Modelling and forecasting incidents in cellular networks

Daniel Almeida
Instituto Superior Técnico / INESC-ID
University of Lisbon
Lisbon, Portugal
daniel.andre.almeida@tecnico.ulisboa.pt

Luis M. Correia
Instituto Superior Técnico / INESC-ID
University of Lisbon
Lisbon, Portugal
luis.m.correia@tecnico.ulisboa.pt

***Abstract*—This paper addresses the opportunity to study the number of incidents in an operator's network, NOS, depending on the meteorological factors. One developed a statistical study which correlates the number of incidents with weather factors in each region of Portugal. It is also developed a forecasting study, in order to better predict the number of incidents with a focus on the peak day. One concludes that each region has a different behaviour regarding the weather variable that is most related to the number of incidents, leading to better results when using data from Regions. Regarding the forecasting study, the best results appear when applying the NARX Neural Network. However, the output of this work cannot be applied in a real operation, due to the lack of corrected peaks hit. Though, it is the first step into a study to be implemented in the real world.**

***Keyword*—Alarms, Incidents, Faults, Correlation, Forecasting, Neural Networks.**

I. INTRODUCTION

The popularity of mobile devices has been increasing, as well as the demand for mobile communication technology, explained by the growth of low-cost cell phones and the improvements in network coverage and capacity. As the communications systems became more complex, the task of identifying and correcting faults in the network has turned into a critical task of a network management.

A fault that could interfere with the services provided by the operator is costly. The detection of these before users can suffer from service degradation is a needed requirement for a proper communication system. Since it is not possible to avoid all faults in these systems, their detection and correction are essential.

In fault management, there are some basic concepts essential to introduce. However, there is no standard in naming these notions: The Alarm, is an exceptional condition occurred. The Root Cause indicates the origin of the abnormal condition. The Incident (also referred as fault and root cause), is a malfunction of the system that could trigger several alarms. Finally, the Ticket, that in most cases are also referred as an alarm is the notification received by the network administrator.

These failures have a significant impact on both operator and customers. For the first, it must provide excellent service to its clients. For the second, wants to be able to call when wanted, and to make it correctly. To achieve this, the operator needs to organize its workforce to respond in a quick way to the incidents that may exist. By knowing additional information about incidents, the operator saves time and money due to the workforce organization and with the customers' satisfaction.

A prediction of the number of incidents is information even more relevant to the operators, due to the importance of critical decision inside the organization. Besides, it is also important the study of the peaks of incidents, which are days of an unusual quantity of incidents.

The major relevance of this study relies on the fact that exists insufficient information about the importance of weather in the occurrence of incidents in a telecommunication network.

The goal of this work is to accomplish a statistical study to relate the meteorological variables with the incidents, as well as a forecast of the number of incidents. The first study will be realized by measuring the importance of meteorological factors and planned works in the occurrence of incidents. The second study relies in the forecasting on the number of faults. One accomplishes both studies using the data of each region of Portugal in separate, and then, using the data compiled.

This study was done in collaboration with NOS, a network operator in Portugal. The conclusions of this work are intended to give additional information to the operators, to be possible an understanding of how the meteorological factors and the planned works affect the number of incidents in NOS's network.

In Section II one introduces the fundamental concepts regarding this work. Section III presents the description of the dataset, as well as the processing needed to this data. One also shows the statistical and forecasting implementation, as well as the assessment of these methods. In Section IV one demonstrates the scenarios used in work, with its analysis. Are also presented the results of the statistical and forecasting implementation. Finally, in Section V, the conclusion is shown.

II. FUNDAMENTAL CONCEPTS AND STATE OF THE ART

A. GSM, UMTS and LTE [1][2][3][4][5][6]

The need for using the same radio access network either in GSM and UMTS leads towards one architecture which can be efficiently integrated into a single UMTS multi-radio network. This architecture is composed by the GERAN and UTRAN, responsible for all radio-related functionalities at GSM and UMTS, respectively. The remaining elements are the Core

Network, responsible for switching and routeing calls, the User Equipment for UMTS, and Mobile Stations, for GSM, being the interface that connects the user to the rest of the network.

GSM standard is based on a Multi-Carrier, TDMA and FDD modes. A frame is subdivided into eight full slots, and one slot is equal to the one-time slot on one frequency, is the data transmitted in one slot denoted as a burst.

WCDMA is used as the radio interface of UMTS, and it is a wideband DS-CDMA system. In this system, user information bits are spread over a wide bandwidth by multiplying this user information with chips.

Regarding LTE, with the commitment for packet switched services optimization, and improvements in the user bit rates led the discussion for System Architecture Evolution. Some of the evolutions were made to involve fewer nodes to reduce latency and improve performance. These improvements were, for example, that the CS part of the network disappeared, and that LTE only support PS services.

The downlink multiple access is based on the OFDMA, and the uplink access relies on the SC-FDMA. This technique for radio transmission and reception is a powerful way to minimize the problems of fading and Inter-Symbol Interference.

B. Modelling alarms and incidents

An alarm notification can be described, according to [7], with a set of valuable information that the network administrators are alerted, to prevent the service outage or degradation, described by five conditions: Resource, Alarm Type, Time, Severity and Information.

The Severity parameter is used to range the malfunctions. These ranges are Critical, Major, Minor and Warning, organized from the most to the least severe, as stated in [8]. The Critical parameter indicates the need for an immediate corrective action, and the Warning it is only the detection of a potential service in fault. One can also describe them as cleared and indeterminate.

