A Data-Mining- BasedMethodology for Transmission Expansion Planning

Judite Ferreira, Sérgio Ramos, Zita Vale, and João Soares

In recent decades, all over the world, competition in the electric power sec- tor has deeply changed the way this sector's agents play their roles. In most countries, electric process deregulation was conducted in stages, beginning with the clients of higher voltage levels and with larger electricity consumption, and later extended to all electrical consumers. The sector liberalization and the operation of competitive electricity markets were expected to lower prices and improve quality of service, leading to greater consumer satisfaction.

Transmission and distribution remain noncompetitive business areas, due to the large infrastructure investments required. However, the industry has yet to clearly establish the best business model for transmission in a competitive environment.[1] After generation, the electricity needs to be delivered to the electrical system nodes where demand requires it, taking into consideration transmission constraints and electrical losses. If the amount of power flowing through a certain line is close to or surpasses the safety limits, then cheap but distant generation might have to be replaced by more expensive closer generation to reduce the exceeded power flows. In a congested area, the optimal price of electricity rises to the marginal cost of the local generation or to the level needed to ration demand to the amount of available electricity. Even without congestion, some power will be lost in the transmission system through heat dissipation, so prices reflect that it is more expensive to supply electricity at the far end of a heavily loaded line than close to an electric power generation.

Locational marginal pricing (LMP), resulting from bidding competition, represents electrical and economical values at nodes or in areas that may provide economical indicator signals to the market agents. This article proposes a data-mining-based methodology that helps characterize zonal prices in real power transmission networks.[2,3] To test our methodology, we used an LMP database from the California Independent System Operator for 2009 to identify economical zones. (CAISO is a nonprofit public benefit corporation charged with operating the majority of California's high-voltage wholesale power grid.) To group the buses into typical classes that represent a set of buses with

the approximate LMP value, we used two-step and k-means clustering algorithms. By analyzing the various LMP components, our goal was to extract knowledge to support the ISO in investment and network-expansion planning.[4]

Transmission Pricing

Typically, the three approaches for transmission pricing are uniform, nodal, and zonal. Uniform pricing is common in interconnected transmission networks without structural congestion problems.[2] For transmission networks with long transmission lines corresponding to large distances between supply and demand areas, nodal prices are adequate. Nodal prices consider the marginal cost of transmission losses and the cost of extra generation that must supply the demand increment, if transmission congestion exists. Zonal pricing gathers nodes into zones that are bounded by potential constraint interfaces. Each of these zones has its own LMP. The purpose is to encourage generators to locate themselves within the boundaries of the high-priced zones, relieving the flow constraints in the congested interfaces between zones.[1]

Given the network constraints, the electricity price reveals the marginal operating costs of electricity production, the transmission congestion charges, and the costs of losses. Customers will pay producers their local price. LMP gives clear signals for the production and consumption of electrical energy and for the construction of new generation and transmission facilities, assuring the correct economic signals in the marketplace.

LMP reflects marginal price variation with time and location based on transmission congestion.[5] It also lets us model detailed power flows on specific lines and provide users with individual nodal pricing, whereas the zonal representation does not involve monitoring individual lines and assumes that all prices are the same within each zone. LMP can be used to define zonal boundaries and to decide whether any zone should be merged with another zone or split into new zones.[6]

Locational Marginal Pricing

The last few years of operation have demonstrated some shortcomings of the CAISO's original zonal congestion management. CAISO has been redesigning its forward scheduling and congestion management procedures.[7] This market redesign and technology upgrade (MRTU) undertakes congestion management using a detailed model of the transmission grid instead of the previous zonal approach.

During recent years, the resulting operational and cost impacts have become progressively higher as new power generation has come online in congested areas of the transmission network. The new design addressed these issues in order to improve network reliability and efficient utilization of California's transmission and generation facilities by producing more transparent price signals.

The proposed method uses these components: LMP energy, LMP loss, and LMP congestion. *LMP energy* corresponds to the energy price and is the same for all buses. The *LMP loss* component reflects the costs of losses and can be positive or negative, depending on the direction of flow away from or toward the reference bus.[8]

Transmission losses along a line are proportional to the square of the power flow. However, the actual losses at each system line change from moment to moment, depending on the power flows within the entire system.[8] The losses factors can be calculated with respect to the reference bus to evaluate the LMP loss component. Nevertheless, the loss factor for moving power between any two buses at a given instant will remain the same, and it can be calculated by taking the difference between the loss factors at the two buses. Thus, both the price differential and loss factor differential must be considered in the cost of moving power from one bus to another.[8]

The *LMP congestion* component accounts for the costs of congestion measured between a certain node and a reference node. This component can be positive or negative, depending whether serving additional demand increases or reduces congestion, respectively.

In competitive electricity markets, LMP gives participants important pricing signals because the effects of transmission losses and binding constraints are embedded in it. Although LMP provides valuable information at each location, it does not provide a detailed description in terms of contribution terms. The LMP components, on the other hand, show the explicit price decomposition into contribution components and thus are better market signals.[2]

In order to evaluate LMP and solve congestion situations, an economic dispatch optimization problem for a snapshot of time t is first considered:

$$\text{Min} \sum_i c_i * Pg_i \qquad (1)$$

subject to

$$\sum_i Pg_i = \sum_i Pd_i \qquad (\lambda)$$

$$Pl^{\min} \le |P_{i-k}| = |Pl| = \left| \sum_l A_{i,l} (Pg_i - Pd_i) \right| \le Pl^{\max}, \forall_l \qquad (\mu_l)$$

$$Pg_i^{\min} \le Pg_i \le Pg_i^{\max} \ \forall_i \qquad (\gamma_i^{\min}, \gamma_i^{\max})$$

(2)

where c_i($/MWh) is the energy bid price of generator i ($/MWh), λ is the shadow prices associated with equality constraints, μ_i is the shadow prices associated with the transmission constraint for line l, and $(\gamma_i^{\min}, \gamma_i^{\max})$ is the shadow prices associated with the generation constraints. Pg_i is the output of generator i (MW), Pd_i is the demand level at bus i (MW), and Pl is the power flow in line l (MW). $Pl^{\min}$ and $Pl^{\max}$ are the minimum and maximum limits of power flow in line l, due to stability and thermal constraints. $A_{i,l}$ is the power flow's sensitivity on line l due to injection at bus i.

