Short-term load forecasting using time series clustering

Ana Martins
Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal
e–mail: ana.martins@isel.pt
ORCID: 0000-0003-3733-6619

João Lagarto
Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal
INESC-ID, Portugal
e–mail: joao.lagarto@isel.pt
ORCID: 0000-0002-7047-6210

Hiren Canacsinh
Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal
e–mail: hiren.canacsinh@isel.pt

Francisco Reis
Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal
e–mail: freis@deea.isel.ipl.pt

Margarida G. M. S. Cardoso
ISCTE-IUL; BRU-IUL, Lisbon, Portugal
e–mail: margarida.cardoso@iscte-iul.pt
ORCID: 0000-0001-6239-7283

ABSTRACT

Short-term load forecasting plays a major role in energy planning. Its accuracy has a direct impact on the way power systems are operated and managed.
We propose a new Clustering-based Similar Pattern Forecasting algorithm (CSPF) for short-term load forecasting. It resorts to a K-Medoids clustering algorithm to identify load patterns and to the COMB distance to capture differences between time series. Clusters' labels are then used to identify similar sequences of days. Temperature information is also considered in the day-ahead load forecasting, resorting to the K-Nearest Neighbor approach.
CSPF algorithm is intended to provide the aggregate forecast of Portugal's national load, for the next day, with a 15-minute discretization, based on data from the Portuguese Transport Network Operator (TSO). CSPF forecasting performance, as evaluated by RMSE, MAE, and MAPE metrics, outperforms three alternative/baseline methods, suggesting that the proposed approach is promising in similar applications.

KEYWORDS

Clustering time series, Distance measures, Load pattern, Sequence Pattern, Similar Pattern Method, Short-term load forecasting.

1. INTRODUCTION

Short-term load forecasting (STLF) can be defined as the forecast of load with a time horizon varying from one day to two weeks (Hong and Shahidehpour 2015), and it is fundamental for several operational processes used by the electrical industry. Among these processes are the economic dispatch of generators, unit commitment, security assessment, and maintenance plans (Ilic et al. 2013). Accurate load forecasts are key for unit commitment because, on the one hand, an overestimation of the load might lead to the start-up of more production units than required, supplying more reserve than needed; on the other hand, an underestimation could lead to a low level of spinning reserve, rendering the power system vulnerable to failures (Fan and Hyndman 2012). Also, in deregulated electricity markets, it is of utmost importance for market participants to have an accurate load forecast, since profits and market shares can be compromised by forecasting errors (Ilic et al. 2013; Fan and Chen 2006). Even at a disaggregated level, with the increasing importance of smart grids, load forecasting can be key for demand side management (DSM) activities, such as load control and voltage regulation (Hong 2010).

Given the volatile, non-linear, and non-stationary nature of the load time series, as well as the diversity of factors that influence it, namely, meteorological (e.g. temperature), calendar (e.g. working and non-working days), and random factors, different techniques and approaches have been applied to STLF which can be found in many literature reviews (Hong and Shahidehpour 2015; Kuster et al. 2017).

Similar Pattern methods are a specific approach to the STLF problems. Their goal is to find similar daily load patterns in the historical dataset and, within these selected similar days, obtain a prediction by using an aggregation measure or some Machine Learning algorithm (Fallah et al. 2019). Different approaches and alternative similarity measures for identifying load patterns have been proposed in the literature (Fallah et al. 2019).

In this paper, a new STLF method is proposed: the Clustering-based Similar Pattern Forecasting (CSPF) method. It implements a Similar Pattern approach (so referred to by the typology of methods proposed by Fallah et al. (2019)) and is a two-step procedure: 1) CSPF first conducts a clustering analysis on daily loads time-series resorting to the use of K-Medoids, capitalizing on its ability to rely on medoids and to use different distance measures; 2) CSPF identifies sequences of days (allocated to clusters previously identified) and then proposes the use of the K-Nearest Neighbour algorithm as an instrument to filter the referred sequences, resorting to temperature data. The filtered days are finally used to provide the target day forecast.

The main contributions of this work may be summarized as follows:

- In step 1) of CSPF, we propose an innovative use of K-Medoids clustering analysis which is based on COMB distance (Cardoso et al. 2021, Cardoso and Martins in press). COMB is a convex combination of four (normalized) distance measures that offer complementary perspectives on the differences between two-time series: the Euclidean distance which captures differences in scale; a Pearson correlation-based measure that takes into account linear increasing and decreasing trends over time; a Periodogram based measure that expresses the dissimilarities between frequencies or cyclical components of the series; and a distance between estimated Autocorrelation structures, comparing the series in terms of their dependence on past observations.
- In step 2) of CSPF, we proposed a CSPF temperature-based filtering process which is expected to provide improved forecasts, since the sequences of days

that are considered in the final forecast are not only exhibiting similar consumption profiles, but also similar temperature profiles.
- Furthermore, the target days are categorized (e.g. considering special holidays) so that similar sequences are considered as a base for forecasting precede days of the same type.

The paper is organized as follows. In section 2, a literature review is conducted comprising STLF methods with a special focus on the Similar Pattern-based approaches to STLF. Then, in section 3, the new Clustering-based Similar Pattern Forecasting (CSPF) method is presented. Afterward, in section 4, a case study applied to the Portuguese power system illustrates the proposed approach together with a comparative analysis between CSPF, daily seasonal Naïve method, Pattern Sequence-based Forecasting (PSF) method, and a Semi-Parametric Additive method. Finally, we present some conclusions and directions for further research in section 5.

2. LITERATURE REVIEW

Among the different methods devoted to Short-term load forecasting (STLF), Statistical methods, such as Linear Regression methods or Semi-Parametric additive models, are commonly used. Machine Learning techniques have also been widely used for STLF. In the following literature review, after a brief examination of the several approaches used to STLF, we specifically focus on a particular Machine Learning approach - the Similar Pattern approach- which inspires our contribution in the domain.

2.1 Short-term load forecasting methods

The methods used for load forecasting are very diverse and include Statistical methods as well as Machine Learning techniques. However, despite the multiplicity of methodologies, there is no consensus on which one is the best (Hong and Fan 2016). The best methodology depends on the specific application at hand and the characteristics of the data. Currently, hybrid methods, that combine various methodologies and learning strategies, are generally viewed as enhancers of successful approaches in the field of forecasting and STLF in particular.

Among the statistical methods, Multiple Linear Regression methods have been applied to perform STLF (Hong 2010; Ružic 2003; Wang et al. 2016; Charlton. and Singleton 2014). In Ružic et al. (2003), the model's parameters are estimated using a set of days with loads and weather conditions similar to the ones expected in the target day (Euclidean distance is used to access this similarity). In Hong (2010), the relationship between load and temperature is modeled by a third-degree polynomial. In Wang (2016), to model the recency of the effect of temperature on load, the previous model is complemented with lagged temperature values and moving averages of daily temperature. In Charlton and Singleton (2014), a regression model of load as a function of temperature and day of the year is refined by combining models from multiple weather stations, removal of outliers, and analysis of public holidays.

Also, Semi-Parametric Additive models, which allow accommodating the nonlinear relationship between temperature and demand, and the autocorrelation of model residues, have been used in STLF – (Fan and Hyndman 2012; Goude et al. 2014; Gaillard et al. 2016). The authors in Fan and Hyndman (2012) develop a Semi-Parametric Additive model for the logarithmic demand. Cubic Splines are used to model temperature and annual load effects. A Bootstrap method is also proposed to

obtain prediction intervals. In Goude et al. (2014), the authors suggest an approach based on generalized additive models that estimate the relationship between load and temperature, calendar variables, and others, where the temperature is modeled by exponential smoothing. In Gaillard et al. (2016), a quantile generalized additive model is fitted in a load forecasting approach that, firstly, produces temperature scenarios that then are used in a probabilistic forecasting load model.

