Working paper

2024-02

Statistics and Econometrics
ISSN 2387-0303

Clustering and forecasting of day-ahead electricity supply curves using a market-based distance

Zehang Li, Andrés M. Alonso, Antonio Elías and

Juan M. Morales

uc3m | Universidad Carlos III de Madrid

Clustering and forecasting of day-ahead electricity supply curves using a market-based distance

Zehang Li[a,*], Andrés M. Alonso[a], Antonio Elías[b], Juan M. Morales[b]

[a] *Department of Statistics, Universidad Carlos III de Madrid, C. Madrid, 126, Getafe, 28903, Madrid, Spain*
[b] *Department of Applied Mathematics, Universidad de Málaga, Avda. Cervantes, 2, Malaga, 29071, Andalucía, Spain*

Abstract

Gathering knowledge of supply curves in electricity markets is critical to both energy producers and regulators. Indeed, power producers strategically plan their generation of electricity considering various scenarios to maximize profit, leveraging the characteristics of these curves. For their part, regulators need to forecast the supply curves to monitor the market's performance and identify market distortions. However, the prevailing approaches in the technical literature for analyzing, clustering, and predicting these curves are based on structural assumptions that electricity supply curves do not satisfy in practice, namely, boundedness and smoothness. Furthermore, any attempt to satisfactorily cluster the supply curves observed in a market must take into account the market's specific features.

Against this background, this article introduces a hierarchical clustering method based on a novel weighted-distance that is specially tailored to non-bounded and non-smooth supply curves and embeds information on the price distribution of offers, thus overcoming the drawbacks of conventional clustering techniques. Once the clusters have been obtained, a supervised classification procedure is used to characterize them as a function of relevant market variables. Additionally, the proposed distance is used in a learning procedure by which explanatory information is exploited to forecast the supply curves in a day-ahead electricity market. This procedure combines the idea of nearest neighbors with a machine-learning method. The prediction performance of our proposal is extensively evaluated and compared against two nearest-neighbor benchmarks and existing competing methods. To this end, supply curves from the markets of Spain, Pennsylvania-New Jersey-Maryland (PJM), and West Australia are considered.

Keywords: Clustering, Forecasting, Supply curve, Electricity market

*Corresponding author
Email address: zehang.li@alumnos.uc3m.es (Zehang Li)

1. Introduction

The electricity market in many countries allows electricity producers to offer bids at different prices, generally related to their marginal costs. After the system operator forms the supply curve by bids from all participants, each participant will be remunerated to the marginal prices, the intersections of the supply and demand curves.

The economic and technical particularities of the electricity supply industry, together with the high level of market concentration and oligopolistic conditions that still prevail today in some countries (see, e.g., [25] for a critical and quantitative analysis on market concentration in European countries), facilitate the exercise of market power in electricity markets [28]. Consequently, the knowledge of the aggregate supply curve of the market proves to be very valuable to both power producers (with the ability and willingness to exercise market power) and to regulators, although evidently for opposite reasons [31, 42].

On the one hand, clustering and characterizing electricity supply curves are vital in the power generation business to maximize profit in varying scenarios. From a technical point of view, the analysis of general curves has been intensively studied in the well-known branch of statistics called Functional Data Analysis (FDA) [39]. Based on the premise that each observation is a smooth function on a continuous common domain $[a, b]$, FDA has produced a wealth of clustering methods ([24]), sparking a currently active area of research [33]. Nonetheless, to our knowledge, none of these methods is specifically suited to the nature of the supply curves in the the electricity industry: step curves defined on an unbounded domain. Functional data is often recorded at some discrete common observation points of the domain and some clustering methods work directly with the raw discrete observations or require the curves to be evaluated on a bounded common grid [see, e.g., 10, 11]. Others first approximate the curves using a finite basis set of functions or are based on dimension reduction techniques assuming that the curves are smooth functions rather than step functions [see, e.g., 1, 38, 30, 26, 13, 16]. Finally, distance-based FDA methods [see, e.g., 43, 17, 14, 23] use clustering algorithms that rely on specific distances for FDA that do not cover the case of unbounded domains or have the flexibility to include additional information on the bidding distribution of the market. For its part, the literature focused on energy has not extensively studied the problem of clustering energy supply curves. On the contrary, clustering approaches have been proposed for household load curves with a focus on the shape of the curves ([2], [15], [27] and [44]). However, it is important to note that load curves have a completely different nature than energy supply curves: load curves are functions defined on all 24 hours of the day, typically recorded for a common grid (hourly, half-hourly, etc.) whereas energy supply curves are step functions of prices on an unbounded domain of quantities, these quantities not being recorded on a common grid but determined by the aggregate generation bids ordered in increasing price (merit order supply curve). This hampers the use of these methods, which have been successfully applied to load curves, to energy supply curves, because they require common grids and a bounded domain as the FDA

approach discussed above.

On the other hand, methods for prediction that can provide reliable and realistic future supply curves are also important to reduce market uncertainty in decision-making. However, the technical literature on such methods is scarce. Much more effort has been put into the prediction of the market equilibrium price, as in [45, 35, 41, 34, 20]. In particular, [45] focuses on electricity price forecasting by modeling the supply and demand separately. Then the market equilibrium is estimated by the intersection between the predicted demand and supply, referring to this procedure as the X-Model: it is based on discretizing the prices into 16 classes, which are subsequently modeled as autoregressive processes estimated by Ordinary Least Squares with LASSO penalty. They achieve good prediction results by including lags and weekday dummy variables as exogenous information for the equilibrium price, but little attention has been paid to the accuracy of the individual supply or demand curves. In [35], a Bayesian inference approach based on Markov Chain Monte Carlo and sequential Monte Carlo techniques is proposed to infer the supply curve of an electricity market from the observed clearing prices and quantities. Therefore, their approach does not require market participants' historical bids or marginal costs. In contrast, they fail to account for the time dynamics of the supply curves, i.e., they seek to estimate an "average" supply curve. [41] and [34] approach the prediction problem from the FDA standpoint using Functional Time Series models able to capture the temporal evolution. Specifically, [41] proposes a nonparametric Functional Autorregresive Model (NPFAR) that leads to a statistically significant improvement in the forecasting accuracy in the Italian Market. In the same vein, [34] considers a double-seasonal SARMAHX functional model able to characterize the temporal daily and weekly dependency and able to incorporate exogenous variables. Although these two methods provide promising prediction results, these FDA approaches require assuming a common bounded domain for the supply curves and, more importantly, their predicted curves are smooth functions. This might not impact performance metrics, but it does produce unrealistic smooth supply curve shapes with a potential information loss for some purposes.

This article aims to address the drawbacks mentioned above by proposing a distance-based hierarchical clustering method and a distance-based prediction method for supply curves. To measure the dissimilarity between the curves, we propose and use a weighted distance that is particularly suitable for the characteristics of the offers on the electricity market, i.e., a distance that 1) solves the problem of the unbounded domain, 2) does not assume any smoothness of the functions, and 3) provides the flexibility to incorporate market information. Specifically, we propose to take into account the distribution of the prices of the offers to amplify the differences in the historically most frequent price intervals. Thus, by giving greater weight to the prices that are most frequent in the offers, we focus on the points where the offer curves present their characteristic jumps. Furthermore, unlike the competing procedures presented in the technical literature to forecast electricity supply curves, our distance-based prediction method does not assume smoothness, which is a restrictive condition far from the very nature of these curves.

The method of clustering is applied to the complete set of supply curves from the day-ahead market (43 848 hourly curves for the years 2016–2020 in the Spanish market; 43 824 hourly curves for 2018–2022 in the PJM market; 87 696 half-hourly curves for 2016–2020 in the West Australian market). Once we have obtained the clusters of the curves, it is essential to understand how they were formed and to determine which variables explain these clusters. This description of the clusters obtained is carried out using a supervised classification procedure (Random Forest) with the cluster labels as the response variable and market and temporal variables as the explanatory variables. We have considered market variables such as predictions of electricity generation through wind, solar, and nuclear energy as well as the forecast demand. We have also considered time variables such as the hour, weekday, month, year, and a dummy variable for national holidays. We remark that these variables were only used in the posterior analysis to explain the main characteristics of the resulting groups, not in the hierarchical cluster procedure directly.

The forecasting exercise takes a period of four years as the training sample and the last year as the test set. The method of prediction makes use of the distance matrix in the training set to predict the distances in the test set. These predictions are used to select the closest curve, which will be the predicted curve. We perform a rolling window exercise for a day-ahead forecast, that is, for a given day, D, we predict the curves of day $D+1$. After that, the training window is extended by one day. The prediction error is the weighted distance between the true curve and its prediction. The different approaches to the prediction are compared through the mean of the prediction errors in the test set.

The rest of this paper is organised as follows. Section 2 discusses the weighted distances we use to emphasize different properties in our case study. Section 3 presents the clustering results and the characterization of the obtained clusters by market and temporal features for the Spanish market. Similar analyses and results are shown in Appendices A and B for the PJM and West Australian markets, respectively. Section 4 presents the prediction procedure and the results of the prediction exercise for the three markets. Finally, in Section 5, we draw conclusions based on performance.

2. The Definition of Distance

Define $t \in [1, T]$ as the index of the hourly supply curves, where T is the number of observed curves. A supply curve is obtained by ordering all the offers received from lowest to highest price and accumulating the quantities of the offers. That is, if $\mathcal{O}_t = \{(p_1, q_1), (p_2, q_2), \ldots, (p_{n_t}, q_{n_t})\}$ is the set of price–quantity pairs, where for simplicity we will assume that prices are ordered, then the supply curve will be obtained from the representation of the set of pairs of price–cumulative quantity, $\overline{\mathcal{O}_t} = \{(p_1, Q_1), (p_2, Q_2), \ldots, (p_{n_t}, Q_{n_t})\}$, where $Q_i = \sum_{j=1}^{i} q_i$.

In this paper, we define the supply curve at hour t as the function $C_t(p) : [0, \infty] \to \mathbb{R}^+$ which is a non-decreasing step function of the price p, having steps in the positions corresponding to the ordered prices of the offers. We follow this representation because all the curves are defined in the interval $[0, \infty]$

whereas the most frequent representation, price as a function of quantity, can have different definition intervals, since the maximum cumulative quantity, Q_{n_t}, varies across hours.

In Figure 1, we show two examples of supply curves where we can see some of their most notable characteristics: step functions with different numbers of steps located at different positions. Moreover, it is clear that to define an appropriate distance between curves of this type we have to develop a way to deal with the (infinite) difference in the last step of both curves. In the next section, we introduce a weighted distance that resolves this inconvenience.

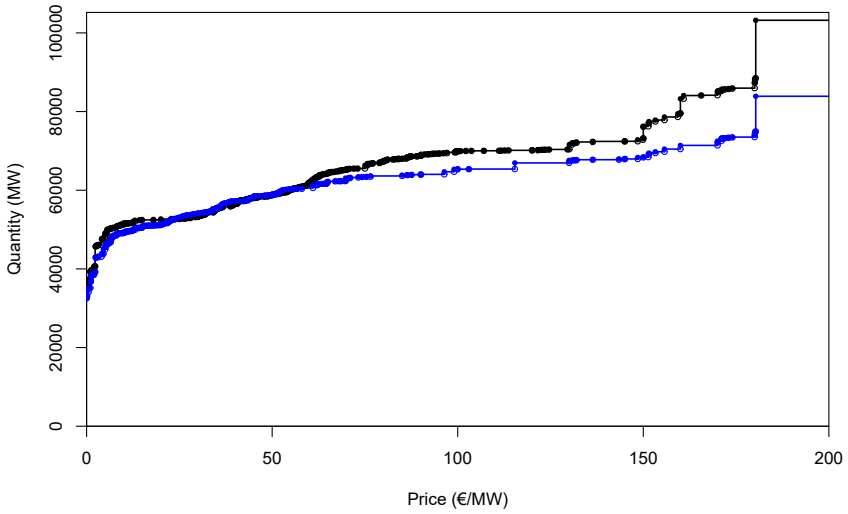


Figure 1: Supply curves for 12 May 2016, 06:00 (blue) and 4 February 2017, 01:00 (black) for the Spanish market. The considered curves are close in the interval of P=[0,75] and depart from each other afterward. In this situation, a classical L2 distance would consider these curves as being far from each other even though they are close in the important region. Moreover, it is incapable of addressing the infinite difference of the last step.