After solving congestion, we can calculate the standard locational price for location i and time t:

$$\mathrm{LMP}_i = \mathrm{LMP}^{\mathrm{energy}} + \mathrm{LMP}_i^{\mathrm{loss}} + \mathrm{LMP}_i^{\mathrm{cong}} \quad (3)$$

where LMP_i is the LMP at bus i ($/MWh), $\mathrm{LMP}^{\mathrm{energy}}$ is the system's marginal energy price ($/MWh), $\mathrm{LMP}_i^{\mathrm{loss}}$ is the marginal loss price at bus i ($/MWh), and $\mathrm{LMP}_i^{\mathrm{cong}}$ is the marginal congestion price at bus i ($/MWh).

To calculate the last two components in Equation 3, the delivery factor (DF) and generation shift delivery factor (GSDF) are required:

$$\mathrm{LMP}_i^{\mathrm{loss}} = (\mathrm{DF}_i - 1)\,\mathrm{LMP}^{\mathrm{energy}} \quad (4)$$

$$\mathrm{LMP}_i^{\mathrm{cong}} = -\sum_{l=1}^{k} \mathrm{GSDF}_l * \mu_l \quad (5)$$

where DF_i is the delivery factor at bus i; GSDF_l is the generation shift factor at line l, representing the sensitivity of the power flow on line l to a change of net injection at each bus; and k is the set of congested transmission lines.

The penalty factor (PF) associated with any bus on the transmission system is the increase required in injection at that bus to supply an increase in withdrawn at the system reference bus with all other bus net injections held constant.[9] We can calculate the penalty factor for bus i as

$$\mathrm{PF}_i = \frac{1}{\left(1 - \frac{\partial P_{\mathrm{Loss}}}{\partial P_i}\right)} \quad (6)$$

where $\left(-\frac{\partial P_{\mathrm{Loss}}}{\partial P_i}\right)$ is the incremental transmission loss. This is calculated by

$$\frac{\partial P_{\mathrm{Loss}}}{\partial P_i} = \frac{\partial P}{\partial P_i}\left(\sum_{l=1}^{nl} P_l^2 \times R_l\right) \quad (7)$$

We can reformulate Equation 7 as follows:

$$\frac{\partial P_{\mathrm{Loss}}}{\partial P_i} = \frac{\partial}{\partial P_i}\left(\sum_{l=1}^{nl}\left(\sum_{i} A_{i,l} \times P_i\right)^2 \times R_l\right) \quad (8)$$

$$\frac{\partial P_{\mathrm{Loss}}}{\partial P_i} = \frac{\partial}{\partial P_i}\left(2 \times \sum_{l=1}^{nl} A_{i,l} \times \left(\sum_{l=1}^{nl} A_{i,l} \times P_i\right) \times R_l\right) \quad (9)$$

where P_i is the net injection at bus i (MW), P_l is the power flow in line l, nl is the number of lines l, n_i is the number of buses i, and R_l is the line l resistance.

In the marginal loss pricing formulation, the delivery factors are needed in addition to the penalty factors. We calculate the DF of bus i as

$$\mathrm{DF}_i = \left(\frac{1}{\mathrm{PF}_i}\right) = \left(1 - \frac{\partial P_{\mathrm{Loss}}}{\partial P_i}\right) \quad (10)$$

(See related works for more information about the analytical LMP formation.[10,11])

Clustering Algorithms

The clustering process is extremely useful for discovering groups and identifying interesting distributions in the underlying data. Nowadays, companies have a lot of information embedded in huge databases. The extraction of knowledge from these data has spurred a tremendous interest in discovering interesting data distributions and patterns leading to intense activity in the area of knowledge discovery in databases (KDD).[12]

The data mining process involves using algorithms to discover patterns among the data following a similarity criterion. The recognition operations of patterns are based on unsupervised-learning techniques.[13]

Clustering can be considered the most important unsupervised-learning problem. Therefore, similar to every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. The main goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. The decision of how to choose what constitutes a good clustering is not always based on a "best" criterion that is independent of the final aim of the clustering. Consequently, the users must define this criterion in such a way that the result of the clustering process will suit their needs.

The clustering process must group the data set in clusters so that records in the same group are highly similar and differ from objects in other clusters.

Before applying the clustering algorithms, we must determine the number of clusters that should result from the database under study. This information is not known a priori, and there might not be a definitive answer concerning the best number of clusters. For instance, when using the k-means clustering algorithm, we can estimate that k clusters can

be obtained from the database using cross-validation. Data mining involves an analyst searching for useful structures and relations among the data, usually without any strong a priori expectations of what it might find. In practice, the analyst usually does not know in advance how many clusters there might be in the sample. For that reason, some programs implement a cross-validation algorithm to automatically determine the number of clusters in the data. However, the number of clusters can be found by using a distance measure among data points. Thus, the distance measure represents an important component of a clustering algorithm. If the components of the data instance vectors are all in the same physical units, the simple Euclidean distance metric could successfully group similar data instances. However, even in this case, the Euclidean distance can sometimes be misleading; in those cases, others measures can be used, such as the Minkowski metric for higher-dimensional data.[14]

We used two different clustering algorithms to obtain the typical LMP diagrams for the proposed case study: two-step and k-means.

