Date of presentation July 20, 2021.

Smart monitoring and unsupervised analysis on Electrical Power Grid Network's operational variables

Andrés García Domínguez[1]

[1]Instituto Católico de las Artes e Industrias, Comillas Pontifical University, Madrid 28015 ESP

Corresponding author: Andrés García Domínguez (e-mail: Andres_garcia@alu.comillas.edu).

This work was supported through a 2020 – 2021 Research Scholarship within the Connected Industry Chair in Comillas Pontifical University.

ABSTRACT In collaboration with the Connected Industry Chair, this project starts with the operational variables from the Transformer Centers in the city of Malaga to develop an analytical tool for value-added information extraction on data aggregation. The main outcome is the development of an online dashboard application based on the Plotly framework, enabling cross-platform visualization and performing deep analysis by means of dimensionality reduction techniques into a reduced latent space, then followed by a subsequent clustering prediction.

INDEX TERMS *Dash Application, Dimensionality Reduction, Outlier Detection*

I. INTRODUCTION

Nowadays, there is a growing interest in developing **advanced control systems** over the electricity power grid that can streamline maintenance and operation of critical infrastructures in favor of the well-being of a country. As electricity demand grows steadily, the complexity of the network increases in scale and nature, and interoperation between assets becomes critical, efficiency and required capacity of the control systems that underlie the architecture become a limiting factor.

Among aspects that cause interest in the revolution that is taking place in the field, is the control and supervision of the system and, specifically, the responsible one for the **monitoring, control and protection** of the substations. Traditionally isolated functions are being merged to the point of being based on a smaller number of devices, but forming a system enhanced by probabilistic models, the laws of inference on latent states from measurable variables or online visualizations to perform descriptive, predictive analysis and ultimately, prescriptive.

The focus on this type of model is on the rise, which is leading to the creation of more and more robust models, as well as real-time processing applications for the inspection and construction of alarm systems around them.

This type of intelligent descriptive algorithms are of particular importance for asset analysis in other methodologies such as vibration testing, non-invasive inspection or infrared temperature [3], are today being used for defect detection in transformer substations with great results and their use in asset maintenance and wear prevention, and ultimately failure, is very prevalent.

Power transformers are at the **core of electrical distribution and transmission systems**. However, they are constantly at the mercy of electrical, mechanical, thermal and environmental stresses. On the other hand, power transformers are today one of the most critical and costly components of the electrical system, in some cases accounting for 60% of the total investment. Due to the **upfront investment** involved before even starting commercial activity, **monitoring and maintaining the condition of these are fundamental tasks in the field**.

II. PROJECT DEFINITION

Advanced Data Analytics and Artificial Intelligence are evolving at breakneck speed, yet there is still a long way to go. Currently, many methods emerge from the contributions born in Applied Statistics, but in many occasions, it is the part of the **implementation** where companies need a greater support. This project's aim is to **respond to the needs of Smart City Málaga** (hereinafter also called SCM) to develop a **multiplatform web-based model in real time providing information of high added value** and making use of intelligent techniques. Thus, it is intended to create a basis for future models aimed at solving this type of problem or similar ones, from the assistance to personnel destined for the physical place, to the execution of alarm managers or detection of anomalies.

In terms of objectives, they have been agreed and specified in the following list:

- Achieve latency and interaction levels as of a web service (less than 1 second).
- Allow automatic generation reports over network congestion and grid health.
- Create a model as robust as possible and agnostic to the data and platform: avoid quality loss in the face of implicit or explicit variations in the input data, and that the program does not depend on the device, operating system or version of the viewer.
- Leverage the analysis and conclusions of the study on intelligent tools for reducing dimensions (less than 15% error with respect to the reconstructed space).

It is appropriate to comment on the alignment of these business and operational metrics with the Industry 4.0 mantras. For the sake of modernization, digitization and simulation of the real world in the digital infrastructure, this project makes sense in particular focusing on the industrial and innovative value it brings to the environment of such scale as energy

III. DESCRIPTION OF THE MODEL

The goal of the project is to develop a digital product built to allow the monitoring and temporal evolution of electrical congestion within a given geographic area. As a real-time interaction tool with flexible functionality on different type of devices, the required result conditions were clear and based on three differentiated axes: **ease of use**, **interpretability of the results** and **stability of the page's integrity** (both front-end as back-end).

A. DATA EXPLORATORY ANALYSIS

The first section is reserved for the **preliminary qualitative and quantitative study** carried out on the starting variables. With the help of three different data evaluation methods (Sweetviz [4], Autoviz [5], Lux [6] libraries), the relationships that bind the behavior of certain physical variables with others are assessed and quantified, as well as the distribution of observations within each of them. the regression line turns negative, and Reactive Power is localized mainly around zero, but inductive is always positive and inductive is more spread in variance.