Weighted distances

Given two curves, C_t and C_s, like those shown in Figure 2, an intuitive definition of distance is the area between the curves. That is, in our case, the sum of the areas of rectangles with a width of $|p_{i+1} - p_i|$ and the height $|q_{t,i} - q_{s,i}|$, where p_i and p_{i+1} are the neighbouring jump points considering both curves and $q_{t,i} = C_t(p_i)$ and $q_{s,i} = C_s(p_i)$ are the corresponding quantities. But, as we have mentioned, there is a problem at the right ends of the curves that would take the distance to infinity. Notice that the last rectangle starts at the maximum bid price considering both curves and ends at infinity. That maximum bid price may depend on the specific pair of curves considered.

This divergence can be addressed by introducing a weight function approaching zero as the price increases. In this paper, we will use the following weighted ℓ_2 distance

$$d(C_t, C_s)^2 = \int_0^{+\infty} |C_t(p) - C_s(p)|^2 W(p) dp$$
$$= \sum_{i=1}^{n} \int_{p_i}^{p_{i+1}} |C_t(p) - C_s(p)|^2 W(p) dp, \quad (1)$$

where C_t and C_s are the curves at times t and s, $W(p)$ is a non-negative weight function and p_i are the prices where C_t and/or C_s have a step. The set of those prices $\{p_1, p_2, ..., p_n, p_{n+1}\}$ satisfies the conditions $0 = p_1 < p_2 < \ldots < p_n < p_{n+1} = +\infty$. It should be noted that the above definition of distance can be adapted to consider negative prices by changing the limits of the integral.

It is clear that the selection of $W(p)$ is crucial for the distance (1). There are many possibilities for $W(p)$. For instance, taking $W(p) = 1$ at a given bounded interval $(\underline{p}, \overline{p})$ and $W(p) = 0$ otherwise, will focus on the differences in that interval. However, as Figure 3 shows, the distribution of the prices of the offers is far from the above uniform distribution defined by support $(\underline{p}, \overline{p})$. We have preferred to give a weight proportional to the frequency of the prices in the offers. The intent is to emphasize distances where curves typically have steps. A higher frequency around a price means that a greater number of curves have bids at those prices. Therefore, using a weight proportional to the frequency stresses the differences where the curves have steps. A good approximation of the distribution of all prices at jump points can be obtained by a mixture of Gaussian distributions. Figure 3 also shows the obtained mixture, which has two components that have means, standard deviations, and component weights of (43.93573, 51.01591), (26.119500, 9.863402) and, (0.7208744 0.2791256), respectively. Figure 3 uses the Spanish market dataset. Figure A.1 in Appendix A and Figure B.1 in Appendix B are for the PJM and West Australian markets, respectively.

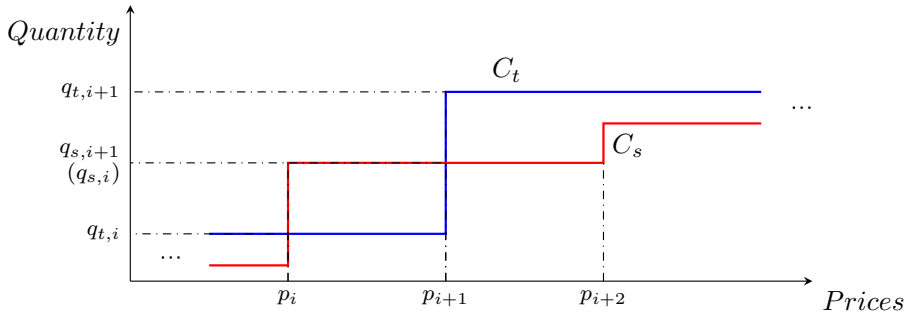


Figure 2: The distance between C_t and C_s in interval $[p_i, p_{i+1}]$ is the area of rectangle with width $|p_i - p_{i+1}|$ and height $|q_{t,i} - q_{s,i}|$. Similarly, the distance in interval $[p_{i+1}, p_{i+2}]$ can be calculated as $|p_{i+1} - p_{i+2}| * |q_{t,i+1} - q_{s,i+1}|$. This definition allows us to calculate the distance by aggregating the areas of all rectangles surrounded by C_t and C_s.

3. Hierarchical clustering with average linkage of the original curves

In this section, we aim to cluster and interpret the main features of the resulting groups of supply curves from the Spanish day-ahead market in the period 2016 to 2020. Similar analyses and results are shown in Appendices A and B for the PJM and West Australian markets, respectively. The definition proposed in Section 2 allows us to apply distance-based clustering methods to a set of step curves defined on an unbounded domain. However, the problem we tackle here presents the additional challenge of the amount of data: It involves dealing with $T = 43,848$ hourly supply curves, each one built from approximately 520 bids on average. To deal with this, we tested several clustering implementations and concluded that Hierarchical Clustering (HC) with average linkage has the best performance/efficiency trade-off for our problem. The remainder of this section discusses this selection and presents the main results. Then, subsections 3.1 and 3.2 provide interpretability and insights into the obtained clusters using supervised classification techniques.

As was mentioned before, the most insightful clustering structure was obtained by applying hierarchical clustering to the $43,848 \times 43,848$ distance matrix, $\boldsymbol{d}$, resulting in the dendrogram in Figure 4. This finding suggests a small number of groups (see the long top clades and sorted bottom leaves) and highlights that the clustering process merges small groups at high levels, which indicates that the outliers are grouped into small clusters far away from the primary clusters (for instance, see the red group at the left in Figure 4). Following ideas

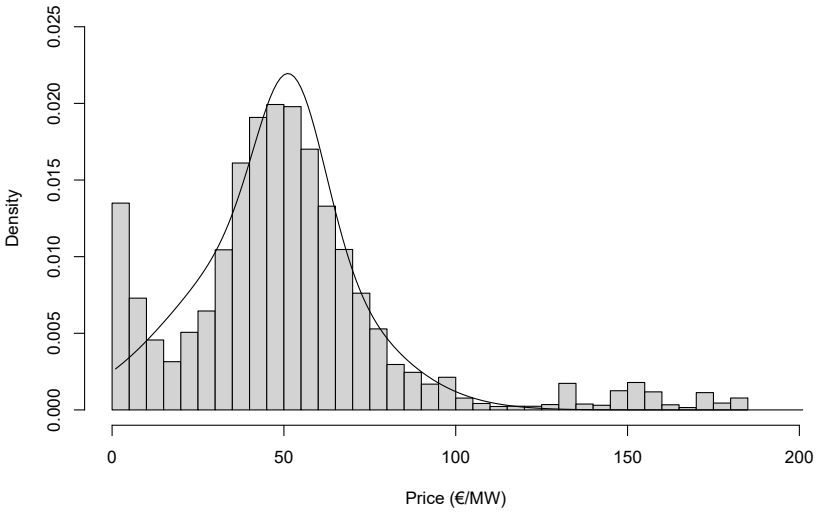


Figure 3: Histogram of the distribution of the prices of the offers and density of Gaussian mixture model with two components. Spanish market, 2016–2020.

from [21], to detect these clusters of outliers, we impose a cut at 0.5 (this value corresponds to the 99th percentile of the distribution of distances where the links of the dendrogram are made) in the first step, where the leaf nodes of the main structure are already formed into large-size groups while outliers gather in ones of small size. The resulting cluster sizes are summarised in Table 1. The curves located in clusters 8–42 were discarded from the following steps for being potential outliers. It is important to note that these clusters have an insignificant number of members and account for only 0.5% of the total number of curves. Finally, HC suggests four clusters, as can be concluded from the average silhouette criterion [40] plotted in Figure 5.

Although there are a wide variety of available clustering algorithms, HC with average linkage combines the ability to cluster high-dimensional data in an interpretable number of groups in a reasonable computational time. Other methods, such as density-based clustering, clustering combined with dimension reduction techniques, Partitions Around Medoids, and Leader Algorithm, resulted in impractical results for the Spanish Day-ahead Energy Supply Curves.

For example, the density-based clustering method highlights the existence of one single group in our data without having great gains in terms of computational time. In particular, we applied the OPTICS algorithm [5], a computationally efficient density-based clustering method that searches through the distance matrix $\boldsymbol{d}$ to detect dense areas and reduce the complexity of the problem. However, the computational improvement is at the cost of requiring extra RAM for a tree-based index [5]. In fact, in our experience with a computing platform with 32GB of RAM, the OPTICS time performance does not surpass that of our preferred HC, which is in line with the discussion of [7]. The same disappointing result was obtained by combining dimension reduction techniques (such as multidimensional scaling) and classical multivariate clustering methods, failing to agglomerate the curves into a reasonable number of clusters.

Another unsuccessfully considered method was Partitions Around Medoids (PAM) [29], another clustering method based on the matrix of distances $\boldsymbol{d}$. In contrast to HC, it has the drawback of requiring a predetermined number of clusters, and the outliers shown in Table 1 make it difficult to determine the number of clusters.

To improve scalability, one-pass strategies such as the Leader Algorithm were also considered [22, 32, 36, 7]. This idea consists of randomly selecting initial leaders and iteratively agglomerating curves into different clusters. However, we found, using our data, that this strategy is in general not robust to the presence of outliers or tiny groups, and that the final output is highly dependent on the initial selection of the leaders.

With regard to the linkage of HC, the selection was based on the exhaustive simulation exercise in [18], which studied the performance of a wide palette of HC setups under a large variety of synthetic functional processes, concluding that this configuration is recommended when the functions under analysis are not generated by Fourier basis expansions and a small number of clusters, of different sizes, are expected.

In the next section, we characterize the main characteristics of the four

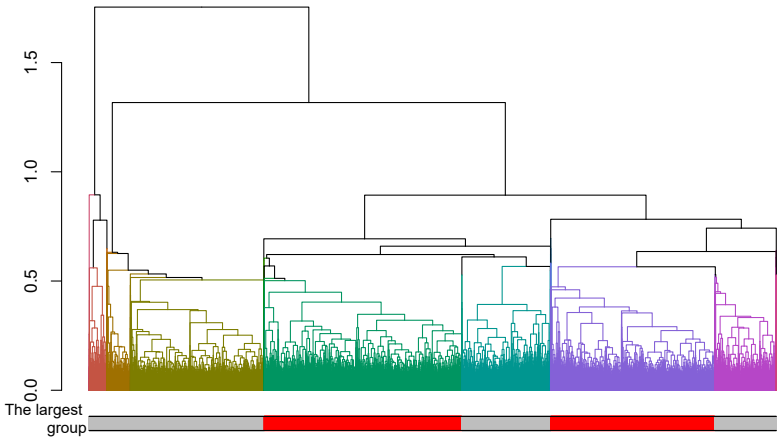


Figure 4: The dendrogram before discarding small groups. By cutting it at a height of 0.5, 42 groups were generated with sizes presented in Table 1. The colour of a branch is that of the cluster to which it belongs. The two largest clusters are highlighted by red blocks in the colour bar. Spanish market, 2016–2020.

obtained groups. To do this, we use a supervised classification procedure with the cluster labels of the four groups of the curves as the target variable. This approach has been used by [4] when classifying the time series of electricity demand by domestic customers. The popular Random Forest (RF) classification procedure is employed by us. Of course, other supervised classification procedures can be used. We have chosen RF for its versatility and simplicity. Operating in a supervised manner, an RF model classifies the labels (in our case, the previously assigned cluster) using the explanatory features. The set of explanatory features may vary, depending on the order of training timing t_{train} and publish time t_{pub} of curves. If the curve of the hour t is unknown until t_{pub}, i.e., $t_{pub} > t_{train}$, the model includes the day-ahead cluster labels, `cluster_24lag`, predicted as the most recently observed label based on the assumption of high temporal dependency on the historical curves. Otherwise, when $t_{pub} \leq t_{train}$, extra geometrical features can be included in the RF and obtained from the already observed curve. Subsection 3.1 presents the model in the former situation, termed the non-concurrent model. Subsection 3.2 explores the concurrent model, which handles the latter case. The design of these two models is two-fold; first, they provide insight into the main drivers that explain the clusters and, second, although this is not exploited in this paper, they could be used to forecast the cluster of a new incoming curve.

