PROBABILISTIC METHODS IN POWER DISTRIBUTION ELECTRICAL NETWORKS

João Nuno TAVARES* (jntavar@fc.up.pt), João Pedro PEDROSO** (jpp@fc.up.pt), Nilson JUNE* (nilsonjune@fc.up.pt), Pedro Miguel CRUZ* (up200506513@fc.up.pt), Sónia GOUVEIA# (sonia.gouveia@ua.pt),
Ana Lopes+ (ana.lopes@edp.pt), Miguel Freitas+ (miguel.freitas@edp.pt), Ricardo Prata+ (ricardo.prata@edp.pt), Luís Oliveira+ (luis.oliveira@edp.pt)

* Faculdade de Ciências and Centro de Matemática da Universidade do Porto (CMUP), Portugal; ** Faculdade de Ciências and INESC TEC, Universidade do Porto, Portugal; # Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Center for R&D in Mathematics and Applications (CIDMA), Universidade de Aveiro, Portugal; + EDP distribuição, Portugal

ABSTRACT

Energy distribution networks require customized and sophisticated methodologies for their analysis and simulation. In particular, the characterization of their probabilistic behaviour is mandatory for a realistic planning aiming network development or optimization. This paper introduces a framework for the analysis of distribution networks and the characterization of their stochastic behaviour, developed in the scope of a successful R&D collaboration between EDP Distribuição and Universidade do Porto, Portugal. The paper includes the description of the developed framework and methods, implementation strategies and illustrative results, obtained from real data of power consumption in secondary substations distributed over Portugal.

INTRODUCTION

The distribution network of the global energy company EDP Distribuição (EDP is the acronym for Energias de Portugal) has several geographically distributed nodes. These include supply nodes (e.g. hydro-electric power plants, wind energy, photovoltaic energy and solar panels), private producers) and consumption nodes (associated with domestic, companies, services and others consumers). As expected, this distributed network has a strong stochastic component that requires new and sophisticated methodologies for its analysis and simulation. If the network/system can be considered in (quasi) stationary state, as happens when transient regimes occur on small time scales when compared to the characteristic scale of observation, one can think of calculating the state of the network at the t-snapshot, by solving a system of algebraic equations. However, for network planning, it is imperative to calculate the state of the system response for all possible combinations of the inputs, which is obviously prohibitive. One way to overcome this shortcoming is to characterize the probabilistic behaviour of the network. This can be achieved by associating a time dependent probability density function (pdf) of the power consumption to each distribution node, that reflects all possible states of the node and their probability of occurrence as a function of time. Then, with the information in all nodes, it is possible to infer about the network state in any spatial location but, most importantly, it is possible to carry out simulation studies to assess the impact of changing the network configuration associated with hypothetical production and consumption scenarios.

This paper presents a framework for the analysis of energy distribution networks, developed in the scope of an ongoing successful R&D collaboration between EDP Distribuição and Universidade do Porto, Portugal. The proposed method begins by assuming that a time series of power consumption is associated to each node of the network, specified by its spatial coordinates. The variability of the power consumption at each node (at a given time instant) is then characterized by a power consumption pdf which is obtained from similar temporal consumption behaviours (obtained automatically from data driven clustering analysis). In addition to the network nodes, auxiliary nodes are introduced in order to calculate the aggregated pdf, i.e. the pdf of the algebraic sum of the power consumptions associated with the network neighbour nodes (random variables). The resulting aggregated pdf allows to answer questions related to network risk disruption, and network (re)dimensioning when new classified nodes are introduced. E.g. it enables to estimate the risk associated to different network development strategies, resulting from different production and consumption scenarios (which reflect different energy policies).

While the developed framework is still being tested in power consumption networks with a realistic dimension, future work includes to model the consumption and the production in the low voltage networks in order to integrate the behaviour of the low voltage network in the global analysis of low, medium and high voltage networks.