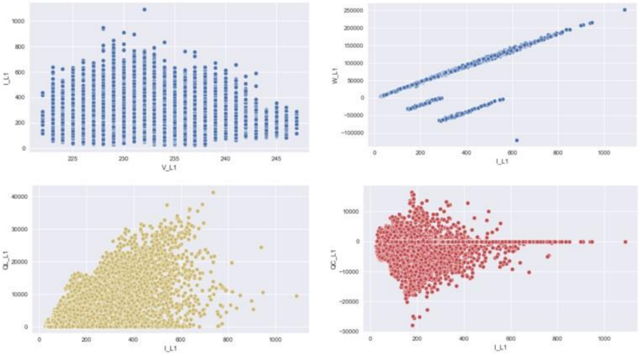


FIGURE 1. **Data Exploratory Analysis Example with Autoviz. Most significant observations are the quantization of voltage measure to the voltage unit, the correlation existing between Active Power and current, segmented and decreased with respect to current by an offset every time**

Based on the conclusions drawn (in particular, evaluation of the quality of the data, nature of the values and correlations between variables), a **preprocessing guide** (cleaning, substitution of values, formatting and normalization) and **selection of attributes** are decided, that ultimately replicate in the import of data inside the tool.

B. DESCRIPTIVE VISUALIZATION TOOLS

The second section translates the generation of results in an intermediate stage of the project's chronological evolution, where a basis is provided to define the share of the tool in charge of the most descriptive part. Beyond choosing one type of graph instead of others, above all it lays the foundations of the global structure for the Dash application template, a key factor for the progression of the work. First, the schematic design that encompasses the graphs of interest was addressed and that ultimately translates to the visualization – style – interactivity world (HTML – CSS – JavaScript).

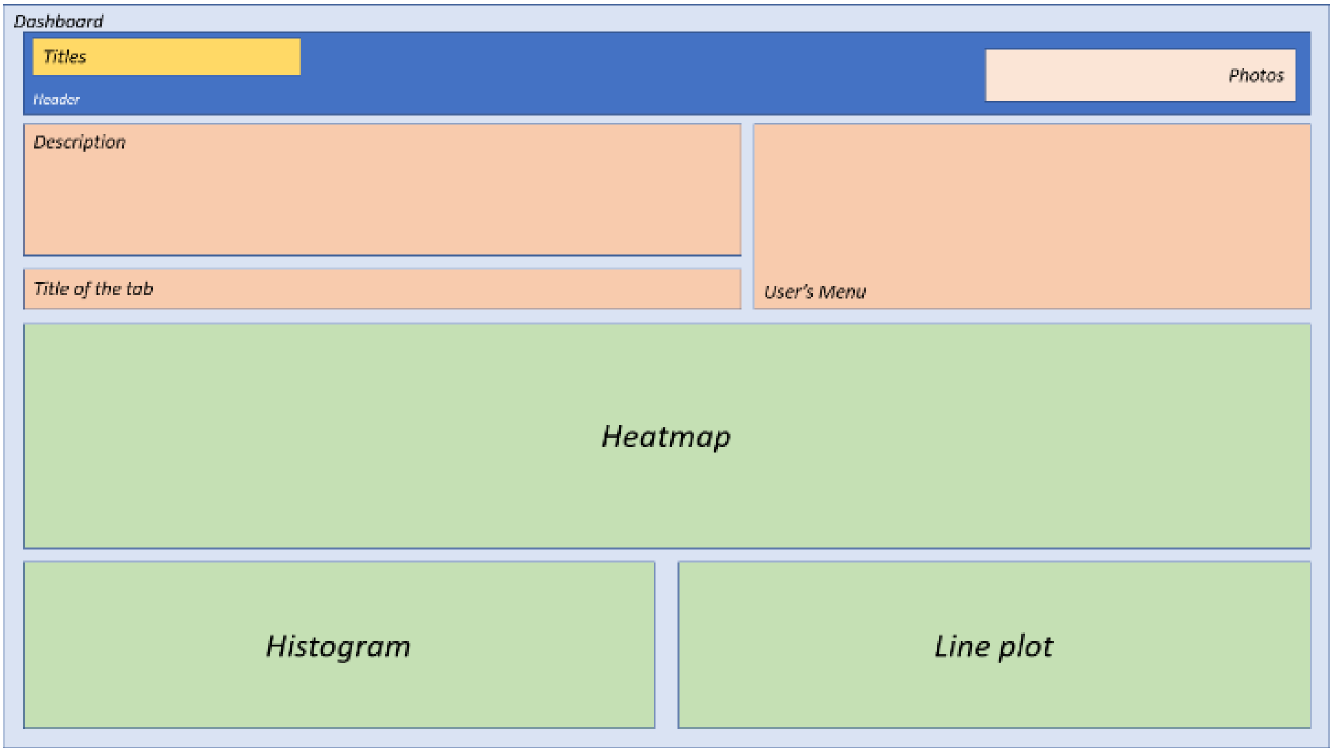


FIGURE 2. Dashboard's conceptual scheme. Blue elements are esthetic, salmon-colored are mostly user's input fields and green ones are resulting graphs after processing of the app, rendered in real-time. The influence from one should be resented by the other, and update should be coordinated with user's interactivity intentions.

