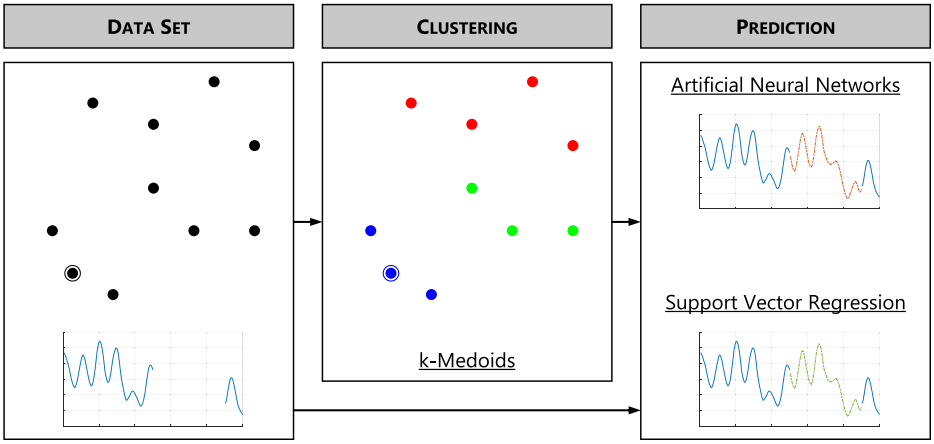

Fig. 1. Outline of the proposed methodology.

2. **Output:** Set of k clusters that minimizes the sum of the dissimilarities of all objects to their nearest medoid.
3. Algorithm:
   (a) Randomly select k objects from D as the initial medoids;
   (b) Assign each remaining object o in D to the closest medoid m using the similarity measure;
   (c) For each medoid m and object o associated, compute the total cost of swapping old medoid m with the currently selected non-medoid object o – select the medoids that result in the lowest cost of the configuration;
   (d) Repeat steps b and c until no change in the medoids.

As seen in Ref. [51], k-medoids is not without its limitations. One major limitation is that it can be heavily influenced by the random initial medoids, possibly generating distinct clusters with each initialization. Additionally, as with several other clustering methods, a proper value for the number of clusters k may be hard to determine.

Regarding the chosen similarity measure, DTW is a well-known technique to find an optimal alignment between two given time-dependent sequences under certain restrictions [55]. Intuitively, DTW (Fig. 2) allows time series to be locally stretched or shrunk before applying the base distance measure [56]. DTW has been applied in several disciplines, such as bioinformatics [57], chemical engineering [58], robotics [59], speech recognition [60,61] and human motion [62–64], among others.

Formally [56,66], given two time series Q and P, where $Q = q_1, q_2, \ldots, q_i, \ldots, q_n$ and $P = p_1, p_2, \ldots, p_j, \ldots, p_m$, DTW can be used to align these two sequences by computing an $n \times m$ matrix where the (i^{th}, j^{th}) element is given by the squared distance between points q_i and p_j.

$$d(q_i, p_j) = (q_i - p_j)^2 \tag{1}$$

By finding the warping path W of the matrix that minimizes the total cumulative distances between Q and P, DTW is found:

$$DTW(Q, P) = \min\left\{\sqrt{\sum_{k=1}^{K} w_k}\right. \tag{2}$$

where w_k is defined as the k^{th} element in the warping path W that characterizes the mapping between Q and P. The path W is acquired by evaluating Equation (3) using dynamic programming, where $\gamma(q_i, p_j)$ defines the cumulative distance as the distance $d(q_i, p_j)$ in the current cell and the minimum of the cumulative distances of the adjacent elements.

$$\gamma(q_i, p_j) = d(q_i, p_j) + \min\left\{\begin{matrix}\gamma(q_{i-1}, p_j) \\ \gamma(q_i, p_{j-1}) \\ \gamma(q_{i-1}, p_{j-1})\end{matrix}\right\} \tag{3}$$

As seen in Ref. [66], several constraints are used to restrict the possible warping paths in order to not only prevent pathological warping, but also slightly speed the calculations. In fact, although DTW tends to offer a greater accuracy than other distance measures in time series analysis, it may also drastically increase the computation time because of its quadratic time complexity [67]. As such, to minimize the computation time of DTW, three distinct method categories exist [68]:

- Constraints – Restricts the cells that are assessed in the cost matrix;
- Data Abstraction – Applies DTW to a lower-resolution representation of the data and maps the acquired result on the full resolution cost matrix;
- Indexing – Decreases the number of times that DTW is performed by application of lower bounding functions.

Constraints is the most popular method. In fact, in Ref. [66] the authors state that virtually all practitioners constrain the warping path in a global sense by limiting its reach r from the diagonal, so that i and j in $w_k = (i, j)$ are constrained to $j - r \leq i \leq j + r$. The resulting matrix subset, also called warping window or band, defines the cells that the warping path can take [69,70]. As seen in Fig. 3, which depicts two popular global constraints – the Sakoe-Chiba Band [71] and the Itakura Parallelogram [72] – the reach r is dependent on the constraint type.

However, in this article the lower bounding measure LB Keogh [69] was applied. LB Keogh builds upon the constraints methods, using the defined reach r to create two new sequences (U for upper and L for lower) that encapsulate any input time series Q:

$$U_i = \max(q_{i-r} : q_{i+r}) \tag{4}$$

$$L_i = \min(q_{i-r} : q_{i+r}) \tag{5}$$

Using U and L, LB Keogh can be defined as follows:

$$LB\_Keogh(Q, P) = \sqrt{\sum_{i=1}^{n}\begin{cases}(p_i - U_i)^2, & if\ c_i > L_i \\ (p_i - L_i)^2, & if\ c_i < L_i \\ 0, & otherwise\end{cases}} \tag{6}$$

Regarding the size of the warping window attained from the r value, Ratanamahatana and Keogh [41] identify that most practitioners create a warping window of 10% width. However, the authors demonstrate that this value is too large, indicating maximum accuracies to smaller percentages. Through experimentation, in this study the authors decided to use a window size of 1%.

