Processing and Visualization of Large Data Sets for Power System Stability Analysis

1st Ferriol Falip Torras
Facultat d'Informàtica de Barcelona
Universitat Politècnica de Catalunya
Barcelona, Spain
ferriol.falip@estudiantat.upc.edu

2nd Francesca Rossi
Departament d'Enginyeria Elèctrica
Universitat Politècnica de Catalunya
Barcelona, Spain
francesca.rossi@upc.edu

3rd Alexandre i Gracia Calvo
Departament d'Enginyeria Elèctrica
Universitat Politècnica de Catalunya
Barcelona, Spain
alexandre.gracia@upc.edu

4th Eduardo Prieto Araujo
Departament d'Enginyeria Elèctrica
Universitat Politècnica de Catalunya
Barcelona, Spain
eduardo.prieto-araujo@upc.edu

***Index Terms*—Data Visualization, Power System Stability, Machine Learning**

I. Introduction

The digitalization and monitoring of processes have led to the existence of large databases that systematically and continuously collect information. Sectors as diverse as e-commerce, healthcare, or transportation are just a few examples of the central role that data play in today's society.

In this context, data visualization emerges as a key tool for understanding and extracting useful knowledge from these large volumes of information. Representing data visually facilitates the identification of patterns, trends, and relationships that would be difficult to detect through conventional numerical or statistical analysis.

The electricity sector is no exception to this transformation. The digitalization and monitoring of energy generation, transmission, and distribution systems have enabled the collection of large amounts of data during grid operation. However, power systems are typically operated conservatively, prioritizing system security and stability under potentially risky situations. Consequently, failure or instability events are relatively rare, making it difficult to obtain sufficient real-world data to analyze system behavior in detail under such scenarios. To address this limitation, a common strategy is to generate synthetic data through simulations that reproduce diverse operating points and explore system responses under adverse conditions.

Within this context, this paper focuses on a database containing an extensive set of operating points of the power grid, randomly sampled within its feasible operating space. For each point, system stability is assessed, resulting in a large and diverse dataset that enables the exploration of different visualization methods with the aim of extracting relevant insights about system stability.

The main objective of this work is to apply data visualization and analysis techniques to explore and identify patterns that may be significant. Through this approach, the aim is to determine which factors have the most relevant influence on grid stability, in order to better understand system behavior and provide a solid foundation for future research directions or potential improvements in grid operation and management.

II. Methodology

The main objective of this work is to apply data visualization and analysis techniques to explore and identify patterns that may prove significant. Through this approach, the aim is to determine which factors exert the most relevant influence on grid stability, in order to better understand system behavior and provide a solid basis for future research directions or potential improvements in grid operation and management.

The main challenge arises from the size of the dataset under study, which includes many variables and a large number of samples. The task, therefore, consists of extracting meaningful insights from the dataset using existing visualization and data analysis techniques.

This section describes the methodological pipeline followed to analyze the stability dataset and evaluate the suitability of different machine-learning techniques for extracting meaningful patterns. The overall process consists of four stages: data preprocessing, dimensionality reduction, clustering, and visualization.

A. Data Preprocessing

The dataset contains a large number of operating points of the electrical network, each described by multiple numerical and categorical variables, along with a binary stability label. All numerical variables were standardized to zero mean and unit variance:

$$x' = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ denote the feature mean and standard deviation, respectively. Standardization is crucial to ensure that all variables contribute equally to the subsequent learning algorithms.

Categorical attributes, such as generator control modes, were encoded numerically when required. No missing values were present, thus no imputation procedures were applied.

B. Dimensionality Reduction

Given the high dimensionality of the dataset, three dimensionality-reduction techniques were used to obtain low-dimensional representations suitable for pattern exploration and visualization.

1) Principal Component Analysis (PCA): PCA is a linear technique that computes orthogonal directions (principal components) maximizing the variance of the projected data [1]. The transformation is given by:

$$Z = XW, \tag{2}$$

where X is the standardized data matrix and W the matrix of eigenvectors of the covariance matrix $X^\top X$. PCA provides a baseline representation and offers insight into the dominant modes of variation in the dataset.

2) t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is a nonlinear embedding method designed to preserve local neighborhood relationships [2]. It models pairwise similarities in the high-dimensional and low-dimensional spaces using conditional probability distributions and minimizes their Kullback–Leibler (KL) divergence:

$$\mathrm{KL}(P \,\|\, Q) = \sum_{i \neq j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right). \tag{3}$$

Although computationally intensive, t-SNE frequently reveals meaningful cluster structures hidden in high-dimensional data.

