COMPUTATIONAL TOOL TO IMPROVE THE INFORMATION'S QUALITY OF THE DSO'S GEOGRAPHIC DATABASE (BDGD) FOR REGULATORY PURPOSES

Davi Mantovani RICCI
Daimon – Brazil
davi.ricci@daimon.com.br

Paulo Henrique BAUMANN
Daimon – Brazil
paulo@daimon.com.br

Fabio ROMERO
Daimon – Brazil
fabio.romero@daimon.com.br

André MEFFE
Daimon – Brazil
andre@daimon.com.br

Armando H. S. G. JESUS
Equatorial Group – Brazil
armando.jesus@cemar-ma.com.br

Eliezer S. OLIVEIRA
Equatorial Group – Brazil
eliezer.oliveira@cemar-ma.com.br

ABSTRACT

This work presents a computational tool whose purpose is to identify anomaly in the information registered in the database of a distribution system operator. This tool has as methodology the application of multiple heuristics to identify incorrect data and estimate the most credible scenario for the context.

INTRODUCTION

This work presents the Software BIAT (BDGD Integrity Assurance Tool), result of the Research & Development Project "Development of a Robust System for Monitoring and Updating the Georeferenced Database of a Distribution System Operator (DSO)" of ANEEL (Brazilian Electricity Regulatory Agency) proposed by the Equatorial Group and executed by Daimon Engineering and Systems. The purpose of this software is to detect anomalies in the DSOs' database of assets and consumers, named Geographic Database of the Distribution System (BDGD). As required by ANEEL, the DSOs must send the BDGD annually to update the SIG-R (Regulatory Geographic Information System). The SIG-R was created with the objective of improving the methodologies of supervision and regulation of the generation, transmission and distribution of electrical energy in Brazil.

The software consists of a set of algorithms based on an intelligent search engine structured in premises that guarantee the operation of the electrical arrangement. This mechanism is described by a hierarchical structural analysis, analogous to a bottom-up approach, that initiate in elements of less complexity, whose respective sets constitute elements of greater complexity, until the complete conception of the electric power distribution system. As the main tool for the development of the algorithms that constitute the Software, the Linear Graph Theory [1] was used. From this, it is possible to divide the elements that make up the network into edges elements and vertex elements. The edge elements, intuitively, are thus classified as being connected to two vertices, and in the network they are responsible for transferring power. In the case, these are the links in the medium and low voltage networks, and transformers, switches and series capacitors for the medium voltage network. On the other hand, vertex elements are summarized to buses with equipment that absorb or inject part of the power transferred by the edge elements.

On medium voltage network buses there are power transformation stations (responsible for feeding the low voltage blocks), the primary stations (responsible for feeding medium voltage consumers), shunt capacitors, reactors and generators. In the low voltage network's buses only consumer units are connected.

The energy flow is established between the consumer and the supplier when there is sufficient minimum resource for the interaction of the electric and magnetic fields within the operation's limits of the source and the load that provides minimum power dispersion. The lack of one of the phases or even of the neutral is enough so that the load-source connection resource be insufficient for the requested electric power transfer, preventing it from being transformed into other energetic manifestations to fulfill the respective design tasks, which would render the system inoperative. The occurrence of a partial disconnection, described by the lack of one or a couple of phases requested by a bus or even the lack of the neutral, is due to the incorrect phasing along the edge elements.

Power distribution networks are divided into two types of operational philosophies. The first one, called radial network, has its structure described by a tree-type graph, which is a connected graph characterized by the existence of a single possible path between two vertices. The second also falls into the class of connected graphs, but it can contain cycles. A cycle is a simple and closed chain, which is any sequence of adjacent edges connecting two vertices. If the sequence of edges has no repeating elements, the chain is said to be simple. The medium voltage network fits the first type, so if it contains some cycle in the register, this is indicated as cadastral inconsistency. On the other hand, the low voltage networks from Equatorial's DSOs are represented by connected graphs with or without cycles.

Within the presented context, it was assumed as initial premise that the electric power distribution system is operative, that is, it contains sufficient minimum resources so that electricity can flow between the source and the consumer unit and fulfill its objective. Therefore, all registered information (that does not damage the base's integrity) that describes some characteristic that makes the distribution system inoperative was assumed as cadastral inconsistency.