2.2. Artificial Neural Networks

ANNs are a well-known technique in machine learning, which have been used since their development in the years of 1960. Nowadays, as other machine learning algorithms, they are applied in many fields, such as medicine [73], weather forecast [74,75], meteorological pollution [76] and finance [77,78], leveraging their capacity to learn about different systems using an intuitive approach. The idea behind the algorithm is to simulate the neurons of the animal brain, creating a network. Each neuron is a mathematical model (or function) that represents the real world. The network between these neurons is composed by a weighted sum, followed or not by an activation function [79]. ANNs

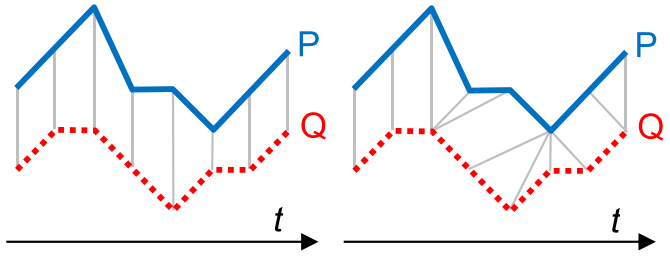

Fig. 2. Sample series alignment using the Euclidean distance (left) and DTW (right). Adapted from Ref. [65].

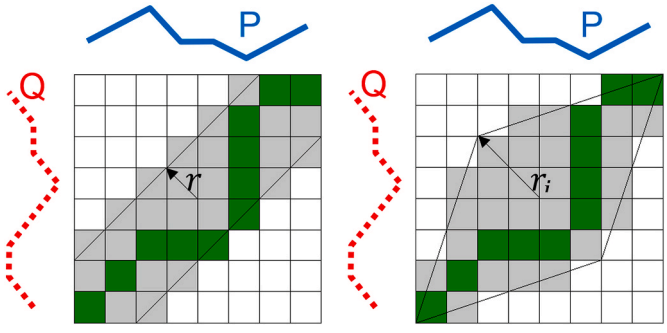

Fig. 3. Constraints: Sakoe-chiba band (left) and itakura parallelogram (right).

work like a black box, not requiring detailed information about the inputs and being able to obtain the relationships between the input parameters [34]. Another critical aspect about ANNs is their ability to handle large and complex amounts of data, with interrelated features [80].

A neural network contains three types of layers: input, hidden and output layer. The first one receives the features values, which are used to "teach" the system. Usually, the number of input nodes in an input layer is equal to the number of features in the system. The hidden layer applies some transformations to the input values inside the network. In the hidden layer, the actual processing is done via a system of weighted connections; there may be one or more hidden layers. The last one is the output layer, which is connected to the previous hidden layer: it returns an output value that corresponds to the response variable's prediction. Fig. 4 shows the structure of a generic artificial neural network. Details of the mathematical formulation are explained in Ref. [81]. Each hidden layer has several neurons related to the previous and/or to the next layer. The number of layers and neurons defines the hyperparameters of the ANN [81]. An ANN with more than one hidden layer is called a multilayer network.

In the last few decades, many architectures and training algorithms have been proposed for ANNs. This paper will focus on the multilayer perceptron (MLP) with back-propagation [82]. The perceptron model was introduced in Ref. [83] and its algorithm is detailed in Ref. [84]. The MLP model presents one or more layers as hidden. According to Ref. [85], the utilization of one hidden layer can represent any continuous function, while the usage of two allows the approximation of any function. The typical implementation of the MLP algorithm uses a fully connected network, with all the neurons connected to those in the previous and next layers. The training algorithm is known as back-propagation and consists of a two-step iteration, forward and backward. In the first step, each input object is presented to the network. Then, each neuron calculates its own output and uses the result in the next layer. In the end, the output is determined. This value is then back-propagated in the network, adjusting the weights in each layer. The speed of this adjustment is determined by a momentum [86].

There are several ways to optimize the hyperparameters of an ANN; however, in complex systems, these optimizations are not easy or feasible to determine because of the high number of parameters or data. In Ref. [84], some approaches are described, such as the empirical, meta-heuristics, pruning and constructive. Bishop [81] also proposes a Bayesian approach to determine these parameters, introducing a Gaussian approximation.

ANN capacity for time series forecasting is studied in Ref. [87], with the differentiation between some models and their characteristics, including techniques to achieve the best choice of the network parameters.

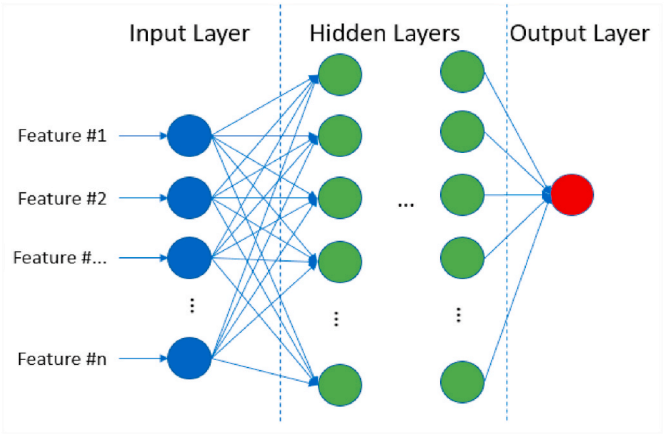

Fig. 4. Structure of a feedforward neural network.

2.3. Support Vector Regression

SVR algorithms represent an extension of support vector machines (SVM) to regression problems, thus preserving the property of sparseness [81,88]. Unlike conventional approaches to regression, they account for not only the approximation error to the data, but also the generalization of the model [30]. Therefore, these algorithms are expected to present a better performance than other techniques that are more prone to overfitting [29].

In this paper, ε-insensitive SVR is implemented: the goal is to find a function f of the predictor variables x that deviates from the observed responses y by a value no greater than ε [88]. As such a function might not be defined, slack variables ξ_n and ξ_n^* are introduced, allowing regression errors to exist up to the values of $\varepsilon + \xi_n$ and $\varepsilon + \xi_n^*$, at each point n of the training set. A constant $C > 0$ controls the penalty imposed on observations lying outside the ε-margin, thus determining the regularization of the model, i.e. the trade-off between the smoothness of f (given by the norm of the vector of weights, w) and the extent to which deviations larger than ε are tolerated [88]. The optimization problem is then formulated as follows:

$$\text{Minimize } \frac{1}{2}w^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*)$$

$$\text{Subject to } \begin{cases} y_n - \langle w, x_n\rangle - b \le \varepsilon + \xi_n \\ \langle w, x_n\rangle - y_n + b \le \varepsilon + \xi_n^* \\ \xi_n, \xi_n^* \ge 0 \end{cases} \tag{7}$$

Nonlinear regression models are addressed by introducing a kernel function that maps the input vectors into a high-dimensional space. This operation requires the dual formulation of the problem, which is detailed, for instance, in Ref. [88].