Machine Learning techniques, in the field of Artificial Intelligence, do not generally require determining explicit complex functional relationships while dealing with non-linearities of time series modeling (Metaxiotis et al. 2003). However, these techniques, typically, do not allow to fully understand the relationships between load and its determinants. Among the most used Machine Learning techniques are the Artificial Neural Networks (ANN) – (Ilic et al. 2013; Azadeh et al. 2014; Fan and Chen 2006; Sharifzadeh et al. 2019; Mohandes 2002; Cheng and Wei 2010; Dedinec et al. 2016; Heydari et al. 2020). In Ilic et al. (2013), a feed-forward multi-layer perceptron ANN is used to perform STLF of a Serbian utility. The ANN is used with a preprocessing unit that allows reducing the size of the input space, thus, improving the training time and the generalization capability of the ANN. Azadeh et al. (2014) also use ANN that captures seasonal features of the load to forecast the Iranian electricity market load. In Sharifzadeh et al. (2019), conventional ANN, Support Vector Machines (SVM), and Gaussian process regression are used to predict wind and solar power, and demand. From the models used, only ANN successfully performed the forecasting of demand. The use of SVM for load forecast has the advantage of achieving higher generalization performance since it tends to avoid over-fitting (Fan and Chen 2006). Considering Fan and Chen (2006), a hybrid network with Self-Organizing Map (SOM) and SVM is used. The SOM network clusters the input data into subsets that then are used in the SVM to predict the next day's load profile. The studies of Sharifzadeh et al. (2019) and Mohandes (2002) are also examples of the use of SVM to perform STLF. With a different approach, Cheng and Wei (2010) use an Adaptive-Network-based Fuzzy Inference System (ANFIS) to forecast the regional electricity load in Taiwan. Firstly, the authors incorporate the one-step ahead method into the ANFIS model. Then, to improve the forecasting capability, they use an adaptive forecasting model to modify the forecast produced by the ANFIS model. In its turn, Dedinec et al. (2016) applies a deep belief network constituted by multiple layers of restricted Boltzmann machines to forecast the Macedonian hourly electricity consumption from 2008 to 2014. The authors use a layer-by-layer unsupervised training procedure to train previously the initial values of the weights in the network, then use a supervised back-propagation training method to fine-tune the parameters. Heydari et al. (2020) propose a combined model that includes a mixed data model based on variational mode decomposition and, a combination of a generalized regression neural network and gravitational search algorithm used as a feature selection model to select the best features of different load and price forecasting signals. The combined model is tested with data from the Pennsylvania-New Jersey-Maryland (PJM) and Spanish power markets, as well as from the real load of the Favignana Island power grid.

2.2 Similar Pattern methods

Similar Patterns methods address the heterogeneity of the data first, commonly relying on some measure of distance, providing a preliminary data analysis that can potentially precede and be incorporated in several Machine Learning and Pattern Recognition algorithms (Duch 2000). In time series forecasting, these methods generally resort to measures of similarity between sections – e.g., seasonal cycles - of the historical data.

Regarding STLF, the load time series data are divided into daily cycles with length n (e.g., n=24 for hourly data, n=96 for 15-minute interval data) and the goal is to find similar daily load patterns within the historical dataset. Considering the selected similar days, a prediction can then be obtained by using an aggregation measure or some Machine Learning algorithm. The authors of Fallah et al. (2019) present a review of Similar Pattern methods including different techniques and alternative similarity measures for identifying load patterns.

The simpler approach resorts to searching, in the historical dataset, the most similar days – similar weather, day of the week, and date – to the forecast day. For example, Chen et al. (2010) propose the identification of similar days, considering the weekday index, the day-of-year neighboring, and weather conditions – wind-chill temperature, air temperature, wind speed, humidex, and dew-points. Days with similar weather conditions are selected by minimizing the Euclidean distance of the weather conditions between the target day and historical days with the same weekday and time of the year. In Mu et al. (2010), each day of the historical dataset is described by a vector of indexes stating the impact of several factors, namely weather conditions, the weekday, and special holidays. The similarity measure between two days is the cosine of the angle between the corresponding vectors. The load forecasting is then a weighted average of daily loads in which the larger weights express higher similarity between daily characteristics. In Mandal et al. (2006), it is proposed an ANN where a weighted Euclidean distance is used for selecting similar days using load deviations and load slope deviations between forecast day and historical days and temperature deviations. The weights are determined using the least squares regression model. The selection of similar days is limited to the same season where the target day is included.

Clustering time series for pattern discovery aims to determine a set of patterns that most accurately represent the original data set, in a way that every time series data can be identified with one of the patterns discovered (Iglesias and Kastner 2013). The authors of Zheng et al. (2017) propose an approach for similar days selection using a weighted Euclidean distance and resorting to the K-Means clustering procedure. Weights considered refer to features and are obtained through an extreme gradient boosting algorithm (Xgboost). Features included referring to climate factors, day type (e.g. weekend or weekday), and also the day-ahead peak load.

In Martinez-Alvarez et al. (2010), it is proposed an approach called Pattern Sequence-based Forecasting (PSF). First, PSF relies on the K-Means algorithm (using Euclidean distance) to cluster the daily (normalized) load data. The selection of the number of clusters results from voting of three clustering validity indices – the average of Silhouette, the Dunn index, and the Davies-Bouldin index. Afterward, the pattern sequences are extracted i.e., days before the forecast day are labeled according to the cluster they belong to. Finally, all the sequences in the historical data that match the sequence referring to the target day are considered for prediction.

In Jin et al. (2015), the authors use a cluster pattern sequence approach and ANN techniques for STLF. In Jin et al. (2015) work, SOM is used to cluster daily load time series and each cluster label is represented by its unique topological coordinates yielded by the algorithm. Considering the pattern sequences of days (represented by their coordinates), an ANN is trained to predict the pair of coordinates of the day to forecast.

The present work capitalizes on the Similar Patterns methods general approach by first dealing with the time-series heterogeneity. Although the consumption of electric energy presents annual and weekly seasonal behavior, not all days of the same type present a similar daily load profile. For example, a working day after a holiday, or

between a holiday and a weekend, does not have the same energy consumption as a normal weekday. This is important, particularly in countries such as Portugal where working days that fall in between a holiday and a weekend have distinct load profiles of the same day of the week that falls between a weekend and a working day. E.g., a Monday that precedes a Tuesday that is a public holiday has a different load profile than a regular Monday. Furthermore, the definition of yearly seasonal effects is not clear, especially in the spring and autumn periods. Thus, for discovering groups of days with similar load profiles we resort to a Cluster analysis.

Also, it follows the general two-step approach proposed by Martinez-Alvarez et al. (2010), by conducting clustering of daily time series first and then extracting similar sequences of days. However, we try to address the following specific issues: i) the need to consider centroids (means of time-series) for clustering, which occurs in K-Means as well as in SOM; ii) the consideration of a very specific measure of dissimilarity between time-series, which can bias the way differences between time-series are viewed and is a common practice (e.g. by using Euclidean distance emphasis is placed on differences in scale); iii) the consideration of sequences of days without including relevant information on the target days' type (e.g. special holidays).

We, therefore, propose the use of K-Medoids to conduct the clustering analysis since it does not resort to centroids but to medoids (a specifically observed time series that can be viewed as the representant of a cluster); furthermore, K-Medoids allow the incorporation of diverse distance measures and thus we can use a convex combination of four distance measures (Euclidean, Pearson-based, Periodogram-based and Autocorrelation-based) in an attempt to capture different features of time series. Also, in the second phase of the method, the forecasting phase, we take into consideration the type of day to forecast (weekdays, holidays, and special holidays) when choosing the days that have a similar pattern sequence of days previous to the target day.

Finally, the forecast of the target day load also considers the temperature profiles of the days within the extracted similar sequences of days.