The origin of these alarms can be considered into five categories, according to [8]. Communication, associated with the procedures to carry information. Quality of Service, related to the degradation of the quality of service. Processing, associated with a software or processing. Equipment, identified with an equipment fault, and finally, Environmental, related to the condition in which the equipment resides.

According to [9], the primary sources of these failures are the human errors and acts of nature. In a study, [9] describes that 28% of downtime was caused due to human errors, with 150 outages in two years. Regarding acts of nature, the values decrease to 18% of downtime and 32 outages. One can conclude that despite acts of nature have fewer outages, the severity is highly superior, leading to a higher period of downtime.

By the use of a management centre, the flux of alarms can be correlated, reducing the amount of information presented to the network administrators. This processing is achieved by removing redundant information or filtering low-priority alarms, according to [10].

C. Failure Prediction approach

To make the best system prediction, one needs to characterise the data used. As reported by [11], the arrival of faults can be considered as a statistical process in time, shown as an ordered time series. This time series, as stated in [12], is multivariate (multiple data for one period) and stochastic, since the future results can only be estimated. One can also describe the time series is non-stationary due to the high-level of daily fluctuations. Concerning the linearity, it is annotated that the variables in the system could be both linear and nonlinear.

Linear time series are well described by ARIMA, while nonlinear time series more adequately describes neural networks (NN). The efficiency of NN is strongly dependent on the inputs.

To better understand the association of the number of incidents with weather variables, one needs to comprehend the relationship between these variables. A possible method to quantify this relation is by the calculation of the correlation coefficient. This coefficient is typically calculated by three methods: Pearson [13], Spearman [14] and Kendall's τ [15]. These methods outcomes a number between -1 and +1, expressing how closely the two variables are related. The ±1 shows a perfect relationship and 0 indicated no connection.

The Pearson correlation is, according to [13], ideal if the data follow a bivariate normal distribution, being a method vulnerable to data deviation of any kind. According to [16], the Pearson coefficient r is calculated by (1).

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \ \sum(Y_i - \bar{Y})^2}} \tag{1}$$

where:

- X_i: Represents the dataset X in index i;
- Y_i: Represents the dataset Y in index i;
- $\bar{X}$: Represents the mean of dataset X;
- $\bar{Y}$: Represents the mean of dataset Y.

Spearman method provides a distribution-free measure of correlation between two variables. In the Spearman coefficient, the ranks of the sorted values determine the result. Thus, the data is first sorted and then computed, by (2), according to [14].

$$r_s = 1 - \frac{6 \ \sum\left(R_p - T_p\right)^2}{n^3 - n} \tag{2}$$

where:

- R_p: Represents the rank of dataset X in position p;
- T_p: Represents the rank of dataset Y in position p;
- n: Represents the number of ranks.

The Kendall's τ method, quite similar to the Spearman method, also measures the range of increasing or decreasing relationships between the pairs of variables monotonically. According to [15], one can define the concordant pairs by (3).

$$Y_i < Y_j \ if \ X_i < X_j \ \vee \ Y_i > Y_j \ if \ X_i > X_j \ \vee \ (X_i - X_j)(Y_i - Y_j) > 0 \tag{3}$$

where:

- X_j: Represents the dataset X in index $j \neq i$;

- Y_j: Represents the dataset Y in index $j \neq i$.

Similarly, one can define the discordant pairs by (4).

$$Y_i < Y_j \; if \; X_i > X_j \; \vee \; Y_i > Y_j \; if \; X_i < X_j \; \vee \; (X_i - X_j)(Y_i - Y_j) < 0 \tag{4}$$

To calculate the rank correlation, one uses (5).

$$\tau = \frac{2\,(P - Q)}{l(l-1)} \tag{5}$$

where:

- P: Represents the number of concordant pairs;
- Q: Represents the number of discordant pairs;
- l: Represents the actual size of the sample.

To forecast the number of incidents, one starts with the simplest method to accomplish this study, the linear regression, one generically defined by (6), according to [17].

$$y_{predictand} = b_0 + m_1 x_1 + m_2 x_2 + \cdots + m_k x_k \tag{6}$$

where:

- $y_{predictand}$: Predictand;
- b_0: Regression constant;
- m_k: Regression coefficient;
- x_k: Predictor.

In a specific case, when representing the number of incidents versus two variables, one uses a variation of the regression equation, described by (7).

$$y_{surface} = b_0 + m_1 x_1 + m_2 x_2 + m_3 x_1 x_2 \tag{7}$$

where:

- $y_{surface}$: Regression surface.

However, in more complex cases, another forecasting models are used. For example, in data similar to the one studied in this work, NN presents reliable results. The most widely used neural network is the multi-layer perceptron (MLP), one defined by (8), according to [18].

$$y_{MLP} = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\,(\beta_{0j} + \sum_{i=1}^{u} \beta_{ij} Y_{t-1}) + \varepsilon_t, \forall t \tag{8}$$

where:

- y_{MLP}: Output of MLP;
- Y_{t-1}: Inputs of MLP;
- α_0, β_{0j}: Bias term;
- u: Number of inputs;
- q: Number hidden nodes;
- α_j,β_{ij}: Connection Weights;
- ε_t: Random Shock.