After going through several iterations to achieve the desired preprocessing and the ideal service for the geographical representation of the heat map, all performed on the Google Colab platform, a research work is done around the available objects in Dash (essentially Dash_html_components, Dash_core_components, Dash_bootstrap_components). Based on these, the proposed general structure, inspired on the best practices for application development, is as follows: 1) Import Python **libraries and frameworks**, 2) **Process the starting information**, 3) Build the **application layout**, with input and output elements inside and 4) Specify at the end the functions (called **callbacks**) that **describe existing interactions on the web page** in response of user events.

The result of the first version, mainly descriptive, is similar to the first tab Figure 4 and Figure 5.

C. TOOL'S ANALYTICAL CORE

The third section fully confronts the half of the tool in charge of the deep analysis work, and presents the exhaustive analysis work that was performed around the two machine learning problems previously evoked: **dimensional reduction**, and **outlier detection**. Supporting the technical decisions on objective results and subjective considerations with respect to user experience, it details the blocks that compose this volume as well as clarifies the justification of the solution and the value provided.

Although at first the use case was approached as a **classificatory task to predict the generating Transformer Substation** from the observed measurements, it is eventually modified to mutate into an **unsupervised outlier detection algorithm**. Ingesting the TC selection, it is interesting for the stakeholder parties to evaluate the possibility of an outlier occurring given the rest of the most frequent operating points in that TC.

In short, an outlier is defined as a measurement that **diverges from a general pattern** of a sample and, in order to fulfill the technical challenge, the original problem is first dimensionally reduced and, in a second step, the internal patterns of the distribution are inferred with the help of unsupervised algorithms. First, **five main attributes (Voltage, Current, Active Power and Reactive Power, Inductive and Capacitive)** are selected from the total of more than 160,000 records. Once filtered, different reduction models are evaluated: **PCA** (Principal Component Analysis), **AE** (Auto-Encoder, in Simple, Multilayer and Nonlinear versions) and **t-SNE** (t-Distributed Stochastic Neighbor) based on the ability of a classifier (such as a Random Forest Classifier) trained on the encodings of each model to associate measurement with its CT.