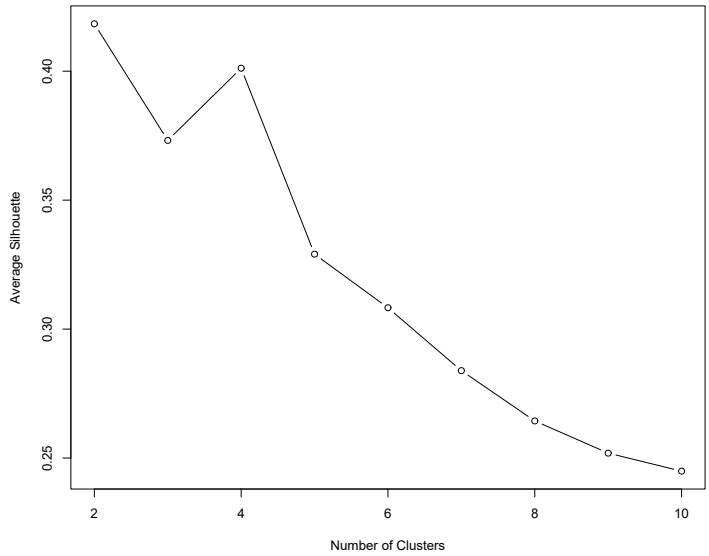


Figure 5: The average silhouettes at two and four clusters are approximately equivalent. We use four to obtain more diversity in terms of the behaviour of the explanatory variables.

Table 1: Cluster sizes before discarding small groups by cutting the dendrogram at height = 0.5. Spanish market, 2016–2020.

Clusters	**Member amounts**	**Clusters**	**Member amounts**	**Clusters**	**Member amounts**
1	1476	15	4	29	2
2	8457	16	6	30	4
3	5631	17	2	31	4
4	12596	18	4	32	4
5	10460	19	6	33	10
6	3887	20	4	34	4
7	1111	21	5	35	2
8	13	22	6	36	5
9	18	23	5	37	2
10	42	24	7	38	10
11	4	25	4	39	6
12	7	26	6	40	2
13	6	27	4	41	10
14	8	28	2	42	2

3.1. Non-Concurrent model: Market variables, temporal variables, and lagged labels.

Table 2 presents the variables involved in the characterization process using the non-concurrent model. These features span from January 2, 2016, to December 31, 2020, to align with the 24-hour lagged variable `cluster_24lag`, which is the cluster of the curve 24 hours before. This lagged variable aims to take into account temporal dependency and data accessibility in reality. One challenge in acquiring the `cluster_24lag` variable is the absence of labels from curves discarded as outliers. It was addressed by assigning those outliers to the cluster whose centroid is the closest, ensuring that every curve has a cluster label history. Additionally, other variables are included with different intentions. Temporal variables aim to capture the seasonality and dynamics throughout the year, during different seasons, and during different days of the week. Market variables aim to capture the market conditions, such as renewable generation, nuclear generation, and demand. Finally, the adjusted seasonal GDP is included as a global economic control variable, which is particularly useful to capture the general market activity during the pandemic. As to the setting of the hyper-parameters of the RF model, we update the uniform cutoff values of each cluster with its samples' percentage to avoid the impact of imbalanced clusters. The out-of-bag error of the final fitted RF model is 8.44%. That is, we correctly classified more than 90% of the curves. This error decreases to 0.005% by fitting the non-concurrent model with the West Australian market features. Remarkably, the error reaches 0% with data from the PJM market. It should be noted that only two clusters were selected for the PJM and Western Australian markets, so classification is easier than for the Spanish market, where four clusters were selected. Keeping the number of clusters as four, the errors are 8.84% in the West Australian market and 0.81% in the PJM market. Figure 6 shows the ranking of the importance of the variables under different criteria. The amount of wind power generation and cluster labels with 24-hour lags perform the most significant roles in terms of the measure of Mean Decrease accuracy (MDA) and Mean Decrease Gini (MDG), respectively. It is worth noticing the consistency of the clusters of a curve and its 'ancestor' at the same hour of the previous day (see Figure 7b). For the first three groups, the majority of members ($> 75\%$) remain in the same groups as in the previous 24 hours. The fourth group seems to disprove this observation but this group shows a very different mean value of `gen_wind` than the others (see Figure 7a)), which suggests the high impact of generation from wind, and that wind power varies from one day to another. Then it is reasonable to find that the curves within the fourth group have lagged labels indicating other clusters.

3.2. Concurrent model: Including geometrical features of the supply curve

In this subsection, we will discuss another situation in which we train a concurrent RF model once we obtain the supply curves for a day. The convenience of this model is that geometrical features, i.e. quantity at price zero, the maximum quantity, and price, can be extracted from each known curve. Introducing these

Table 2: Description of the explanatory variables for the Spanish market.

	Variable name	Description
Temporal variables	hour	Integer. The hour of day of the supply curve;
	month	Factor. The month of the supply curve;
	year	Factor. The year of the supply curve;
	weekday	Factor. The day of the week of the supply curve;
	holiday	Factor. If a supply curve is for a public holiday;
Market variables	gen_Solar[a]	Numeric. Predicted solar power generation (including solar photovoltaic and solar thermal generation);
	gen_Wind[a]	Numeric. Predicted wind power generation;
	gen_Nuclear[a]	Numeric. Predicted nuclear power generation;
	pred_demand[a]	Numeric. The predicted electricity demand for the hour of the supply curve;
	adjusted_GDP[b]	Numeric. Seasonally adjusted GDP;
Lagged label	cluster_24lag	Factor. The clusters with 24-hour lags.

Source:

[a] `https://www.esios.ree.es/en/generation-and-consumption`

[b] `https://www.ine.es/en/`

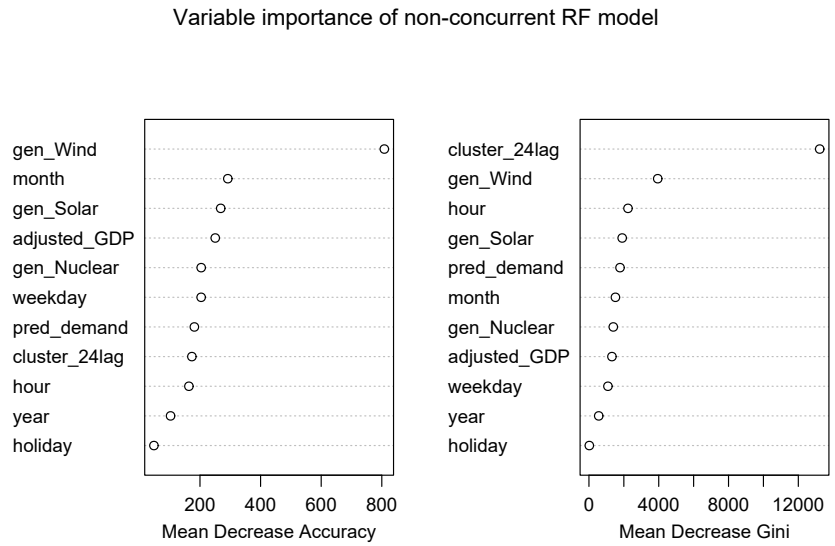
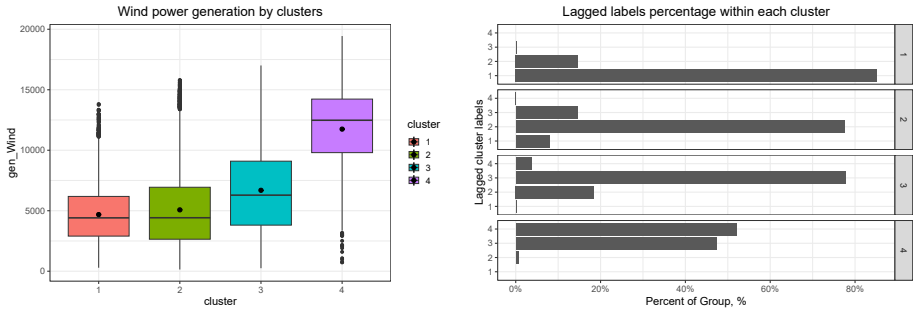


Figure 6: Analysis of the importance of the variables in the non-concurrent model. Spanish market, 2016–2020.



(a) Boxplot of amount of wind power generation by cluster.

(b) The percentages of lagged labels in each cluster.

Figure 7: Descriptive analysis of wind power generation and lagged labels by cluster. Spanish market, 2016–2020.

variables is useful to capture the general steepness, magnitude, and shape of the curves. With these additional geometrical features, the OOB of the concurrent RF model decreases to 5.73% while the hyper-parameters remain the same as in the non-concurrent case. In the case of two clusters, the OOB errors reach 0% and 0.002% in the PJM and the West Australian markets, respectively. When selecting four clusters, the models for the PJM and West Australian markets surpass the Spanish market model as they reduce the OOB errors to 0.51% and 4.78%, respectively.

Focusing on the Spanish market, again, the lagged labels variable ranks at the top under the criterion of MDG. The other most significant variable according to the MDA criterion is the month, which is distributed approximately uniformly within the first three groups but centralizes itself in winter and spring in the fourth cluster (see Figure 8). Zero-price quantities are also of great importance in the model for both measures, MDA and MDG. The fourth cluster shows again that it is different, with the highest average value (see Figure 9b). Bearing in mind the high wind power generation of this cluster, the weather conditions of winter/spring months are likely to contribute to the large 'free' wind power production.

The two criteria share two moderately significant variables: the hour and the maximum quantity. Figure 9d shows that all the curves in the first cluster are from the early hours of the day, whereas the third and fourth clusters present a high percentage of late hours. The maximum quantity follows a similar pattern as the quantity at zero price, for which the average increases across clusters. The remaining geometrical feature, the maximum price, ranks at the bottom of the variables' importance, as the defined weight function extends the price domain to infinity.

4. Forecasting day-ahead electricity supply curves

A simple k-NN-based forecasting procedure consists of looking for past days that have behaved similarly to day D and taking the curves of the following day for those similar days as predictions. In fact, we use $k = 1$ since averaging or a linear combination of supply curves produces a much smoother curve with more steps than the actual curves. Table 3 shows summary statistics for the number of steps in the original curves and in the curves resulting from using one, three, and five nearest neighbors. It is clear that taking three or five neighbors, which are commonly used values, returns predictions that are much smoother with many more steps than the real curves. This implies that if we used 3-NN or 5-NN, we would obtain predictions that look different from the real curves.

The 1-NN procedure is formulated as follows:

- Given the H supply curves of the day D, $\mathbf{C}(t) = \{C_{t-H-1}, C_{t-H-2}, \ldots, C_t\}$, we want to predict the H supply curves of the day $D+1$, $\mathbf{C}(t+H) = \{C_{t+1}, C_{t+2}, \ldots, C_{t+H}\}$.
  - $H = 24$ for the Spanish and PJM markets and $H = 48$ for the Australian market.