Two-Step clustering algorithm

This clustering method is based on a scalable cluster analysis algorithm designed to handle very large data sets. It can handle continuous and categorical variables or attributes and requires only one data pass. It involves two steps:

1. Precluster the records into many small subclusters.
2. Cluster the subclusters resulting from the precluster step into the desired number of clusters.

The precluster step uses a sequential clustering approach. It scans the data records one by one and decides if the current record should merge with the previously formed clusters or start a new cluster based on the distance criterion. The procedure is implemented by constructing a modified cluster feature tree. The tree consists of levels of nodes, and each node contains a number of entries. A leaf entry represents a final subcluster. If the cluster feature tree grows beyond an allowed maximum size, the cluster feature tree is rebuilt based on the existing cluster feature tree by increasing the threshold distance criterion. The rebuilt cluster feature tree is smaller and hence has space for new input records. This process continues until a complete data pass is finished.[15] All records falling in the same entry can be collectively represented by the entry's cluster feature.

When a new record is added to an entry, the new cluster feature can be computed from this new record and the old cluster feature without knowing the individual records in the entry. The cluster feature tree might depend on the input order of the cases or records. To minimize the order effect, it is necessary to randomly order the cases.

The cluster step takes subclusters resulting from the precluster step as input and then groups them into the desired number of clusters using an agglomerative hierarchical clustering method—in this case, single-linkage clustering. In fact, the distance between one cluster and another cluster is considered equal to the shortest distance from any member of one cluster to any member of the other cluster. (This algorithm is known as the closest neighbor.) In single-link hierarchical clustering, each step merges the two clusters with the two closest members that have the smallest distance. In general, the larger the number of subclusters produced by the precluster step, the more accurate the final result. However, too many subclusters will slow down the clustering in the second step.

A distance measure is needed in both the precluster and cluster steps. Two distance measures are available: the log-likelihood distance measure that can handle both continuous and categorical variables and the Euclidean distance, which is only applied if all variables are continuous.

K-Means algorithm

The k-means algorithm is one of the most common, simplest unsupervised-learning algorithms that solve the well-known clustering problem.[16] The name of the algorithm comes from representing each of the k clusters by the weighted average of its points, called the cluster center. In other words, k-means clustering is a method of cluster analysis that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It requires the initialization of the number of clusters and starts by assigning k cluster centers randomly selected from the pattern set. Then, it proceeds by assigning each pattern from the initial set to the nearest cluster center and recounts the center using the current cluster memberships, until reaching the convergence criterion. This algorithm has the advantage of clear geometrical and statistical meaning, but it only works conveniently with numerical attributes and it is sensible to outliers.

The interactive optimization process should undertake the following steps:

1. Choose the number of clusters k.
2. Randomly generate k clusters and determine the cluster centers, or

directly generate k random points as cluster points.
3. Assign each point to the nearest cluster center, where *nearest* is defined with respect to one of the distance measures.
4. Recompute the new cluster centers.
5. Repeat the two previous steps until some convergence criterion is met (usually that the assignment has not changed).

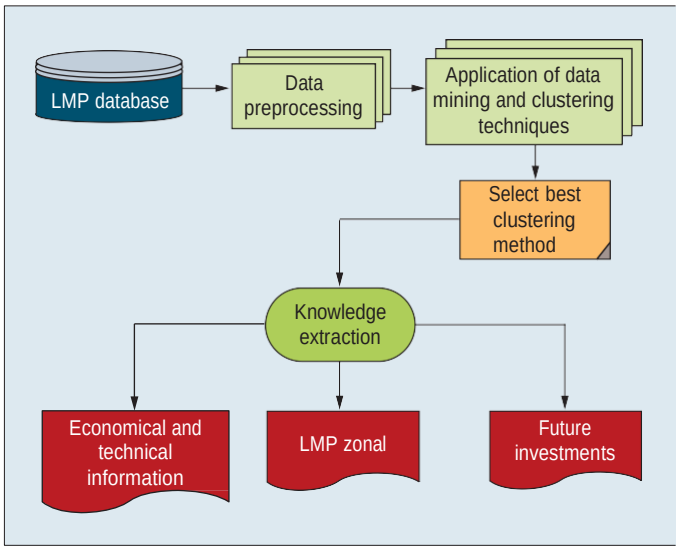

Figure 1. Data-mining methodology. We applied the two-step and k-means clustering algorithms and evaluated their performance using adequacy measures.

The main advantages of this algorithm are its simplicity and speed, which allow it to run on large data sets. The disadvantage is that it does not yield the same result with each run because the resulting clusters depend on the initial assignments—that is, the algorithm is sensitive to the initial partition chosen. If this is not well chosen, the algorithm can converge to a local minimum of the objective function.

To be able to get a good initial partition so that the algorithm converges to the global minimum, different variants of the method may be used. Instead of the random generation of the clusters or the clusters centers, as in step 2, different techniques such as simulated annealing or genetic algorithms can be used to solve the problem of optimizing the initial partition, which prevents the convergence to a local minimum.