Support Vector Machines (SVM) is also a method to describe time series. The primary objective is to find a decision rule capable of selecting a subset of training data, as stated in [18]. One can mathematically define the SVM in two cases, first when the data is linearly separable, and the second into nonlinearly separable data. The first is defined by (9), according to [18].

$$\left.\begin{array}{ll} Minimize & \frac{1}{2}\|w\|^2 \\ Subject\ to & Y_i(w^t X_i + b) \geq 1;\ \forall i = 1,2,\ldots,v \end{array}\right\} \tag{9}$$

where:

- w: Weight vector;
- (Y_i, X_i): Input-Output pair;
- b: Bias term;
- v: Number of vectors.

Regarding the case where the data is nonlinearly separable, e.g. XOR classification, one is mathematically defined by (10), according to [18].

$$\left.\begin{array}{ll} Minimize & \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{N} \xi_i\right) \\ Subject\ t & Y_i(w^t X_i + b) \geq 1 - \xi;\ \forall i = 1,2,\ldots,v \wedge \xi_i \geq 0 \end{array}\right\} \tag{10}$$

Where:

- ξ: Slack variables;
- C: Regularization constant.

Regarding Bayesian Networks, [19] refers it as a convergence of Artificial Intelligence and Statistics, due to the creation of a probabilistic model, that can be used to query possible outcomes from the input data. One can define a Bayes network by (11), according to [20].

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v=1}^{n} P(X_v = x_v | X_{v+1} = x_{v+1}, \ldots, X_n = x_n) \tag{11}$$

where:

- X_n: Dataset in index n;
- X_v: Dataset in index v.

Another method is the Nearest Neighbours, which determine a point in a dataset that is nearest to a query point, according to [21]. This is accomplished by examining the distribution of the distance between the query and data points. The identification of the distance is made by an evaluation of the number of points that are longer a factor of the distance between the query point. It is typically used the Euclidean distance, defined by (12), to proceed the classification, as stated in [22].

$$d(X, Y) = \sqrt{\sum_{i} (X_i - Y_i)^2} \tag{12}$$

where:

- $d(X, Y)$: Euclidean distance between two points.

In a study, [12] concludes that the best results to forecast faults appeared when applying dynamic models. One,

Nonlinear Autoregressive Network with Exogenous Inputs (NARX), demonstrates superior performance. According to [23], one formalizes NARX by (13).

$$y(t) = \psi(u(t-n_u), \dots, u(t-1), u(t), y(t-n_y), \dots, y(t-1)) \quad (13)$$

where:

- $u(t)$: Input of the network at time t;
- $y(t)$: Output of the network at time t;
- n_u: Input order;
- n_y: Output order;
- ψ : Nonlinear function.

According to [24], when the function $\psi(.)$ is approximate by an MLP, the resulting system is the NARX.

To evaluate the accuracy of the models, one uses the Mean Squared Error (MSE), being one of the most common performance measure used, defined by (14), as stated in [18].

$$\overline{\varepsilon^2} = \frac{1}{l}\sum_{t=1}^{l}(Y_t - f_t)^2 \quad (14)$$

where:

- Y_t: Represents the actual value;
- f_t: Represents the forecasted value.

Regarding the performance of linear regression, the coefficient of determination gives one simple fit indicator, R^2. According to [25] this provides a reasonable and rapid model fit indication and can be computed by (15).

$$R^2 = 1 - \frac{\sum_{i=1}^{l}(X_i - \widehat{X}_i)^2}{\sum_{i=1}^{l}(X_i - \bar{X})^2} \quad (15)$$

where:

- $\widehat{X}_i$: Predicted value for x_i.

Another method to define the performance is the standard deviation. According to [17], this method defines the square root of the average squared difference between the data points and their sample mean.

D. State of the art

With the increased use of the mobile phones, not just to make phone calls but even more to access the Internet, the network is growing in complexity, and then, producing more alarms. The minimization of failures with proper design and preventive maintenance is becoming more critical. The opportunity of acting preventively and proactively enhances the motivation of using incidents prediction to understand possible failures, to be possible a quick reaction of solving these problems.

Some methods in cellular networks for correlating the alarms are studied. In [26], a framework to reduce the number of alarms to the network administrators is presented. This framework consists of the representation of the systems and devices as nodes in a graph, and then, if a device fails, by the traversing the graph, it is possible to reach the node which caused the fault.

In [27], it is done a prediction of the expected number of failures in the network, by modelling of the number of failures as a time series. By applying a statistical method, some elements such as lightning or rainfall are identified as the most significant predictable causes of faults. Using the results from [12], it is proposed the used of the NARX as the most likely network for predicting quantities of reported failures in complex systems.

Another use of NARX network is the study of turbines, as possible to see in [28]. The use of this kind of network was decided due to the capability of capturing dynamics of complicated systems, as in the case of the gas turbines.

The influence of weather is extensively studied in other areas. One of them is the effect in electrical distribution networks. Reference [29] present us the study of the network's outages in electric facilities, mainly related to weather factors.

The other major area of study about the influence of weather on the people's lives are the health issues. Reference [30], for example, presents the study of how the weather affects the mood of the citizens. Another example is presented in [31], similar to [30], with a correlation study about the weather variables regarding patient's headache.

III. Dataset and Implementation Description

A. Data Description

The data used can be divided into two groups: Incidents and Meteorological. The first, provided by the Portuguese operator NOS [32], is consisted by two files. The first congregate the information from the incidents and the second the base stations' localization. The meteorological data contains information received by Weather Underground [33] and from *Instituto Português do Mar e Atmosfera* (IPMA) [34]. The files received by the first covers the meteorological information except for electric discharges. Regarding the files received by IPMA, these are representative of one of the three meteorological stations capable of collecting electrical discharges.