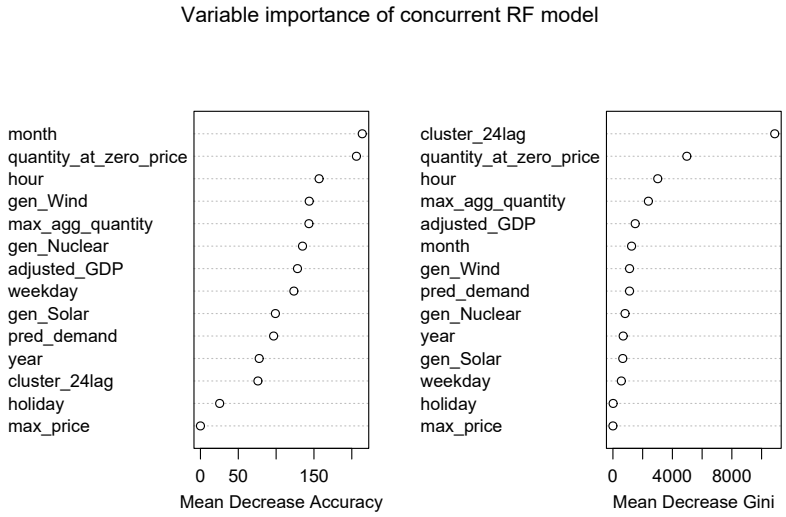
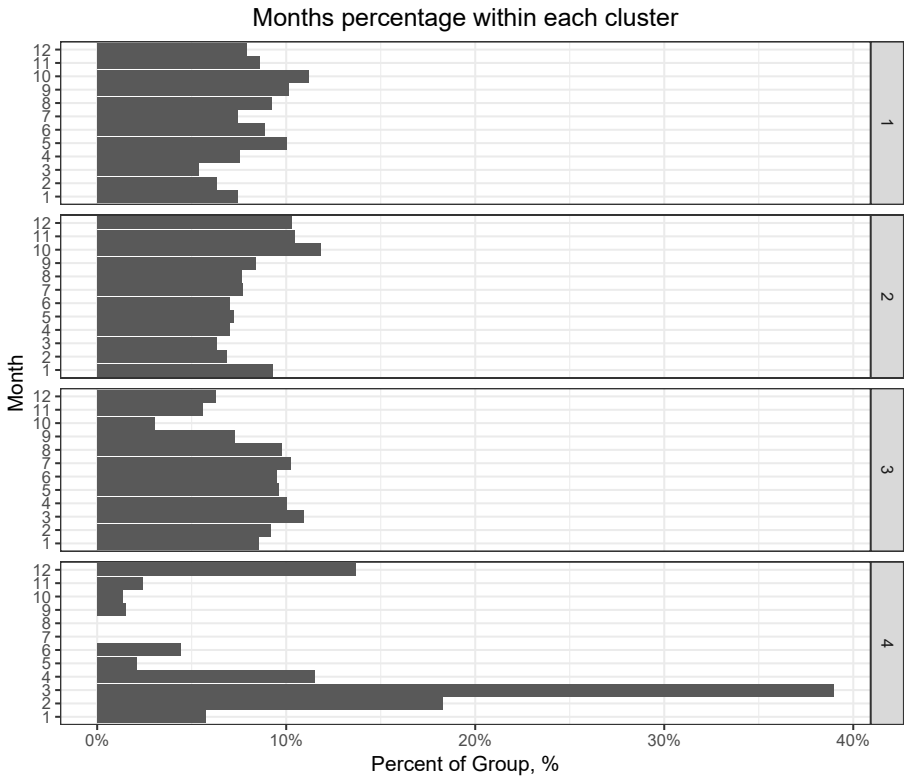


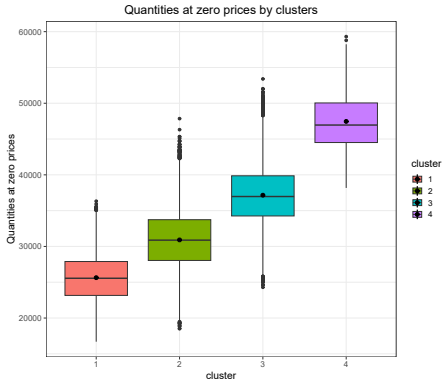
Figure 8: Analysis of the importance of the variables in the concurrent model. Spanish market, 2016–2020.

Spanish market					
Curves	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Original	303	419	547	594	677
1-NN	290	415	548	597	688
3-NN	901	1240	1649	1781	1999
5-NN	1470	2050	2751	2960	3287
PJM market					
Curves	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Original	529	576	597	619	669
1-NN	531	576	598	619	704
3-NN	1601	1729	1793	1857	2000
5-NN	2671	2882	2989	3093	3328
West Australian market					
Curves	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Original	8	23	29	36	68
1-NN	9	23	29	36	65
3-NN	31	72	86	105	182
5-NN	54	121	143	173	300

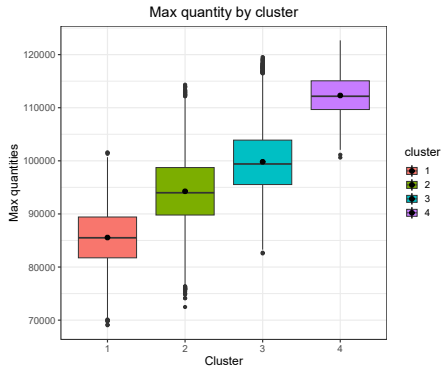
Table 3: Summary statistics of the number of steps in the original curves and in the curves resulting from using one, three, and five nearest neighbors for the three considered markets.



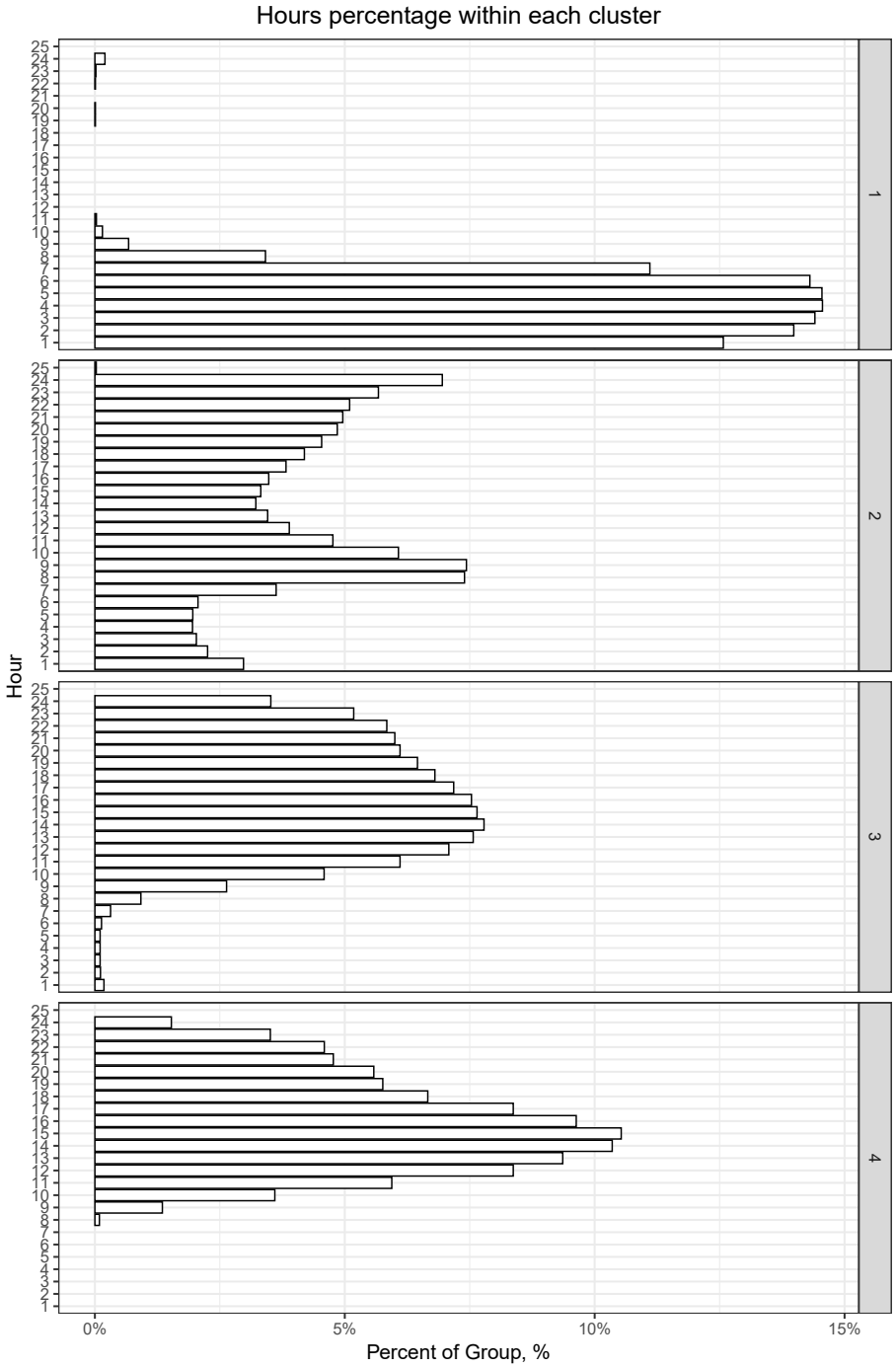
(a) The percentage of months in each cluster.



(b) Boxplot of quantity at zero price by cluster.



(c) Boxplot of maximum quantity by cluster.



(d) The hour percentage in each cluster. The 25th hour exists due to the daylight saving hour.

Figure 9: Descriptive analysis of month, hour, quantity at zero price and maximum quantity by cluster. Spanish market, 2016–2020.

- Let $s^* = \arg\min_{\{s \leq t-H\}} d(\mathbf{C}(t), \mathbf{C}(s))$, that is, $\mathbf{C}(s^*)$ are the H closest consecutive curves to $\mathbf{C}(t)$. The prediction is obtained by $\mathbf{P}(t+H) = \mathbf{C}(s^*+H) = \{C_{s^*+1}, C_{s^*+2}, \ldots, C_{s^*+H}\}$.

In the previous procedure, the distance, d, between the sets of curves $\mathbf{C}(t)$ and $\mathbf{C}(s)$ would remain to be defined. In our experiments, we used the following choices: (B1) the sum of the distances at different hours, $d(\mathbf{C}(t), \mathbf{C}(s)) = \sum_{i=0}^{H-1} d(C_{t-i}, C_{s-i})$, and (B2) the maximum daily distance, $d(\mathbf{C}(t), \mathbf{C}(s)) = \max\{d(C_{t-i}, C_{s-i}) : 0 \leq i \leq H-1\}$. These are our benchmark models.

The study of these different distances leads us to the question of what distance is important for the prediction, $d(\mathbf{C}(t), \mathbf{C}(s))$ or $d(\mathbf{C}(t+H), \mathbf{P}(t+H))$? Obviously, the second of these distances is more relevant because it measures the prediction error. However, this distance cannot be calculated a priori because $\mathbf{C}(t+H)$ is not known at the time of prediction. Then, the question is whether we can learn from $d(\mathbf{C}(t), \mathbf{C}(s))$ to predict $d(\mathbf{C}(t+H), \mathbf{C}(s+H)$. Note that the prediction of $\mathbf{C}(t+H)$ by the 1-NN procedure will be $\mathbf{P}(t+H) = \mathbf{C}(s+H)$.

In what follows we present a distance-based learning procedure. We simplify the problem and predict $d(C_{t+h}, C_{s+h})$, that is, the distance between the curves of hour h. For two times in the training sample, s and t with $s \leq t-H$, we can calculate the following distances:

$$\begin{matrix} d(C_t, C_s), d(C_{t-1}, C_{s-1}), \ldots, d(C_{t-H-1}, C_{s-H-1}) \\ d(C_t, C_{s+h}), d(C_{t-1}, C_{s+h}), \ldots, d(C_{t-H-1}, C_{s+h}) \end{matrix} .$$

These $2H$ variables will be the input of a machine learning procedure to predict the distance $d(C_{t+h}, C_{s+h})$. Figure 10 illustrates this procedure for all horizons (A) and for a specific lag (B). For the case of all horizons, we predict the sum of the distances, $\sum_{h=1}^{H} d(C_{t+h}, C_{s+h})$. Algorithm 1 presents the distance-based learning procedure for a generic machine learning model. It is clear that model "g" in step 2 of the algorithm can be any model with a numerical response variable and numerical explanatory variables. In Tables 4–6, we present the results with two tree-based models: Random Forest (RF) and eXtreme Gradient Boosting (XGB) [8].