In the context of time series and climate models, several applications of SVR were already found in the literature, either in the prediction of daily maximum temperatures [29,30], wind speed [35] and meteorological pollution [76], or in the development of precipitation models [32,33]. Regarding the predictors, two major approaches can be identified. In Ref. [30,32], they consist on records of climatic parameters that are simultaneous with the ones being estimated, thus presuming independence between observations. In the selection of the predictors, some prior knowledge about the potentially explanatory variables was involved, and statistical tests were also performed. On the contrary, in Ref. [29,35] the prediction of a climatic parameter is based on n previous records of that feature, n representing the order of the model, defined through experimentation.

In all of these five applications, a radial basis function was used as a kernel [30,35,76], or identified as the one providing better results [29,32]. Methodologies for selecting the hyperparameters present slight variations; nevertheless, cross-validation [35] and grid search algorithms [29,32] were considered and implemented. The generalization error of the models was evaluated in a test set corresponding to 10–33% of the available records.

3. Data set

The data set considered in this study is publicly available on the Internet [89]. It comprises hourly records of 17 climatic parameters, obtained from 122 weather stations in southeastern Brazil. The records refer to a period beginning on a variable date (between May 24, 2000 and June 23, 2016) and ending on September 30, 2016 (Fig. 5). A starting point of January 1, 2008 was defined, with the subsequent stage of data preparation focused on only 96 weather stations.

As stated in the introduction section of this article, weather data frequently presents incorrect or null information. As such, with the purpose of using the present dataset for training machine learning

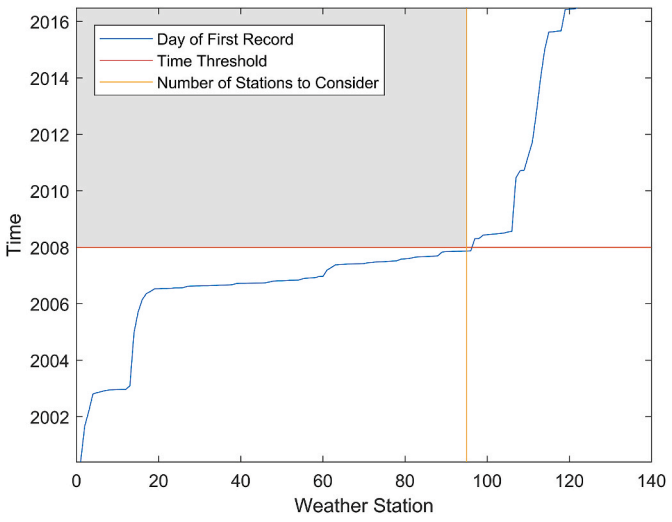

Fig. 5. Date of first record available for each weather station (the shaded area represents the period and weather stations considered in this study).

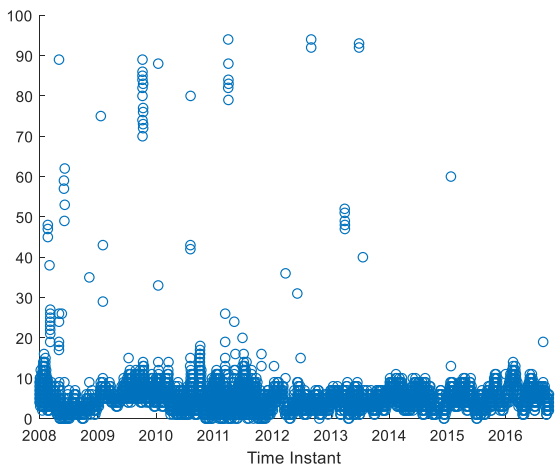

Fig. 7. Total number of weather stations with only missing or zero records per time instant.

algorithms, a cleaning stage was prepared. Initially, missing values or zero records without physical significance were identified. The "solar radiation" feature, for instance, presented missing values during the night period; they were then filled with zeros, since records at the end of each day and at the beginning of the following were also already zero. Furthermore, features with more than 10% of missing values were deleted from the data set; this led to the exclusion of "precipitation" (ratio of 87%) and "wind speed" (11%), along with "wind direction" and "wind gust".

The next step consisted on identifying hourly records in which all the climatic parameters were missing or zero; Figs. 6 and 7 represent their distribution in both space and time. Concerning the latter, it becomes apparent that these instants have a uniform distribution over the period of analysis; for some of them, almost all the stations are completely missing. Therefore, on the following stage of this study, the threshold in Fig. 6 was established, defining a subset of 34 weather stations in which the lack of data was less severe. In these stations, hourly records with only zero values were then excluded. Finally, the same procedure was adopted for hourly records with at least one missing value.

The threshold in Fig. 6 allowed the maximization of the amount of available data. In fact, if a higher number of time instants without any record were considered as a limit, the number of weather stations in the resulting subset would have been larger, as the total of empty lines to delete. On the contrary, a more demanding criterion should have led to a reduced number of weather stations in the subset, albeit with more complete records. Fig. 8 reveals that 34 is the optimal number of weather stations for the sake of maximizing the amount of available data.

As a result of this methodology, the data set considered in the following applications comprises 44,925 simultaneous hourly records in 34 weather stations, without any missing values and referring to a period between January 1, 2008 and September 30, 2016. Thirteen climatic parameters are thus available: air pressure [hPa]; maximum and minimum air pressure for the last hour [hPa]; solar radiation [kJ/m^2]; air temperature [°C]; dew point temperature [°C]; maximum and minimum temperature for the last hour [°C]; maximum and minimum dew point temperature for the last hour [°C]; relative humidity [%] and maximum and minimum relative humidity for the last hour [%].

The resulting hourly records are representative of different time instants along the day and each year, so the data set obtained is expected to be unbiased. All the data was then properly normalized.

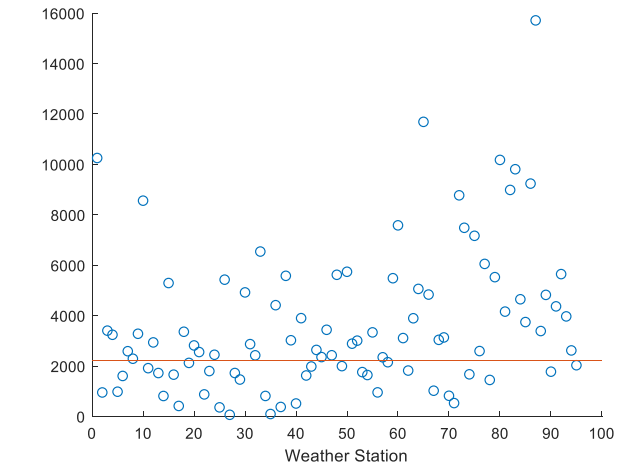

Fig. 6. Total number of time instants with only missing or zero records per weather station.