EXPERIMENTAL DATA

The experimental data used in this work consists of discrete time series of power consumption (kVA) during one year of monitoring (366 days). The time series have a 15-minutes temporal resolution and therefore, the daily period is composed of 96 observations while the annual period has 35.136 observations. Figure 1 displays 3 examples of daily power consumption, showing the large variability of daily patterns with respect to temporal evolution along the day as well as overall average and standard deviation of the power consumption.

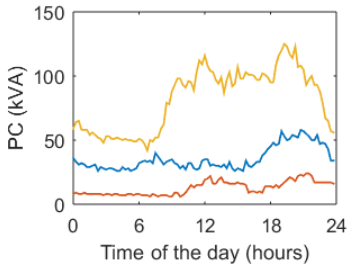

Fig.1 Illustrative example of 3 days of power consumption, randomly selected from the available data.

The data was acquired by EDP Distribuição smartgrid devices, namely through the data concentrators installed at the Secondary Substations (SS). The study focused the universe subset of 42.673 SS i.e. the network nodes spatially distributed over the Portuguese territory.

METHODOLOGIES

The general outline of the developed framework and the interconnection between its functionalities is outlined in Figure 2. In this diagram, the functionalities are represented by blocks and the connectors characterize the execution order and the input / output relation between procedures.

The *first procedure* is data management and storage, which includes the creation of the database with the power consumption data, the preprocessing of the data (e.g. interpolation of missing data for SS with a minimum number of annual observations) and daily segmentation.

The variability of the power consumption at a given node and at a given time was assessed by Clustering analysis, the *second procedure*. Here, the hierarchical clustering process resulted in daily clusters whose data allowed to obtain centroids (i.e. the average pattern of daily consumption in that cluster) and to build a pdf of the power consumption for each time instant of the day.

The daily clusters allowed to estimate the daily pattern of a new SS or of an SS that was not completed by interpolation. This constituted the *third procedure* which was based on classification analysis of the general characteristics of the new SS (e.g. "Municipality" and "Nominal Voltage in kV"). At the end of the previously described procedures, there is one empirical pdf of power consumption associated with each cluster and at each time instant.

Finally, the *fourth procedure* consisted of the aggregation step. Here, the pdf on each aggregation node of the network is obtained by convolution of the pdfs associated with the corresponding leafs of that node. The aggregation operation assumes independence and finite moments of the empirical pdf's (1st and 2nd order), and it is efficiently implemented via Fast Fourier Transform.

Data management and storage

Due to the enormous size of the data that had to be processed, a careful implementation of the databases was required. The first step was to process the raw files obtained by the company, which occupy about 300 GB for the one-year period under analysis. The information contained therein is used to fill a database, aiming at allowing quick access to the data of each SS and each day of the year via a SQL interface. The second step, consisted on dealing with missing data. Missing values (which often arise, e.g. due to communication issues) have been filled using linear interpolation for the time series exhibiting at least a minimum number of annual observations. Finally, each yearly series was divided into segments of 24 hours each (i.e. 96 observations). These preprocessed data have been used in the remainder of this work.