The selection of these variables is motivated by the high temporal dependence between the curves, both in the short term and in lags that are multiples of H (see, for instance, [19]). Of course, other variables can be used. The choice of tree-based procedures is based on its versatility in modelling non-linear relations and interactions between variables. The same procedure will be carried out for the different prediction horizons, $h = 1, 2, \ldots, H$, that is, H models are trained. These models allow us to make predictions for the next day using the predictions of the distances for the different forecast horizons.

Tables 4–6 show the means of the weighted distances between the real curve and its prediction in the test sample (8,784 hours of the year 2020 for the Spanish market, 8,760 hours of the year 2022 for the PJM market and 17,568 half-hours of the West Australian market). In the tables, 1-NN-B1 and 1-NN-B2 correspond to the benchmark models using the sum and the maximum of the daily distances,

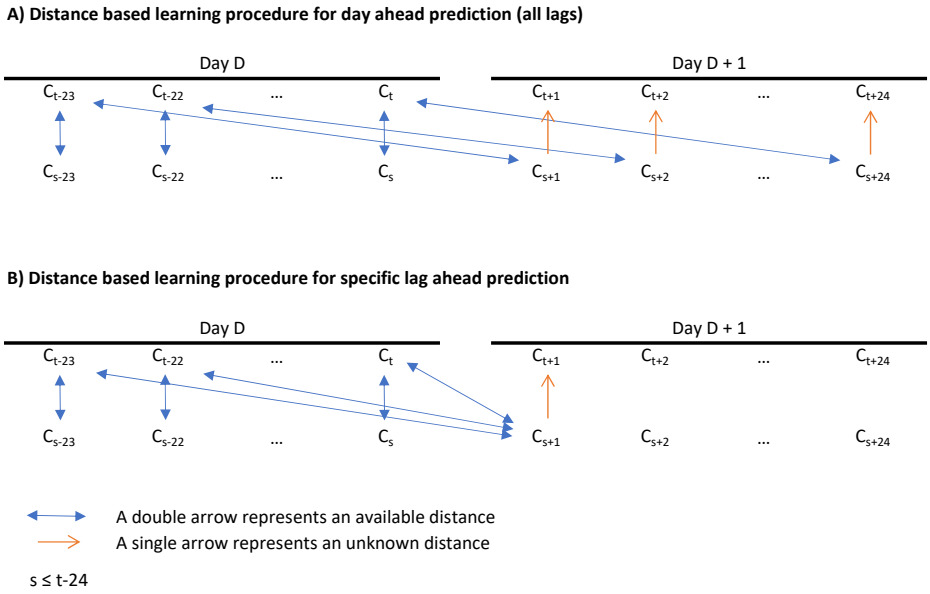


Figure 10: Distance-based learning procedure representation.

respectively; 1-NN+RF (1-NN+XGB) corresponds to the combination of 1-NN with the random forest (extreme gradient boosting) procedure trained for all horizons represented in Figure 10(A); and 1-NN+RF(h) (1-NN+XGB(h)) with the random forest (extreme gradient boosting) procedure trained for horizon h represented in Figure 10(B). We have also incorporated the results of two classic methods for function prediction: (FAR) functional autoregression where the response and explanatory variables are the current and lagged functions, respectively [9], and (FPC+ARIMA), which is a two-step procedure that obtains the functional principal components and fits seasonal ARIMA models to the time series of the scores [6]. These procedures have been used in the prediction of supply curves by [41] and [37]. The number of principal components was chosen so that 99% of the variability observed in the set of curves was explained. Two, four, and six components were chosen for the Spanish, PJM, and West Australian markets, respectively. To compare predictions at the same time horizon in the different markets, we select the prediction horizons $h = 1$, 12 and 24 for the Spanish and PJM markets (hourly data), and $h = 2$, 24 and 48 for the West Australian market (half-hourly data).

First, for the Spanish market, we see that the proposed procedure when trained for all horizons (1-NN+RF and 1-NN+XGB) improves the two benchmark models at horizons, $h = 1$ and 12. There are small differences for $h = 24$ between the proposals and the two benchmark procedures. Second, when trained for a single horizon (1-NN+RF(h) and 1-NN+XGB(h)), the proposed procedure

Algorithm 1: Distance-based learning procedure.

Input: A $T \times T$ distance matrix, $\boldsymbol{d}$ with elements $d_{t,s} = d(C_t, C_s)$. Forecasting horizon h.

1 For any pair of indices t and s such that $H \leq s \leq t - H \leq T - 2H$ make up the following vector:

$$\boldsymbol{X}(t,s) = \big(d_{t,s}, d_{t-1,s-1}, \cdots, d_{t-H-1,s-H-1}, d_{t,s+h}, d_{t-1,s+h}, \cdots, d_{t-H-1,s+h}\big),$$

and scalar:

$$Y(t,s) = d(C_{t+h}, C_{s+h}).$$

2 Train a model $g(\cdot)$ such that

$$Y(t,s) \sim g(\boldsymbol{X}(t,s)).$$

3 For $t = T$, and all s such that $H \leq s \leq T - H$, obtain

$$\widehat{Y}(T,s) = g(\boldsymbol{X}(T,s)).$$

4 Select the index with the minimum predicted distance:

$$s^* = \arg \min_{H \leq s \leq t-H} \widehat{Y}(T,s).$$

5 The forecast for C_{T+h} is C_{s^*+h}.

improves the benchmarks at all prediction horizons so it is concluded that training for each prediction horizon separately is significant. The reduction in the mean of the prediction error is notable for $h = 1$ since it is reduced by almost half, while for $h = 12$ and 24 it is close to 20%. When we compare 1-NN with the two procedures that assume smoothness (FAR and FPC+ARIMA), we see quite similar results, although these alternatives obtain slightly better results for $h = 1$ and 12, and worse for $h = 24$. For the PJM and the West Australian markets, both 1-NN+RF and 1-NN+RF(h) (1-NN+XGB and 1-NN+XGB(h)) outperform the two benchmarks. Specifically, for 1-NN+RF(h), the reductions in the mean prediction error with respect to the best benchmark are greater than 10% (PJM) and 15% (West Australian). We also observe that procedures based on nearest neighbors significantly improve the FAR and FPC+ARIMA procedures in these two markets. Generally, RF-based versions obtain slightly better results than those based on XGB, although the differences between the two are small.

The difference in performance between distance-based methods and methods that assume smooth functions can be explained by two reasons: (i) the different numbers of steps and heights of the supply curves of the three markets and (ii) the dispersion of prices and the variability of the curves. That is, the greater the number of steps and the lower the height, the better the smooth approximation

Method	h=1	h=12	h=24
FAR	0.1240 (0.0004)	0.2676 (0.0005)	0.2928 (0.0006)
FPC+ARIMA	0.1251 (0.0004)	0.2528 (0.0005)	0.2668 (0.0006)
1-NN-B1	0.2525 (0.0006)	0.3450 (0.0007)	0.3261 (0.0008)
1-NN-B2	0.2689 (0.0006)	0.3703 (0.0008)	0.3434 (0.0008)
1-NN+XGB	0.2270 (0.0006)	0.3050 (0.0006)	0.3610 (0.0008)
1-NN+XGB(h)	0.1580 (0.0004)	0.3350 (0.0007)	0.3080 (0.0006)
1-NN+RF	0.1929 (0.0005)	0.2902 (0.0006)	0.3368 (0.0007)
1-NN+RF(h)	0.1328 (0.0003)	0.2714 (0.0005)	0.2617 (0.0005)

Table 4: Means of the weighted distances between the actual curve and its prediction in the Spanish market (standard errors in parentheses).

Method	h=1	h=12	h=24
FAR	0.1555 (0.0001)	0.2006 (0.0001)	0.1941 (0.0001)
FPC+ARIMA	0.1552 (0.0001)	0.1807 (0.0001)	0.1618 (0.0001)
1-NN-B1	0.0685 (0.0001)	0.0731 (0.0001)	0.0696 (0.0001)
1-NN-B2	0.0711 (0.0001)	0.0789 (0.0002)	0.0740 (0.0002)
1-NN+XGB	0.0673 (0.0001)	0.0679 (0.0001)	0.0654 (0.0001)
1-NN+XGB(h)	0.0650 (0.0001)	0.0679 (0.0001)	0.0650 (0.0001)
1-NN+RF	0.0616 (0.0001)	0.0633 (0.0001)	0.0617 (0.0001)
1-NN+RF(h)	0.0610 (0.0001)	0.0624 (0.0001)	0.0611 (0.0001)

Table 5: Means of the weighted distances between the actual curve and its prediction in the PJM market (standard errors in parentheses).

Method	h=2	h=24	h=48
FAR	0.2894 (0.0002)	0.3348 (0.0002)	0.3189 (0.0002)
FPC+ARIMA	0.3540 (0.0005)	0.4285 (0.0005)	0.3521 (0.0004)
1-NN-B1	0.1983 (0.0002)	0.1786 (0.0002)	0.2100 (0.0002)
1-NN-B2	0.2056 (0.0002)	0.1906 (0.0003)	0.2228 (0.0003)
1-NN+XGB	0.1825 (0.0002)	0.1616 (0.0002)	0.2058 (0.0002)
1-NN+XGB (h)	0.1696 (0.0002)	0.1560 (0.0002)	0.1926 (0.0002)
1-NN+RF	0.1766 (0.0002)	0.1581 (0.0002)	0.1931 (0.0002)
1-NN+RF(h)	0.1595 (0.0002)	0.1512 (0.0002)	0.1770 (0.0002)

Table 6: Means of the weighted distances between the actual curve and its prediction in the West Australian market (standard errors in parentheses).

of the supply curve, and the less the variability of the curves, the easier the prediction of that market will be.

In fact, we observe that the Spanish and PJM markets have supply curves

with more steps (517.5 and 598.3 on average, respectively) and a lower step height (0.0062 and 0.0059 on average, respectively). This contrasts with the Western Australian market, which has supply curves with the lowest number of steps (29.75 on average) and the greatest height between steps (0.0314 on average), with distance-based methods being much superior for this market given that they preserve the stepped nature of the curves.

Regarding the dispersion of prices, for the periods considered, the Spanish market has prices in the range [0, 180.3], while the PJM and West Australian markets have prices in the ranges [0, 2000] and [0, 604], respectively. We also observe that functional principal components (FPC) necessary to explain 99% of the variability are much lower in the Spanish market than in the other markets. In fact, the variability explained by the first component in the three markets was: 96.4%, 57.6%, and 89.2%. This explains why the Spanish market is easier to predict than the PJM market although its curves have a similar number of steps and heights of steps.

5. Conclusions

In this paper, we clustered supply curves in the Spanish day-ahead market, the Pennsylvania–New Jersey–Maryland market, and the West Australian market, using a weighted distance that takes into account the offer's price distribution. We have found four main clusters for the Spanish market that are explained mainly by temporal variables such as month, day of the week, and hour, and market variables such as global system demand and amount of power generated by different technologies (wind, solar, and nuclear). Depending on the availability of concurrent curves when training, we developed models that integrate the explanatory variables (i.e. market variables, temporal variables, and lagged labels) with/without geometric features, as the latter cannot be captured until all concurrent curves are at hand. The random forest model fitted with geometric information has the lowest out-of-bag (OOB) error of 5.73%. This suggests that the months, quantities at zero prices, and the labels of the 24-hour-ahead ancestor shape the profiles of the curves most powerfully. In the absence of concurrent information, the RF model does not use geometric features and hence characterizes the clusters with an OOB error of 8.44%. The most important variables change to wind power generation and 24-hour lagged labels. That the use of concurrent features reduces the errors is consistent across the other two markets as well when aggregating into four clusters. The error decreases from 8.84% to 4.78% in the West Australian market and from 0.81% to 0.51% in the PJM market. If the number of clusters is set to two, as suggested by the average silhouette, model misclassifications are close to zero in the PJM and Western Australian markets.

The availability of the distance matrix allows us to propose a forecasting procedure based on the distances between the curves of the previous day and the curves in the training set. We carried out a prediction exercise for all hours (half hours) of the year in the test set, where we evaluated the behavior of the proposed procedures and compared it with two benchmark procedures (based on

near neighbors) and two procedures that assume that the functions are smooth. The proposed procedure has a competitive behavior in the three considered markets.