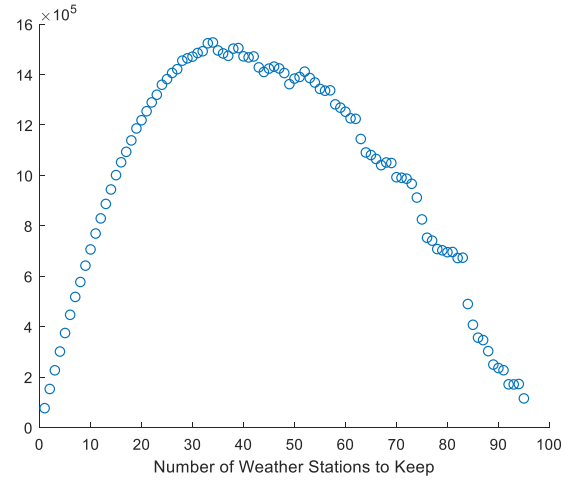

Fig. 8. Total number of lines in the matrix of the data set, for different subsets of weather stations.

4. Results and discussion

4.1. Clustering

Fig. 9 shows the distribution of the studied weather stations in southeastern Brazil, coloured in accordance to their respective cluster. These clusters were acquired through k-medoids, using DTW and LB Keogh. From the values considered reasonable for the number of clusters ($k \in \{2, 3, \ldots, 12\}$), $k = 5$ was selected as the value that best describes the data set. This value was acquired by performing 40 random initialization of k-medoids for each k-value, computing the within-cluster sum of squares for each initialization [90,91] and applying the elbow method [92,93] (Fig. 10).

By furthering the cluster analysis, it was possible to identify a linkage between the acquired clusters and the weather stations elevation. In fact, although the stations elevation variable was not used for clustering (as described in the data set preparation), it is reasonable to relate the clusters in Fig. 9 to the elevations seen in Fig. 11.

4.2. Prediction

The regression models for air temperature developed in this section concern the seven weather stations belonging to the cluster nearer the sea (identification numbers 303, 306, 307, 348, 372, 376 and 388). To investigate the utility of including information relative to the cluster in the model, three types of problems, which vary in the definition of the predictors, are considered here (Table 1); in all these problems, the predictors consist on records of climatic parameters that are simultaneous with the ones being estimated. Furthermore, both ANN and SVR models are built up, allowing the comparison of their performances.

The generalization error of these models was evaluated in a test set corresponding to 25% of the records, the first 75% being used for training and validation. Throughout this study, the performance metric adopted was the Mean Squared Error (MSE); its definition can be found, for instance, in Ref. [35].

4.2.1. Artificial Neural Networks

To build up the ANNs and improve its hyperparameters, a combination of empirical and constructive methods was used in this paper. Therefore, the system was trained with one, two and three intermediate layers in each problem. It was then possible to compare the influence of increasing the number of layers and the extent of overfitting promoted by the third layer. The number of neurons in the network started with the number of predictors in the dataset (n) and was increased until $2n + 1$; this number was then repeated in the next intermediate layers. The activation function considered was the hyperbolic tangent. In summary, ANNs were built using six, seven, 13, 25, 91 or 181 neurons, with one, two or three hidden layers. The number of repetitions also varied between 100, 200 and 300.

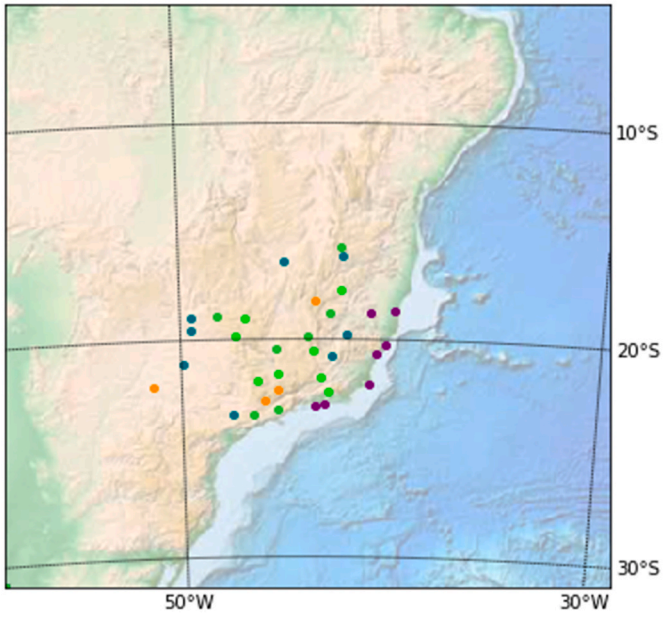

Fig. 9. Weather stations in southeastern Brazil, color-coded by their respective cluster for $k = 5$.

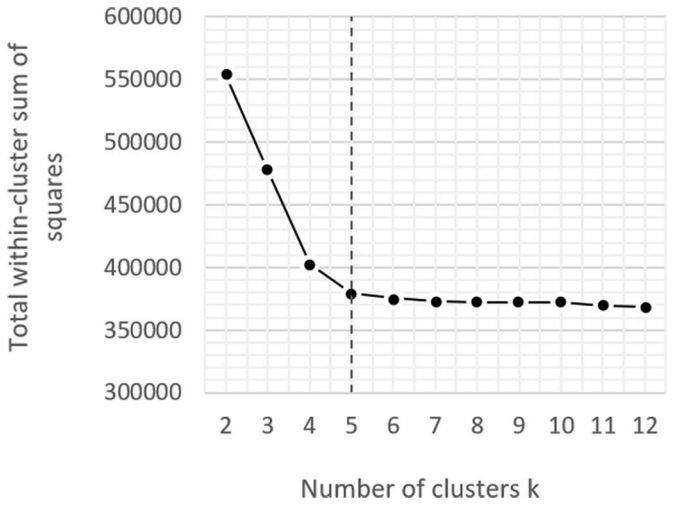

Fig. 10. Elbow method.