The Theory of Graphs and related algorithms was not enough to correctly identify the cadastral inconsistencies. Among those that fit this, it is important to highlight the incompatibility between the distribution transformer (responsible for feeding the blocks of the low voltage network) and associated loads. For this it was necessary

to elaborate a decision algorithm based on reliability weights provided by the software user, since the analyzed database is the only reference for the computational tool to discern which information is correct or not. Therefore, although the inconsistency is visible to an analysis of high abstraction, its true location was confusing when intrinsically verified.

TRANSFORMER, LOW VOLTAGE GRID AND LOAD ACCORDANCE

In order to identify some partial disconnection between the distribution transformer and the load, some facts about the feeder transformer must be in accordance with its downstream grid and its load. When the set of variables that represents these facts are incoherent, one can conclude that some of them might contain errors. If this is the case, the goal is to find out which variables are responsible for the misalignment and to correct them.

The variables are: (1) total consumption of associated consumer units, (2) transformer type given its nominal power and standardized values, (3) transformer line voltage, (4) connection type of the secondary side of the transformer (SST), (5) number of phases of the first downstream link and (6) maximum number of phases of the transformer's consumers.

The nominal power of the transformer must be sufficient to supply the associated consuming units, in addition to being fitted to one of the standard power values [2]. The possible types of distribution transformers for the networks covered by the software are: single-phase or three-phase. For these, the secondary can be single-phase or two-phase, or three-phase respectively. As a consequence of these configurations two different line voltages are possible, this is because the phase voltages of the two-phase secondary are offset by 180º each other and the phase voltages of a three-phase secondary by 120º. That said, the line voltage used by the load and offered by the source must be the same. In the load-source alignment analysis, the number of phases of the first link downstream of the transformer must match the number of phases used by the consumer units and offered by the transformer. Lastly, the number of phases offered by the secondary must be in accordance with the number of phases requested by the consumers. Given these variables, there are three possible operational hypotheses for low voltage networks:

TABLE 1. EXPECTED VALUES FOR EACH OF THE 3 CONFIGURATIONS

Config.	A	B	C
1	$< E_{max,1\phi}$	$< E_{max,1\phi}$	$<, \geq E_{max,1\phi}$
2	1ϕ	1ϕ	3ϕ
3	-	440V	380V
4	1ϕ	2ϕ	3ϕ
5	1	2	3
6	1	2	3

"Configuration A":

- Load energy consumption lower than the maximum monthly consumption for a single-phase transformer $E_{max,1\phi}$ (it assumes a power factor and a load factor).
- Nominal power is a normalized value for single-phase transformers.
- Single-phase connection on the SST.
- One phase in the first link downstream of the transformer.
- Number of phases used by consumer units, as a whole, equal to one.

"Configuration B":

- Load energy consumption lower than $E_{max,1\phi}$ the maximum monthly consumption for a single-phase transformer.
- Nominal power confers with the normalized value of single-phase transformer.
- Secondary line voltage is 440V.
- Two-phase connection on the SST.
- Two phases in the first link downstream of the transformer.
- Number of phases used by consumer units, as a whole, equal to two.

"Configuration C":

- Nominal power is a normalized value of three-phase transformer.
- Line voltage in the secondary is 380V.
- Three-phase connection on the SST.
- Three phases in the first link downstream of the transformer.
- Number of phases used by consumer units, as a whole, equal to three.

There are six reliability weights, one for each variable, restrained in the interval $0.5 \leq c_i \leq 1$. They indicate the probability of the information contained in the respective variable is correct.

The likelihood of the configuration is calculated using equation (1) in which P_i is the participation of the attribute in the respective configuration. This, in turn, is calculated through the equation (2). The configuration with the highest likelihood is taken as the truth. Unsound parameters to this configuration are classified as incorrect.

$$L_j = \prod_{i=1}^{N} P_i = P_1 \cdot P_2 \cdots P_N \tag{1}$$

$$P_i = \begin{cases} c_i, \text{ If "}i\text{" fits in "}j\text{"} \\ (1 - c_i), \text{ If "}i\text{" does not fit in "}j\text{"} \end{cases} \tag{2}$$

After correcting such inconsistencies, one last data is verified: the coherence between the phase configuration of the transformer and the phase configuration of its loads. Loads with phases different from the ones at the secondary of the transformer are adjusted.