LMp clustering

We must choose the clustering method that produces the best data partition. For this work, we used measuring indices to check the quality of the data partition. We then tested two measures of adequacy: the mean index adequacy (MIA) and clustering dispersion indicator (CDI).[17]

We defined the distances according the following equations to assist the formulation of the adequacy measure. The distance between two diagrams is

$$d(li,lj) = \sqrt{\frac{1}{H} \times \sum_{h=1}^{H} (li(h) - lj(h))^2} \qquad (11)$$

where H represents the number of records read and li and lj represent the LMP on bus i and on bus j, respectively.

The distance between a representative diagram and the center of a set of diagrams is

$$d(r^{(k)}, L^{(k)}) = \sqrt{\frac{1}{n^{(k)}} \sum^{n^{(k)}} d^2(r^{(k)}, l^{(m)})} \qquad (12)$$

where $n^{(k)}$ represents the number of diagrams outside cluster k.

Consider a set of M load diagrams separated in k classes with $k = 1, \ldots, K$, where K is the total number of clusters and each class is formed by a subset $C^{(k)}$ of diagrams, and where $r^{(k)}$ is a pattern assigned to cluster k.

The MIA is defined by

$$\mathrm{MIA} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})} \qquad (13)$$

The MIA depends on the average of the mean distances between each pattern assigned to the cluster and its center. The CDI index depends on the distance between the LMP diagrams in the same cluster and (inversely) on the distance between the class representatives diagrams.[17] In this case,

$$\mathrm{CDI} = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{2.n^{(k)}} \sum_{n=1}^{n^{(k)}} d^2(l^{(m)}, C^{(k)}) \right|}}{\sqrt{\frac{1}{2K} \sum_{k=1}^{K} d^2(r^{(k)}, R)}} \qquad (14)$$

R is the set of the class representative diagrams.

Using both distances (Equations 13 and 14), it is possible to define performance measures to evaluate the clustering tools. A good clustering tool can well separate classes of LMP values and assure that those LMP diagrams assigned to the same class are highly similar.

Case Study

To characterize zonal prices and support transmission expansion planning,[18,19] we applied our data-mining-based methodology to the CAISO's 2009 data. Figure 1 presents the methodology that has been implemented.

To group the buses in clusters, suggesting a partition in terms of the similarity of the LMP value, we used two different clustering algorithms

and evaluated their performance using two adequacy measures. Specifically, we measured cluster compactness (MIA) and cluster separation (CDI).[17]

The CAISO real database information we used for this case study is available online from CAISO Market Oasis site (http://oasis.caiso.com). The collected data concerns the day-ahead CAISO market. The transmission network used is formed by 3,458 buses and the LMP for each bus was recorded with a cadence of one hour for all of 2009. For our analysis, we obtained 121,168,320 records. This database includes the values of total LMP, LMP energy, LMP loss, and LMP congest.

Starting from the typical LMP's profile obtained, we were able to extract useful knowledge namely, the identification of zonal price, economical indicators, and zonal congestion management.

Using this 2009 data, our goal was to obtain relevant knowledge about congestion management that might lead to an improvement of the planning network investments as well as identify areas in need of additional power generation plants.

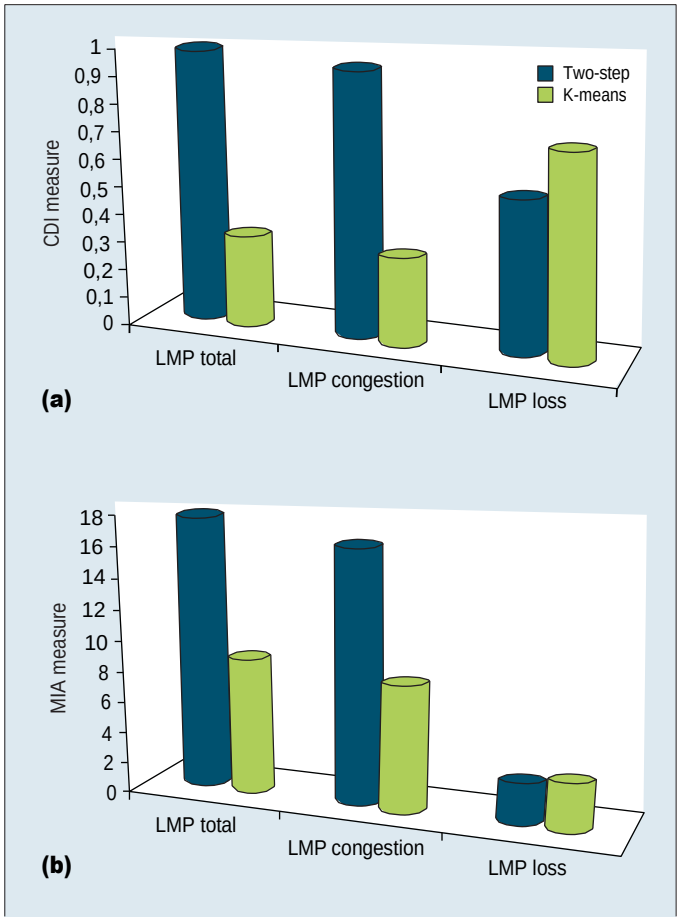

Figure 2. Algorithm performance. (a) Clustering dispersion indicator and (b) mean index adequacy performance.

Data preprocessing

Before proceeding with the use of data mining techniques, we needed to handle the data selected in the previous phase, clean it, and prepare it with the data mining algorithms.[19]

There are always problems with data that require undertaking a previous data-cleaning phase to detect and correct bad data in any data mining process. Initially, there were more than 4,000 buses to be analyzed during the data-cleaning phase, the number of buses was reduced to 3,458 because our study only considered the buses with fields that were completely filled at all hours for the entire year. We discarded all the transmission nodes that did not have all this information available from the sample under study.

clustering

For this study, we tested the two-step and k-means clustering algorithms. The algorithm that produces the smallest MIA and CDI indexes values prevailed in terms of performance of partition. Indeed, the smallest value of MIA indicates more compact clusters.