3. THE PROPOSED ALGORITHM

The proposed Clustering-based Similar Pattern Forecasting algorithm (CSPF) is intended to provide the aggregate forecast of Portugal's national load, for the next day, with a 15-minute discretization, based on data from the Portuguese Transport Network Operator (TSO). The CSPF method is a two-step approach illustrated in Fig. 1:

Step 1) A clustering algorithm resorting to COMB distance (a combination of diverse distance measures) is used for discovering clusters of days (n periods long time series) exhibiting similar load patterns. Thus, each day a cluster label is allocated, and load pattern sequences are formed. Then, days with the same type of day as the target day to forecast and exhibiting similar sequences of clusters labels in the previous days are selected.

Step 2) Among the days selected in step 1), we consider their temperature profile and implement a search for nearest neighbors. The load forecasting is then obtained based on these neighbors' days' loads. In the following, a more detailed explanation will be given.

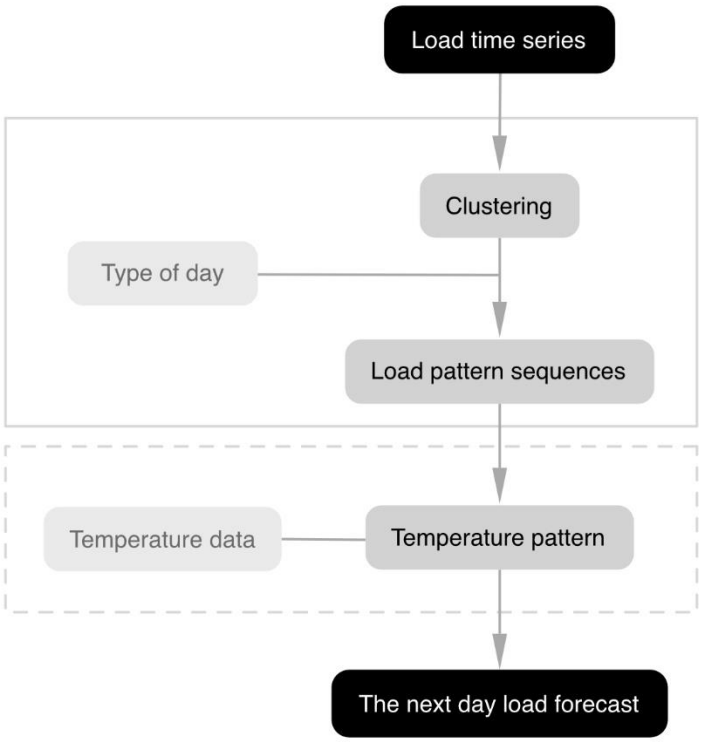

Fig. 1 The CSPF algorithm overview

3.1 Clustering

Consider that the load time series data is divided into N daily cycles represented by $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$, where $\boldsymbol{x}_d = (x_{d,1}, x_{d,2}, \dots, x_{d,n}), (d = 1, \dots, N)$ represents a daily load data with n periods – for example for hourly data $n = 24$, for 15-minute interval data, $n = 96$.

To cluster load time-series data and constitute well-separated groups of days, with each cluster including days having similar load profiles, we adopt the K-Medoids algorithm, Kaufman and Rousseeuw (2009). K-Medoids aims at the minimization (for all clusters) of the distance between time-series belonging to a cluster from the cluster's Medoid i.e. a time-series that exhibits the smallest distance to all the other elements of the cluster. It is somewhat more flexible in terms of cluster shapes and more robust to outliers and noise than K-Means. Also, by considering a Medoid (a member of the data set), it overcomes the need to determine a Centroid, based on an averaging of different series, which can be a problematic issue.

Furthermore, the K-Medoids capacity of dealing with several distance measures is a critical aspect of our approach. We resort to the COMB distance (Cardoso et al. 2021) a convex combination of four (normalized) distance measures: the Euclidean distance, d_{Eucl}, captures differences in values; a Pearson correlation based distance, $d_{Pearson}$, emphasize differences in trends; the Euclidean distance between periodograms, captures differences in cyclical behaviors and the Euclidean distance between estimated autocorrelation functions stresses the differences regarding the dependence on past observations.

As Pearson-based distance, we consider the rooted normalized one-minus-correlation distance measure proposed by Rodrigues (2008):

$$d_{Pearson} = \sqrt{\frac{1 - \rho_{\boldsymbol{x}_{d_1}, \boldsymbol{x}_{d_2}}}{2}}, \quad (1)$$

with $\rho_{\boldsymbol{x}_{d_1}, \boldsymbol{x}_{d_2}}$ representing the Pearson correlation between the load time series $\boldsymbol{x}_{d_1}$ and $\boldsymbol{x}_{d_2}$, at days d_1 and d_2, respectively. This distance is invariant to scale and $0 \leq d_{RNOMC} \leq 1$.

The Euclidean distance between $\boldsymbol{x}_{d_1}$ and $\boldsymbol{x}_{d_2}$ is a one-to-one measure that considers the closeness of the observations indexed in time (e.g. Montero and Vilar 2014).

$$d_{Eucl} = \left(\sum_{t=1}^{n} \left(\boldsymbol{x}_{d_1,t} - \boldsymbol{x}_{d_2,t}\right)^2\right)^{1/2}. \quad (2)$$

The Euclidean distance between the periodograms, (Caiado et al. 2006), is also adopted, expressing the contribution of the various frequencies or cyclical components to the variability of the daily load series. Thus, we consider this distance between $P_{\boldsymbol{x}_{d_1}}(w_j) = (1/n)\left|\sum_{t=1}^{n} \boldsymbol{x}_{d_1,t} e^{-itw_j}\right|^2$ and $P_{\boldsymbol{x}_{d_2}}(w_j) = (1/n)\left|\sum_{t=1}^{n} \boldsymbol{x}_{d_2,t} e^{-itw_j}\right|^2$ the periodograms' for $\boldsymbol{x}_{d_1}$ and $\boldsymbol{x}_{d_2}$, respectively, at frequencies $w_j = 2\pi j/n, j = 1,2,\ldots,[n/2]$ (where $[n/2]$ is the largest integer less or equal to $n/2$).

Finally, we consider the estimated autocorrelations functions $ACF(\boldsymbol{x}_{d_1})$ and $ACF(\boldsymbol{x}_{d_2})$ that represent the autocorrelations functions of $\boldsymbol{x}_{d_1}$ and $\boldsymbol{x}_{d_2}$, respectively, and adopt the Euclidean distance between these estimated functions, comparing the series in terms of their dependence on past observations, (Montero and Vilar 2014).
Before combining the distances, each one of the distances is normalized using a min-max transformation,

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \quad (3)$$

where x represents a distance measure, and $min(x)$ and $max(x)$ are the minimum and maximum of x, thus guaranteeing normalized values range from 0 to 1. Then, a convex combination of the four (normalized) distances referred is considered in the clustering procedure (Cardoso et al. 2021).
For determining the best number of clusters, the K-Medoids algorithm is used considering a range for the number of clusters. For each of these solutions, four measures are calculated: Average Silhouette (Kaufman and Rousseeuw 2009), Calinski and Harabasz (Calinski and Harabasz 1974), Dunn modified index (Bezdek and Pal 1998) and the relative improvement or rate of change in within clusters' variation between two successive solutions (with k − 1 and k clusters). A higher value of each of these indices suggests a better clustering solution, that is a solution with more compact and well separated clusters. All indices' values are normalized, using (3), and then, for each candidate number of clusters, a summated indicator of all indices is calculated with its maximum value indicating the best number of clusters, according to these indices.
For the implementation of K-Medoids we use R package "cluster" (Maechler et al. 2013). All distance measures are implemented in the R package "TSclust" (Montero and Vilar 2014). The cohesion-separation measures are all implemented in the "fpc" R package (Hennig 2020). These are auxiliary packages for the R implementation of CSPF.

