Transparently mining data from a medium-voltage distribution network: a prognostic-diagnostic analysis

Matteo Nisi
Department of Electronics and Telecommunications
Politecnico di Torino, Italy
m.nisi@studenti.polito.it

Daniela Renga
Department of Electronics and Telecommunications
Politecnico di Torino, Italy
daniela.renga@polito.it

Daniele Apiletti
Department of Control and Computer Engineering
Politecnico di Torino, Italy
daniele.apiletti@polito.it

Danilo Giordano
Department of Control and Computer Engineering
Politecnico di Torino, Italy
danilo.giordano@.polito.it

Tao Huang
Department of Energy
Politecnico di Torino, Italy
tao.huang@polito.it

Yang Zhang
Department of Energy
Politecnico di Torino, Italy
yang.zhang@polito.it

Marco Mellia
Department of Electronics and Telecommunications
Politecnico di Torino, Italy
marco.mellia@polito.it

Elena Baralis
Department of Control and Computer Engineering
Politecnico di Torino, Italy
elena.baralis@polito.it

ABSTRACT

With the shift from the traditional electric grid to the smart grid paradigm, huge amounts of data are collected during system operations. Data analytics become of fundamental importance in power networks to enable predictive maintenance, to perform effective diagnosis, and to reduce related expenditures. The final goal is to improve the electric service efficiency and reliability to the benefit of both the citizens and the grid operators themselves.

This paper considers a dataset collected over 6 years in a real-world medium-voltage distribution network by the Supervisory Control And Data Acquisition (SCADA) system. A transparent, exploratory, and exhaustive data-mining approach, based on association rule extraction, is applied to automatically identify correlations among SCADA events occurring before and after specific service interruptions, i.e., distribution network faults of interest. Therefore, both the prognostic and the diagnostic potentials of the dataset are investigated with respect to the occurrence of permanent service interruptions. Our results highlight a limited predictive capability of the available set of SCADA events, while they can be effectively exploited for diagnostic purposes.

1 INTRODUCTION

Electric grid operators welcome predictive maintenance to avoid the costs of scheduled inspections and reactive maintenance interventions. To this aim, datasets describing the electric grid operations, with historical data about failures and alarm signals, are under investigation. Although this data has been collected for different purposes, companies are interested in determining their predictive maintenance capability: to reduce management costs, to speed up intervention-time, and to improve efficiency and reliability.

For our study, we rely on a big data dataset spanning over 6 years, collected by a leading Italian electric grid operator. The dataset describes the operations of a medium-voltage distribution network in northeastern Italy, and it records events and failure through the Supervisory Control And Data Acquisition (SCADA) system. Our aim is to assess whether this dataset could be exploited to (i) predict future electric network failures (predictive maintenance) and/or (ii) effectively diagnose the failures after it is reported by the maintenance system. Since the predictive capability of such dataset, and the capability to model system degradation, are unknown, we address the predictive task by means of an exploratory predictive maintenance analysis. To this aim, two exploratory approaches are applied: a statistical data characterisation approach, and a transparent exhaustive method based on association rule mining. The latter, automatically extracts all correlations, above specific thresholds, among SCADA events occurring before each fault of interest (prognostic), and separately, after the faults (diagnostic). Quality metrics are exploited to highlight the most meaningful correlations. Finally, human-readable patterns describing such correlations are investigated.

To the best of our knowledge, our work is the first study that investigates both the prognostic and diagnostic capabilities of a real-world historical dataset collected by a Supervisory Control and Data Acquisition (SCADA) system in an electric grid, with respect to the occurrence of severe service interruptions. Thanks to the application of an exhaustive analysis methodology, by extracting association rules among faults and events, we addressed the issue of providing smart grid operators an assessment of the exploitation potential of currently available datasets for predictive maintenance and diagnosis. The proposed methodology can be applied to similar datasets from any grid operator.

2 DATASET

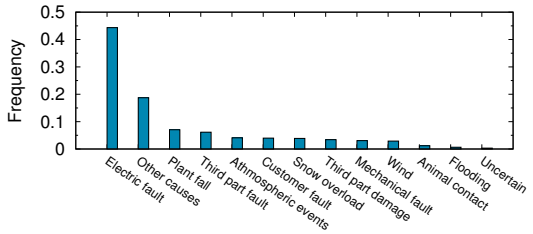
The dataset under analysis contains events recorded by the SCADA system of a leading Italian grid operator, on its medium-voltage distribution network. The dataset is recorded over a period of 6 years (2010-2016), covering two northeastern Italian regions (Veneto and Friuli-Venezia-Giulia). The dataset is characterised by 3,901 faults of interest, 30 different affected components, 153,094 general SCADA events of network operations. The SCADA events are divided into 67 different event types, with the generic failure event type accounting 79,833 events. The faults of interest correspond to those: (i) lasting more than 180 seconds, (ii) with the location in the network identified, and (iii) with the cause determined. These events are named Permanent Service Interruptions (SIPs), tagged with a cause among 45 different reasons and linked to one among the 30 affected components.