We are currently studying extensions of the proposed clustering and forecasting procedure to incorporate covariates and other distances applied in the literature, such as [3] and [12]. As to the covariates, we will consider the measurements of wind speed and solar radiation in a grid distributed in the application area of each market using the availability of these data at Free Open-Source Weather API. The problem we face is converting this spatially distributed meteorological information into useful covariates to predict supply curves. [3] and [12] propose combining distances and the selection of optimal weights to improve the performance of supervised and unsupervised classification procedures, respectively. We will study whether their proposals are extendable to the prediction of functions.

Acknowledgements

The authors gratefully acknowledge financial support from the Spanish government through the Ministry of Science and Innovation projects PID2019-108311GB-I00, PID2020-115460GB-I00, and PID2022-138114NB-I00. Andrés M. Alonso has been a beneficiary of the Google Cloud Research Credits Program carrying out part of the calculations on this platform. Antonio Elías was supported under PAIDI 2020 funded by the Junta de Andalucía and the European Social Fund. The work of Juan M. Morales has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement No 755705).

Appendix A. Hierarchical clustering with average linkage of the original curves of PJM market

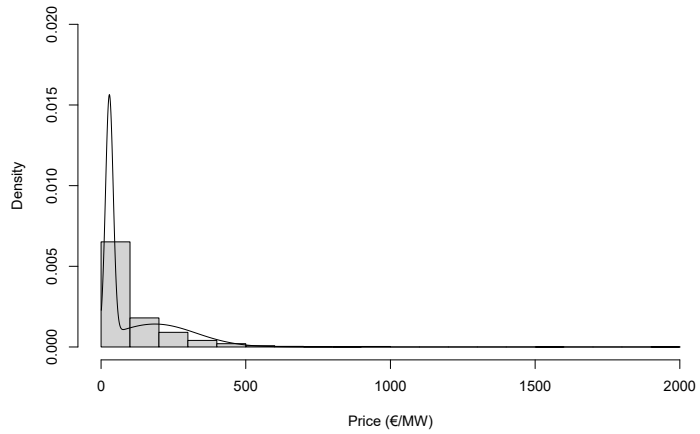


Figure A.1: Histogram of the distribution of the prices of the offers and density of the Gaussian mixture model with two components. PJM market, 2018–2022.

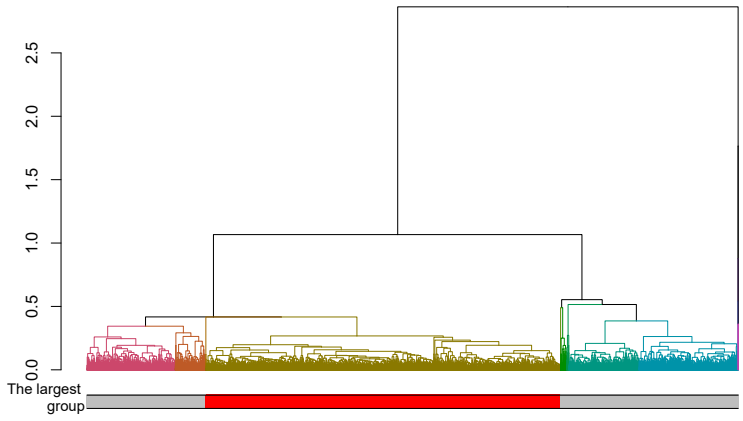


Figure A.2: The dendrogram before discarding small groups. A height of 0.3 was used to cut 14 groups with sizes presented in Table A.1. The colour of a branch is that of the cluster to which it belongs. The largest cluster is highlighted by the red block of the colour bar. PJM market, 2018–2022.

Table A.1: Cluster sizes before discarding small groups by cutting the dendrogram at height = 0.3. PJM market, 2018–2022.

Clusters	**Member amounts**	**Clusters**	**Member amounts**
1	418	8	1
2	2032	9	1
3	5939	10	6674
4	23854	11	1
5	1	12	4709
6	48	13	144
7	1	14	1

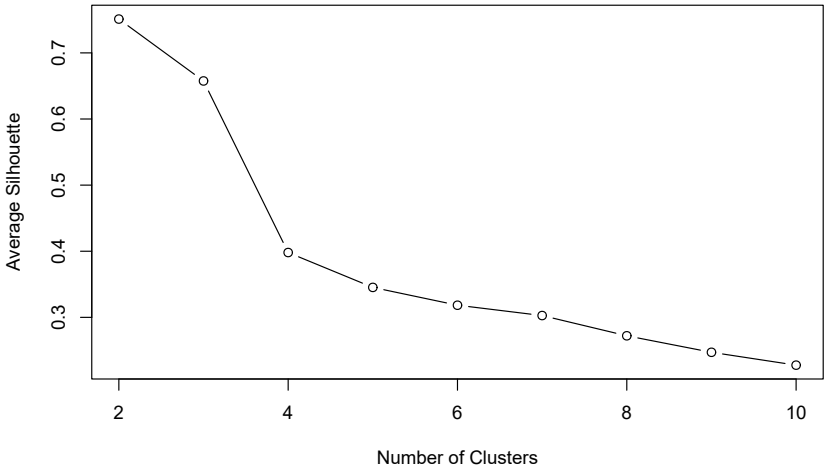


Figure A.3: The average silhouette peaks at two. PJM market, 2018–2022.

Table A.2: Descriptions of the explanatory variables for the PJM market.

	Variable name	Description
Temporal variables	hour	Integer. The hour of day of the supply curve;
	month	Factor. The month of the supply curve;
	year	Factor. The year of the supply curve;
	weekday	Factor. The day of the week for the supply curve;
	holiday	Factor. If a supply curve is for a public holiday;
Market variables	gen_Solar[a]	Numeric. Real solar power generation;
	gen_Wind[a]	Numeric. Real wind power generation;
	gen_Nuclear[a]	Numeric. Real nuclear power generation;
	pred_demand_T5[b]	Numeric. The predicted electricity demand of the hour of the supply curve, the prediction is performed at 05:45 of the previous day;
	pred_demand_T9[b]	Numeric. The predicted electricity demand of the hour of the supply curve, the prediction is performed at 09:45 of the previous day;
	pred_demand_T11[b]	Numeric. The predicted electricity demand of the hour of the supply curve, the prediction is performed at 11:45 of the previous day;
	pred_demand_T17[b]	Numeric. The predicted electricity demand of the hour of the supply curve, the prediction is performed at 17:45 of the previous day;
	pred_demand_T23[b]	Numeric. The predicted electricity demand of the hour of the supply curve, the prediction is performed at 23:45 of the previous day;
	real_GDP[c]	Numeric. Quarterly real GDP;
Lagged label	cluster_24lag	Factor. The clusters with 24-hour lags.

Source:

[a] `https://dataminer2.pjm.com/feed/gen_by_fuel/definition`

[b] `https://dataminer2.pjm.com/feed/load_frcstd_hist/definition`

[c] `https://apps.bea.gov/regional/`

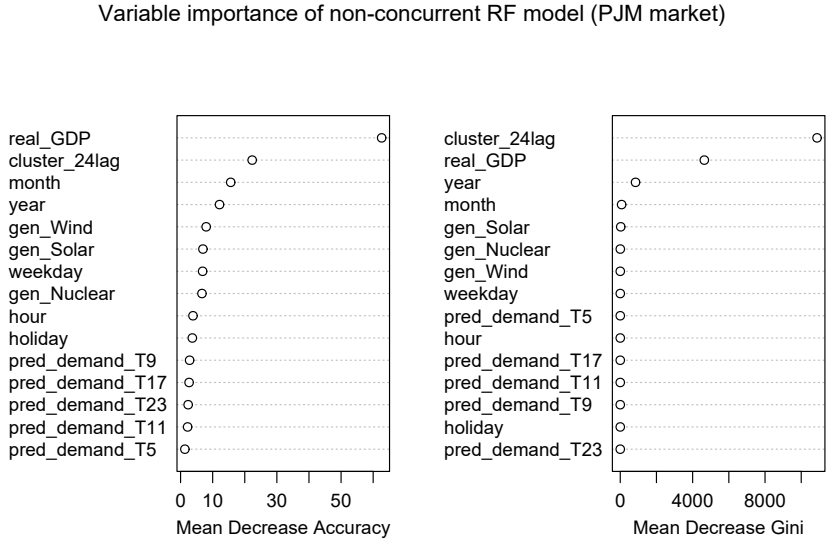
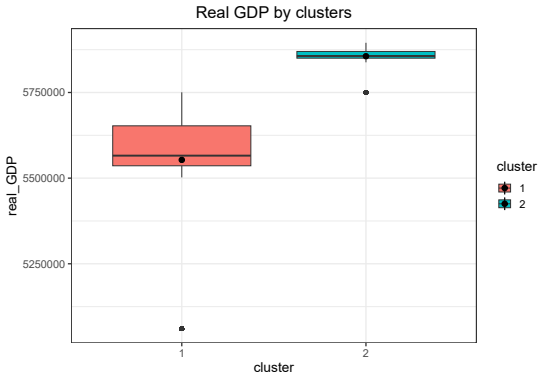
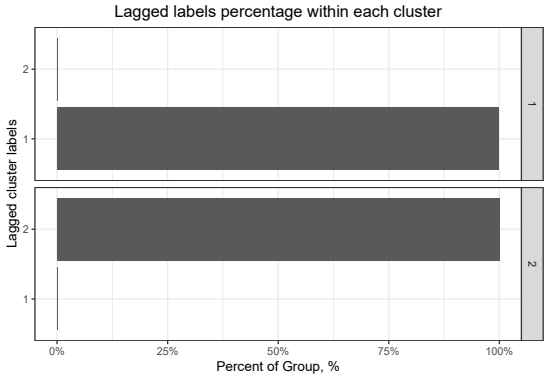


Figure A.4: Analysis of the importance of the variables in the non-concurrent model. PJM market, 2018–2022.



(a) The boxplot of real GDP by cluster. PJM market, 2018–2022.



(b) The lagged labels percentage in each cluster.

Figure A.5: Descriptive analysis of real GDP and lagged labels by cluster. PJM market, 2018–2022.