3.2 Load pattern sequences

The goal of this phase is to select the days in the historical dataset that have load patterns in the preceding days that are similar to the load patterns (sequences of clusters labels) in the days prior to target day. Also, these selected days must be of the same type as the target day to forecast. For each day d $(d = 1, \dots, N)$ is known the type of day, wk_d, which includes the days of the week (Sunday, Monday, …, Saturday), a Holiday category, and where some especial holidays can also be considered.

The load pattern sequences procedure is summarized in Fig. *2* The CSPF algorithm

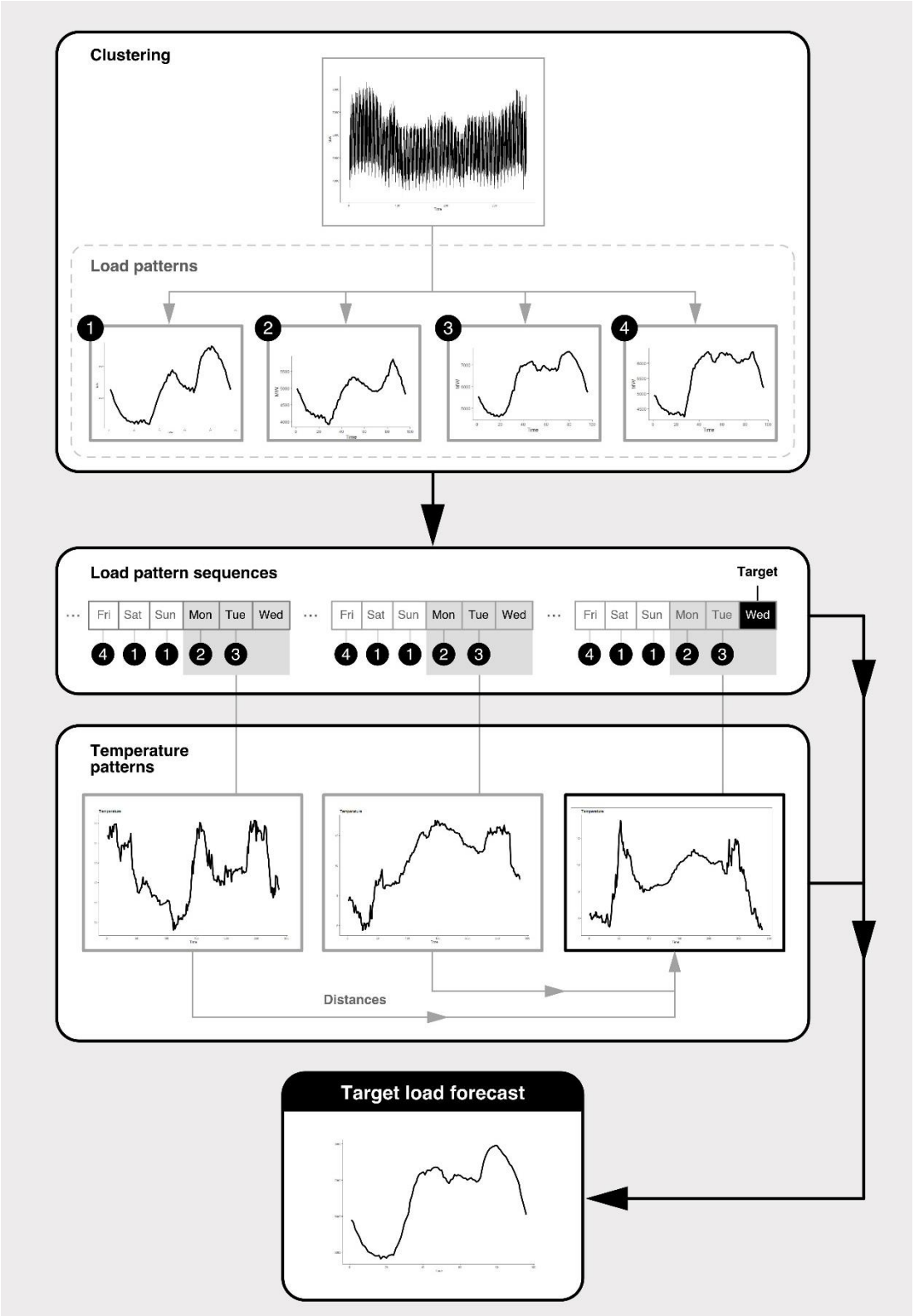

Fig. 2 The CSPF algorithm

As result of the clustering procedure a sequence of labeled days is obtained, $L_1, L_2, \dots, L_N$ where L is the label of the cluster of the day d $(d = 1, \dots, N)$. Consider also the sequence of labels of the p days immediately before the target day to forecast: $L^* = L_{N-p+1}, \dots, L_{N-3}, L_{N-2}, L_{N-1}, L_N$. The objective is to search in the dataset for all

the sequences equal to L^* that also are followed by a day of type wk_{N+1} (type of day corresponding to target). Finally, all the days immediately after the selected sequences are kept. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N_1}$ be the load of these selected N_1 days which have a load profile in the previous days like the load profile of the days before the day to forecast and also are of the same type of day as the target day.

If in the data available there is not any day that fulfills these two conditions, the number p of previous days is reduced (subject to $N_1 \geq 1$).

3.3 Temperature pattern

Having selected the N_1 days, that are of the same type as the target day and have similar load profile (clusters' sequence) in the preceding days as the days before the target day, we now conduct a filtering process according to information available on temperature. Let $\boldsymbol{T}_1, \boldsymbol{T}_2, \dots, \boldsymbol{T}_N$, $\boldsymbol{T}_d (d = 1, \dots, N)$ represent the temperature time series data in daily cycles, where in each day the temperature is recorded in m intervals, $\boldsymbol{T}_d = \left(T_{d,1}, T_{d,2}, \dots, T_{d,m}\right), d = 1, \dots, N$.

Let P be a proportion of the number of N_1 days, identified in the previous step, a parameter to be set by the analyst. Then, the $N_2 = round(P * N_1, 0)$ filtered days to keep will provide the ground for forecasting the target day.

The selection of these N_2 days is conditional to temperature patterns: we consider the temperature of each of the N_1 days and also of the q days preceding them, that is, the time series with $m * (q + 1)$ observations

$\boldsymbol{T'}_d = (\boldsymbol{T}_{d-q+1}, \dots, \boldsymbol{T}_{d-1}, \boldsymbol{T}_d)$, $d = 1, \dots, N_1$. Then, we measure the distances $d(\boldsymbol{T}^*, \boldsymbol{T'}_d)$ between $\boldsymbol{T'}_d$ and the temperatures time series referring to the forecast day $\boldsymbol{T}^* = (\boldsymbol{T}_{N-q+1}, \dots, \boldsymbol{T}_N, \widehat{\boldsymbol{T}}_{N+1})$. Finally, we keep the N_2 nearest neighbor's days according to temperature.

Note that for evaluation purposes we consider a test dataset in which $\boldsymbol{x}_{N+1}$ and $\boldsymbol{T}_{N+1}$ are known and $\widehat{\boldsymbol{T}}_{N+1} = \boldsymbol{T}_{N+1}$.