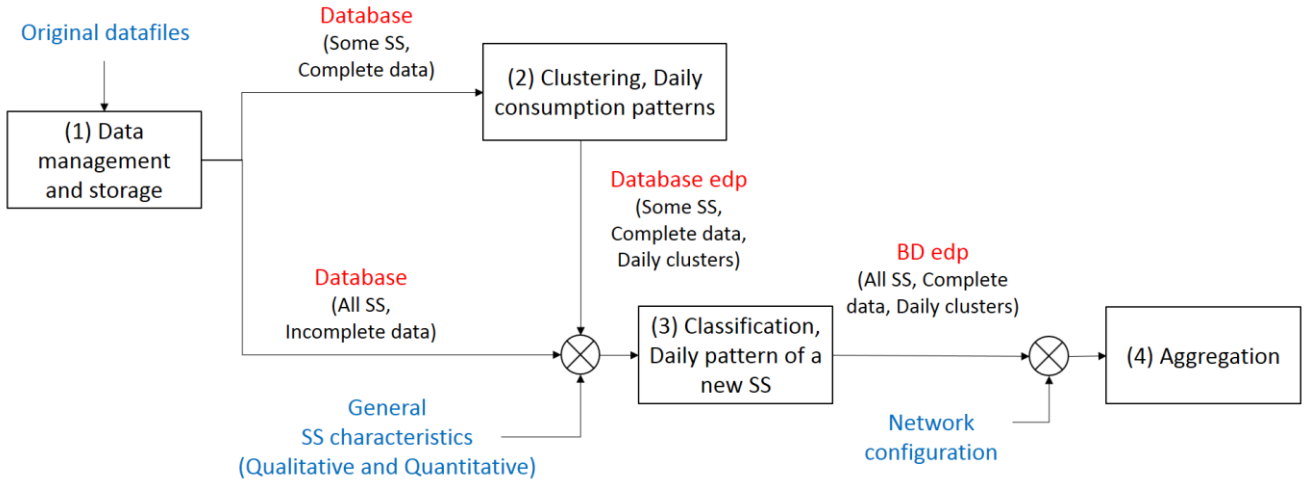

Fig.2 General structure of the framework and interconnection between the procedures developed in this project. Each block represents a feature and the main entries of each block are highlighted in blue.

Daily consumption patterns by clustering analysis

The daily consumption patterns were obtained in two steps. First, each daily segment of the original database, represented by a vector $x = (x_1, \dots, x_{96})$ was normalized according to $x' = (x - m(x))/s(x)$, where $m(x)$ and $s(x)$ are, respectively, the mean and the standard deviation of x. As illustrated in Figure 3, the normalization highlights the shape of the daily pattern while reducing the differences between segments due to differences in constant daily mean and daily variability.

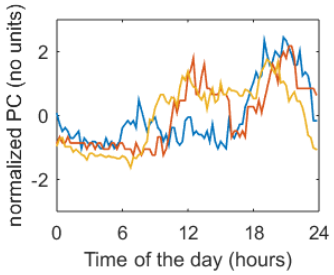

Fig.3 Standardized daily power consumption. Same data as in Figure 1.

Second, the set of normalized daily segments was then subjected to an agglomerative hierarchical clustering procedure by Ward's method [1]. In this work, we considered an efficient implementation of this procedure by a two-level process, in order to avoid the computation of a similarity matrix of large dimensions. At level 1, the clustering procedure was applied to each day of the year. The similarity between daily segments was measured by the sum of squared errors between the 96 observations. In this analysis, 50 clusters were produced for each of the 366 available days, in a total of 18.300 clusters obtained at level 1. Then, level 2 considered a similar clustering procedure but based on the centroids of the 18.300 clusters obtained at level 1. The centroid for each cluster was obtained as the average of the normalized segments within that cluster. Finally, the similarity between daily centroids was measured by the sum of squared errors between the 96 observations weighted by the number of original daily segments associated with each centroid.

The hierarchical clustering procedure allows to obtain a dendrogram, i.e. a tree diagram that illustrates the hierarchical organization of the several daily centroids. Then one has to choose the optimal number of final clusters, K. Despite the diversity of clustering algorithms, intuitively there are two important characteristics to consider for the choice of optimal K: *compactness*, which expresses how similar the cluster elements are to each other, and *separability*, which evaluates how distinct the clusters are from each other. As high compactness and high separability is desirable, a good clustering is characterized by small intra-cluster distances and by large inter-clusters distances. An accomplishment of these objectives is embodied in the Davies-Bouldin, Silhouette and Calinski-Harabaz indexes, which can be used simultaneously to choose the optimal number of final clusters K. Figure 4 shows the three indexes as a function of the number of clusters. Note that optimal K in Davies-Bouldin index minimizes the index whereas the optimal K maximizes the value for the remaining indexes. Therefore, $K = 7$ was considered as an optimal number and the final clusters were labelled as {a,b,c,d,e,f,g}.