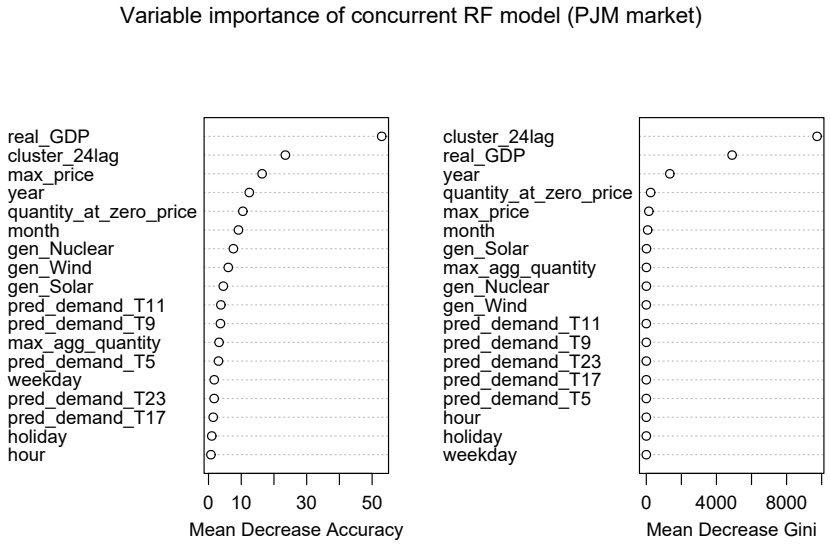
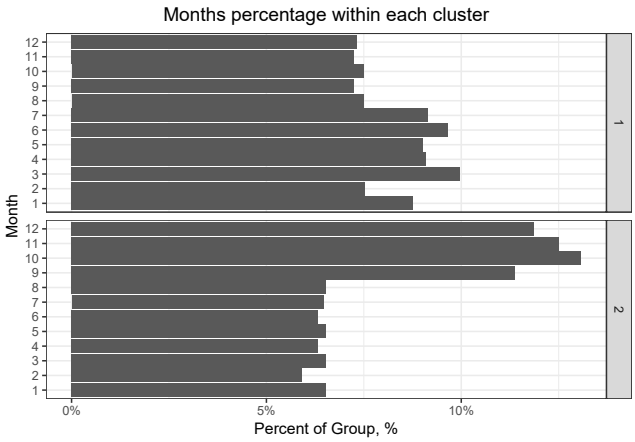
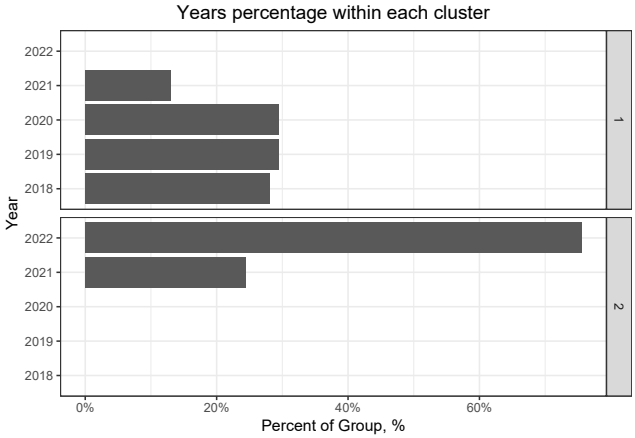


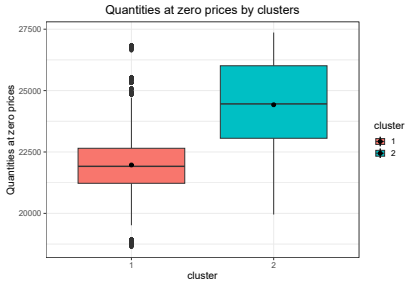
Figure A.6: Analysis of the importance of the variables in the concurrent model. PJM market, 2018–2022.



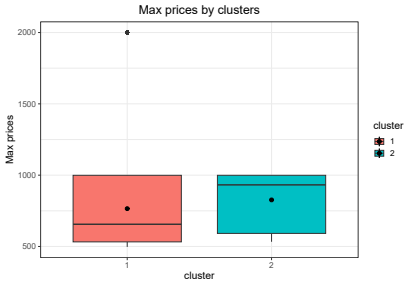
(a) The percentage of months in each cluster.



(b) The percentage of years in each cluster.



(c) Boxplot of quantity at zero price by cluster.



(d) Boxplot of maximum price by cluster.

Figure A.7: Descriptive analysis of month, year, quantity at zero price and maximum price by cluster. PJM market, 2018–2022.

Appendix B. Hierarchical clustering with average linkage of the original curves of the West Australian market

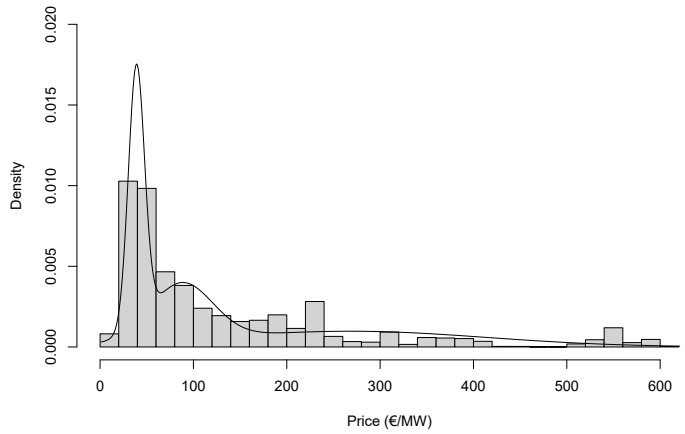


Figure B.1: Histogram of the distribution of the prices of the offers and density of Gaussian mixture model with three components. West Australian market, 2016–2020.

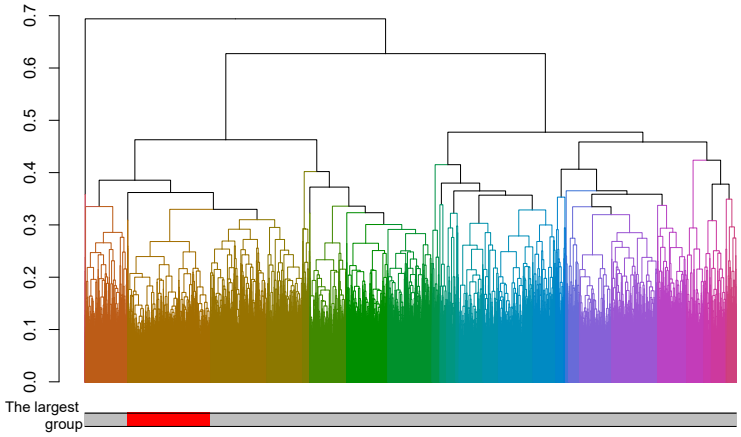


Figure B.2: Dendrogram before discarding small groups. Cutting at the height of 0.3 generated 38 groups with sizes presented in Table B.1. The colour of a branch is that of the cluster to which it belongs. The largest cluster is highlighted by the red block of the colour bar. West Australian market, 2016–2020.

Table B.1: Cluster sizes before discarding small groups by cutting the dendrogram at height = 0.3. West Australian market, 2016–2020.

Clusters	**Member amounts**	**Clusters**	**Member amounts**	**Clusters**	**Member amounts**
1	1442	14	261	27	5964
2	1291	15	429	28	13
3	1032	16	4301	29	6
4	113	17	3214	30	5543
5	1189	18	2064	31	17
6	2318	19	7635	32	943
7	209	20	4788	33	12
8	1931	21	5628	34	82
9	1003	22	11132	35	979
10	6179	23	12	36	452
11	832	24	4961	37	3931
12	3075	25	6	38	4
13	4686	26	19		

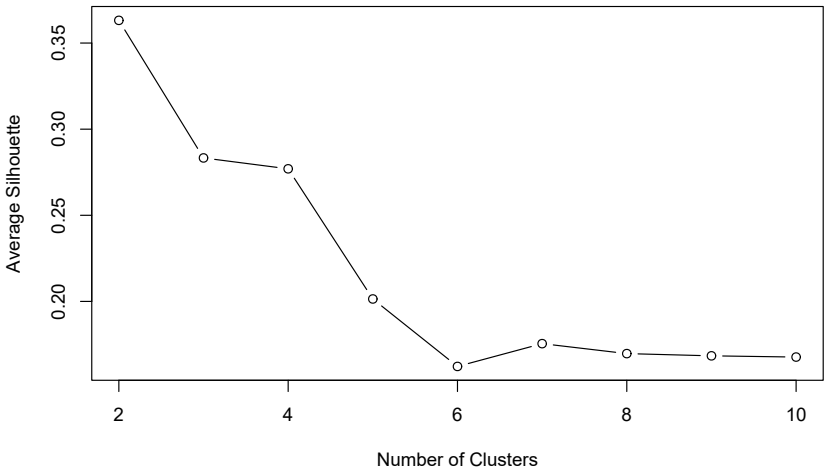


Figure B.3: The average silhouette peaks at two. West Australian market, 2016–2020.

Table B.2: Descriptions of explanatory variables for the West Australian market.

	Variable name	Description
Temporal variables	interval_number	Integer. The semi-hour trading interval for the supply curve (e.g. the interval commencing at 8:00 am is marked as 1, and 9:30 am as 4.);
	month	Factor. The month of the supply curve;
	year	Factor. The year of the supply curve;
	weekday	Factor. The day of the week of the supply curve;
	holiday	Factor. If a supply curve is for a public holiday;
Market variables	generation[a]	Numeric. Real total sent out generation;
	pred_demand[a]	Numeric. The predicted electricity demand of the hour of the supply curve;
	annual_GDP[b]	Numeric. Annual real GDP (Chain volume measures);
Lagged label	cluster_24lag	Factor. The clusters with 24-hour lags.

Source:

[a] `https://aemo.com.au/energy-systems/electricity/wholesale-electricity-market-wem/data-wem/market-data-wa`

[b] `https://www.abs.gov.au/statistics/economy/national-accounts`

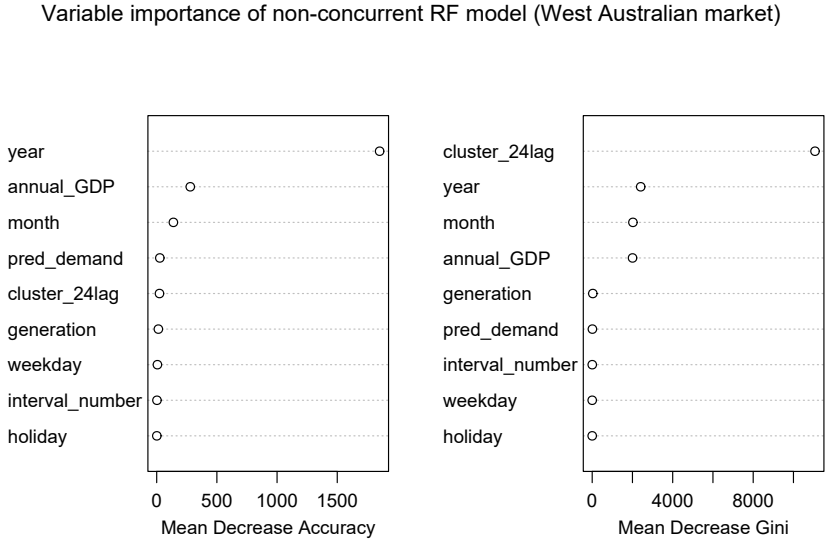
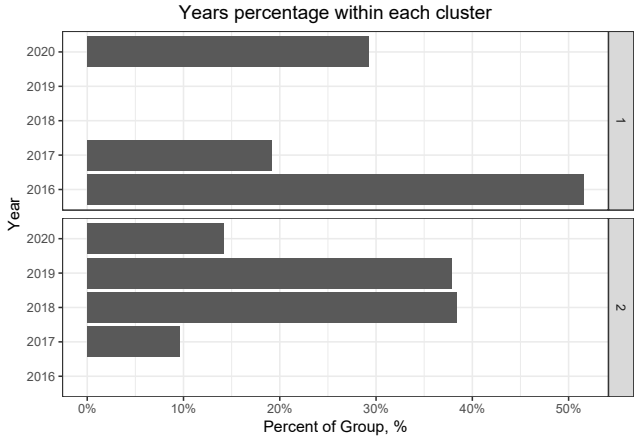
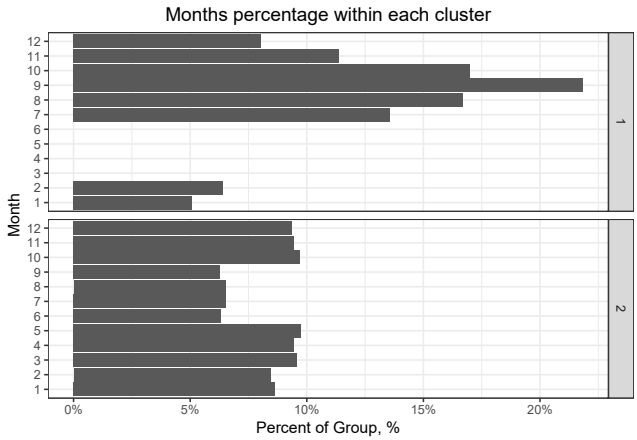


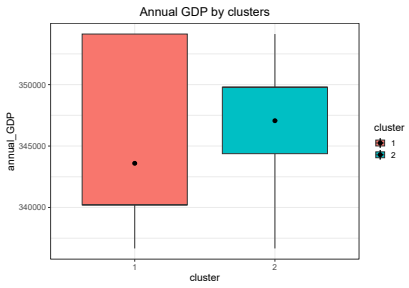
Figure B.4: Analysis of the importance of the variables in the non-concurrent model. West Australian market, 2016–2020.