Figure 2 presents the results obtained for MIA and CDI.

Figure 2 shows that the k-means clustering algorithm yielded a better partition based on total LMP and LMP congestion. A better partition was obtained for LMP loss using the two-step algorithm, but this advantage is not very significant. Based on these results, we chose the k-means algorithm for the clustering procedure.

Before applying the clustering algorithms, we needed to define the number of clusters. By using the measurements indices in Equations 13 and 14, we compared the quality partition for 10, 20, 30, and 40 clusters. Our results show that up to 30 clusters, the two indices were significantly decreased, indicating improvement of partition. We obtained no significant change for more than 30 clusters, so we choose 30 clusters for further testing. (Our preliminary work also supported this choice.[16]) The MIA index decreases significantly when the number of clusters increases until 30. After this, MIA decreases slightly (with the increase of the number of clusters).[16]

Figure 3 gives the LMP energy diagram that resulted from the 2009 CAISO transmission network. Each point reflects the LMP energy value per hour and for each node. Typically, the energy component is defined as the cost to serve the next increment of demand at the specific node, considering that this increment can be produced from the least

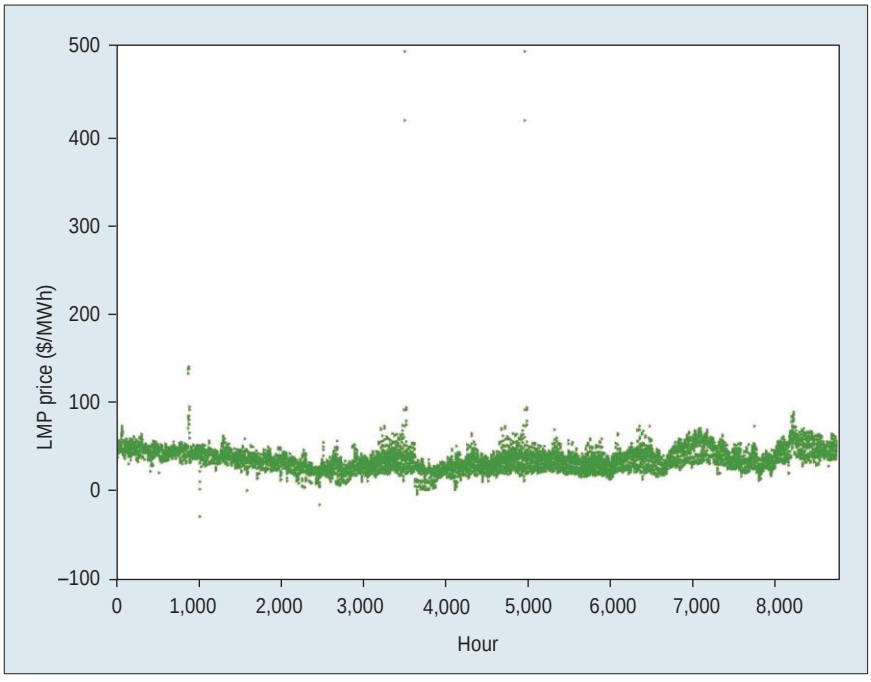

Figure 3. LMP energy diagram. Each point reflects the LMP energy value per hour and for each node.

expensive generating unit in the system that still has available capacity. This means that the LMP energy component reflects the marginal cost of providing energy from a designated reference location. Thus, for a certain hour, all nodes will assume the same LMP energy value.

Figure 4 shows the LMP loss that resulted from applying the k-means clustering algorithm to the CAISO data. Each curve represents a typical annual LMP loss diagram for a certain set of transmission nodes. Several clusters have a negative LMP loss component—namely shadows {2, 3, 5, 7}—meaning that serving additional demand reduces transmission losses. On the contrary, for the clusters located in the {1, 4, 6, 8} shadows,