3.4 The next day load forecast

The goal of this last step of the algorithm is to predict the forecast for day N+1, $\widehat{\boldsymbol{x}}_{N+1} = \left(\hat{x}_{N+1,1}, \hat{x}_{N+1,2}, \dots, \hat{x}_{N+1,n}\right)$ using the load of the selected N_2 days, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N_2}$. For this end we compute a weighted mean of loads $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N_2}$

$$\widehat{\boldsymbol{x}}_{N+1,i} = \sum_{d=1}^{N_2} \frac{d_{1.0}(\boldsymbol{T}^*{}_i, \boldsymbol{T'}_{1,i})}{\sum_{d=1}^{N_2} d_{1.0}(\boldsymbol{T}^*{}_i, \boldsymbol{T'}_{d,i})} x_{d,i} \ , i = 1, \dots, n \tag{4}$$

were the weights considered, $d_{1.0}(\boldsymbol{T}^*, \boldsymbol{T'}_d)$, are the distances $d(\boldsymbol{T}^*, \boldsymbol{T'}_d)$ transformed,

$$d_{1.0}(\boldsymbol{T}^*, \boldsymbol{T'}_d) = 1 - d(\boldsymbol{T}^*, \boldsymbol{T'}_d)_{norm} \tag{5}$$

with $d(\boldsymbol{T}^*, \boldsymbol{T'}_d)_{norm}$ defined by (3), such that values near zero indicate higher distances and thus less weight in the forecasting.

3.5 Forecasting accuracy

For evaluating the forecasting accuracy, we resort to three measures most commonly used in the literature (e.g., Hyndman and Athanasopoulos 2021) namely, RMSE (Root Mean Square Error), MAE (Mean Absolut Error) and MAPE (Mean Absolute Percentage Error). The forecast errors in period t, e_t, is defined as the difference

between the observations and the corresponding forecasted values, $e_t = x_t - \hat{x}_t$. Considering the daily load data with n periods, the accuracy measures are defined by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{6}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{T} \left|\frac{e_t}{x_t}\right|. \tag{7}$$

Both RMSE and MAE are on the same scale as the data and the MAPE is unit-free.

4. CASE STUDY

The proposed approach is applied to the years 2014-2017 time series data of the Portuguese Transmission System Operator (TSO) including load (referred to as emission which includes the losses) and temperature data, both in 15-minutes intervals. The data were obtained through the operators' website. These data are used to obtain the day-ahead load forecast with the discretization of 96 periods. For implementation reasons, due to different winter and summer times, the raw data have one missing hour (Daylight Saving Time), which was imputed by the average of the two nearest hour data, and also a redundant hour data that was removed.

4.1 Data analysis

Load time series are volatile, non-linear, and non-stationary and depend on multiple factors, namely, meteorological (e.g., temperature), calendar (e.g., holidays, weekends, working days), network topology (e.g., load shifting), and random noise.
The electrical load time series data is presented in Fig. *3* where the annual seasonality is clear. The empirical autocorrelation function is exhibited in Fig. *4* where we can realize the daily and weekly variation. Moreover, the electricity consumption depends on the type of weekday as can be seen in Fig. *5*, with a larger difference between weekend and non-weekend days.
As load and temperature data are in 15-minutes intervals, $n = m = 96$. From the available dataset, the year 2017 dataset is separated and considered for testing the forecasting procedure. Thus, the remaining dataset, $N = 1096$ days with $n = m = 96$ are considered for learning with the proposed CSPF algorithm.

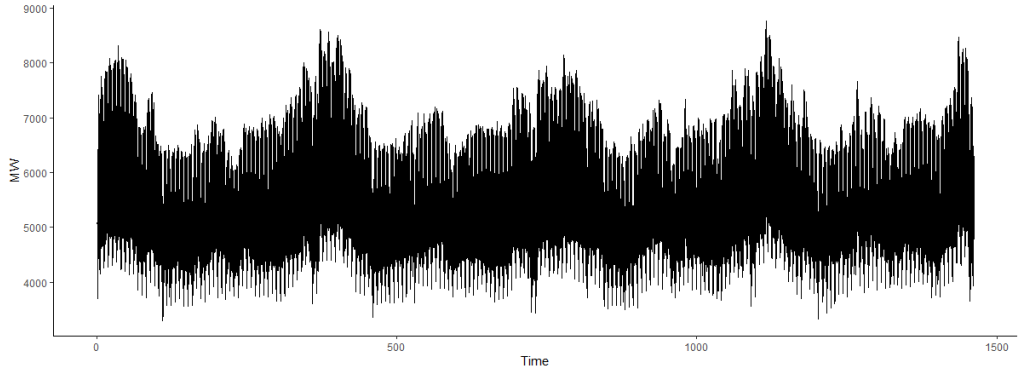

Fig. 3 The load time series for the 2014-2017 period.

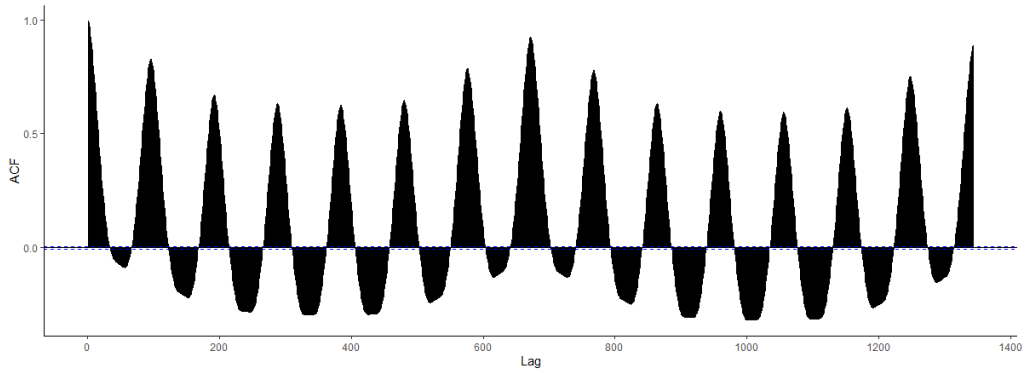

Fig. 4 The load time series ACF.

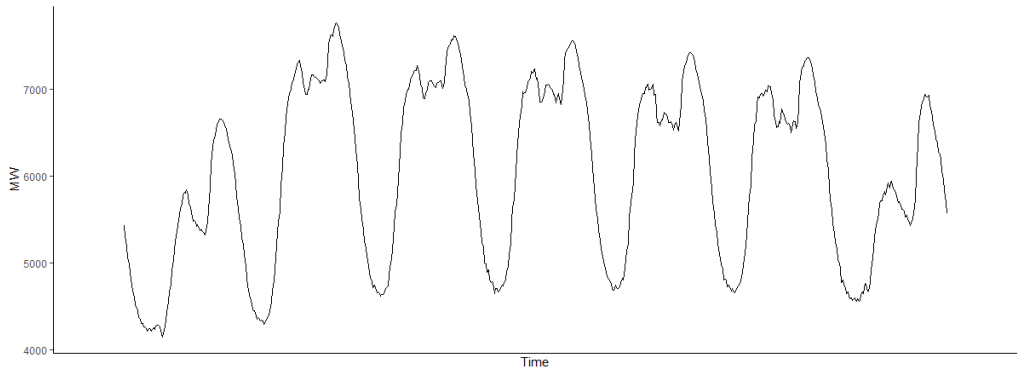

Fig. 5 Weekly load data: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday.

The non-linearity relationship between load and temperature is presented in Fig. ***6*** where it is also exhibited its dependence on the hour of the day.

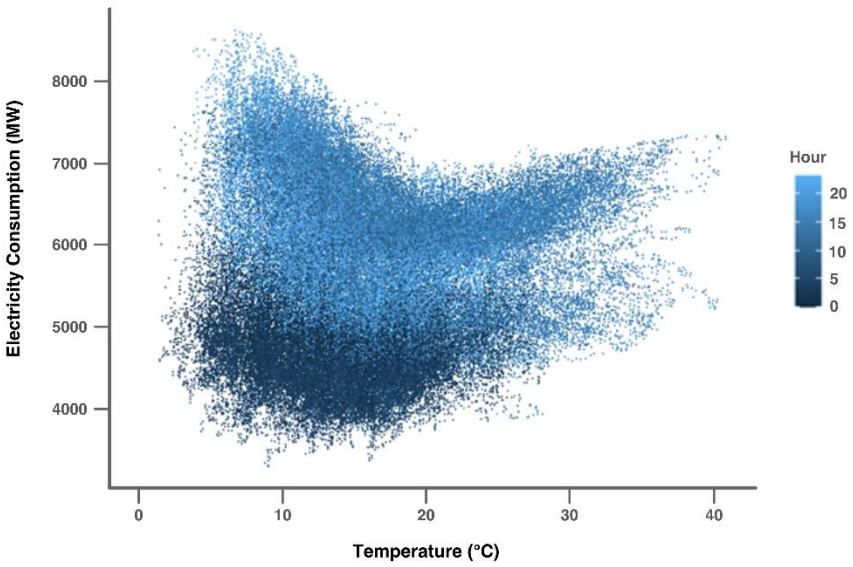

Fig. 6 The relationship between electricity consumption and temperature.