To obtain the meteorological data from Portugal, one uses the API for developers, provided by Weather Underground. One developed an application based on JSON to receive the information, incorporated with a Python application to save the data in an Excel file.

B. Data processing

Since it is used data from multiple sources, one needs to structure the several steps to organize and process the data from each source to obtain the final file. This processing comprises several procedures, including the removal of information not needed to structure the information, one described in Fig. 1.

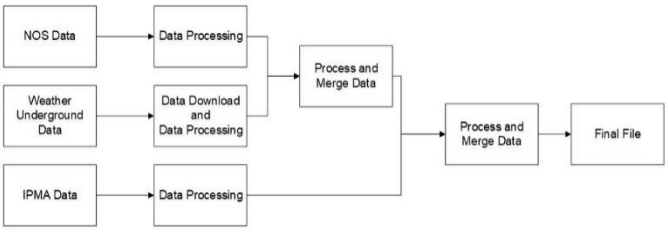

Fig. 1. Steps for processing the data.

Two major objectives compose this thesis, the statistical and forecasting studies. The first implements a study on the

influence of each meteorological factors in the occurrence of incidents. The second is the study of different datasets and methods to predict the number of incidents. One described in Fig. 2 both implementations and the studies accomplished.

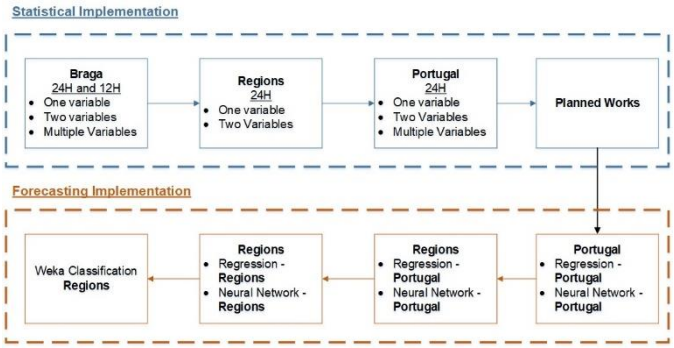

Fig. 2. Schematization of processes.

The first step for the statistical study is the calculation of the correlation coefficient. However, to achieve this, one needs several intermediate steps. To start, one needs to understand the physical location of each incident, achieved by an application which relates the incidents' file with the base stations' location.

Knowing where the incident occurred, one needs to relate this information to the meteorological data. To organize the data, and for the 12-hour interval, one needs to establish the maximum values of each variable in this interval.

Regarding the electrical discharges data, the first step is to discover the location of the region of each discharge, which appears with GPS coordinates. To accomplish this, one uses an API of Google Maps. The last step is the calculation of the maximum value of intensity and the number of discharges during the time intervals studied.

Having the information about weather variables processed, one needs to finalize the processing in incidents file. The first step is the calculation of the number of incidents by day and merging the weather data. One uses the same application to calculate the planned works incidents.

The last step necessary is to relate the information between electrical discharges and incidents by the time that each one occurred, in both files. With this last step, one has an organized file to process the remaining studies. The same procedure is done for each region of Portugal. Then, this data is compiled to form the Portugal file.

C. *Statistical Implementation*

The first step for the statistical tests of the number of incidents is their relationship versus one weather variable. This relationship is studied in two parameters: Correlation and Regression, both presented in Section II.C. The first method is achieved by the use of SciPy [35] libraries on Python. Regarding the regression method, is completed by the utilization of Excel chart tool. One represents a scatterplot and then, calculates the regression equation and determination coefficient.

Regarding the two variables study, one accomplishes the simulation using Matlab, using the example presented in [36], and using (7). To have a method to compare the results, one calculated the maximum value that the regression surface has.

Finally, for the more than two variables study, one uses the SPSS software to calculate the regression coefficients for each variable. For this study, it is used the actual values for each meteorological factor.

D. *Forecasting implementation*

The simplest procedure to implement a prediction is using a regression equation. To complete the forecasting, it is used the equations provided by the SPSS. Then, to assess this method, one uses Excel to assess each equation. This procedure is presented in Fig. 3.

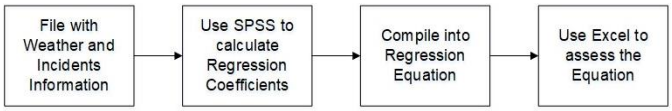

Fig. 3. Regression Equation process.

To apply the NARX network, one uses the Matlab Neural Network Toolbox, where this network is implemented. To use this method, it is needed the division of the data into three subsets, as described in [37]. These subsets are the training, validation and test. The training part, which represents 70% of the data, is used to computing the gradient and updates the network weights and biases. The validation subset, representing 15% of the data, is when the validations errors are being monitored during the training part to validate the training. Finally, the test subset, representing 15% of the data, is used to compare different models, plotting the errors during the training process. The data is randomly divided into each of these three subsets at each time the network is trained, leading to different outputs each time the network is trained and validated.

NARX is trained using a second-order algorithm, the Levenberg-Marquardt algorithm, due to the increased training speed compared to other algorithms. The inputs to train and validate the network can be divided into two files, one referring to the objective of the network, the number of incidents. The second, input, related to the several weather variables.

The next step is to set the size of the network, where it is possible to change the number of neurons and the delay. To achieve the best network, one trains the network with several numbers of neurons and delays. After each training, it is necessary to consider the MSE, that is different in each training, of the three subsets. Due to this, it is necessary several pieces of training until reaching the optimal network. One described in Fig. 4 the representation of the network in training.