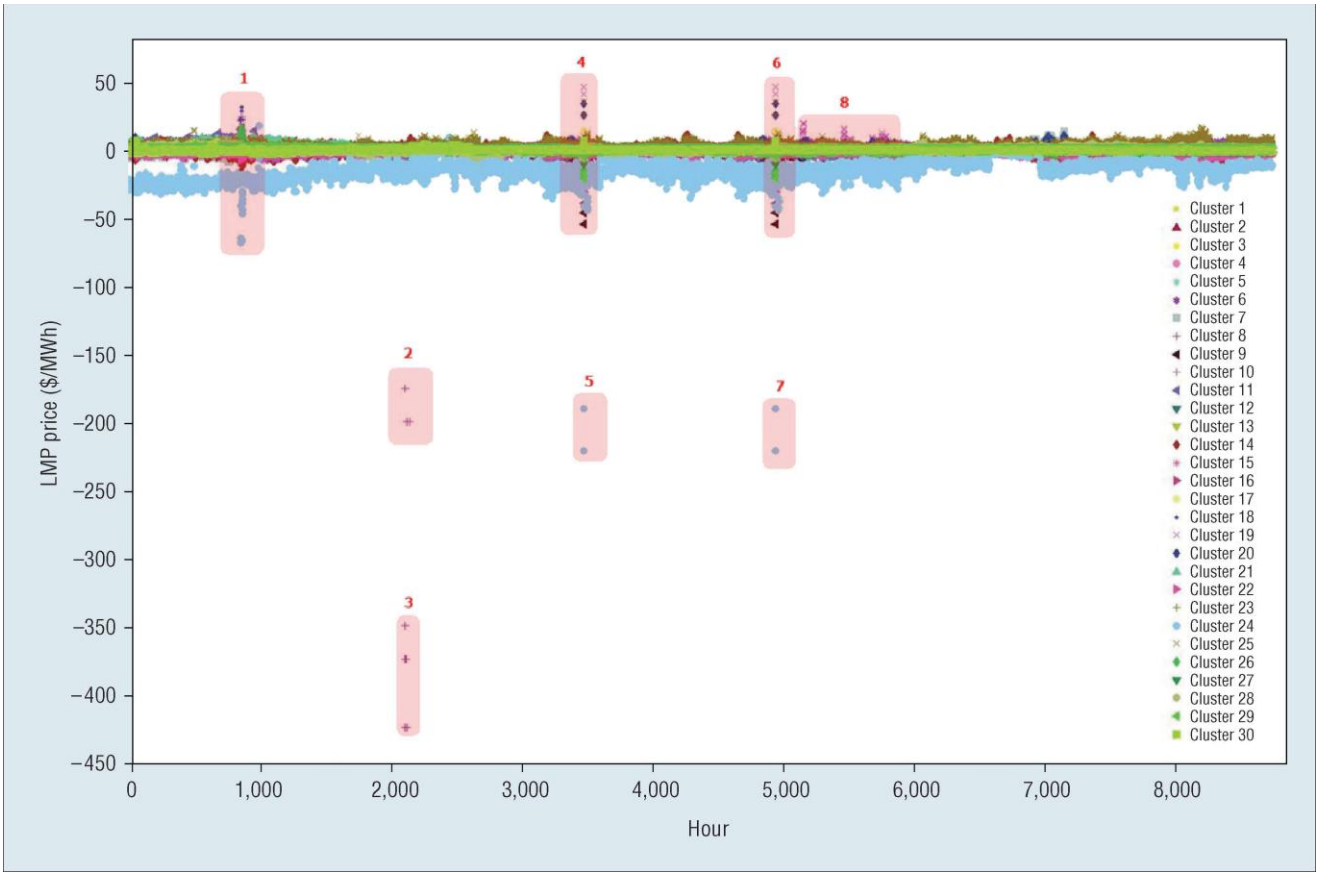

Figure 4. LMP loss diagram. Each curve represents a typical annual LMP loss diagram for a certain set of transmission nodes. Based on these results, we can identify where it would eventually be more profitable to locate distributed generation to reduce LMP loss.

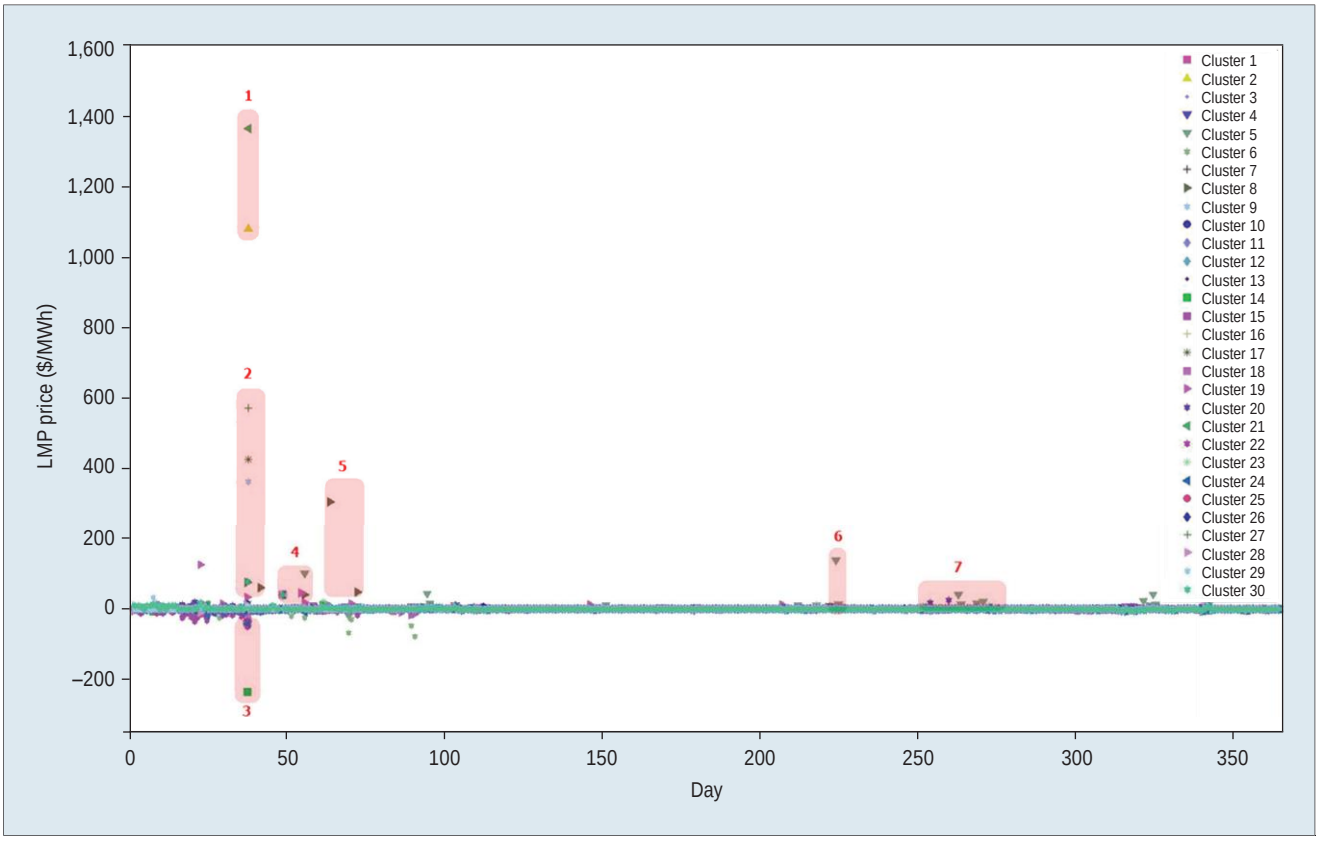

Figure 5. LMP congestion diagram. Positive and negative values for this LMP component reflect what happens, system wide, by serving an additional increment of demand at a specific location from the reference bus.