4.2 Parametrization and CSPF results

Several empirical experiments and experts' consultations were considered to tune the algorithm parameters.

For the Portuguese data we consider three special holidays - Christmas, New Year, and Carnival.

For the clustering of daily load data, the four distances were given the same (uniform) weights.

The clustering results indicate two very well-separated daily pattern groups. The characterization of these clusters indicate that Group 1 contains almost all weekend days and Group 2 the weekdays – Fig 7. It is worth noting that Group 1 also includes several days before, after, or between holidays, days that have a similar load profile to weekend days. In addition, in Group 2, there are several days that, despite being working days, have a load profile similar to weekend days, and this similarity was captured by the Cluster analysis.

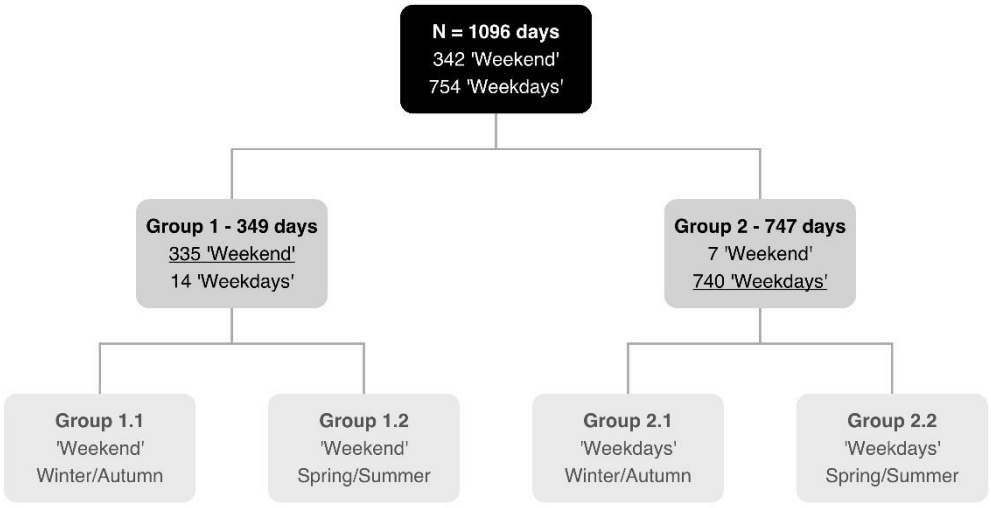

Fig 7 The clustering results

In order to go further in the categorization of the daily load cycles, we cluster once again the observations within each group. Each of the groups resulting from the first clustering procedure is then divided into two groups – Fig 7 and Table 1.

Table 1 Clusters' characterization by month

	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Out	Nov	Dec
Group 1.1	31	25	7	0	0	0	0	0	0	4	28	35
Group 1.2	0	0	22	31	31	28	26	33	24	24	0	0
Group 2.1	62	58	47	6	0	0	0	0	0	3	45	54
Group 2.2	0	2	17	53	62	62	67	60	66	62	17	4

Clusters obtained capture also the differences regarding the season of the year, while uncovering the different daily load patterns - Fig 8 The clusters medoids

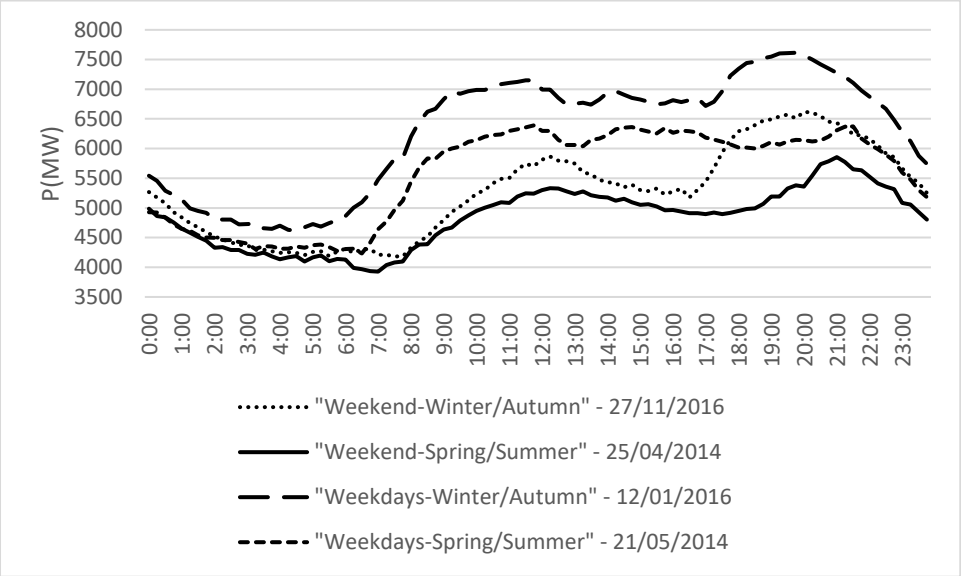

Fig 8 The clusters medoids

For the load pattern sequence search, we considered a window with the last five days, $p = 5$.

For the filtering based on temperature profiles, only the Euclidean distance was considered. This was decided as a result of various empirical experiences carried out, where the value itself of the temperature was revealed to be more important than other movements in the daily profile.

For the temperature pattern, only the two days preceding the target day were considered, i.e. $q = 2$. Finally, $P = 0.2$, that is 20% of the most similar days, according to temperature, was considered to estimate the day-ahead load.

The algorithm was applied to predict the day-ahead load corresponding to 96 periods of 15 minutes. The results of forecasting accuracy of CSPF method are summarized in Worth noticing is the fact that the maximum MAPE obtained across all remaining months is 4.8%.

The best forecasting performance were achieved in May, August and November, with MAPE values between 2.7% and 2.8%. Forecasting in December proved to be the most difficult task, with greater forecasting errors. It is worth mentioning that in Portugal this month has many holidays turning searching similar patterns more difficult due to historical dataset limitations. Worth noticing is the fact that the maximum MAPE obtained across all remaining months is 4.8%.

4.3 Comparative Performance

For comparison purposes we resort to the following approaches:

A. The Seasonal Naïve (**SN**) method is one of the simplest benchmark methods for seasonal data. Considering the daily seasonality, each forecast is equal to the last observed value for the same period of the day.
B. An algorithm from the same family of Similar Patterns based methods: the Pattern Sequence-based Forecasting (**PSF**) algorithm (Martinez-Alvarez et al. 2010) with R implementation in package "PSF" (Bokde et al. 2016).
C. A Semi-Parametric Additive (**SPA**) method to forecast half-hourly electricity demand, implemented in R - Hyndman and Athanasopoulos (2021). SPA deals with multiple seasonality using harmonic regression. The type of day – working or non-working day is also considered. The temperature is modeled with a piece-wise linear function. Finally, an ARIMA model is selected using the AICc criterion.