We briefly characterise the dataset by analysing the distribution of SIPs causes and types of SCADA events.

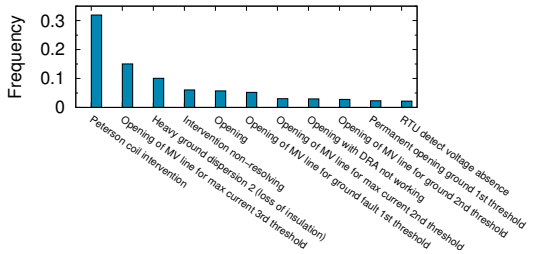
Figure 1a reports the probability distribution of the most frequent causes of SIPs among the 45 available: the top 4 causes account 75% of the SIPs, with “electric fault” being the most

frequent cause (45%). More than 20% of SIPs are due to natural causes, such as: weather issues, plant falls, snow overload, wind, and animal contact. All these causes are unpredictable without contextual knowledge outside the electrical grid operational events. Furthermore, another 20% of SIPs are due to unknown "other causes" (second most frequent value).

Figure 1b reports the probability distribution of the most common SCADA events types. The distribution is skewed, with about 75% of SCADA events belonging to just 6 different types, and with the most frequent one with a frequency above 30%.



(a) Causes of faults (SIPs)



(b) Types of SCADA events

Figure 1: Frequency distribution of the values of (a) causes of faults and (b) types of SCADA events.

3 PROGNOSTIC-DIAGNOSTIC APPROACH

Since this work aims at investigating both the prognostic and diagnostic potential of SCADA events with respect to SIPs, we focus on the analysis of those events occurring both before and after a SIP, in the same portion of the network, under the assumption that the time and space correlations might capture causalities of the system.

3.1 Pre-Fault and After-Fault Windows

In the time dimension, we define a time window preceding the occurrence of a SIP, denoted as *Pre-Fault Window* (PFW), and a time window immediately following the SIP, denoted as *After-Fault Window* (AFW). In the space dimension, we consider only SCADA events observed in the same portion of the network where the SIP occurs, i.e., reported by the same feeder as origin of the collected data, since according to the domain experts they are more likely to be correlated to the considered SIP.

Considering that the grid operator is interested in predicting future SIPs occurring within the next month at most, the time windows are defined with the following variable lengths: 1-7-30 days for PFW, and 1 hour, 1 day or 7 days for AFW. These values result from wider preliminary analyses, with the aim of capturing behaviours of the distribution network at different time scales of interest for domain experts of the electric grid company.

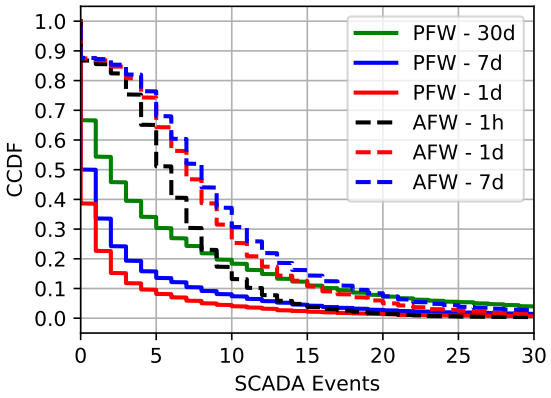


Figure 2: Complementary Cumulative Distribution Function (CCDF) of the number of SCADA events registered during various lengths of PFWs and AFWs.

Figure 2 reports the Complementary Cumulative Distribution Function (CCDF) of the number of SCADA events registered during the PFWs (continuous curves) and the AFWs (dotted curves). Comparing the CCDFs of PFWs and AFWs, in almost 90% of the AFWs at least one SCADA event is observed, even within 1 hour; instead, in 50% of the 7-day PFWs and in 60% of the 1-day PFWs, no SCADA events are registered at all. Furthermore, PFW curves show a more gradual descent with respect to the AFW: SCADA events are more likely to follow a SIP rather than preceding the fault of interest. This data-driven intuition is also confirmed by domain knowledge: many types of SCADA events are known to be triggered by a SIP.