serving additional demand increases transmission losses, which corresponds to a positive loss component. Based on these results, we can identify where it would eventually be more profitable to locate distributed generation to reduce this LMP component. It is possible to identify in each moment which buses belong to a certain cluster. For instance, cluster 19 (shadow 4 in Figure 4) includes the buses 135, 274, 367, 698, 699, 774, 1218, 1629, and 2131, which are characterized by a higher LMP loss component.

All transmission lines have losses due to the lines impedance, which means that the further the demand is from the supply, the higher will losses be. Therefore, the further demand is from the reference bus, the greater impact will be noticed in the LMP loss component.

The knowledge concerning the LMP loss historical behavior can help identify the most appropriate place to locate distributed production and characterize and quantify the losses.

The LMP congestion component at a certain node accounts for congestion costs, as measured between that node and a reference node. LMP congestion can be a positive or negative value. When serving additional demand increases congestion, LMP congestion will be positive, but if serving additional demand reduces congestion, this LMP component will be a negative value. Finally, if none of the transmission power lines reaches the threshold of their capacity limit, the congestion component will be zero at all system nodes.

Figure 5 represents the LMP congestion diagram. There are positive and negative values for this LMP component that reflect what happens, system wide, by serving an additional increment of demand at a specific location from the reference bus. In the 37th day (shadow 3 in Figure 5), cluster 3, which consists of 104 buses, presents a negative LMP congestion. The buses belonging to the clusters that are included in the shadows {1, 2, 4, 5, 6, 7} reflect that serving additional demand increases congestion.

Knowledge about the impact of LMP congestion in a power transmission network for a whole year might help us identify points where it is desirable to make additional investment by expanding or improving existing power transmission lines. Important information that can be extracted from this study is the identification of the best locations for distributed generation. In fact, distributed generation can help decongest some parts of the network. In this case, it should be

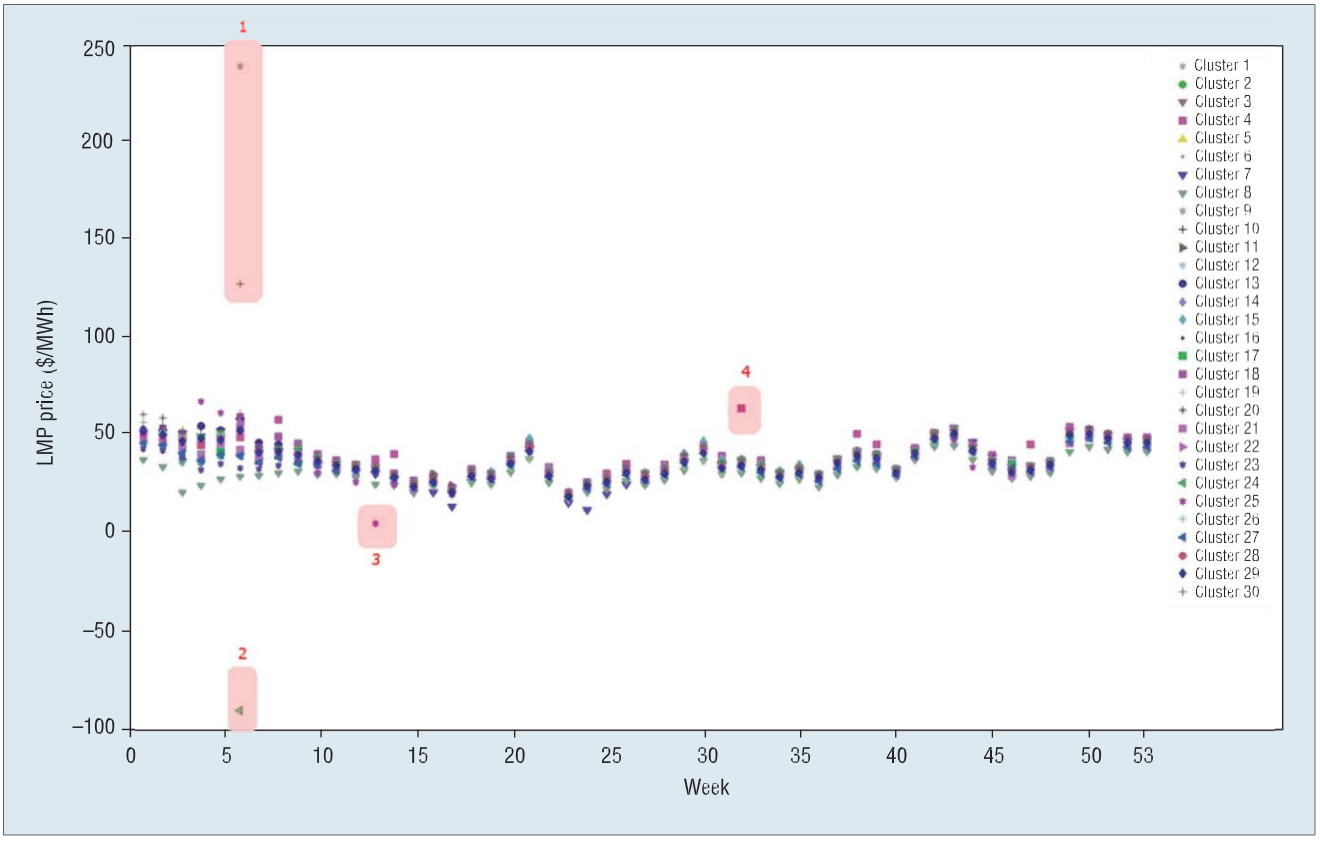

Figure 6. LMP total diagram. LMP total is the sum of the energy, congestion, and loss components.