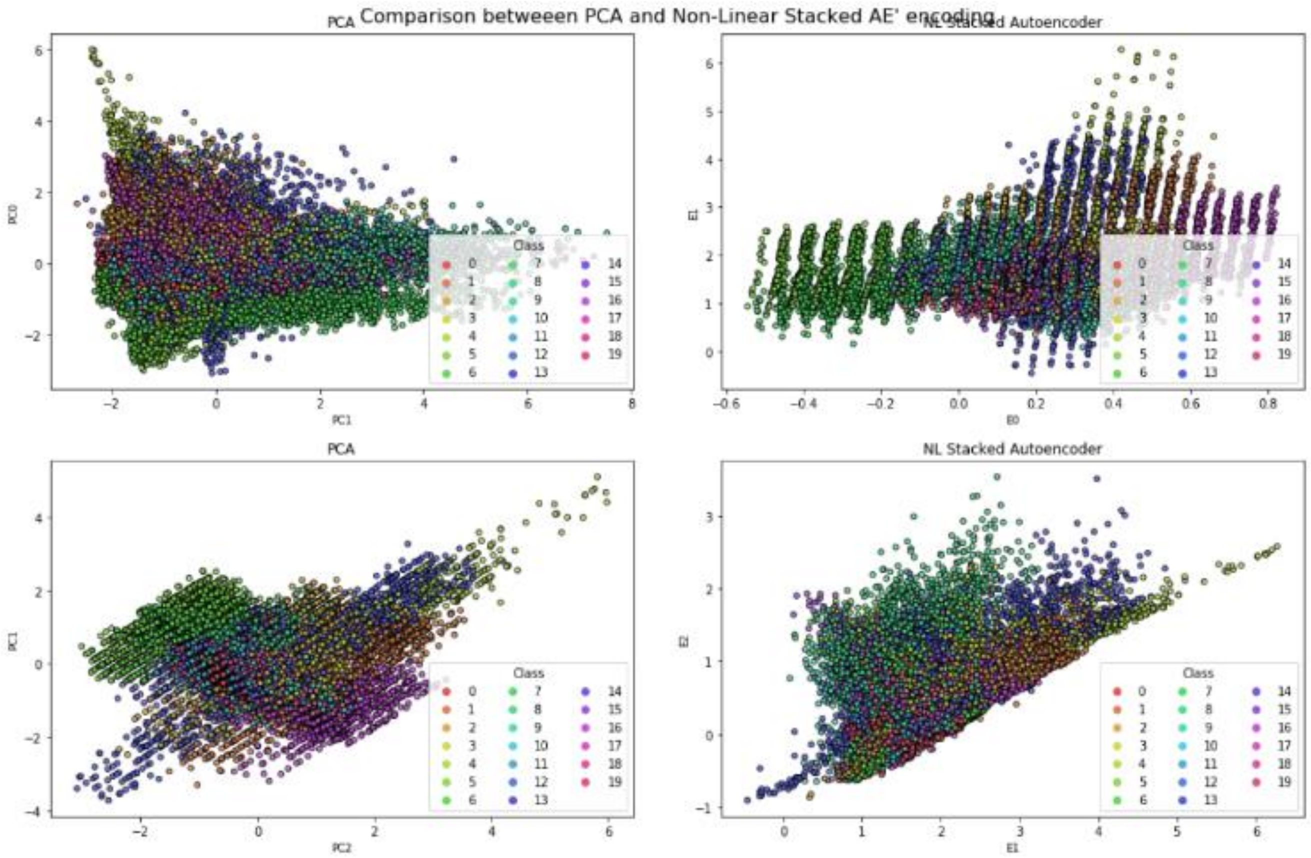


FIGURE 3. Codings for PCA and Non-Linear Stacked AE, paired in the first and second, then third and second component. It is shocking to see this time the behavior of the discontinuous stress values better reflected in the NL_St_AE model between the first and second node but more so between PC1 and PC2. The differences increase, the universe of choice for hyperparameter fitting grows and with it, the complexity of the model, which tests new methods for dimensional reduction.

Considering that the **problem is not linearly separable** but the regions of points are **localized**, the classifier does not stand out for its success rate, but it does provide an **objective and orderly measure** (evaluated on a **Validation** and **Test dataset**) from which to estimate the quality of the dimensional reduction. This metric, coupled with the training efficiency of the algorithm, constitutes the objective function that the model choice is intended to maximize.

Although the results of this analysis are developed in 4. Results, it is worth mentioning that the algorithm finally chosen is the **PCA (n=3).**

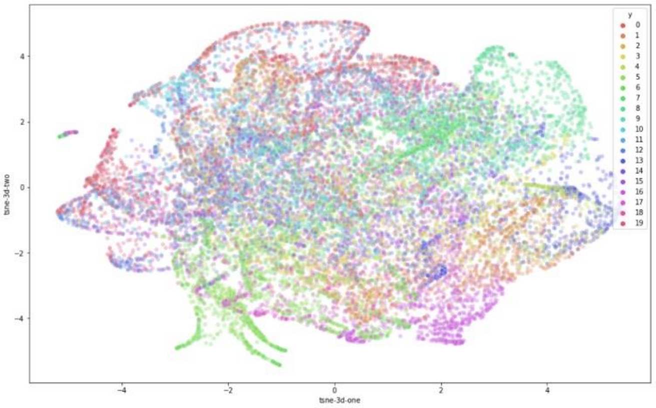


FIGURE 4. Codings from t-SNE. This representation of first 20.000 points is the outcome model the input set as well as possible, with the mapping such that the correspondence between the two distributions is minimally distorted. The way in which this technique is computed is cumbersome and therefore has serious performance limitations.

Once the PCA encodings are established, the **cumulated variance** captured by the projection is evaluated (which must be high enough to ensure the consistency of the results) and the **interpretability of the eigenvectors** is analyzed through their loadings (in order to comment on whether the translation between both spaces is straight or complex). The latter allow the transformation from one world to the other, which benefits the result in two ways: on the one hand, it allows the visualization of the analyses within the app itself, and it maintains a more information-dense dataset with which to predict outliers for a given TC.

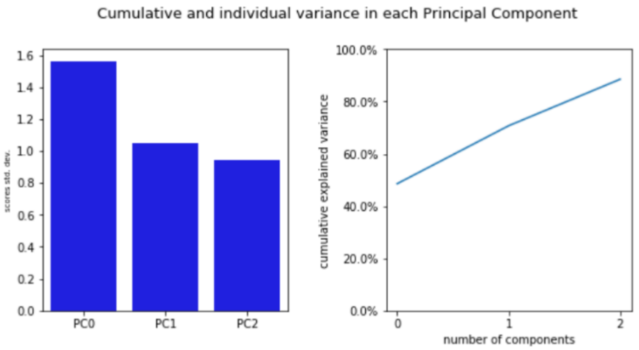


FIGURE 5. Loadings standard deviation and Explained Variance with three first PC. As visible, the ordering responds to the functionality principle of PCA maximization from the covariance matrix and, following the chart result on the right side, more than three quarters of the information is hold in the two first PC, with more than 88% in the accumulated variance on the three.

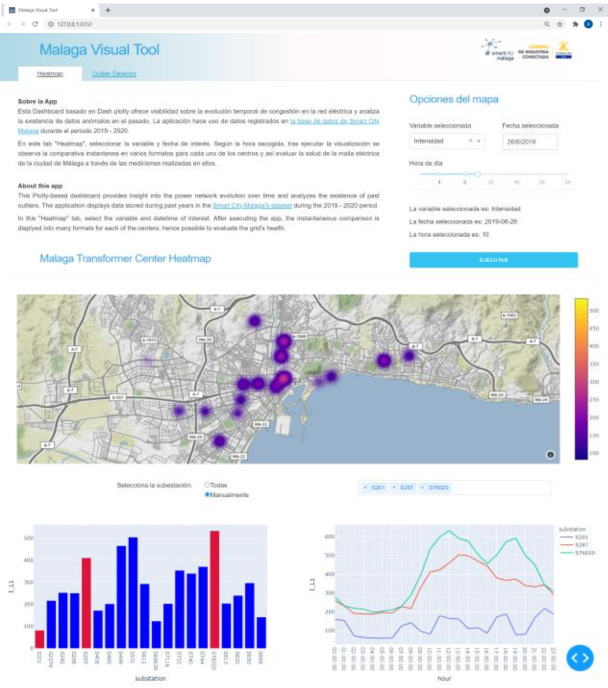
In the report, a taxonomy of the solutions and problems families in this field of diverse applications such as fraud detection, medical inspection or network cybersecurity. With the help of the **Scikit-learn** model implementation, the results of the **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise), **OPTICS** (Ordering Points To Identify the Clustering Structure) and **LOF** (Local Outlier Factor) models are compared. The comments on the theoretical review underline the benefits and risks of each, the use cases to which they are best suited and, for the practical section, analysis of the detections are carried out and materialize the visualization graphs that later come into play within the application.