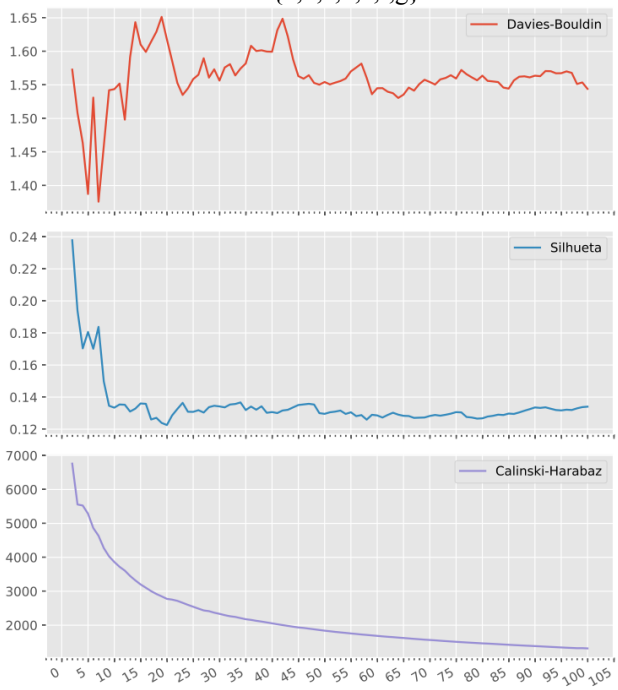

Fig.4 Clustering quality indexes: Davies-Bouldin, Silhouette and Calinski-Harabaz indexes as a function of the final number of clusters.

Figure 5 presents the results of the clustering procedure, showing 4 daily patterns out of the final $K = 7$. As expected, the daily patterns within each cluster are quite similar whereas there is a visible difference between the average daily patterns depicted from each cluster.

Daily consumption patterns of a SS with missing data by classification analysis

The original data included SS without the minimum number of annual observations and these time series were not considered in the clustering procedure. Therefore, the daily consumption patterns associated with those SS were not obtained by the previously described clustering procedure. Instead, the daily patterns for those SS were obtained by an inverse classification scheme.

This inverse classification procedure considered both qualitative (class labels) and quantitative characteristics of an SS to obtain the most likely daily cluster (from a to g) associated with that SS. Class labels included "District" and "Municipality" where the SS is installed, as well as "Nominal Voltage in kW", among other qualitative characteristics. On the other hand, the quantitative

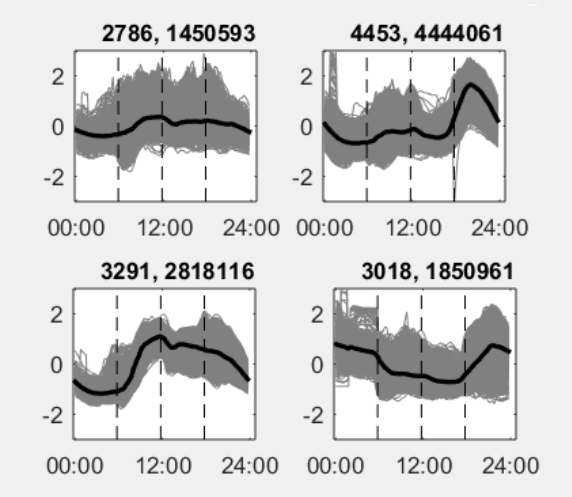

Fig.5 Level 2 clustering results showing 4 clusters out of 7 final clusters. The daily consumption centroids obtained at level 1 are represented in grey and the level 2 centroid associated to each final cluster is represented in black. The numbers on top represent, respectively, the number of level 1 centroids and of original segments in each cluster.

characteristics included "Installed power (kVA)" and "Number of single phase consumers", in a total of s quantitative characteristics which were assembled into a vector $c \in R^s$. The most likely daily cluster associated with a new SS (from the set of final clusters {a,b,c,d,e,f,g}, see previous section) was obtained from the distribution of daily clusters produced with the k=50 nearest SS (by evaluating the Euclidean distance in R^s) from the set of SS with the same class labels as the new SS. The optimal number of 50 nearest SS was obtained after a sensitivity analysis on the results by changing the number of nearest SS and observing that the most likely daily cluster did not change (see Figure 6 for an illustrative example).