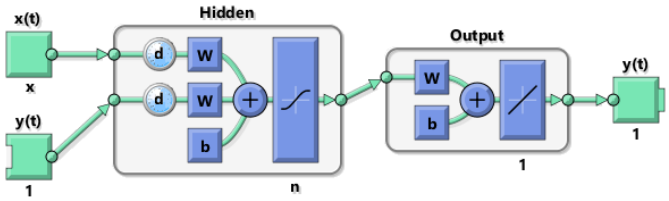

Fig. 4. Training Network.

The remaining forecasting simulation is implemented by the Weka [38] software. Weka collects a set of machine learning algorithms for predictive and classification tasks. In this work, one only used the classification option.

The first step for the Weka classification is the division per classes of each file, according to the number of incidents that occurred. One divides the file into three categories, A, B and C. The A class refers to the days with a low quantity of incidents, the B class the intermediate days, and finally, the C class represents the days of an abnormal number of incidents.

To accomplish the training, one uses the option of Weka to cross-validation the data ten times. This means that the data is divided into ten equal parts, and then, uses nine of these parts to realize the training and the last section to test. This is done ten times, using every time a different part to test. The performance of each method is thus, the mean of all simulations. One described in Fig. 5 the process of Weka classification.

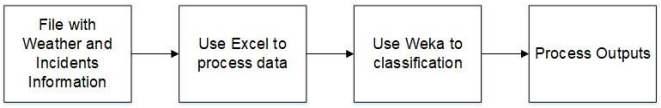

Fig. 5. Weka process.

E. Forecasting assessment

To assess the forecasting methods, one uses two approaches. The first is the study of the results that the methods calculate. The second is the study of peaks of incidents, to recognize how these methods behave when the severity of incidents increases.

For the Weka classification, one uses only the data from each Region. To assess this classification, one accomplishes three studies. The first is the percentage of corrected classification, in a global way. The second is the amount of corrected and false peaks detected. The last one is the percentage of corrected classification by each of the three classes.

Concerning the study of NN and the regression, one calculates the MSE to assess each method. Regarding the study of the peaks of incidents, one calculates the mean and the maximum error together with the peaks that the method hits.

One of the forecasting evaluation uses the error of false peaks, considered between the number of false peaks and the number of corrected forecasted peaks, calculated by (16).

$$e_{fpeaks\,[\%]} = \frac{f_{peaks}}{c_{peaks} + f_{peaks}} \times 100 \quad (16)$$

where:

- e_{fpeaks}: Error between false and total forecasted peaks;
- f_{peaks}: Quantity of false peaks;
- c_{peaks}: Quantity of corrected peaks.

IV. RESULTS ANALYSIS

A. Scenario Description

The scenario is composed by a dataset from 1st January of 2016 until 28th of February 2017 in Continental Portugal. The scenario was decided in a meeting with NOS [32]. One describes in Table 1 some information about the incidents dataset.

After a brief analysis of the data, one concludes that the number of incidents is directly related to the number of base station's sectors in each region. However, it does not mean that with the increase of base stations, the number of incidents also increase. To better understand this behaviour, one conducted a study to relate the ratio of incidents per base station.

Table 1. Information about dataset used

Incidents	Planned Works	Incidents in study	BS sectors	Incidents per BS sectors	Incidents per 1000 inhabitants
40037	1190	16238	6081	2.67	1.58

One of the conclusions drawn is that the two regions with more incidents and base stations, Lisbon and Porto, are the ones with fewer incidents per base station. This can be explained due to the importance of those regions by the operator. On the other position are Portalegre and Évora. This can be explained due to the position of some base stations in locations sometimes unprotected from natural elements.

Another conclusion that can be drawn is that the number of incidents per region size in Lisbon and Porto is significantly higher than the rest of the regions. One explained because these regions are hugely populated with people and base stations. On the other hand, Beja is the region with less ratio, explained by its dimension. For the inhabitants' study, the conclusions are similar, with Lisbon and Porto represented as some of the best regions for this ratio. However, Braga represents the best region. After a brief processing of the incidents file, one can relate some of the cause-effect in the occurrence of incidents.

Table 2. First processing in cause-effect problems.

Cause	Effect	Examples
Planned Work	• Interruption.	• Equipment replacement; • Tests.
Severe weather	• Energy supply; • Interruption; • Perturbation.	• Electric board; • Air conditioner; • Power generator.
Equipment	• Perturbation; • Interruption.	• Service quality alarms; • Hardware and software issues.
Infra-structure	• Perturbation; • Interruption.	• Cooling problems; • Vandalization.

B. Scenarios Processing

To test the procedures, one selects a pilot city to test the meteorological information and the data provided. The city chosen is Braga, due to the vast number of incidents during the period of the data. This study is presented into two intervals: 24-hour and 12-hour interval.

The first study is from the linear equations regarding one weather variable, one presented in Table 3, where Temperature is presented by T, Humidity by H, Precipitation by P, Wind Speed by W, Gust Speed by G, Number of Discharges by D and the Maximum Intensity Discharge by I.

One can conclude from Table 3 that the number of incidents is mostly related with D and I. One also calculated the

determination coefficient, but since this value is small for all the variable, one does not present it.

Table 3. Linear Equation for the 24-hour interval.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.02	- 0.05	0.03	0.05	0.06	0.20	0.11
b	2.14	6.83	2.76	0.98	0.85	2.33	1.88

One presents in Table 4 the results of the correlation coefficient study. It is presented a colour scheme by correlation method, where green represents the maximum value of correlation, the red a correlation near zero, and the yellow cells the intermediate ones.