Table 2 CSPF forecasting comparative performance

	RMSE (MW)	MAE (MW)	MAPE (%)		RMSE (MW)	MAE (MW)	MAPE (%)
January				July			
CSPF	**315.6**	**260.3**	**4.0**	**CSPF**	**307.0**	**271.9**	**4.8**
SN	661.2	424.8	7.2	SN	597.9	397.1	7.3
PSF	1541.2	1356.8	20.2	PSF	779.8	624.0	10.3
SPA	470.7	370.0	5.8	SPA	294.6	230.7	4.1
February				August			
CSPF	**417.8**	**270.6**	**4.5**	**CSPF**	**200.3**	**151.7**	**2.7**
SN	557.0	374.2	6.4	SN	416.2	273.2	5.1
PSF	747.6	538.2	9.5	PSF	574.8	417.6	8.0
SPA	368.4	272.1	4.4	SPA	269.6	187.8	3.4
March				September			
CSPF	**395.8**	**249.3**	**4.3**	**CSPF**	**267.3**	**216.4**	**3.8**
SN	567.7	368.4	6.7	SN	560.1	350.4	6.6
PSF	558.1	435.8	7.7	PSF	568.0	374.2	7.3
SPA	319.1	248.8	4.3	SPA	291.6	242.7	4.6
April				October			
CSPF	**291.5**	**172.1**	**3.4**	**CSPF**	**283.3**	**197.3**	**3.6**
SN	572.5	404.4	8.1	SN	595.0	393.8	7.6
PSF	694.8	553.0	11.4	PSF	942.38	745.3	12.7
SPA	355.5	254.0	4.8	SPA	280.9	219.5	4.2
May				November			
CSPF	**280.7**	**141.5**	**2.7**	**CSPF**	**223.5**	**162.5**	**2.8**
SN	544.2	349.1	7.0	SN	599.4	395.9	7.4
PSF	557.2	486.65	9.1	PSF	605.8	434.0	8.2
SPA	322.2	230.4	4.4	SPA	344.7	252.7	4.4
June				December			
CSPF	**330.0**	**258.0**	**4.0**	**CSPF**	**426.38**	**313.3**	**5.1**
SN	602.1	392.5	7.2	SN	611.8	408.1	6.9
PSF	584.4	411.4	7.6	PSF	851.8	687.1	10.8
SPA	345.8	262.1	4.6	SPA	617.5	471.4	7.7

Considering the monthly results presented in Table 2, referring to the three metrics of forecasting errors, RMSE, MAE and MAPE, the CSPF method presents the best results overall: in eight months (excluding February, March, July and October) the CSPF achieves the lowest values on all three metrics.

To infer from these 12 months' data, we first conduct three Friedman tests (Siegel and Castellan 1988) to compare the performance of CSPF forecasting with the referred methods (SN, PSF and SPA). Results obtained are presented in Table 3, referring to Pairwise Comparisons (note that p-values values have been adjusted by the Bonferroni correction for multiple tests). Considering a 0.05 significance level, according to the Friedman tests' results, and in the context of the four methods considered, CSFP approach exhibits significant differences with all approaches except with SPA; we can also point

out that SPA and SN approaches, as well as SN and PSF, do not show significant differences.

Since, for the task at hand, the main competitor of CSPF is the SPA approach, we further focus on these two methods to better understand their comparative performance. Results from the Related-Samples Wilcoxon Signed Rank Test (Siegel and Castellan 1988) are in Table 4. According to them, in terms of MAE and MAE metrics, the CSPF performance significantly surpasses the SPA approach performance; also, if we consider a higher significance level, 0.1, we could state the same referring to the RMSE metric.

The results obtained show that, for the application considered, the CSPF method compares favorably with the baseline methods SN, PSF and SPA, thus being a promising approach for STLF.

Table 3 CSPF comparative performance: Friedman tests' Pairwise Comparisons

	RMSE		MAE		MAPE	
	Test Statistic	Adj. p-value	Test Statistic	Adj. p-value	Test Statistic	Adj. p-value
CSPF- SN	-1.750	**0.005**	-1.750	**0.005**	-1.708	**0.007**
CSPF - PSF	-2.500	**0.000**	-2.833	**0.000**	-2.792	**0.000**
CSPF - SPA	-0.417	1.000	-0.750	0.928	-0.667	1.000
SPA - SN	1.333	0.068	1.000	0.347	1.042	0.289
SPA - PSF	-2.083	**0.000**	-2.083	**0.000**	-2.125	**0.000**
SN -PSF	-0.750	0.928	-1.083	0.239	-1.083	0.239

Table 4- CSPF comparative performance with SPA: Wilcoxon tests' results

	SPA-CSPF differences					
	Positive	Negative	Ties	Test Statistic	Standardized Test Statistic	p-value
RMSE	8	4	0	60.0	1.647	0.099
MAE	10	2	0	70.0	2.432	0.015
MAPE	9	2	1	60.5	2.447	0.014

5. CONCLUSIONS AND FURTHER RESEARCH

In this paper, we propose a new Clustering-based Similar Pattern Forecasting algorithm (CSPF), for short-term load forecasting. CSPF is a two-step approach. In **Step 1)** we address the heterogeneity of the historical data, using a clustering algorithm – K-Medoids - and resort to COMB distance, a combination of various distance measures to capture different aspects of the time series dissimilarities: values (Euclidean distance), trends (Pearson based distance), cyclical behaviors (Euclidean distance between periodograms) and autocorrelation structures (Euclidean distance between estimated autocorrelation functions); to each day is then allocated a cluster label and load pattern sequences are considered those precede days of the same type (weekday, holidays and special holidays) as the target day. In **Step 2)** among the sequences of days previously determined, we consider subsequences revealing similar temperature profiles as the temperatures estimated for the target and preceding days,

including a search for Nearest Neighbors sequences. The load forecasting is then obtained based on these neighbor-days' loads.

The proposed algorithm integrates a Similar Pattern approach with expert's knowledge that is mapped to its parametrization - e.g., deciding which types of days to consider or which percentage of load sequences to retain, based on the temperature criterion.

The CSPF approach was applied to three years time series data in 15-minutes resolution of the Portuguese Transmission System Operator. Considering the year 2017, the load forecasts obtained for the 96 periods of the day-ahead exhibit very good indicators of performance (e.g., monthly MAPE in the range [2.7% - 5.1%]) when compared to an alternative Pattern Sequence-based Forecasting (PSF) (e.g., monthly MAPE in the range [7.3% - 20.2%]). After investigating the comparative performance of CSPF with three baseline methods – the Seasonal Naïve (SN) method, the Pattern Sequence-based Forecasting (PSF) algorithm, and a Semi-Parametric Additive (SPA) method - we conclude the proposed method shows a significant advantage for the task at hand.

In future research, the proposed method should be used for different data sets, namely with longer time series. We will also further investigate the algorithm parametrization so that it gains (informed) autonomy.

ACKNOWLEDGMENT

This work was supported by Instituto Politécnico Lisboa (IPL) with reference IPL/2020/ELForcast_ISEL and Fundação para a Ciência e a Tecnologia, grants UIDB/00315/2020 and UIDB/50021/2020.

We thank Tiago G. S. Chambel Cardoso for the paper Figures design.