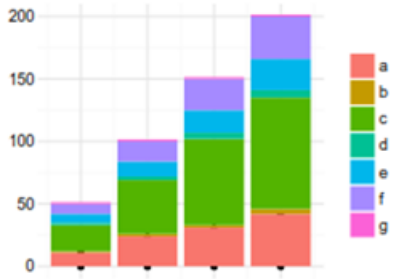

Fig.6 Distribution of the number of nearest neighbours per daily cluster for several values of the nearest neighbours k = 50, 100, 150, 200. Illustrative example for one SS at a given day, randomly selected from the available data.

Probability density function at aggregation nodes

For each final cluster and time instant, we cross-cut the segments included in the cluster, record the obtained values and produce a normalized histogram of power consumption. This histogram is the instantaneous (empirical) pdf at that time instant, associated with that final cluster, as those represented in Figure 7.

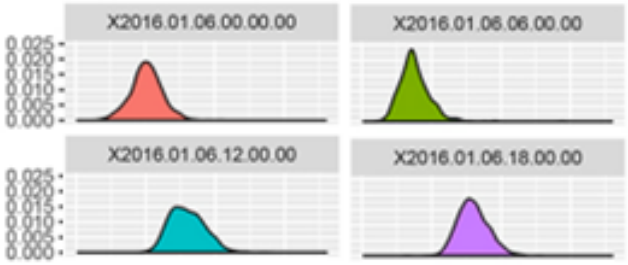

Fig.7 Illustrative example of instantaneous pdfs from one cluster on January 6, 2016 at 00:00, 06:00, 12:00 and 18:00. Illustrative example randomly selected from the available data.

The aggregation procedure is based on the network configuration. Let's consider a network with a tree topology, whose root is a Primary Substation (r = PS), and whose leaf nodes are either generator nodes or charge nodes. The power consumption pattern at each node (generator or load) is described, at each day, by a word of the alphabet A = {a, b, c, d, e, f, g}, where each letter represents a normalized cluster. For a given day, each leaf node of the tree (whose root is r = PS) has a certain SS associated. Then, the following procedure is applied in three steps.

First, the normalized pdf at each node is converted back in kVA, by multiplication per $s(x)$ and by sum per $m(x)$, being $s(x)$ and $m(x)$ respectively the original standard deviation and mean of the power consumption associated with that SS at a given day.

Second, the pdfs returned by denormalization are then defined at regular intervals, where the number of regular intervals depend on the scaling factors and, thus, the number of regular intervals is not necessarily the same for all SS.

Third, the convolution of the pdfs is computed. In this stage, we considered 4 different algorithms for the implementation of the convolution, and we choose for the fastest algorithm, which involves Fast Fourier Transform operations and products. The values of the convolutionable pdfs must be defined in a regular grid, common to the various pdfs to be convoluted. Therefore, each denormalized pdf is resampled by linear interpolation to a grid of thickness $\delta > 0$, a user-defined input parameter of the algorithm. The sensitivity analysis of the results according to the value of δ indicated that $\delta = 0.1$ allowed to achieve good results in terms of precision and performance. Finally, the denormalized and resampled pdfs were convoluted by Fast Fourier Transform.

Figure 8 illustrates the aggregation process for a network toy model of small size. Once a maximal tree whose root r is a PS is rebuilt, the aggregation operation in the aggregator nodes (i.e. nodes that have more than one child

node) is accomplished by convolution of the snapshots associated to each child node. More concretely, the operation starts from the leaf-nodes of the tree, which are generator / load nodes with their respective snapshots, and raise a level for parents who have more than one leaf node as a child. Each of these parents is an aggregator node whose instantaneous pdf is obtained by the convolution of the instantaneous (denormalized) pdfs of the corresponding children. The aggregation proceeds in this systematic way, rising from level to level in the hierarchy of nodes, until reaching the last aggregator node (i.e. the node r = PS). As expected, the higher is the number of parcels in the convolution, the higher is the similarity of the aggregated pdf with respect to a normal distribution (consequence of the Central Limit Theorem).