Table 4. Correlation results from 24-hour interval data.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
Pearson	0.08	-0.13	0.10	0.28	0.31	0.54	0.32
Spearman	0.06	-0.08	0.14	0.19	0.20	0.30	0.26
Kendall Tau	0.04	-0.06	0.12	0.14	0.16	0.21	0.20

One can conclude the same as before, where D and I are the variables mostly related with the number of incidents. The next step is the accomplishing of the same study, but using the 12-hour interval. One described the linear equation in Table 5.

Table 5. Linear equation for 12-hour interval.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.02	- 0.03	0.01	0.04	0.04	0.14	0.04
b	1.45	4.25	1.97	0.91	0.83	2.50	3.73

From Table 5 one draws that D continues as the variable with a higher slope, an equal result as the 24-hour study. The conclusions of the determination coefficient are the same as previously. Regarding the correlation study, from now on, one only presents the Spearman result, presenting in Table 6 is presented the results from 12-hour simulation.

Table 6. Correlations results

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
Spearman	0.10	-0.14	0.12	0.21	0.21	0.30	0.15

The results maintain, with D as the mostly related variable, and T with the lowest value, the remaining conclusions are nearly the same as before. The next study is regarding the pair of variables. In this simulation, one uses two weather variables and study the relationship between the number of incidents, as described in Section III.C. One accomplishes this study using the two time intervals.

From the two variables study, one concludes that, excluding the pair with electrical discharges, W and H is the most severe case. The results are equal from both intervals. With this information, one uses only the 24-hour interval for the remaining studies, since the results are nearly equal from both intervals.

C. Regions Analysis

The first phase is the analysis with one weather variable with the number of incidents. To better perceive the results from the Spearman coefficient, one presents the information of the most and least severe variable represented at each region in Fig. 6.

a) Most severe variable b) Least severe variable

Fig. 6. The most and least severe variable at each region.

One can conclude that from the north of Portugal, P is the variable which most appear. However, W and G also have meaningful results. For the least variable, T is the dominant variable. Regarding the centre of Portugal, one cannot draw great conclusions. However, P and T are the variables most present as the most severe. For the least severe, one also cannot draw any conclusion. However, W and G appear often in the interior of this area. Finally, for the south, T is the most relevant variable in the occurrence of incidents. For the less severe variable, there is no major conclusion. However, it is curious to relate that Beja has P has the most severe variable and T as the least severe variable. An explanation for this fact could be that the base stations in this area are prepared for the elevated temperatures since it is normal all over the year.

The next step is the analysis, as done before, for the pairs of weather variables for each region. Using the same subdivision as before, one can conclude that for the north, P and T is the pair most related. It is curious to relate that P is again present in this result. For the least severe, T and H is the pair. Again, T appears as one of the least severe variables. Regarding the centre, one cannot take any reliable conclusion for the most and least severe pair. Lastly, for the south of Portugal, W and P represents the pair most severe in two regions. Regarding the least severe pair of variables, also two regions have T and H as this pair.

D. Portugal analysis

With the statistics done in each region, it is important to understand the statistics of using the Portugal data. For the linear

regression equation and from the correlation coefficients, one can conclude that the most correlated variable is I. For this case, the least related variable is H, having almost 0 as the correlation coefficient. One presents in Table 7 the Spearman coefficient.

Regarding the results for the pair of variables study, one can conclude that W and P are the most related one. T and H, by other hand, appears as the least related.

Table 7. Correlations results from Portugal.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
Spearman	0.04	<0.01	0.09	0.19	0.14	0.07	0.21

Regarding the planned works study, one compared the quantity of planned work with the incidents caused by other causes. One concludes that the number of incidents by planned works is in a much smaller quantity than the others. The same results appear in the number maximum of incidents occurred in one day. These results can be explained because the planned works are prepared, and for this reason, it is controlled the severity of this work. Since the number of planned works is not significative, this parameter will not be further evaluated.

E. Forecasting

Using the equations from the Portugal data, one concludes that the equation with the best result reaches an MSE of 9.24 when using all the weather variables.

Regarding the NARX results, one accomplishes a study with twelve networks with varied sizes and delays, to understand which have the best result. One also simulated the network without the variable H, since its correlation is approximately zero. One concluded that the best result appears when using 40 neurons, 3 as delay and without the variable H.

One accomplishes the study for the entire outputs, and for the peaks of incidents. One considered a peak the top 5% of incidents occurred. In Table 8 is presented a comparison between the Linear Regression and NARX with Portugal data.

Table 8. Comparison between NN and Regression in Portugal study.

	Mean Error	MSE	Cor. Peaks	N. Peaks	Cor. Peaks [%]	F. Peaks	F. Peaks [%]
NN	2	8.85	12	230	5.2	9	43
Reg	2	9.24	3		1.3	2	40

One concludes that there is no major difference between the mean value of incidents error using both methods. The main differences appear when studying the peaks. NN hits more peaks, but also increases the number of false peaks. Despite this conclusion, one can settle that these results are insufficient, with only 5% of corrected peaks in the best scenario.

The next study is using the NN and the regression equation previously trained with Portugal data, but using the region data to forecast. One presents in Table 9 the average results for each region study together with the total peaks and the corrected ones.

One can conclude that the regression equation obtains better results than NN, both in the general outputs as in peaks study. However, regarding peaks study, the values of false peaks are superior to the previous study, with nearly 90% of false peaks in NN. Again, these results are unsatisfactory, with only 7% of corrected peaks and a huge percentage of false peaks.