TOTAL OR PARTIAL DISCONNECTION

The partial disconnection is identified by the lack of phase or neutral for some bus. This is caused by incorrect registration of some edge element that completes the source-load path, so that it either contains insufficient number of phases or a phase configuration incompatible with the source and the load (eg. ABN instead of BCN).
On the other hand, the total disconnection occurs when there are no links connecting one or more buses to its source

Identification of partial or total disconnection

The identification of the disconnection, whether partial or total, is done by obtaining an equivalent graph of the system, on which the Breadth First Search (BFS) algorithm is applied. This checks whether the graph is connected or not. If it is not, it points all the buses that do not have access to the source bus, which can be isolated or interconnected.
To verify total disconnection, the graph constructed by the links of the network is analysed. Conversely, it is necessary to evaluate a graph for each phase individually to assess the presence of partial disconnection.

Correction of partial disconnection

The correction of the partial disconnection is made by changing the phase configuration of the link indicated as responsible for the inconsistency to the most appropriate.

Correction of total disconnection

A totally disconnected structure varies from a simple bus to a set of interconnected vertex elements and edge elements. The reconnection procedure, in general terms, is to create a new link between the most suitable pair of buses capable of solving the issue. One of the buses of the pair belongs to the disconnected structure, whereas the other is directly or indirectly connected to the source. For this, the following characteristics are taken into account:

- Distance between the buses
- Electrical compatibility of the buses
- Arrangements of streets surrounding the candidate link

For the first characteristic, distance between the buses, the algorithm Kdtree [3] is used to identify the nearest buses from the respective geographical coordinates in the base.
This algorithm is defined by a set of procedures, which processes the coordinates of the elements and the distances between them, constructing a data structure represented by a binary tree (connected graph characterized by the existence of a single possible path between two vertices). In this tree, each vertex stores the coordinate of a device, so that through another set of simple procedures, it is possible to browse for nearby elements. The navigation starts at the root of the binary tree, which at each node has at most the extension for two other nodes. This feature speeds up the search process, since at each level of the data structure there are only two options, of which only one is true.

It should be noted that, along with the discarded branch, the branches derived from it are also eliminated, being able to reduce in half the number of possible nodes for each analyzed level if the tree constructed is symmetric (or close to), which avoids an exhaustive sweep of the possible combinations of pairs of buses.
However, the proximity of the buses do not guarantee that these are the most suitable for reconnection. The electric compatibility between these is verified. At last, the new link must be aligned with the arrangements of streets in order to avoid cases where it could cross over a building for example.

MESHED TOPOLOGY IN THE MEDIUM VOLTAGE NETWORK

Due to the operational philosophy adopted by the DSOs involved in this project, its equivalent graph must be a tree, that is, connected and without cycles. A cycle by itself is a graph $G=\{E,V\}$ described by a set of vertices V and a set of edges E. In a connected graph, the number of linearly independent cycles "n_f" is calculated by equation (4), where "e" is the number of edges and "n" is the number of vertices.

$$n_f = e - n + 1 \tag{3}$$

The set of linearly independent cycles of a connected graph is the fundamental set or canonical basis that through the "or exclusive" operation between the meshes constitutes the vector space of cycles of the structure.
Let $A=\{V_A, E_A\}$ and $B=\{V_B, E_B\}$ two cycles, the "or exclusive" operation among them is given by expression (5).

$$A \oplus B = C \rightarrow$$
$$C = \begin{cases} V_C = V_A \cup V_B - V_A \cap V_B \\ E_C = E_A \cup E_B - E_A \cap E_B \end{cases} \tag{4}$$

Thus, from the amount of elements edges and vertices of an analyzed circuit, whose equivalent graph must be connected, the existence of meshes is verified by the equation (3). If $n_f > 0$, then the cycles are identified. For this, the algorithm proposed in [4] was selected, which, from a heuristics of construction of the graph analyzed expansion tree, returns the fundamental cycles. To use this algorithm, the library [5] was inserted in the Software.
Among the elements edges of the medium voltage network, there are the switches. If the identified cycles consist of switches whose operating state is normally closed, one of them is chosed to be opened. But, if there are not any switches, the link with the lowest current is selected for exclusion.

INADEQUATE ELECTRICAL CONDUCTOR

The total connection between a consumer unit and its respective source is the minimum necessary condition for

the steady state power flow, but not enough. This means that the current capacity of the cables that integrate the route must be sufficient for the energy to flow without causing definitive damage to the system. So, through the characteristics previously described, the power flow is calculated through the links of the network. Those classified as overloaded are pointed as unlikely information, rejecting the null hypothesis "The conductor is not inappropriate". The radial configuration of the medium voltage system allows the usage of the Newton-Raphson method for solving the power flow equations. On the other hand, due to the possibility of meshing in the LV network, it is necessary to use the Gauss-Seidel method.