Once again, analysis of results of conclusions, compromises between objectives and conclusions are discussed in detail in the Results. However, in order to continue with the outcome in this summary, it is anticipated that the model used for performance in results and loading times is none other than the **DBSCAN model (eps = 0.55).**

D. INTEGRATION AND INTERACTIVITY

In the last and fourth section, a space is reserved for the **explanation of the integration of the two previous modules**. The development goes beyond the code used in the descriptive section, rather going deeper into the means and methods used to **interlace both sections** and make their use intuitive as if it were a commercial application, and in particular, the privileged and primordial use reserved for **callbacks**.

Callbacks are **Python functions automatically called** by Dash **whenever the property of some component changes**. When the input property mutates, the callback decorator function is automatically invoked and Dash launches the execution of the function with the new input value to return the new output. This is called **Reactive Programming** and, unless the value is manually overwritten, it enables application interactivity. Advanced callbacks are nonetheless able to take several inputs or outputs in the same function, include smooth transitions to changes in graphics... and are at the origin of the application, from the generation of heat maps or 3D renderings where the user has the power to select the view that best suits him, to the discrimination of styles or the tab system implemented to navigate from the descriptive to the analytical panel, and vice versa.



Sobre la App
Esta Dashboard basado en Dash plotly ofrece visibilidad sobre la evolución temporal de congestión en la red eléctrica y analiza la existencia de datos anómalos en el pasado. La aplicación hace uso de datos registrados en la base de datos de Smart City Malaga durante el periodo 2019 - 2020.

En este tab "Heatmap", seleccionar la variable y fecha de interés. Según la hora escogida, tras ejecutar la visualización se observa la comparativa instantanea en varios formatos para cada uno de los centros y así evaluar la salud de la malla eléctrica de la ciudad de Málaga a través de las mediciones realizadas en ellos.

About this app
This Plotly-based dashboard provides insight into the power network evolution over time and analyzes the existence of past outliers. The application displays data stored during past years in the Smart City Malaga's dataset during the 2019 - 2020 period.

In this "Heatmap" tab, select the variable and datetime of interest. After executing the app, the instantaneous comparison is diaplyed into many formats for each of the centers, hence possible to evaluate the grid's health.

Opciones del mapa

Variable seleccionada: Intensidad
Fecha seleccionada: 26/6/2019
Hora de día

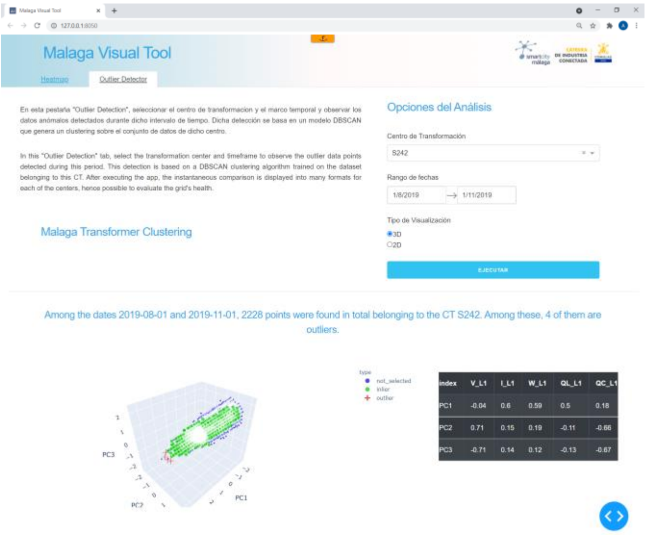
La variable seleccionada es: Intensidad
La fecha seleccionada es: 2019-06-26
La hora seleccionada es: 10

Malaga Transformer Center Heatmap

Selecciona la subestación: ○ Todas ● Manualmente
S201 S287 S76020

FIGURE 6. First tab presentation of the Malaga_Virtual_Tool. This display is the first one appearing when running the script from a local virtual environment, being hosted locally (127.0.0.1).