Finally, the 1-hour AFW curve shows a steeper descend than the longer-lasting AFWs, but with the same starting (leftmost) values: most SCADA events are typically observed within the first hour after a SIP, and then few events are collected after 1 or 7 days. On the contrary, the curves of the 7-day AFW and the 30-day AFW show larger differences, since few events are collected in the immediately preceding days of a SIP. Most SCADA events occurring before a SIP are registered in the previous 1-7 days. Although few additional events are observed considering a 30-day-PFW, we also note that a higher number of SCADA events in the PFW correlates with a higher probability of registering another non-permanent service interruption during the same PFW (results missing due to space limitations, partially discussed in Section 3.2), so a significant portion of the 30-day-PFW events could be ideally associated to AFWs of those minor service interruptions.

All considerations tend to suggest a limited *prognostic* potential of the SCADA events with respect to SIPs due to fewer events, more time-unrelated, also considering the high variety of SCADA event types. Conversely, the *diagnostic* exploitation seems better supported by more data, nearer to the event of interest.

3.2 Inter-Fault Window

We define *Inter-Fault Window* the time interval between two consecutive faults on the same portion of the network, denoted as IFW. The aim of such analysis is to determine how many events following a SIP, i.e., in its AFW and inherently diagnostic, are also included in a PFW before another SIP, thus being modelled also as prognostic features. Both SIPs and other minor Service

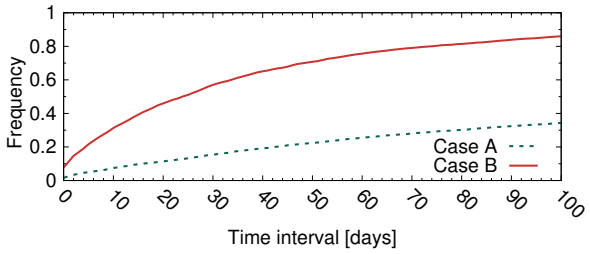


Figure 3: IFW lengths of various types of faults.

Interruptions generate diagnostic SCADA events in their AFWs, hence different IFWs can be defined, depending on the type of faults considered (SIPs only or all Service Interruptions). Figure 3 shows the probability distribution of the duration of two types of IFWs:

- Case A (dotted green curve): IFW between each pair of consecutive SIPs.
- Case B (continuous red curve): IFW between each registered SIP and the immediately preceding Service Interruption of any type (either SIP or not).

In 80% of cases, the IFW between two consecutive SIPs lasts more than 40 days, and there is only a 7% probability that two SIPs are separated by an interval of less than 7 days (Case A). Hence, with a 7-day PFW, we limit the interference of AFWs of other SIPs into the PFW of the current SIP under analysis, by guaranteeing that prognostic and diagnostic events are kept separate for different SIPs.

However, in Case B, the duration of the IFW between a SIP and the immediately preceding Service Interruption lasts up to 30 days in almost 60% of the cases, with the probability of having an IFW shorter than 7 days risen to 26%, three-fold with respect to Case A. Hence, there exist SCADA events registered during a PFW preceding a SIP that are generated as a consequence, i.e., in the AFW, of a previously occurring Service Interruption.

3.3 Challenges

From the time-window-based data characterization, the following takeaways can be identified:

- 60% of the SIPs have no SCADA events in their 7-day PFW.
- 10% of the SIPs have no SCADA events in their 1-day AFW.
- Most diagnostic events occur in the 1-hour AFW.
- Many apparently-prognostic events occur more then 1 week before the SIP (PFW), however, they include events generated as a consequence of other minor faults, i.e., they are in the AFW of non-permanent Service Interruptions, in 60% of the cases for a 30-day PFW, and in 26% of cases for a 7-day PFW.

4 RULE MINING

To address challenges identified in Section 3.3, we exploited a transparent, exhaustive and exploratory data mining approach: association rule mining. The technique and its evaluation metrics, as required by the scope of the current work, are defined as follows.

4.1 Association Rule Extraction

Let $\mathcal{D}$ be a dataset whose generic record r consists of a set of co-occurring events, i.e., events that occur in the same time window. Each event, also called *item*, is a couple *(attribute, value)*. In the current work, the *attribute* is either a SCADA event type, or an alleged cause, or a failed component, and the *value* is 1 if that attribute is true in the time window under exam (e.g., the SCADA event is present, the component failed, or the specific cause was determined), or 0 otherwise. Note that a SCADA event might represent another SIP or a minor fault occurring before or after the analyzed SIP. An *itemset* I is a set of co-occurring events, failed components, and alleged causes among the records r in the dataset $\mathcal{D}$. Such set of items I in a PFW or, separately, in an AFW constitutes the input feature vector of the rule mining extraction.

