ACER
European Union Agency for the Cooperation of Energy Regulators

ACER's Decision on the alternative bidding zone configurations to be considered in the bidding zone review process for the Baltic region

ANNEX II

Description of the clustering algorithms

21 December 2023

1. Introduction to clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data analysis and a common technique for statistical data analysis, used in many fields. For the purpose of defining bidding zone configurations, the set of objects consists of network nodes, each of them represented by a selected feature. There are in principle different variables that can be used as clustering feature, e.g. nodal prices, Power Transfer Distribution Factors (PTDFs), social welfare or shadow prices. For the purpose of this Decision on alternative configurations, using nodal prices as clustering feature was found adequate; it allows to derive indicators that are easy to understand and can be used as proxies for economic efficiency, in line with the objectives of the Electricity Regulation.

Clustering can be achieved by various algorithms that may differ significantly in what constitutes a cluster and how efficiently clusters are determined. Given that all algorithms tested for ACER's Decision on alternative bidding zone configurations use the same feature, i.e. the time series of nodal prices, their shared objective is to minimise price dispersion within each bidding zone. This is in line with the objectives of the Electricity Regulation, as explained in the main document of ACER's Decision.

In the following, the clustering algorithms considered in the analysis are presented and described.

2. The clustering algorithms considered in the analysis

Three clustering algorithms were considered for the purpose of defining bidding zone configurations:

- Constrained k-means clustering;
- Spectral clustering with constrained k-means; and
- Spectral clustering with Constrained Deterministic Iterative Refinement Clustering (CDIRC).

Each of them is further described below.

To avoid that the clustering algorithms could identify extremely small bidding zones, e.g. smaller than a city, which would unlikely be implemented, a threshold that refers to the minimum number of nodes comprised in a 'new' BZ was introduced as a constraint[1].

2.1. Constrained k-means clustering

k-means is a clustering algorithm aiming at partitioning a set of objects into a predefined number of clusters, in which each observation belongs to the cluster with the nearest mean

[1] This minimum threshold depends on the number of bidding zones, as follows: 10% for two BZs, 9% for three BZs, 8% for four BZs and 7% for five BZs. Such a constraint should not have a relevant impact on the delineation of BZs per se.

(cluster centroid). The k-means algorithm minimises the total intra-group variance of the considered feature, according to equation (1) below:

$$\min\{E\} = \min\left\{\sum_{i=1}^{k}\sum_{x\in C_i}\left(d(x,\ z_i)\right)^2\right\} \qquad (1)$$

where z_i is the centroid of cluster C_i and $d(x,\ z_i)$ is the Euclidean distance between the features x and z_i. As, in our application, the time series of nodal prices is used as feature, price dispersion within each newly created bidding zone is minimised. The basic version of the k-means algorithm does not include any connectivity constraints. In order to enforce that a bidding zone is only formed by interconnected nodes, a constrained version of the clustering algorithm has been developed.

The algorithm consists of the following steps:

- Initialisation of the centroids for the selected number of clusters:
  - The first centroid is chosen randomly.
  - The next centroid is determined randomly considering a probability distribution based on the squared distance from the first centroid.
  - The other centroids are generated considering a probability distribution based on the squared distance of each point from its closest centroid.
- Successive calculation of the grouping that minimises the variance for the same centroids, and the centroids that minimise the variance for the same grouping, until stabilisation of the variance.

2.2. Spectral clustering with constrained k-means

Contrarily to k-means, spectral clustering considers the information on the interconnectivity of nodes since the beginning of the clustering procedure. It is based on the construction of a node-to-node similarity matrix in which the similarity is based on the features, with null similarity for disconnected nodes. Therefore, also in this case, nodes with null or very limited price differences are confined in the same cluster.

In order to build the similarity matrix, the Euclidean distance between each pair of nodes is first computed. Those distances are then rescaled in the range [0;1] and collected in the symmetric dissimilarity matrix **D**. The similarity matrix **S** is finally computed according to equation (2) below:

$$\mathbf{S} = \mathbf{1} - \mathbf{D} \qquad (2)$$

Spectral clustering is a two-stage procedure that requires to execute another clustering algorithm in the second stage. In this case, k-means is run as second stage.

The algorithm consists of the following steps:

- Formation of the similarity matrix based on the price differences for each pair of nodes and the node interconnections.
- Computation of the eigenvalues and eigenvectors of the similarity matrix.
- Formation of the input matrix to the second stage, composed of a number of eigenvectors (that correspond to the minimum eigenvalues) equal to the selected number of clusters.
- Execution of the k-means clustering algorithm.

2.3. Spectral clustering with CDIRC

This algorithm follows the same procedure as the one above, with the only difference that in the second stage, CDIRC instead of k-means is run.

CDIRC adopts a deterministic iterative refinement clustering (DIRC) procedure to form the clusters, integrated with the check that all nodes in each cluster are interconnected.

The algorithm consists of the following steps:

- Initialisation of the centroids for the selected number of clusters:
    - The first two centroids are determined by calculating the pair of nodes with maximum distance between the features.
    - The other centroids are found one by one by calculating the maximum distance to the previously determined centroids.
    - All nodes are assigned to the cluster with the closest centroid, with update of the centroid.
- Iterative refinement:
    - Iterative process in which all nodes can change cluster if the closest centroid is different from the current one, with update of the centroids, until stabilisation.