In the Figure 6, menu in top right section allows the user to **interact with the app**: choice of **variable selection** among the selected headers (Tension, Current, Power, Outside Temperature), **selection of the date** (among the possible, any between 16-06-2019 and 08-06-2020) as well as **time** in the day. By executing the app, the graphs below it readapts and display, respectively 1) a **Heatmap** with values on geographical coordinate of the TC (with relevant hover label and auto-scaled legend), 2) **Histogram** with values for each CT at that time and 3) **Plot line chart** with evolution of variable for selected CT throughout the day.



En esta pestaña "Outlier Detection", seleccionar el centro de transformacion y el marco temporal y observar los datos anómalos detectados durante dicho intervalo de tiempo. Dicha detección se basa en un modelo DBSCAN que genera un clustering sobre el conjunto de datos de dicho centro.

In this "Outlier Detection" tab, select the transformation center and timeframe to observe the outlier data points detected during this period. This detection is based on a DBSCAN clustering algorithm trained on the dataset belonging to this CT. After executing the app, the instantaneous comparison is displayed into many formats for each of the centers, hence possible to evaluate the grid's health.

Malaga Transformer Clustering

Opciones del Análisis

Centro de Transformación: S242

Rango de fechas: 1/8/2019 → 1/11/2019

Tipo de Visualización: ● 3D ○ 2D

Among the dates 2019-08-01 and 2019-11-01, 2228 points were found in total belonging to the CT S242. Among these, 4 of them are outliers.

index	V_L1	I_L1	W_L1	QL_L1	QC_L1
PC1	-0.04	0.6	0.59	0.5	0.18
PC2	0.71	0.15	0.19	-0.11	-0.66
PC3	-0.71	0.14	0.12	-0.13	-0.67

FIGURE 7. Second tab presentation of the Malaga_Virtual_Tool. This display is the second one, appearing when pressing the "Outlier detection" tab.

In Figure 7, Menu allows a similar selection as previous tab, only that besides selecting the TC, the user is selecting a **date range**. The PCA is applied on all the measurements at once during app initialization, but to make conclusions consistent, DBSCAN is **only trained over** all the data from the selected TC (no matter the date). Then, a conclusion paragraph is displayed at the header of the graph area, announcing the number of points detected for the selected CT during the provided window timeframe, then the **counting of predicted outliers**.

Subsequently, the 2-D or 3-D plot is presented with a strict color legend: blue dots represent **non-selected** registers (belonging to the TC but not within the time limits), green dots for **inlier** (points being effectively labeled as common or expected given the distribution) and red crosses for **outlier** (only those between the dates of the time range). The graph is completely interactive so **zooming**, **screenshooting** or **filtering** classes is fully accessible (for example, extracting a view on the red crossed among green dots after rotating the view, with the axes in the background, is perfectly possible).

E. OPTIMIZATION IN FRONT AND BACK-END

Finally, the fifth chapter ends up listing the **set of operations performed to optimize the use of the tool**, either by improving the result delivered in the output or by using the minimum computational resources so as not to conflict with the fluidity of the website. Among the chosen techniques, we highlight the slight adaptation of figures, despite being more complex, gain flexibility and maintenance over time; the perimeter traced in the face of the selection options available to the user and, last but not least, the adaptation of the template and objects to the size and characteristics of the device's screen, thus abstracting the display of elements of the device and **making the tool cross-platform**.

IV. RESULTS

This section reviews the results of the studies mentioned in the previous chapters, both for the comparison of models involved in dimensional reduction and the candidates for outlier identification.

A. DIMENSIONALITY REDUCTION ANALYSIS

In the analysis of dimensional reduction, the results that the RFC yields on Transformation Centre (TC) prediction based on the codings of each system are assessed, both for the Validation and Test sets. Evaluating both scores is of interest to evaluate the **bias-variance trade-off** of each model, and to know if there is **under or overtraining**.

Given this statement, two phenomena are expected: first, it has already been discussed that the problem is not linearly separable into 2D or 3D, which means that the RFC's hit rates will be far from optimal. **The RFC is a meta-estimator that trains a series of classification trees on various subsets of the training base** and, on new data, aggregates the predictions from each of them to improve the accuracy of the meta-prediction, while better controlling for overtraining (the size faced by each tree is smaller than the total). But despite this, the model is not infallible and will be seen below, but fortunately this is not the purpose of this study, but to be able to choose the model that gives the RFC the best result.

Secondly, it is possible to anticipate that the results of the RFC predictions will always be **worse based on the encodings of the observations than on the original dataset**. Obviously, since the original table has more non-redundant information to start with (five variables, as opposed to three dimensions), its result will be higher than the rest. However, this does not prevent there being some closer than others to this limit, and it is from this ordering that the conclusions emerge.