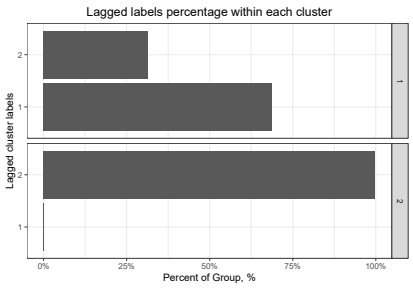
(a) The percentage of a year in each cluster.



(b) The percentage of months in each cluster.



(c) Boxplot of annual GDP by cluster.



(d) The lagged label percentage in each cluster.

Figure B.5: Descriptive analysis of cluster_24lag, year, month, and annual GDP by cluster. West Australian market, 2016–2020.

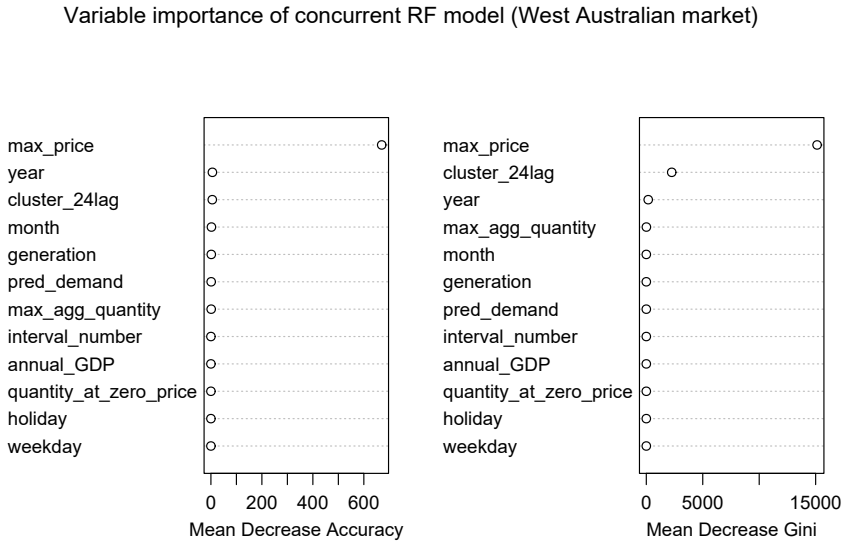


Figure B.6: Analysis of the importance of the variables in the concurrent model. West Australian market, 2016–2020.

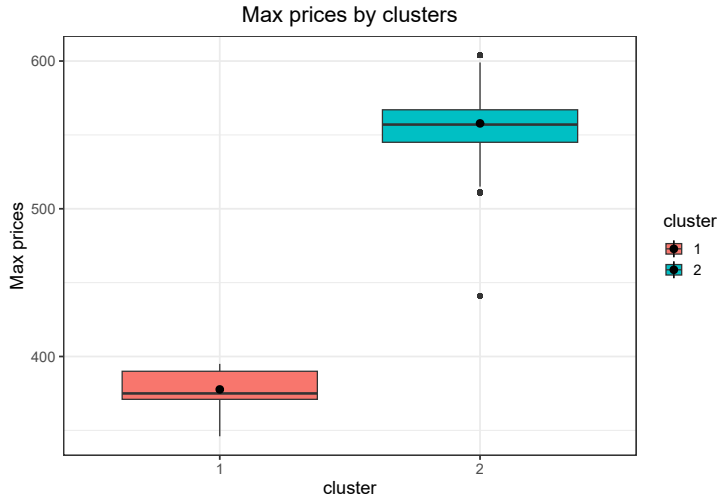


Figure B.7: The boxplot of maximum price by cluster. West Australian market, 2016–2020.

References

[1] Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scandinavian Journal of Statistics 30, 581–595. doi:https://doi.org/10.1111/1467-9469.00350.

[2] AlMahamid, F., Grolinger, K., 2022. Agglomerative hierarchical clustering with dynamic time warping for household load curve clustering, in: 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 241–247. doi:10.1109/CCECE49351.2022.9918481.

[3] Alonso, A.M., Casado, D., Romo, J., 2012. Supervised classification for functional data: A weighted distance approach. Computational Statistics & Data Analysis 56, 2334–2346. doi:https://doi.org/10.1016/j.csda.2012.01.013.

[4] Alonso, A.M., Nogales, F.J., Ruiz, C., 2020. Hierarchical clustering for smart meter electricity loads based on quantile autocovariances. IEEE Transactions on Smart Grid 11, 4522–4530. doi:10.1109/TSG.2020.2991316.

[5] Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J., 1999. OPTICS: Ordering points to identify the clustering structure. SIGMOD Rec. 28, 49–60. doi:10.1145/304181.304187.

[6] Aue, A., Norinho, D.D., Hörmann, S., 2015. On the prediction of stationary functional time series. Journal of the American Statistical Association 110, 378–392. doi:10.1080/01621459.2014.909317.

[7] Béjar-Alonso, J., 2013. Strategies and algorithms for clustering large datasets: A review. URL: `http://hdl.handle.net/2117/23415`.

[8] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review , 1937–1967doi:https://doi.org/10.1007/s10462-020-09896-5.

[9] Bosq, D., 2000. Linear Processes in Function Spaces. Springer-Verlag, New York.

[10] Boullé, M., 2012. Functional data clustering via piecewise constant nonparametric density estimation. Pattern Recognition 45, 4389–4401. doi:10.1016/j.patcog.2012.05.016.

[11] Bouveyron, C., Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis 71, 52–78. doi:10.1016/j.csda.2012.12.008.

[12] Chen, H., Reiss, P.T., Tarpey, T., 2014. Optimally weighted L2 distance for functional data. Biometrics 70, 516–525. doi:https://doi.org/10.1111/biom.12161.

[13] Chiou, J.M., Li, P.L., 2007. Functional Clustering and Identifying Substructures of Longitudinal Data. Journal of the Royal Statistical Society Series B: Statistical Methodology 69, 679–699. doi:10.1111/j.1467-9868.2007.00605.x.

[14] Cuesta-Albertos, J.A., Fraiman, R., 2007. Impartial trimmed k-means for functional data. Computational Statistics & Data Analysis 51, 4864–4877. doi:https://doi.org/10.1016/j.csda.2006.07.011.

[15] Dasgupta, S., Srivastava, A., Cordova, J., Arghandeh, R., 2019. Clustering household electrical load profiles using elastic shape analysis, in: 2019 IEEE Milan PowerTech, IEEE. pp. 1–6. doi:10.1109/PTC.2019.8810883.

[16] Delaigle, A., Hall, P., 2010. Defining probability density for a distribution of random functions. The Annals of Statistics 38, 1171–1193. doi:10.1214/09-AOS741.

[17] Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer-Verlag, New York.

[18] Ferreira, L., Hitchcock, D.B., 2009. A comparison of hierarchical methods for clustering functional data. Communications in Statistics-Simulation and Computation 38, 1925–1949. doi:10.1080/03610910903168603.

[19] Franco-Comas, A.R., Alonso, A.M., 2021. Prediction and classification based on partially observed distances, in: Peña, D., Poncela, P., Ruiz, E. (Eds.), Econometric analysis and big data (In Spanish). Funcas, Madrid, pp. 191–222.

[20] Ghelasi, P., Ziel, F., 2022. Hierarchical forecasting for aggregated curves with an application to day-ahead electricity price auctions. International Journal of Forecasting , In press.doi:10.1016/j.ijforecast.2022.11.004.

[21] Guha, S., Rastogi, R., Shim, K., 1998. CURE: An efficient clustering algorithm for large databases. SIGMOD Rec. 27, 73–84. doi:10.1145/276305.276312.

[22] Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.

[23] Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V., 2013. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. Journal of the Royal Statistical Society. Series C (Applied Statistics) 62, 401–418. URL: http://www.jstor.org/stable/24771812.

[24] Jacques, J., Preda, C., 2014. Functional data clustering: A survey. Advances in Data Analysis and Classification 8, 231–255. doi:10.1007/s11634-013-0158-y.

[25] Jamasb, T., Pollitt, M., 2005. Electricity market reform in the European Union: review of progress toward liberalization & integration. The Energy Journal 26, 11–41. URL: http://www.jstor.org/stable/23297005.

[26] James, G.M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98, 397–408. doi:10.1198/016214503000189.

[27] Jin, Y., Bi, Z., 2018. Power load curve clustering algorithm using fast dynamic time warping and affinity propagation, in: 2018 5th International Conference on Systems and Informatics (ICSAI), IEEE. pp. 1132–1137. doi:10.1109/ICSAI.2018.8599336.

[28] Karthikeyan, S.P., Raglend, I.J., Kothari, D.P., 2013. A review on market power in deregulated electricity market. International Journal of Electrical Power & Energy Systems 48, 139–147. doi:10.1016/j.ijepes.2012.11.024.

[29] Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley.

[30] Kayano, M., Dozono, K., Konishi, S., 2010. Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. Journal of Classification 27, 1432–1343. doi:10.1007/s00357-010-9054-8.

[31] Kroes, N., 2007. More competitive energy markets: Building on the findings of the sector inquiry to shape the right policy solutions. http://europa.eu/rapid/press-release_SPEECH-07-547_en.htm. European Energy Institute, Brussels.

[32] Ling, R.F., 1981. Cluster analysis algorithms for data reduction and classification of objects. Technometrics 23, 417–418. doi:10.1080/00401706.1981.10487693.

[33] Matuk, J., Bharath, K., Chkrebtii, O., Kurtek, S., 2022. Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. Journal of the American Statistical Association 117, 1964–1980. doi:10.1080/01621459.2021.1893179.

[34] Mestre, G., Portela, J., Muñoz San Roque, A., Alonso, E., 2020. Forecasting hourly supply curves in the Italian day-ahead electricity market with a double-seasonal SARMAHX model. International Journal of Electrical Power & Energy Systems 121, 106083. doi:10.1016/j.ijepes.2020.106083.

[35] Mitridati, L., Pinson, P., 2018. A Bayesian inference approach to unveil supply curves in electricity markets. IEEE Transactions on Power Systems 33, 2610–2620. doi:10.1109/TPWRS.2017.2757980.

[36] Patra, B.K., Nandi, S., Viswanath, P., 2011. A distance based clustering method for arbitrary shaped clusters in large datasets. Pattern Recognition 44, 2862–2870. doi:10.1016/j.patcog.2011.04.027.

[37] Pelagatti, M., 2013. Supply function prediction in electricity auctions, in: Grigoletto, M., Lisi, F., Petrone, S. (Eds.), Complex Models and Computational Methods in Statistics. Springer-Verlag, Milano, pp. 203–213. URL: https://doi.org/10.1007/978-88-470-2871-5_16, doi:10.1007/978-88-470-2871-5_16.

[38] Peng, J., Müller, H.G., 2008. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. The Annals of Applied Statistics 2, 1056–1077. doi:10.1214/08-AOAS172.

[39] Ramsay, J., Silverman, B., 2005. Functional Data Analysis. 2nd ed., Springer-Verlag, New York.

[40] Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

[41] Shah, I., Lisi, F., 2020. Forecasting of electricity price through a functional prediction of sale and purchase curves. Journal of Forecasting 39, 242–259. doi:10.1002/for.2624.

[42] Sioshansi, F.P., 2008. Electricity market reform and 'reform of the reforms'. International Journal of Global Energy Issues 29, 3–27. doi:10.1504/IJGEI.2008.016339.

[43] Tarpey, T., Kinateder, K.K.J., 2003. Clustering functional data. Journal of Classification 20, 1432–1343. doi:10.1007/s00357-003-0007-3.

[44] Teeraratkul, T., O'Neill, D., Lall, S., 2018. Shape-based approach to household electric load curve clustering and prediction. IEEE Transactions on Smart Grid 9, 5196–5206. doi:10.1109/TSG.2017.2683461.

[45] Ziel, F., Steinert, R., 2016. Electricity price forecasting using sale and purchase curves: The X-model. Energy Economics 59, 435–454. doi:10.1016/j.eneco.2016.08.008.