The *support count* of an itemset I is the number of records r containing I. The *support* $s(I)$ of an itemset I is the percentage of records r containing I with respect to the total number of records r in the full dataset $\mathcal{D}$. An itemset is *frequent* when its support is greater than or equal to a minimum support threshold $MinSup$.

Association rule mining aims at identifying collections of itemsets (i.e., sets of co-occurring events) that are frequently present in the dataset under analysis, according to statistically relevant metrics. The extracted rules are all and only those adhering to the thresholds of statistical relevance defined as parameters of the mining process, hence being an exhaustive, thus powerful, exploratory approach within the boundaries of the problem formulation (i.e., itemset definition and threshold settings).

Association rules are usually represented in the form $X \rightarrow Y$, where X (rule antecedent) and Y (rule consequent) are disjoint itemsets (i.e., they include different attributes). To identify the most meaningful rules among those extracted by the mining process, quality measures can be exploited as ranking criteria. The following popular quality measures are used in the current work: rule support, confidence, and lift. *Rule support* $s(X, Y)$ is the percentage of records containing both X and Y. It represents the prior probability of $X \cup Y$, i.e., the support of the corresponding itemset $I = X \cup Y$ in the dataset. *Rule confidence* is the conditional probability of finding Y given X. It describes the strength of the implication and is given by $c(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$ [5].

All and only association rules with support and confidence above (or equal to) a support threshold $MinSup$ and a confidence threshold $MinConf$ are to be extracted. Among those surviving the thresholds, a rank based on descending support, confidence and lift values can drive the attention to focus on the most statistically-relevant patterns. The *lift* [5] of a rule $X \rightarrow Y$ measures the (symmetric) correlation between antecedent and consequent, and it is defined as follows.

$$\text{lift(X, Y)} = \frac{\text{c}(X \rightarrow Y)}{\text{s(Y)}} = \frac{\text{s}(X \rightarrow Y)}{\text{s(X)} \cdot \text{s(Y)}} \quad (1)$$

In Equation (1), $c(X \rightarrow Y)$ and $s(X \rightarrow Y)$ are the rule confidence and support; $s(X)$ and $s(Y)$ are the supports of the rule antecedent and consequent, respectively. If lift(X,Y)=1, itemsets X and Y are not correlated, i.e., they are statistically independent. Lift values below 1 show a negative correlation between itemsets X and Y, while values above 1 indicate a positive correlation, with higher lift indicating stronger rules, hence typically more meaningful and interesting correlations.

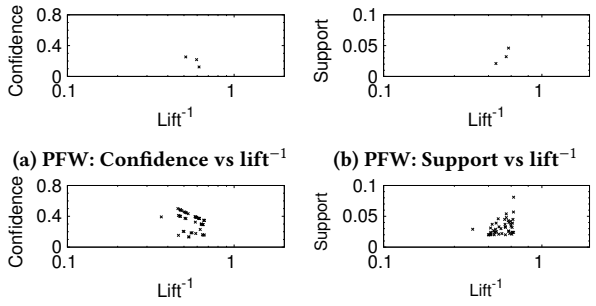
4.2 Rule quality analysis

The analysis of the extracted rules has been performed for various parameter values. Due to space constraints, we report only the most meaningful results based on the rules obtained by (i) setting $MinSup$ 0.02, then focusing on rules (ii) whose lift is higher than 1.5, and (iii) having a cause or component as conclusion.

The number of rules resulting from such selection have been reported in Figures 4a-4b for a 7-day PFW. They are scatter-plotted according to support, confidence and lift values. For comparison, the same results have been reported in Figures 4c-4d for an AFW of 1 day. The *diagnostic* potential (AFW) is confirmed by a larger number of correlations with better quality metrics with respect to the *prognostic* capability (PFW):

- 45 rules extracted in the AFW vs 3 in the PFW.
- 50% max rule confidence in AFW vs 25% in PFW.
- 2.73 max lift value in AFW vs 1.9 in PFW.
- 8% max support in AFW vs 4.5% in PFW.

Eventually, top rules according to lift, confidence and support have been inspected by domain experts from the grid company, allowing to transparently evaluate the correlation model and the prognostic-diagnostic approach.



(a) PFW: Confidence vs lift^{-1} **(b) PFW: Support vs lift^{-1}**

(c) AFW: Confidence vs lift^{-1} **(d) AFW: Support vs lift^{-1}**

Figure 4: Association rules extracted from the 7-day PFW (a-b) and from the 1-day AFW (c-d), with causes or components as conclusion (x-axis in log scale).