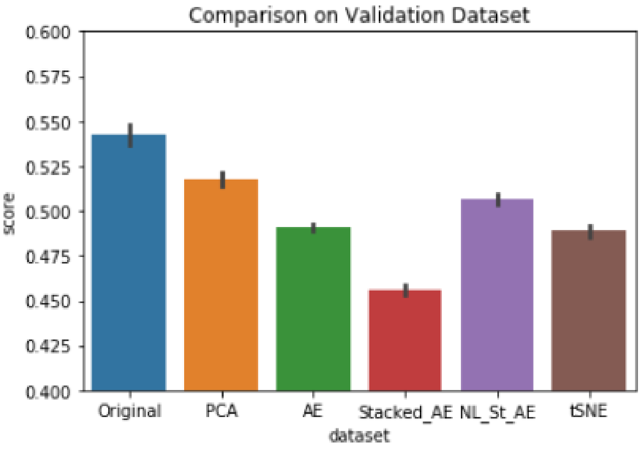


FIGURE 8. RFC Cross-Validation (k=5) results

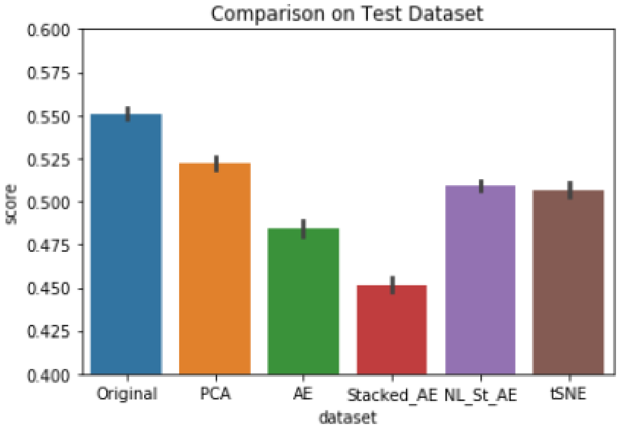
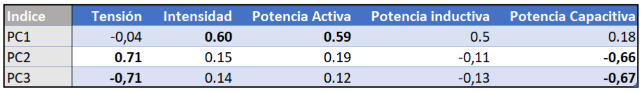


FIGURE 9. RFC Test (k=5) results

Looking at the differences in classification performance between the original space and the different others, discussions are quite conclusive. In terms of efficiency, the fastest models were the PCA (fitting time = 3.13 s) and the simple AE (fitting time = 2.5, optimized with the Keras package).

The conclusion is that, in order to faithfully maintain the point distribution and enable its representation, the best option is to **reduce the dimensions from five to three by means of a PCA**. Once this has been taken as the cornerstone of the transformation, it is convenient to judge the clarity of the eigenvectors and their relation to the starting variables.

TABLE I
PCA EIGENVECTORS' LOADINGS

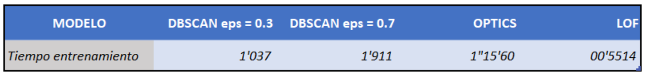


Indice	Tensión	Intensidad	Potencia Activa	Potencia inductiva	Potencia Capacitiva
PC1	-0,04	**0.60**	**0.59**	0.5	0.18
PC2	**0.71**	0.15	0.19	-0,11	**-0,66**
PC3	**-0,71**	0.14	0.12	-0,13	**-0,67**

Essentially, PC1 has a strong dependence on **Current and Active Power** and catches the cases **of overcurrent in the event of load unbalance** (it also accounts for almost 50% of the variance). PC2 emphasizes **high Voltage and low Capacitive Power** values, focusing its attention on cases where the voltage exceeds the network regulation setpoint. Finally, PC3 is the third perpendicular component and focuses on **strong voltage dips** (caused by failures of supply equipment or infrastructure, switching on of large loads, ...).

Finally, the last major block of analysis is the comparison of models for the detection of outliers. For each one, **training times** are measured on a random fraction of the set and their discrimination is analyzed.

TABLE II
TRAINING TIMES FOR DBSCAN, OPTICS AND LOF



MODELO	DBSCAN eps = 0.3	DBSCAN eps = 0.7	OPTICS	LOF
Tiempo entrenamiento	*1'037*	*1'911*	*1"15'60*	*00'5514*