Besides MSE, the correlation coefficient R was also computed to evaluate the results of the predictions. These parameters were used to determine the optimal ANN model for solving the problems defined in Table 1. The network was trained through cross-validation, after dividing the first 75% of the data set in training (60%) and validation (15%).

Weather station 303 was randomly selected to be used in the determination of the hyperparameters that would be extended to the remaining stations of the cluster. Varying these hyperparameters (number of layers, number of neurons by layer and number of repetitions), Figs. 12–14 were obtained for problems 1 to 3, respectively; its analysis support the conclusion that an increase from two to three hidden layers did not provide a great improvement in the MSE of the ANN for problems 1 and 3. In the second problem, the best ANN was the one with three intermediate layers and 13 neurons in each one.

Results obtained for the R-value are consistent with the ones presented for the MSE, revealing an average of 0.98 for the three time series and the optimal set of hyperparameters.

Following these experiments, the best ANN model for problem 1 was considered to present two hidden layers with 13 neurons each, while problem 2 would be addressed by an ANN with three layers and 13 neurons, and problem 3 with two layers and 25 neurons. These sets of hyperparameters were applied for the respective problems in the remaining weather stations of the cluster, with 500 repetitions, given the general increase in the accuracy with the number of repetitions. Resulting MSE in the test set are presented in Table 2.

4.2.2. Support Vector Regression

Twenty-one SVR models were developed, in accordance with the seven weather stations belonging to the cluster nearer the sea and the three types of problems previously discussed. For all of them, a radial basis function was used as a kernel. The computation of the hyperparameters (ε, C and the kernel scale γ) was performed through the Bayesian optimization algorithm, varying the values of these parameters within a predefined range and minimizing the cross-validation error.

The performance of these models in the test set is synthetized in

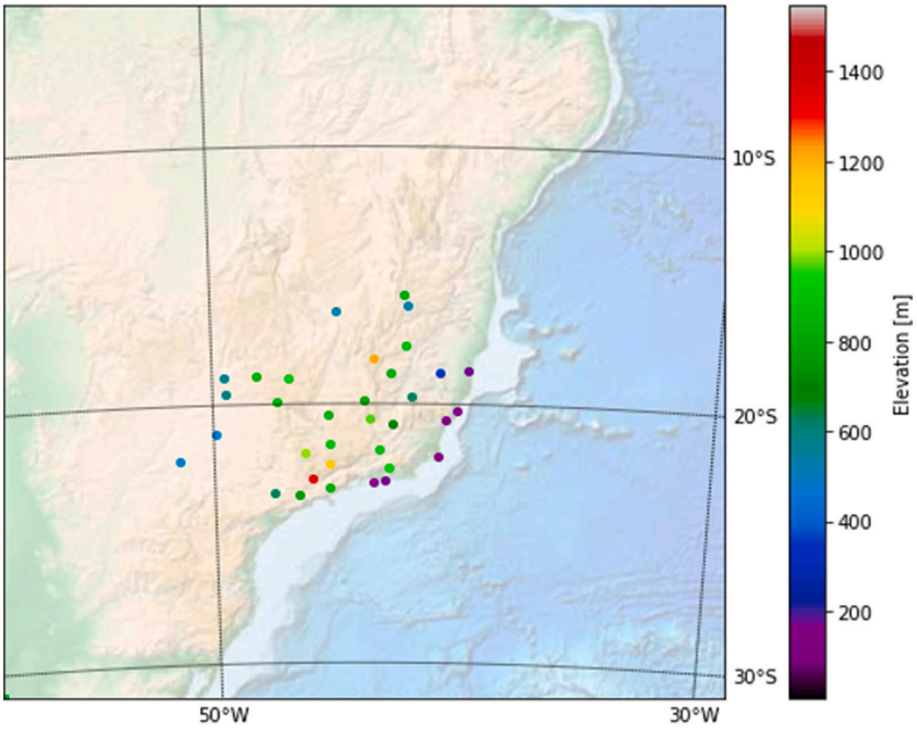

Fig. 11. Weather stations in southeastern Brazil, color-coded by their respective elevation.

Table 2; it is noted that, as for weather station 372 and problem 3, the procedure presented here failed to converge.

4.3. Performance evaluation

The analysis of Table 2 reveals that the best estimates for air temperature in a weather station are obtained with regression models that include the information of simultaneous climatic parameters in that station (problem 2). Incorporating data relative to the cluster largely increases the computational effort, without any improvement in the generalization error. Models of type 1 present the most significant MSE.

Additionally, ANN models seem to outperform SVR in the prediction of air temperatures, as the MSE obtained by implementing the first algorithm are typically smaller, considering all problems and weather stations.

As an illustrative example, Fig. 15 represents the air temperatures registered in the weather station 303 during the month of September 2016, along with the ones estimated in problem 2 by ANN and SVR algorithms.

5. Conclusions

Building energy analysis and simulation tools are an important component in achieving current international energy consumption and efficiency goals. To perform properly, these tools require a building's envelop environment information, which is typically acquired from neighbouring weather stations. Given the frequent malfunctions to which these stations are subject, the current study presented an integrated, machine learning approach, comprised of clustering algorithms and regression models, to rectify incorrect or null values in weather data sets. From the obtained results, the following conclusions are drawn:

- Both ANN and SVR provided accurate air temperature estimates for the sets of predictors in models of type 2 and 3, presenting themselves as effective alternatives when the recorded data is erroneous or missing. The first algorithm slightly outperformed the latter, and

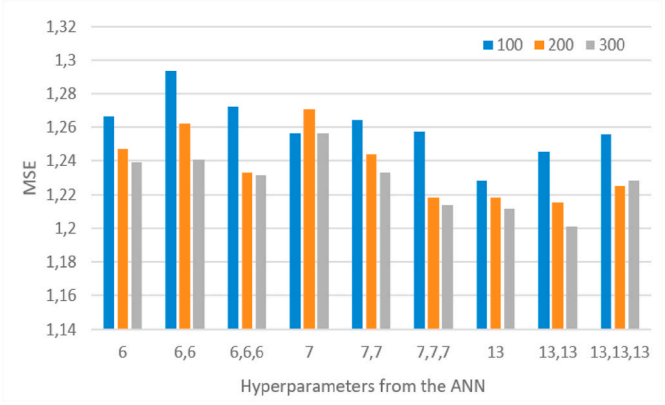

Fig. 12. MSE of ANN models with different hyperparameters, for problem 1 of weather station 303.