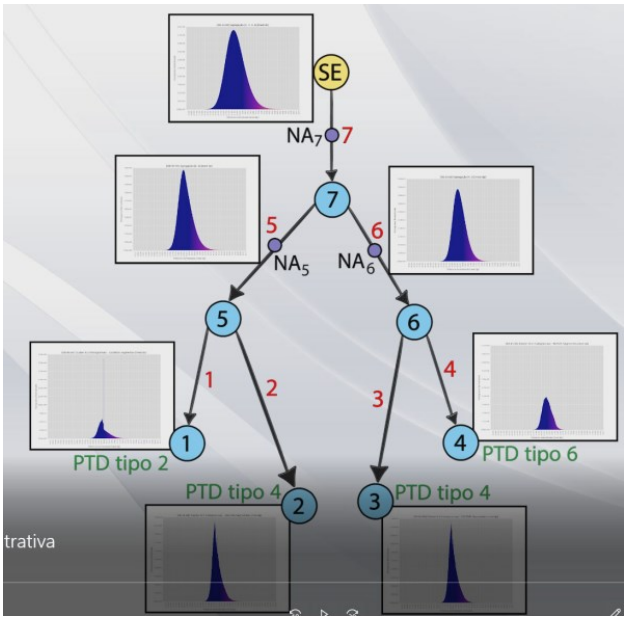

Fig.8 A snapshot of a network toy model with 4 generator/load nodes and 3 aggregation nodes, including the root r = SE = PS.

It should be emphasized that the above described process is only valid when assuming two essential hypotheses. First, the random variables associated to each node-leaf are independent, which allows to calculate the pdf of their sum (aggregation) through the convolution of the nodes-leaf pdfs. The same hypothesis is assumed in the aggregations of "higher" level. Second, the network topology is a tree (that is, there are no cycles) which ensures the uniqueness of the aggregations. In fact, it is possible to prove that the random variables associated to each node of the tree, other than leaf node, assembled in the random vector X, is a linear combination of the random variables associated with leaf nodes, assembled in random vector b, given that X is solution of the system AX = b, where A is the reduced array of node-arc incidence of the tree in question.

CONCLUSIONS

The state of a power network is usually simulated in a deterministic manner, i.e. by setting a constant (and hypothetical) value for the power consumption at a given node and at a given time instant, and by observing the resulting (constant) state in different spatial points of the network. In this work, we describe a methodology that allows to characterize the probability density function for the power consumption, in different nodes of a spatial network and at a given time instant. Therefore, this approach is able to provide the distribution of possible power consumption values besides the constant value of the traditional simulations strategies. This constitutes a data driven approach whose final aim is to provide a support tool for network planning, e.g. allowing the probabilistic simulation of future expansion scenarios for the power network.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Engº Ribeiro da Silva for all the enthusiastic support he gave to this project, and for all his lucid explanations.

FUNDING

This work was supported by EDP Distribuição, Portugal and partially supported by Portuguese funds through the Portuguese Foundation for Science and Technology (FCT), within the following projects: CMUP/UP (UID/MAT/00144/2019, Centro de Matemática da Universidade do Porto, https://cmup.fc.up.pt/), INESC TEC (Institute for Systems and Computer Engineering, Technology and Science, https://www.inesctec.pt), IEETA/UA (UID/EEI/UI0127/2019, Instituto de Engenharia Electrónica e Informática de Aveiro, www.ieeta.pt) and CIDMA/UA (UID/MAT/04106/2019, Centro de I&D em Matemática e Aplicações, www.cidma.mat.ua.pt). S Gouveia acknowledges the individual postdoctoral grant by FCT (ref. SFRH/BPD/87037/2012).

REFERENCES

[1] Murtagh and Legendre, 2014, " Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?", *Journal of Classification* 31, 274-295.

[2] Charu C. Aggarwal, 2015, *Data Mining: The Textbook*, Springer; 2015 ed. Edition.

[3] Charu C. Aggarwal (Editor), Chandan K. Reddy (Editor), 2013, *Data Clustering: Algorithms and Applications*, Chapman and Hall/CRC