REFERENCES

Azadeh A, Ghaderi SF, Sheikhalishahi M, Nokhandan BP (2014) Optimization of Short Load Forecasting in Electricity Market of Iran Using Artificial Neural Networks. Optim Eng 15:485–508. https://doi.org/10.1007/s11081-012-9200-8

Bezdek JC, Pal NR (1998) Some new indexes of cluster validity. IEEE Trans Syst Man Cybern, Part B: Cybernetics 28(3):301–315. https://doi.org/10.1109/3477.678624

Bokde N, Asencio-Cortes G, Martinez-Alvarez F., Bokde MN (2016) Package 'PSF'. URL: https://cran.r-project.org/web/packages/PSF/PSF.pdf. Accessed 13 April 2021

Caiado J, Crato N, Peña D (2006) A periodogram-based metric for time series classification. Computational Statistics & Data Analysis 50(10):2668-2684. https://doi.org/10.1016/j.csda.2005.04.012

Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3(1):1–27

Cardoso MGMS, Martins A, Lagarto J (2021) Combining various dissimilarity measures for clustering electricity market prices. In Estatística: Desafios Transversais às Ciências dos Dados - Atas do XXIV Congresso da Sociedade Portuguesa de Estatística (Paula Milheiro et al. eds), Edições SPE

Cardoso, MGMS, Martins, A (in press) The performance of a combined distance between time series. In Recent Developments in Statistics and Data Science - Proceedings of the XXV Congress of the Portuguese Statistical Society. Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R. and de Carvalho, M. (eds.), Springer

Charlton N, Singleton C (2014) A refined parametric model for short term load forecasting. Int J Forecast 30:364–368. http://dx.doi.org/10.1016/j.ijforecast.2013.07.003

Chen Y, Luh PB, Guan C, Zhao Y, Michel LD, Coolbeth MA, Friedland PB, Rourke SJ (2010) Short-term load forecasting: Similar day-based wavelet neural networks. IEEE Trans Power Syst 25(1):322-330. https://doi.org/10.1109/TPWRS.2009.2030426

Cheng C-H, Wei L-Y (2010) One step-ahead ANFIS time series model for forecasting electricity loads. Optim Eng 11: 303–317. https://doi.org/ 10.1007/s11081-009-9091-5

Dedinec A, Filiposka S, Dedinec A, Kocarev L (2016) Deep belief network based electricity load forecasting: An analysis of Macedonian case. Energy 115:1688-1700. http://dx.doi.org/10.1016/j.energy.2016.07.090

Duch, W (2000) Similarity-based methods: a general framework for classification, approximation and association. Control Cybernetics 29(4):937–968

Fallah SN, Ganjkhani M, Shamshirband S, Chau KW (2019) Computational intelligence on short-term load forecasting: A methodological overview. Energies, 12(3), 393. https://doi.org/10.3390/en12030393

Fan S, Chen L (2006) Short-Term Load Forecasting Based on an Adaptive Hybrid Method. IEEE Trans Power Syst 7(1):392-401. https://doi.org/10.1109/TPWRS.2005.860944

Fan S, Hyndman R (2012) Short-term load forecasting based on a semi-parametric additive model. IEEE Trans Power Syst 7:1-8. https://doi.org/10.1109/TPWRS.2011.2162082

Gaillard P, Goude Y, Nedellec R (2016) Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int J Forecast 32:1038–1050. https://doi.org/10.1016/j.ijforecast.2015.12.001

Goude Y, Nedellec R, Kong N (2014) Local Short and Middle Term Electricity Load Forecasting with Semi-Parametric Additive Models. IEEE Trans on Smart Grid 5(1):440-446. https://doi.org/10.1109/TSG.2013.2278425

Hennig C (2020) Package 'fpc'. URL: http://cran. r-project. org/web/packages/fpc/fpc. Pdf. Accessed 13 April 2021

Heydari A, Nezhad MM, Pirshayan E, Garcia DA, Keynia F, Santoli L (2020) Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm. Applied Energy 277:115503. https://doi.org/10.1016/j.apenergy.2020.115503

Hong T (2010) Short Term Electric Load Forecasting. PhD Thesis. North Carolina State University

Hong T, Fan S (2016) Probabilistic electric load forecasting: A tutorial review. Int J Forecast 32:914-938. https://doi.org/10.1016/j.ijforecast.2015.11.011

Hong T, Shahidehpour M (2015) Load Forecasting Case Study. EISPC. https://pubs.naruc.org/pub.cfm?id=536E10A7-2354-D714-5191-A8AAFE45D626. Accessed 13 April 2021

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 8 April 2022.

Iglesias F, Kastner W (2013) Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies 6(2):579-597. https://doi.org/10.3390/en6020579

1 Ilic S, Selakov A, Vukmirovic S, Erdeljan A, Kulic F (2013) Short-term load
2 forecasting in large scale electrical utility using artificial neural networks. Journal
3 of Scientific & Industrial Research 72:739-745
4 Jin CH, Pok G, Lee Y, Park HW, Kim KD, Yun U, Ryu KH (2015) A SOM clustering
5 pattern sequence-based next symbol prediction method for day-ahead direct
6 electricity load and price forecasting. Energy Convers Manag 90:84-92.
7 http://dx.doi.org/10.1016/j.enconman.2014.11.010
8 Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster
9 analysis. John Wiley & Sons. https://doi.org/10.1002/9780470316801
10 Kuster C, Rezgui Y, Mourshed M (2017) Electrical load forecasting models: A critical
11 systematic review. Sustainable Cities and Society 35:257-270. https://doi.org/
12 10.1016/j.scs.2017.08.009
13 Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K, Studer M (2013) Package
14 'cluster'. URL: https://cran.r-project.org/web/packages/cluster/cluster.pdf.
15 Accessed 13 April 2021
16 Mandal P, Senjyu T, Urasaki N, Funabashi T (2006) A neural network based several-
17 hour-ahead electric load forecasting using similar days approach. Int J Elect Power
18 & Energy Syst 28(6):367-373. https://doi.org/10.1016/j.ijepes. 2005.12.007
19 Martinez-Alvarez F, Troncoso A, Riquelme JC, Ruiz JSA (2010) Energy time series
20 forecasting based on pattern sequence similarity. IEEE Trans Knowledge and
21 Data Eng 23(8):1230-1243. https://doi.org/10.1109/TKDE.2010.227
22 Metaxiotis K, Kagiannas A, Askounis D, Psarras J (2003) Artificial intelligence in
23 short term electric load forecasting: a state-of-the-art survey for the researcher.
24 Energy Convers Manag 44:1525–1534. https://doi.org/10.1016/S0196-
25 8904(02)00148-6
26 Mohandes M (2002) Support vector machines for short-term electrical load
27 forecasting. Int J Energy Res 26:335-345. https://doi.org/10.1002/er.787
28 Montero P, Vilar J (2014) TSclust: An R Package for Time Series Clustering. JSS J
29 Stat Softw 62(1):1–43. https://doi.org/10.18637/jss.v062.i01
30 Mu Q, Wu Y, Pan X, Huang L, Li X (2010) Short-term load forecasting using
31 improved similar days method. In Proceedings of 2010 Asia-Pacific Power and
32 Energy Engineering Conference, Chengdu. https://doi.org/10.1109/APPEEC
33 .2010.5448655
34 Rodrigues P (2008) Hierarchical clustering of time-series data streams. IEEE Trans
35 Knowledge Data Eng 20(5):1–13. https://doi.org/10.1109/TKDE.2007.190727
36 Ružic S, Vuckovic A, Nikolic N (2003) Weather Sensitive Method for Short Term
37 Load Forecasting in Electric Power Utility of Serbia. IEEE Trans Power Syst
38 18(4):1581-1586. https://doi.org/10.1109/TPWRS.2003.811172
39 Sharifzadeh M, Sikinioti-Locka A, Shah N (2019) Machine-learning methods for
40 integrated renewable power generation: A comparative study of artificial neural
41 networks, support vector regression, and Gaussian Process Regression.
42 Renewable and Sustainable Energy Reviews 108:513–538.
43 https://doi.org/10.1016/j.rser.2019.03.040
44 Siegel, S, Castellan, Jr (1988) Nonparametric Statistics for the Behavioral Sciences,
45 2nd edition, McGraw-Hill.
46 Wang P, Liu B, Hong T (2016) Electric load forecasting with recency effect: A big
47 data approach. Int J Forecast 32:585–597. https://doi.org/10.1016/j.ijforecast.
48 2015.09.006

Zheng H, Yuan J, Chen L (2017) Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. Energies 10(8):1168. https://doi.org/10.3390/en10081168