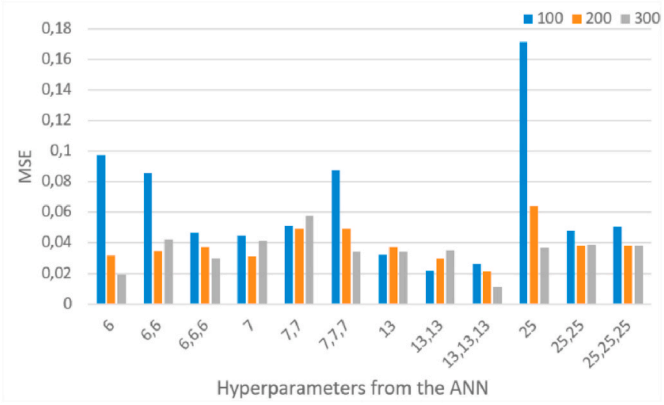

Fig. 13. MSE of ANN models with different hyperparameters, for problem 2 of weather station 303.

Table 1
Types of regression models developed for the response variable air temperature and each of the weather stations in the cluster nearer the sea.

Problem	Predictors	Number of Predictors
1	*air temperature in the remaining weather stations of the cluster*	6
2	*remaining climatic parameters in the weather station*	12
3	*remaining climatic parameters in the weather station and all of the climatic parameters in the remaining weather stations of the cluster*	90

Table 2
Generalization error of the ANN and SVR models developed for each weather station and evaluated in the test set (MSE).

Weather Station	Problem 1		Problem 2		Problem 3	
	ANN	SVR	ANN	SVR	ANN	SVR
303	1.2009	1.3269	0.0115	0.0368	0.0388	0.0694
306	2.0133	2.1106	0.0162	0.0672	0.0622	0.2535
307	1.2856	1.3643	0.0094	0.0073	0.0485	0.0260
348	2.6974	2.7700	0.0117	0.0099	0.0465	0.3805
372	3.5221	3.5572	0.0898	0.0611	0.2199	-
376	1.5738	1.5895	0.0445	0.0744	0.0531	0.1023
388	2.4753	2.5776	0.0089	0.0087	0.0998	0.0574

its hyperparameter optimization process was also faster and more efficient.

- Including data relative to the cluster largely increased the computational effort involved in the regression models, without any improvement of the generalization error. In fact, MSE values for models of type 3 are higher than those of type 2, for all the weather stations considered.
- Nevertheless, the methodology proposed in this research and the attained results demonstrate that the k-medoids algorithm, alongside the DTW similarity measure, is suitable for the automatic mapping of distinct climatic conditions, as evidenced by the five clusters in Fig. 9 and the elevation and proximity to the sea illustrated in Fig. 11. Moreover, the two-step approach of this paper may prove adequate in other weather data sets, which should be assessed in the future.

A natural extension of this work would comprise the development and validation of an online, real-time framework, in which machine learning algorithms classify, identify and correct inconsistent or missing values in continuously recorded weather data.

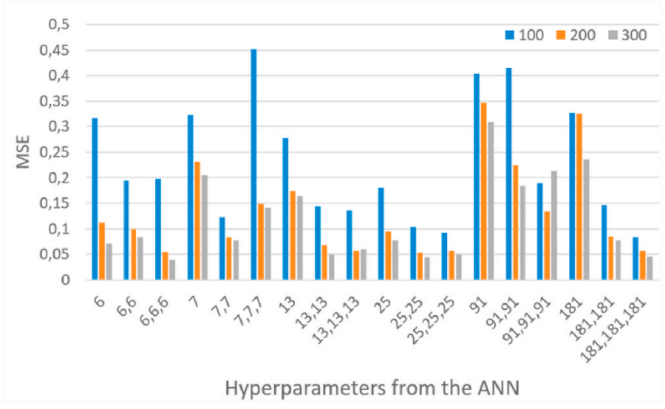

Fig. 14. MSE of ANN models with different hyperparameters, for problem 3 of weather station 303.

Author statement

Luís Sanhudo: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **João Rodrigues:** Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Ênio Vasconcelos Filho:** Methodology, Software, Validation, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

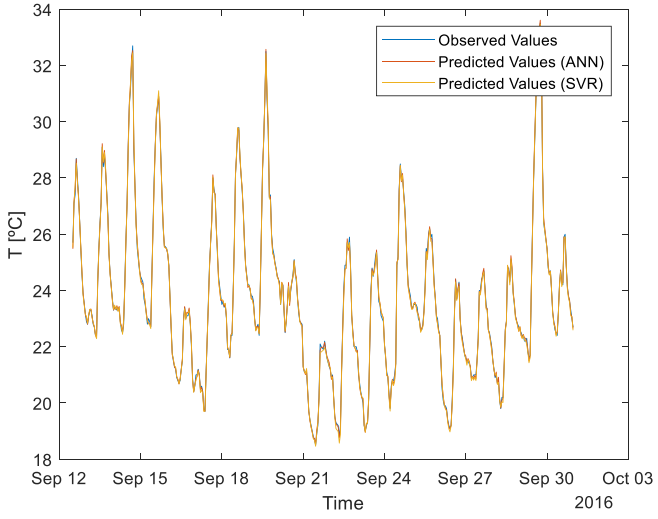

Fig. 15. Air temperatures registered in weather station 303 during September 2016, and corresponding values predicted through ANN and SVR models (problem 2).

Acknowledgements

The first and second authors would like to acknowledge, respectively, the PhD scholarships SFRH/BD/129652/2017 and SFRH/BD/129183/2017 awarded by Fundação para a Ciência e a Tecnologia (FCT).

This work was partially financially supported by UID/ECI/04708/2019 – CONSTRUCT –Instituto de I&D em Estruturas e Construções and UIDB/04234/2020 – CISTER Research Unit, both funded by national funds through the FCT/MCTES (PIDDAC).