ANOMALOUS CONSUMPTION

To detect anomaly in the consumption information, a simple technique was developed to detect outliers in one-dimensional data. The tool is described in the paper [6]

ANALYSIS AND RESULTS

The software can be described as a binary classifier if analyzed for the ability to identify anomaly in the database. From this perspective, this tool classifies information as plausible or implausible.

The inconsistency of the information is associated with the impossibility of the same being found in the field, because if it is real, part of the system is not able to operate normally. Given the implausible information, the Software suggests corrections, which are not necessarily synchronized with reality, because the scope of the tool is only the evaluated base.

Given this, the performance of the software was measured by the ability to classify the information of the database as plausible or implausible, rejecting or not the null hypothesis "This information is not implausible." In this way, the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are quantified. Sensitivity, precision and accuracy were selected as performance statistics. In the context of "Binary Classifiers", sensitivity is the ratio of positives detected (true positives) to total positives (false negatives and true positives). This metric is reduced by the occurrence of false negatives. Precision is defined as the ratio of positives detected (true positives) to the total number of objects classified as positive, whether false or true. Precision is reduced by the occurrence of false positives. The sensitivity and precision were taken as performance metrics due to the tool's goal of minimizing the occurrence of inconsistencies while keeping a low rate of false negatives. As a global meter, accuracy was chosen. This is calculated by the total of hits (true positives or true negatives) on the total of evaluated objects.

Some figures of merit (precision, sensitivity and accuracy) were calculated on samples, whose sizes were estimated assuming 95% of confidence level, sampling error of 10%, and the worst case scenario of prevalence, i.e. 50% for each expected event (data with error or without error). For each class of cadastral inconsistency a suitable sample was extracted. Due to the prevalence of correct information in the database, the samples were obtained through stratified sampling, applying simple random sub-sampling on the information classified as correct by the software and simple random oversampling on the information classified as incorrect, so that the sample be contained within 50% of each type of information.

The verification of the positives and negatives to classify them as true or false was done, for some samples, in the field and for others via SQL queries in the DSO's database. For the samples related to the "transformer, low voltage grid, load accordance" inconsistency, the verification was done gathering real data in the field, since the annotation made by the software is performed according to the most probable scenario estimated from the confidence weights informed by the user. The same was done for the "inadequate electrical conductor" inconsistency due the impossibility to assert the correctness of the results proposed by the software using only information available in the database. For the circuits that did not identify inconsistencies such as "transformer, low voltage grid, load accordance", samples were also made for the "partial disconnection" registry error, which in turn was verified through SQL query. For cases of "partial disconnection" from MV circuits, verification was also made via SQL query. For the "total disconnection" inconsistency, both for LV circuits and for MV circuits the verification was made via SQL query, as well as for the "meshed topology in the medium voltage network" inconsistency, given that it was sufficient to identify a mesh in the circuit in order to classify it as incorrect. For the samples whose verification was done via SQL query, BIAT software had 100% precision, sensitivity and accuracy. For the samples whose check is done in the field, the values for the respective figures of merit are not yet completed, since the teams responsible for the procedure have not yet concluded.

Figure 1 illustrates some information useful for management and analysis of the data quality of the DBMS, as well as for planning actions to mitigate insertion of new errors like the ones pointed out by the tool.