Table 9. Comparison between NN and Regression in Region Study.

	Mean Error	MSE	Cor. Peaks	Total Peaks	Cor. Peaks [%]	F. Peaks [%]
NN	2	9	14	264	5.3	86
Reg	1.9	7.8	19		7.2	67

The next study it is created a regression equation and trained a NARX network for each region, and using the same data to forecast. One presents in Table 10 the complete results.

Table 10. Comparison between NN and Regression in the second region Study.

	Mean Error	MSE	Cor. Peaks	Total Peaks	Cor. Peaks [%]	F. Peaks [%]
NN	3.5	3.6	64	264	24	17
Reg	5.2	4.5	30		11	19

One can conclude that, and as expected, this is the best result until this point. Both methods have the lowest MSE, as well as the increase in the percentage of corrected peaks and the lowest false peaks percentage. Nevertheless, both approaches cannot achieve great performances in the forecasting of the peaks, despite the superior results. One can also refer that this is an average result, meaning that is calculated the mean of all regions results. These results can range from 0% to 60% of corrected peaks or from 0% to 100% in false peaks. This demonstrates the fragility in this model regarding the inputs selection.

The final study is the Weka [38] classification. As described in Section III.D, the first step is the division of dataset into classes. Then, using the four methods of classification, one accomplishes this study. In Table 11 is presented the comparison of the four methods.

Table 11. Comparison of the four classification methods.

	Mean [%]	Mean A [%]	Mean B [%]	Mean C [%]	F. Peaks [%]
Bayes Net.	73	98	4	18	56
MLP	72	92	14	17	62
Near. Neigh.	63	77	24	19	78
SVM	73	99	1	8	22

Observing the global accuracy, both Bayes Network, SVM and MLP have an accuracy very close, leading to being the best methods in this study. For the class A, both SVM and Bayer Network have good results. However, SVM has low accuracy in

class B and C. Nearest Neighbour appears as the best method to classify class B and C but has the most unsatisfactory result in the false peaks. For the false peaks study, the best result appears in the SVM simulation.

V. CONCLUSIONS

The primary goal of this thesis is the study of the number of incidents regarding two main variables: Meteorological and Planned Works. After a brief analysis, one concludes that the incidents caused by the latter are much inferior to other causes. Due to this, one only deepens the study of the meteorological factors. With the possibility of perceiving how the number of incidents is related to the meteorological factors, the paradigm of the network administrators could end up changing. Knowing this information, the operators could better organize their team and network, providing better service to the customers.

Five chapters compose this paper, being the first the Introduction. This chapter contains a summary of the mobile wireless communication evolution, as well as a brief introduction to the incidents thematic. One also presents the motivation and contents behind this work.

In Section II, one provides the description of GSM, UMTS and LTE. Next, it is introduced the definition of incidents, providing the methods to accomplish the statistical and forecasting studies. It is also presented state of the art.

In Section III, it is provided with the description of the dataset used. One also presents the processing of each dataset, to a final file. The organization of the statistical and forecasting studies are introduced, as well as its detailed explanation.

Section IV starts with the description of the scenario in study. A ratio between some variables and the number of incidents is presented. From this study, one can conclude that Portugal does not have a homogenous country in relation to the metrics studied. This is explained due to the vast diversity in the number of inhabitants, size and infrastructures that each region has.

In the first scenario, Braga, one could conclude that the results are similar using the two-time intervals. Due to this, one uses only the 24-hour interval further in work. Regarding the weather study, D appears as most related variable. Regarding the pair of variables, W and H surge as the most related pair.

The following study is from the regions of Portugal. One subdivides the country into three parts, North, Centre and South, to accomplish the one variable study. For the north, one draws that P is the one which appears more often. However, W and G also have an important role. Regarding the least severe variables, T is the one which appears more often. For the centre, one cannot draw great conclusions from both most and least severe variables. However, P and T are the variables which appear more often as the most related, and W and G appear more frequently as the least related. Finally, for the south, T is the most important variable. For the least severe, there are no significant conclusions. Though, it is curious to relate that Beja has P as the most related variable and T as the least related. One could explain this situation since high temperatures are normal in Beja, and the sites could be prepared to deal with this situation.

One completed the same study for the pairs of the variable. For the north, P and T appear as the most related pair. Regarding the least severe pair, T and H are the most present in this area. For the centre, one cannot draw any possible conclusions for both studies. Regarding the south, W and P is the most common pair, appearing in two regions. T and H also appears in two regions as the least severe pairs.

To conclude the statistical study, one studies the data from Portugal. Despite knowing that these results do not represent each region, one completed the study to compare with the remaining ones. I reveal as the most related variable and H as the least related. For the pair of variables, W and P as the most related, T and H as the least related pair. One also accomplishes a study on planned work, concluding that the number of incidents caused by this variable is inferior to the remaining.

The next step is the forecasting implementation. The first study is with the use of the Portugal data to train and forecast. For the regression, one reaches an MSE of 9.24, hitting 1.3% of peaks and reaching 40% of false peaks. For the NN, one reaches an MSE of 8.85, hitting 5.2% of peaks and 43% of false peaks. In this case, NN obtains better results.

The second study, using the equation and NN trained before, one uses the data from each region to forecast. The regression obtains, in average, an MSE of 7.8, hitting 7.3% of peaks and reaching 67% of false peaks. On the other hand, NN reaches an MSE of 9, with 5.3% of peaks and 86% of false peaks. In this case, regression reaches better results, but remain unsatisfactory, reaching nearly 70% of false peaks.