References

[1] E. Recast, Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast), Official Journal of the European Union 18 (6) (2010) 2010.
[2] T.E. Commission, *Climate & Energy Package | Climate Action*. 2016 2016-11-23T07:50+01:00, 2020 [cited 2017 05 June]; Available from: https://ec.europa.eu/clima/policies/strategies/2020_en.
[3] L. Sanhudo, et al., Building information modeling for energy retrofitting–A review, Renew. Sustain. Energy Rev. 89 (2018) 249–260.
[4] Commission, E., Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A policy framework for climate and energy in the period from 2020 up to 2030. 2014, Brussels.
[5] E. Parliement, E. Council, DIRECTIVE (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018, Official Journal of the European Union, 2018 amending directive 2010/31/EU on the energy performance of buildings and directive 2012/27 on energy efficiency.
[6] A. Verbruggen, Retrofit of a century old land-house to a low-energy house, Int. J. Environ. Technol. Manag. 9 (4) (2008) 402–412.
[7] G. Verbeeck, H. Hens, Energy savings in retrofitted dwellings: economically viable? Energy Build. 37 (7) (2005) 747–754.
[8] T. Dixon, et al., Retrofitting commercial office buildings for sustainability: tenants' perspectives, J. Property Invest. Finance 26 (6) (2008) 552–561.
[9] Y. Ham, M. Golparvar-Fard, EPAR: energy Performance Augmented Reality models for identification of building energy performance deviations between actual measurements and simulation results, Energy Build. 63 (2013) 15–28.
[10] S.-M. Yu, Y. Tu, C. Luo, Green Retrofitting Costs and Benefits: a New Research Agenda, National University of Singapore, Singapore, 2011. Institute Real Estate Studies Working Paper Series, the Department of Real Estate.
[11] Z. Ma, et al., Existing building retrofits: methodology and state-of-the-art, Energy Build. 55 (2012) 889–902.
[12] Energy, U.S.D.o., Building Energy Software Directory, 2015.
[13] A. Farzaneh, D. Monfet, D. Forgues, Review of using Building Information Modeling for building energy modeling during the design process, Journal of Building Engineering 23 (2019) 127–135.
[14] F.H. Abanda, A.H. Oti, J.H.M. Tah, Integrating BIM and new rules of measurement for embodied energy and CO2 assessment, Journal of Building Engineering 12 (2017) 288–305.
[15] Z. Pezeshki, A. Soleimani, A. Darabi, Application of BEM and using BIM database for BEM: a review, Journal of Building Engineering 23 (2019) 1–17.
[16] A. Andriamamonjy, D. Saelens, R. Klein, A combined scientometric and conventional literature review to grasp the entire BIM knowledge and its integration with energy simulation, Journal of Building Engineering 22 (2019) 513–527.
[17] A. Rocha, et al., A case study to improve the winter thermal comfort of an existing bus station, Journal of Building Engineering 29 (2020) 101123.
[18] X. Xie, et al., Impact of neighbourhood-scale climate characteristics on building heating demand and night ventilation cooling potential, Renew. Energy 150 (2020) 943–956.
[19] R. Yao, et al., An integrated study of urban microclimates in Chongqing, China: historical weather data, transverse measurement and numerical simulation, Sustainable Cities and Society 14 (2015) 187–199.
[20] P. Eguía, et al., Weather datasets generated using kriging techniques to calibrate building thermal simulations with TRNSYS, Journal of Building Engineering 7 (2016) 78–91.
[21] M. Hosseini, F. Tardy, B. Lee, Cooling and heating energy performance of a building with a variety of roof designs; the effects of future weather data in a cold climate, Journal of Building Engineering 17 (2018) 107–114.
[22] Y. Geng, et al., Building energy performance diagnosis using energy bills and weather data, Energy Build. 172 (2018) 181–191.
[23] Y.J. Huang, D. Crawley, Does it Matter Which Weather Data You Use in Energy Simulations? Lawrence Berkeley National Lab., CA (United States), 1996.
[24] M. Bhandari, S. Shrestha, J. New, Evaluation of weather datasets for building energy simulation, Energy Build. 49 (2012) 109–118.
[25] F.D. Bender, P.C. Sentelhas, Solar radiation models and gridded databases to fill gaps in weather series and to project climate change in Brazil, Advances in Meteorology 2018 (2018).
[26] B. Henn, et al., A comparison of methods for filling gaps in hourly near-surface air temperature data, J. Hydrometeorol. 14 (3) (2012) 929–945.
[27] I. Zahumenský, Guidelines on Quality Control Procedures for Data from Automatic Weather Stations, World Meteorological Organization, Switzerland, 2004.
[28] J. Estévez, P. Gavilán, J.V. Giráldez, Guidelines on validation procedures for meteorological data from automatic weather stations, J. Hydrol. 402 (1) (2011) 144–154.
[29] Y. Radhika, M. Shashi, Atmospheric temperature prediction using support vector machines, International Journal of Computer Theory and Engineering (2009) 55–58.
[30] A. Paniagua-Tineo, et al., Prediction of daily maximum temperature using a support vector regression algorithm, Renew. Energy 36 (11) (2011) 3054–3060.
[31] M. Hossain, et al., Forecasting the weather of Nevada: a deep learning approach. Proceedings of the International Joint Conference on Neural Networks, 2015.
[32] D.A. Sachindra, et al., Statistical downscaling of precipitation using machine learning techniques, Atmos. Res. 212 (2018) 240–258.
[33] Y. Xiang, et al., A SVR–ANN combined model based on ensemble EMD for rainfall prediction, Appl. Soft Comput. 73 (2018) 874–883.
[34] R.C. Deo, M. Şahin, Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia, Atmos. Res. 161 (2015) 65–81.
[35] M.A. Mohandes, et al., Support vector machines for wind speed prediction, Renew. Energy 29 (6) (2004) 939–947.
[36] Q. Hu, R. Zhang, Y. Zhou, Transfer learning for short-term wind speed prediction with deep neural networks, Renew. Energy 85 (2016) 83–95.
[37] S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering–A decade review, Inf. Syst. 53 (2015) 16–38.
[38] X. Wang, et al., A Scalable Method for Time Series Clustering, Monash University, 2004. Technical Report.
[39] H. Zhang, et al., Unsupervised feature extraction for time series clustering using orthogonal wavelet transform, Informatica 30 (3) (2006).
[40] J. Paparrizos, L. Gravano, k-shape: efficient and accurate clustering of time series. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015.
[41] T.W. Liao, Clustering of time series data—a survey, Pattern Recogn. 38 (11) (2005) 1857–1874.
[42] A. Belhadi, et al., Space–time series clustering: algorithms, taxonomy, and case study on urban smart cities, Eng. Appl. Artif. Intell. 95 (2020) 103857.
[43] J. Paparrizos, L. Gravano, Fast and accurate time-series clustering, ACM Trans. Database Syst. 42 (2) (2017) 1–49.
[44] F. Ferstl, et al., Time-hierarchical clustering and visualization of weather forecast ensembles, IEEE Trans. Visual. Comput. Graph. 23 (1) (2017) 831–840.
[45] M. Rana, I. Koprinska, V.G. Agelidis, International Joint Conference on Neural Networks, in: Solar Power Forecasting Using Weather Type Clustering and Ensembles of Neural Networks, IJCNN), 2016.
[46] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, vol. 344, John Wiley & Sons, 2009.
[47] P.-N. Tan, Introduction to Data Mining, 2007. Pearson Education India.
[48] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[49] B. Pardeshi, D. Toshniwal, Advance Computing Conference (IACC), in: Improved K-Medoids Clustering Based on Cluster Validity Index and Object Density, IEEE, 2010. IEEE 2nd International. 2010.
[50] M.J. Rattigan, M. Maier, D. Jensen, Graph clustering with network structure indices. Proceedings of the 24th International Conference on Machine Learning, ACM, 2007.
[51] R. Pratap, et al., An efficient density based improved k-medoids clustering algorithm, IJACSA) International Journal of Advanced Computer Science and Applications 2 (6) (2011).
[52] S. Vishwakarma, P.S. Nair, D.S. Rao, A comparative study of K-means and K-medoid clustering for social media text mining, Int. J. 2 (11) (2017).
[53] T. Velmurugan, Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points, Int. J. Computer Technology & Applications 3 (5) (2012) 1758–1764.
[54] P. Arora, S. Varshney, Analysis of k-means and k-medoids algorithm for big data, Procedia Computer Science 78 (2016) 507–512.
[55] M. Müller, Information Retrieval for Music and Motion, vol. 2, Springer, 2007.
[56] A.W.-C. Fu, et al., Scaling and time warping in time series querying, The VLDB Journal—The International Journal on Very Large Data Bases 17 (4) (2008) 899–921.
[57] J. Aach, G.M. Church, Aligning gene expression time series with time warping algorithms, Bioinformatics 17 (6) (2001) 495–508.
[58] K. Gollmer, C. Posten, Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses, IFAC Proceedings Volumes 28 (12) (1995) 101–106.
[59] M.D. Schmill, T. Oates, P.R. Cohen, Learned models for continuous planning, AISTATS (1999).
[60] L.R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, vol. 14, PTR Prentice Hall, Englewood Cliffs, 1993.
[61] H.F. Silverman, D.P. Morgan, The application of dynamic programming to connected speech recognition, IEEE ASSP Mag. 7 (3) (1990) 6–25.
[62] J. Blackburn, E. Ribeiro, Human motion recognition using isomap and dynamic time warping. Human Motion–Understanding, Modeling, Capture and Animation, Springer, 2007, pp. 285–298.
[63] S. Sempena, N.U. Maulidevi, P.R. Aryan, Electrical Engineering and Informatics (ICEEI), in: Human Action Recognition Using Dynamic Time Warping, IEEE, 2011. International Conference on.