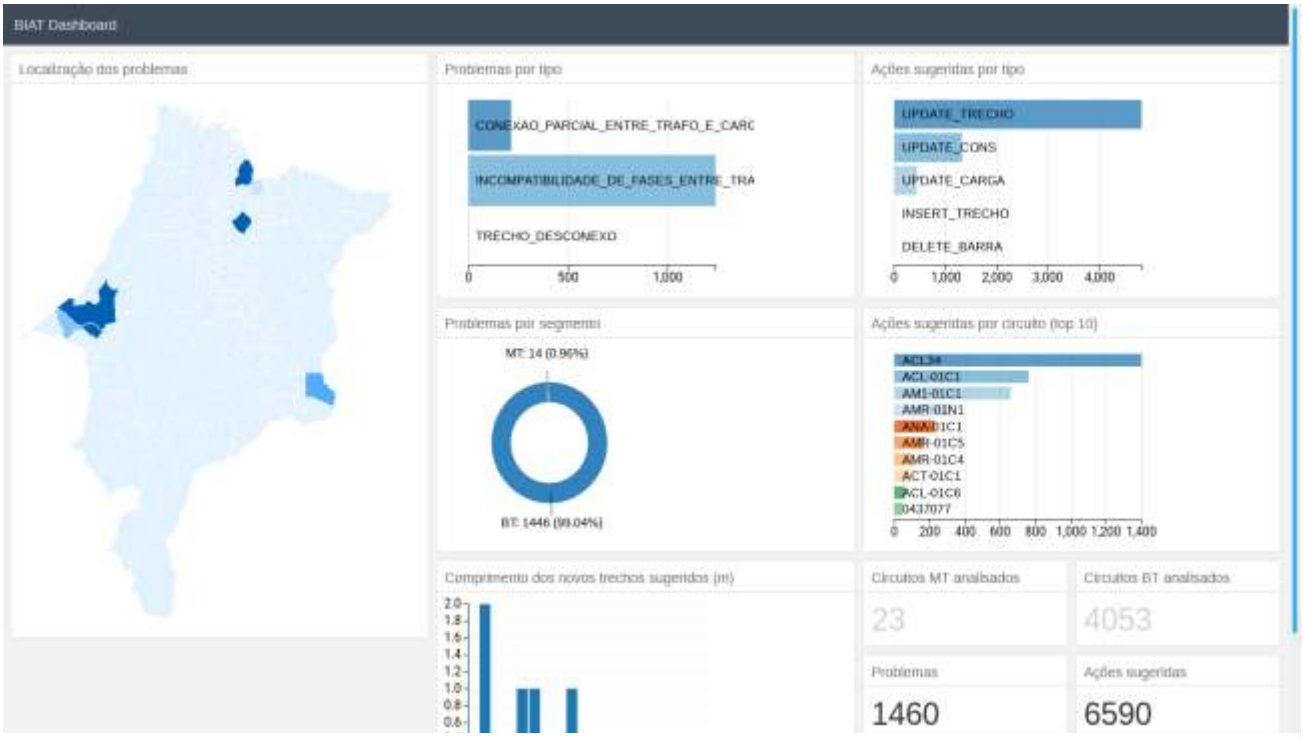

FIGURE 1 – BIAT'S DASHBOARD.

These are presented through a dashboard generated when processing a set of circuits, from low voltage network and medium voltage network, analyzing only a portion of the types of problems. The circuits evaluated are part of 6 substations in Equatorial's concession area. They are: ACL (Açailândia), ACT (Alcântara), ADG (Anjo da Guarda), AM1 (Suprimento Piauí), AMR (Amaranta - Cepisa) and ANA (Anajatuba). There are graphical and intuitive tools to indicated the type errors (in this example, partial connection between transformer and load, incompatibility between phases and disconnected sections) by voltage level (low and medium voltage), besides pointing out repair actions. The tool also geographically indicates the municipality where the errors are located, making it possible to propose accompanying measures and review of the methodologies used to acquire the data.

CONCLUSION

For automation of the process of identification of inconsistency of register in the DSO's database, firstly the errors were characterized by analogous mathematical problems through the use of graphs and similar theory. Subsequently, search algorithms that presented simplicity, effectiveness and efficiency were researched. Using the pre-established classifications and the advantages offered by the selected methodologies, search and correction strategies were developed. This resulted in a simple tool, which is able to identify cadastral inconsistency in the BDGD. Besides identifying them, it is still capable of suggesting fixes with maximum likelihood with the operational philosophy of the energy distributor.

REFERENCES

[1] Reinhard Diestel, *Graph Theory*. New York, NY, USA: Springer-Verlag.

[2] "NBR5440: Transformadores para redes aéreas de distribuição - Requisitos".

[3] *Paper: Second Place Multidimensional Binary Search Trees Used for Associative*. .

[4] K. Paton, "An Algorithm for Finding a Fundamental Set of Cycles of a Graph", *Commun ACM*, vol. 12, nº 9, p. 514–518, set. 1969.

[5] "cycle_basis — NetworkX 1.10 documentation". [Online]. Disponível em: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.cycles.cycle_basis.html. [Acessado: 20-dez-2018].

[6] Davi Mantovani Ricci, Paulo Henrique Baumann, Fabio Romero, André Meffe, Armando H.S.G.Jesus, e Eliezer S. Oliveira, "SIMPLE TECHNIQUE FOR DETECTION OF OUTLIERS IN ONE-DIMENSIONAL NUMERICAL DATA USED FOR POINT OUT ANOMALOUS CONSUMPTION", in *CIRED 2019*, Madrid, Spain, 2019.