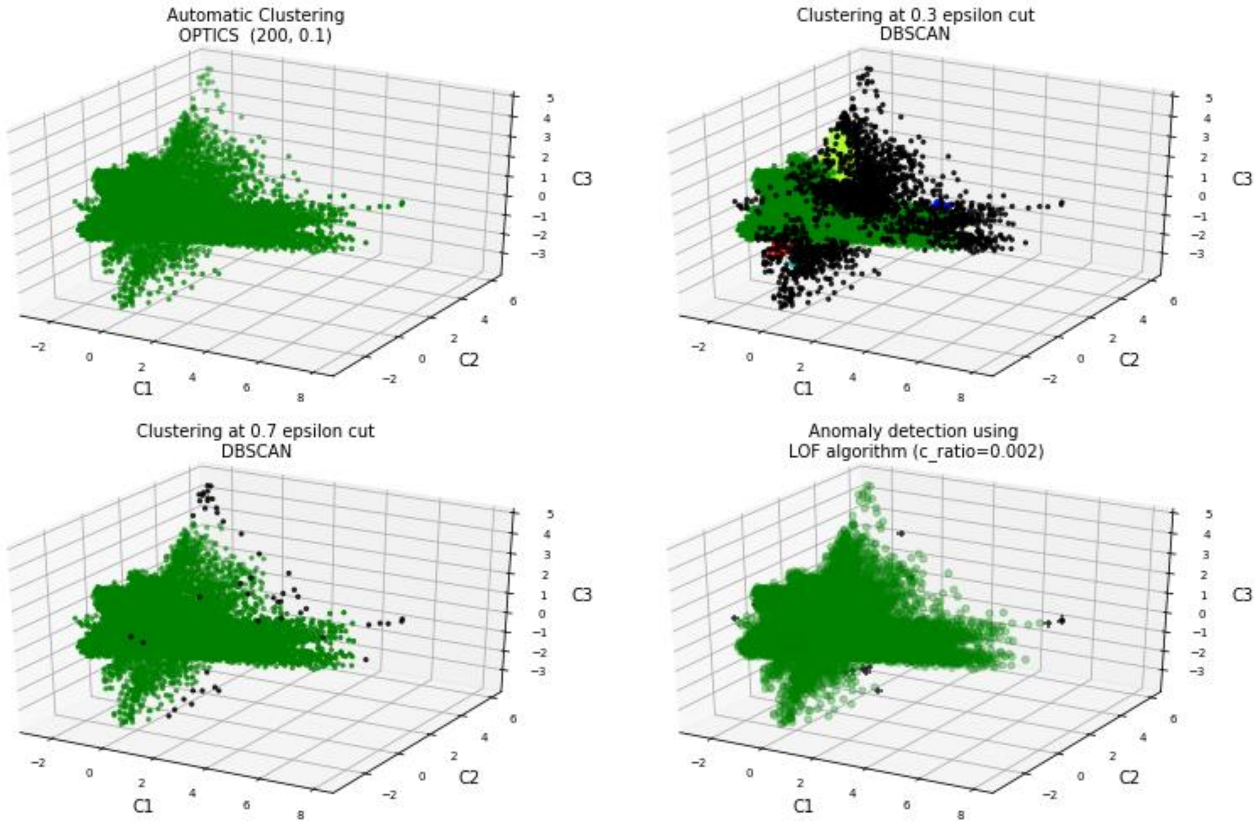


FIGURE 9. Outlier Detection results on dataset cross-TC with from 20.000 samples from OPTICS, DBSCAN (eps = 0.3), DBSCAN (eps=0.7) and LOF models.

Among the four cases, it is the DBSCAN model the one that offers the best performance in terms of ease of use, agility and better adaptability to a wider range of training sizes (which will vary according to the user's selection).

V. CONCLUSIONS

Throughout this project, it has achieved to process, explore, extract information and present it in a **multiplatform, agile and distributable application** for the support and support of maintenance tasks of Transformer Substations in the city of Málaga.

This section includes an extensive guide for the deployment of the application in two Platform as a Service (PaaS) alternatives that would facilitate the distribution of the project, the sharing and alignment of objectives and integration into the company's IT system.

The Dash application is **holistic to data and devices**, and manages to **generate an analytical tool for presenting descriptive and predictive estimations** around observations measured on the power grid lines. The web page latency is **less than one second**, it presents an **intuitive user interface** and responds to the problems encountered today in exploiting large historical databases with records of measured physical variables. The initial objectives are satisfied globally and even functionalities that were not originally programmed are added, within a shorter time frame than initially planned.

With regard to the CIC, this tool is now in a sweet status where thousands of extensions and adaptation variations are open for new collaborators of the Chair who want to extract valuable information with the power of Data Science, very necessary in the XXI century and in the future that lies ahead.

REFERENCES

[1] C. Yao, Z. Zhao, Y. Mi, C. Li, Y. Liao and G. Qian, "Improved Online Monitoring Method for Transformer Winding Deformations Based on the Lissajous Graphical Analysis of Voltage and Current," in IEEE Transactions on Power Delivery, vol. 30, no. 4, pp. 1965-1973, Aug. 2015, doi: 10.1109/TPWRD.2015.2418344. https://ieeexplore.ieee.org/document/7086082

[2] A. Singh and P. Verma, "A review of intelligent diagnostic methods for condition assessment of insulation system in power transformers", 2008 International Conference on Condition Monitoring and Diagnosis, 2008, pp. 1354-1357, doi: 10.1109/CMD.2008.4580520.: https://ieeexplore.ieee.org/document/4580520

[3] United States Department of the Interior, Bureau of Reclamation, "Transformer Diagnostics", Facilities Instructions Standards and Techniques, 2003, colum3 3-31: https://www.usbr.gov/power/data/fist/fist3_31/fist3-31.pdf

[4] Himanshu Sharma, "Automating Exploratory Data Analysis – Part 1", Feb 2021, medium.com: https://medium.com/swlh/automating-exploratory-data-analysis-part-1-f5f2b7d548e5

[5] Himanshu Sharma, "Automating Exploratory Data Analysis – Part 2", Feb 2021, medium.com: https://medium.com/swlh/automating-exploratory-data-analysis-part-2-f03083f42ecf

[6] Himanshu Sharma, "Automating Exploratory Data Analysis – Part 3", Feb 2021, medium.com: https://medium.com/swlh/automating-exploratory-data-analysis-part-3-d04352b83072