[64] S. Celebi, et al., VISAPP vol. 1 (2013). Gesture Recognition Using Skeleton Data with Weighted Dynamic Time Warping.
[65] P. Capitani, P. Ciaccia, in: Efficiently and Accurately Comparing Real-Valued Data Streams, SEBD, 2005.
[66] C.A. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints. Proceedings of the 2004 SIAM International Conference on Data Mining, SIAM, 2004.
[67] Berndt, D.J. and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. in *KDD Workshop*. 1994. Seattle, WA.
[68] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (5) (2007) 561–580.
[69] E. Keogh, C.A. Ratanamahatana, Exact indexing of dynamic time warping, Knowl. Inf. Syst. 7 (3) (2005) 358–386.
[70] T.M. Rath, R. Manmatha, Lower-bounding of Dynamic Time Warping Distances for Multivariate Time Series, University of Massachusetts Amherst Technical Report MM, 2002, p. 40.
[71] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. Speech Signal Process. 26 (1) (1978) 43–49.
[72] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 23 (1) (1975) 67–72.
[73] L. Jena, N.K. Kamila, Distributed data mining classification algorithms for prediction of chronic-kidney-disease, International Journal of Emerging Research in Management &Technology 4 (11) (2015) 110–118.
[74] M. Hossain, et al., Forecasting the weather of Nevada: a deep learning approach. Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE, 2015.
[75] S. Hudnurkar, A. Wanchoo, A. Malhotra, Optimized Artificial Neural Network Model for Day Ahead Maximum Temperature Forecasting, vol. 10, 2015, pp. 40647–40655.
[76] S. Osowski, K. Garanty, Forecasting of the daily meteorological pollution using wavelets and support vector machine, Eng. Appl. Artif. Intell. 20 (6) (2007) 745–755.
[77] A.-S. Chen, M.T. Leung, H. Daouk, Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index, Comput. Oper. Res. 30 (6) (2003) 901–923.
[78] E. Guresen, G. Kayakutlu, T.U. Daim, Using artificial neural network models in stock market index prediction, Expert Syst. Appl. 38 (8) (2011) 10389–10397.
[79] S. Haykin, A comprehensive foundation, Neural Network. 2 (2004) 41.
[80] M. Şahin, Comparison of modelling ANN and ELM to estimate solar radiation over Turkey using NOAA satellite data, Int. J. Rem. Sens. 34 (21) (2013) 7508–7533.
[81] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, Singapore, 2006.
[82] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Internal Representations by Error Propagation, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
[83] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386.
[84] K. Faceli, et al., Inteligência Artificial: Uma abordagem de aprendizado de máquina, 2011.
[85] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of control, signals and systems 2 (4) (1989) 303–314.
[86] D.E. Rumelhart, J.L. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Foundations vol. 1 (1986).
[87] L. Wang, et al., Optimal forecast combination based on neural networks for time series forecasting, Appl. Soft Comput. 66 (2018) 1–17.
[88] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[89] I. Bdmep, Banco de Dados Meteorológicos para Ensino e Pesquisa, 2019. Available from: http://www.inmet.gov.br/projetos/rede/pesquisa/.
[90] W.J. Krzanowski, Y. Lai, A Criterion for Determining the Number of Groups in a Data Set Using Sum-Of-Squares Clustering, Biometrics, 1988, pp. 23–34.
[91] D. Steinley, K-means clustering: a half-century synthesis, Br. J. Math. Stat. Psychol. 59 (1) (2006) 1–34.
[92] T.S. Madhulatha, An Overview on Clustering Methods, 2012 arXiv preprint arXiv: 1205.1117.
[93] T.M. Kodinariya, P.R. Makwana, Review on determining number of cluster in K-means clustering, Int. J. 1 (6) (2013) 90–95.