5 RELATED WORK

With the shift from the traditional electric grid to the Smart Grid paradigm, data analytics and related applications are becoming of fundamental importance in power networks, as shown by the several studies available in the literature focusing on this topic [6, 9]. However, few research efforts have been specifically devoted to predictive maintenance. Some studies aim at performing fault detection in power networks, based on historical weather data mining [7], on extreme learning machine models [10], or on electrical feature extraction techniques [2]. Authors in [4] deploy an effective method to detect faults in smart grids, trading off the need for reducing the huge volume of available collected data, related to the Phasor measurement unit, and the need for keeping critical information. Other studies aim not only to detect faults, but also to further characterise them by identifying and exploiting significant features. Classifiers based on clustering and dissimilarity learning techniques [3] or on feature extraction algorithms [1] are used to analyse massive data to perform fault recognition or distribution fault diagnosis. The deployment of fault detection methods with prognostic purposes is not well investigated in the literature. Authors in [8] aim at reducing the outages in Medium Voltage distribution networks by exploiting rule-based, data mining and clustering techniques to design a method providing diagnostic and prognostic functions for Distribution Automation systems.

6 CONCLUSIONS

The work analysed 6 years of data recorded from a medium-voltage distribution network, with the purpose of estimating both the prognostic and diagnostic potential for severe faults, i.e., permanent service interruptions. Time-window data characterisation and exhaustive rule-mining results confirm the capability of the collected data to support diagnostic tasks, whereas their prognostic potential is limited since only few and poor predictive correlations are present in the data. Future works include wider analyses of the rules for different thresholds and changes into the transactional dataset derived from the raw data to enable the extraction of additional correlations. Finally, further investigations of the predictive capability will be performed by testing the effectiveness of the obtained rules in detecting actual failures.

ACKNOWLEDGMENT

The research leading to these results has been funded by Enel Italia, e-distribuzione, and the SmartData@PoliTO center for Data Science technologies and applications.

REFERENCES

[1] Y. Cai and M. Chow. 2009. Exploratory analysis of massive data for distribution fault diagnosis in smart grids. In *2009 IEEE Power Energy Society General Meeting*. 1–6.

[2] Q. Cui, K. El-Arroudi, and G. Joos. 2017. An effective feature extraction method in pattern recognition based high impedance fault detection. In *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*. 1–6.

[3] Enrico De Santis, Lorenzo Livi, Alireza Sadeghian, and Antonello Rizzi. 2015. Modeling and Recognition of Smart Grid Faults by a Combined Approach of Dissimilarity Learning and One-class Classification. *Neurocomput.* 170, C (Dec. 2015), 368–383.

[4] Huaiguang Jiang, Xiaoxiao Dai, Wenzhong Gao, Jun Zhang, Yingchen Zhang, and Eduard Muljadi. 2016. Spatial-Temporal Synchrophasor Data Characterization and Analytics in Smart Grid Fault Detection, Identification and Impact Causal Analysis. *IEEE Transactions on Smart Grid* 7 (09 2016), 1–1.

[5] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley.

[6] Chunming Tu, Xi He, Zhikang Shuai, and Fei Jiang. 2017. Big data issues in smart grid – A review. *Renewable and Sustainable Energy Reviews* 79 (2017), 1099 – 1107.

[7] Jian Wang. 2016. Early warning method for transmission line galloping based on SVM and AdaBoost bi-level classifiers. *IET Generation, Transmission and Distribution* 10 (November 2016), 3499–3507(8). Issue 14.

[8] Xiaoyu Wang, Stephen McArthur, Scott Strachan, John D. Kirkwood, and Bruce Paisley. 2017. A Data Analytic Approach to Automatic Fault Diagnosis and Prognosis for Distribution Automation. *IEEE Transactions on Smart Grid* PP (05 2017), 1–1. https://doi.org/10.1109/TSG.2017.2707107

[9] Yang Zhang, Tao Huang, and Ettore Francesco Bompard. 2018. Big data analytics in smart grids: a review. *Energy Informatics* 1, 1 (2018), 8.

[10] Y. Zhang, Y. Xu, Z. Y. Dong, Z. Xu, and K. P. Wong. 2017. Intelligent Early Warning of Power System Dynamic Insecurity Risk: Toward Optimal Accuracy-Earliness Tradeoff. *IEEE Transactions on Industrial Informatics* 13, 5 (Oct 2017), 2544–2554.