preferably installed near the nodes that belong to the clusters included in the zones {1, 2, 4, 5, 6, 7} in Figure 5.

The LMP total diagram in Figure 6 reports the general tendency of the LMP value during the year for all clusters. As we explained earlier, the LMP total is formed by the sum of the energy, congestion, and loss components (the last two components may have a negative value). Typically, the LMP total pattern is similar for all buses in the power transmission network. The shadow identified by {1, 4} in Figure 6 represents the positive peaks of LMP total. The shadow illustrated in {2} shows an exceptional case occurred in buses belonging to cluster 24, for which the value of LMP had a deeply negative value. On the other side, between the 10th and the 15th week, cluster 25 presents a LMP value close to zero.

Based on the proposed data min- ing methodology, the knowledge extracted from the LMP historical databases allows the ISO to formulate relevant decisions concerning network planning and investment in electricity generation. In future work, we will compare this methodology with other clustering algorithms. ■

Acknowledgments

We acknowledge the Foundation for Science and Technology (FCT), Knowledge Engineering and Decision Support Center (GECAD), and funding programs for Science, Technology, Innovation and Development with EU participation (FEDER, POCTI, POSI, POCI, POSC, POTDC and COMPETE) for their support of this work.

References

1. G. Rothwell and T. Gómez, *Electricity Economics: Regulation and Deregulation*, IEEE Press, 2003.
2. L. Cao, V. Gorodetsky, and P.A. Mitkas, "Agent Mining: The Synergy of Agents and Data Mining," *IEEE Intelligent Systems*, vol. 24, no. 3, 2009, pp. 64–72.
3. J. Zhang, H. Huang, and J. Wang, "Manifold Learning for Visualizing and Analyzing High-Dimensional Data," *IEEE Intelligent Systems*, vol. 25, no. 4, 2010, pp. 54–61.
4. G. Marreiros et al., "Context-Aware Emotion-Based Model for Group Decision Making," *IEEE Intelligent Systems*, vol. 25, no. 2, 2010, pp. 31–39.
5. J.E. Price, "Market-Based Price Differentials in Zonal and LMP Market Designs," *IEEE Trans. Power Systems*, vol. 22, no. 4, 2007, pp. 1486–1494.
6. J. Ferreira, Z. Vale, and R. Puga, "Nodal Price Simulation in Competitive Electricity Markets," *Proc. 6th Int'l Conf. European Electricity Market* (EEM 09), IEEE Press, 2009, pp. 1–6.
7. J.E. Price, "Market-Based Price Differentials in Zonal and LMP Market Designs," *IEEE Trans. Power Systems*, vol. 22, no. 4, 2007, pp. 1486–1494.

8. L. Liu and A. Zobian, "The Importance of Marginal Loss Pricing in an RTO Environment," *The Electricity J.*, vol. 15, no. 8, 2002, pp. 40–45.
9. F. Li, J. Pan, and H. Chao, "Marginal Loss Calculation in Competitive Electrical Energy Markets," *Proc. 2004 IEEE Int'l Conf. Electric Utility Deregulation, Restructuring and Power Technologies* (DRPT 2004), vol. 1, IEEE Press, 2004, pp. 205–209.
10. M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems: Forecasting, Scheduling and Risk Management*, John Wiley & Sons, 2002.
11. L. Xie, H. Chiang, and S. Li, "Locational Marginal Pricing Under Composite Dynamic Load Models: Formulation and Computation," *Proc. IEEE Power and Energy Soc. General Meeting*, IEEE Press, 2010, pp. 1–8.
12. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Comm. ACM*, vol. 39, no. 11, 1996, pp. 27–34.
13. V. Figueiredo et al., "An Electric Energy Characterization Framework Based on Data Mining Techniques," *IEEE Trans. Power Systems*, vol. 20, no. 2, 2005, pp. 596–602.
14. K.A.J. Doherty, R.G. Adams, and N. Davey, "Non-Euclidean Norms and Data Normalisation," *Proc. European Symp. Artificial Neural Networks* (ESANN 2004), 2004, pp. 181–186; www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es2004-65.pdf.
15. T. Zhang, R. Ramakrishnon, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proc. ACM SIGMOD Conf. Management of Data*, ACM Press, 1996, pp. 103–114.

the AuthorS

16. J. Ferreira et al., "Zonal Prices Analysis Supported by a Data Mining Based Meth- odology" *Proc. IEEE Power and Energy Soc. General Meeting*, IEEE Press, 2010.
17. G. Chicco et al., "Customer Character- ization Options for Improving the Tariff Offer," *IEEE Trans. Power Systems*, vol. 18, no. 1, 2003, pp. 381–387.
18. I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier, 2005.
19. K.J. Cios et al., *Data Mining: A Knowledge Discovery Approach*, Springer, 2007.