3) Uniform Manifold Approximation and Projection (UMAP): UMAP is a nonlinear method based on Riemannian geometry and topological data analysis [3]. The algorithm constructs a weighted graph of nearest neighbors and optimizes a low-dimensional representation by minimizing a cross-entropy objective:

$$\mathcal{L}_{\text{UMAP}} = \sum_{(i,j)} \left[w_{ij} \log(\sigma(d_{ij})) + (1 - w_{ij}) \log(1 - \sigma(d_{ij})) \right], \tag{4}$$

where w_{ij} are graph weights, d_{ij} is the low-dimensional distance, and $\sigma(\cdot)$ is a differentiable approximation of a step function. UMAP is computationally efficient and preserves both local and global structure more effectively than t-SNE.

C. Clustering Methods

After reducing dimensionality, clustering algorithms were applied to evaluate whether the dataset exhibits natural groupings that correspond to stability properties or other operational regimes.

1) k-Means Clustering: k-Means is a centroid-based clustering algorithm that partitions data into k clusters by minimizing the within-cluster sum of squared distances [4]:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2, \tag{5}$$

where C_i denotes the i-th cluster and μ_i its centroid. Several values of k were tested to analyze cluster granularity. Although simple and fast, k-means assumes spherical cluster shapes and may struggle with irregular structures.

2) DBSCAN: DBSCAN is a density-based clustering method that groups points closely packed together while marking isolated points as noise [5]. Two parameters must be selected: the neighborhood radius ε and the minimum number of neighbors minPts. A point p is a core point if:

$$|\{q : \|p - q\| \leq \varepsilon\}| \geq \text{minPts}. \tag{6}$$

DBSCAN is well suited for identifying arbitrarily shaped clusters and is robust to noise, making it appropriate for datasets with potentially complex geometric structures.

D. Visualization and Evaluation

Low-dimensional embeddings from PCA, t-SNE, and UMAP were visualized in two-dimensional scatter plots, with each point colored according to its stability label. These visualizations allow a qualitative assessment of whether the embedding techniques reveal separability between stable and unstable operating points.

Clustering results (from k-means and DBSCAN) were projected onto the same embeddings to examine whether algorithmic clusters align with meaningful system behaviors. Although the analysis is unsupervised, this qualitative evaluation provides insight into the structure of the dataset and the ability of the methods to capture relevant operational patterns of the electrical network.

III. Case Study

The network used to generate the dataset for this project is the NREL-118 system [6]. Figure 1 shows a schematic of the network.

This system consists of 118 buses divided into 3 regions and 53 generators that can operate with different energy sources. Specifically:

- **Synchronous Generators (SG)**: Corresponding to conventional and hydropower units.
- **Converter-Interfaced Generators (CIG)**: Corresponding to solar and wind power units.

Of the 53 generators, 18 are equipped with CIG technology. Furthermore, CIG units can operate under two different control modes, *Grid-following* (GFOL) or *Grid-forming* (GFOR). Although their specific operation will not be detailed here, the presence of one mode or the other affects grid stability differently and is therefore an important parameter to consider.

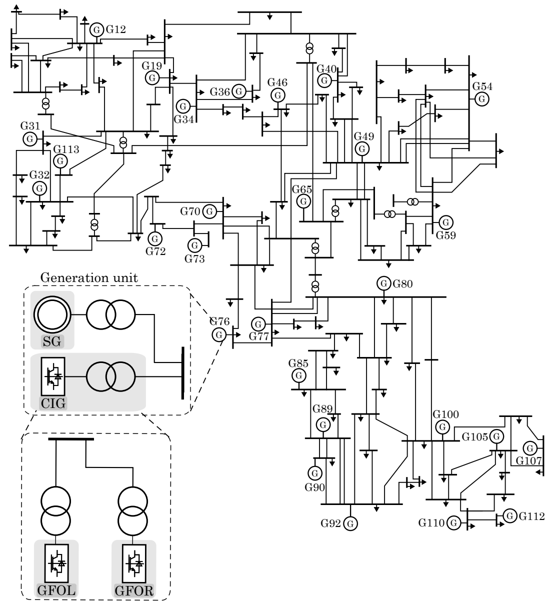

Fig. 1. NREL-118 Network

References

[1] I. T. Jolliffe and J. Cadima, *Principal Component Analysis*. New York, NY: Springer, 2 ed., 2016.

[2] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," in *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS)*, pp. 257–264, 2008.

[3] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[4] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 226–231, 1996.

[6] I. Pena, C. B. Martinez-Anido, and B.-M. Hodge, "An extended ieee 118-bus test system with high renewable penetration," *IEEE Transactions on Power Systems 33.1*, 2017.