The following study is training and forecasting the network with region data. In this case, the regression obtains, in average, an MSE of 4.5, with 11% of corrected peaks and 19% of false peaks. For the NN, has in average an MSE of 3.6, hitting 24% of peaks and 17% of false peaks. These are the best results so far, as expected. The biggest problem is the peaks prediction, where the best result only hits 24%.

To be able a comparison, one uses the Weka software. In this case, the best method to predict peaks is the Nearest Neighbours, with 19% of corrected ones. The problem is that this approach reaches 78% of false peaks. In average, the best methods are the Bayes Network and the SVM.

One can globally conclude that the best results came from training the NARX with the data from each region and use the same data to simulate. However, regarding the peak study, this result is not satisfactory for use in a real scenario, where it hits only 24% of peaks. This can be explained due to the data applied, where it is only used one weather station per region.

Regarding future work, a more profound analysis could be made, since this is a new field of study where it exists very few information about the relationship between incidents and the meteorological variables. It is important to refer that this thesis is part of a new study about this thematic, and one of the main goals is first to introduce and initiate the study for the theme. The first improvement can be applied to the incidents data, where a definition of the incidents caused be severe weather could be pointed. The second improvement is from the weather data, wherein this thesis it is used data from personal weather stations, which sometimes is fallible. Then, the next

improvement is by the use of weather stations nearby the base stations. In this thesis it is only studied the number of incidents, it could be essential to define the severity on each incident. Finally, in this thesis, it is addressed some of the machine learning algorithms, and it could be studied another algorithm.

ACKNOWLEDGMENT (HEADING 5)

The first appreciations go to my thesis supervisor, Prof. Luís M. Correia. It has been a remarkable opportunity to realize this work under his guidance, due to all knowledge and all the valued pieces of advice that professor provided.

A special thanking to Eng. Jorge Seabra and Eng. João Duarte, for all the support and for being the point of contact with NOS.

REFERENCES

[1] H. Holma and A. Toskala, HSDPA/HSUPA for UMTS, Wiley, Chippenham, United Kingdom, 2006.

[2] H. Holma and A. Toskala, WCDMA for UMTS – HSPA Evolution and LTE, Wiley, Chippenham, United Kingdom, 2007.

[3] T. Halonen, J. Romero and J. Melero, GSM, GPRS and EDGE Performance – Evolution Towards 3G/UMTS, Wiley, Chippenham, United Kingdom, 2003.

[4] H. Holma and A. Toskala, LTE for UMTS – Evolution to LTE-Advanced, Wiley, Chippenham, United Kingdom, 2011.

[5] C. Cox, An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications, Wiley, Chichester, United Kingdom, 2012.

[6] S. Sesia, I. Toufik and M. Baker, LTE – The UMTS Long Term Evolution, Wiley, Chippenham, United Kingdom, 2011.

[7] S. Wallin, "Chasing a Definition of "Alarm"", Journal of Network and System Management, Vol.17, No. 4, Dec. 2009, pp. 457-481.

[8] ITU - International Telecommunication Union, X.733: Information Technology – Open Systems Interconnection – Systems Management: Alarm Reporting Function, Recommendation, Geneva, Switzerland, 1992 (https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-X.733-199202-I!!PDF-E&type=items).

[9] R. Kuhn, "Sources of Failure in the Public Switched Telephone Network", IEEE Computer, Vol. 30, No. 4, Apr. 1997, pp. 31-36.

[10] G. Jakobson and M. Weissman, "Real-time telecommunication network management: extending event correlation with temporal constraints" in A.S. Sethi et al. (eds.), Integrated Network Management IV, Chapman and Hall, London, United Kingdom, 1995.

[11] D. Željko and M. Kunstic, "A Comparison of Methods for Fault Prediction in the Broadband Networks", in Proc. SoftCOM 2010 – 18th International Conference on Software, Telecommunications and Computer Networks, Split, Croatia, Sep. 2010.

[12] D. Željko, M. Kunstic and B. Spahija, "Using Temporal Neural Networks to Forecasting of Broadband Network Faults", in Proc. SoftCOM 2011 – 19th International Conference on Software, Telecommunications and Computer Networks, Split, Croatia, Sep. 2011.

[13] C. Reimann, P. Filzmoser, G. Garret and R. Dutter, "Correlation" in John Wiley & Sons, Ltd, Statistical Data Analysis Explained: Applied Environmental Statistics with R, Wiley, Chippenham, United Kingdom, 2008.

[14] J. Zar, "Spearman Rank Correlation" in John Wiley & Sons, Ltd, Encyclopedia of Biostatistics, Wiley, Chippenham, United Kingdom, 2005.

[15] P. Chen and P. Popovich, Correlation – Parametric and Nonparametric Measures, Sage Publications, Thousand Oaks, California, United States of America, 2002.

[16] J. Rodgers and W. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient", The American Statistician, Vol.42, No.1, Feb. 1988, pp. 59-66.

[17] D. Wilks, Statistical Methods in the Atmospheric Sciences, Elsevier, London, United Kingdom, 2006.

[18] R. Adhikari and R. Agrawal, An Introductory Study on Time series Modeling and Forecasting, Lambert Academic Publishing, Saarbrücken, Germany, 2013.

[19] J. Heaton, Bayesian Networks for Predictive Modeling, Forecasting & Futurism newsletter, Society of Actuaries, Illinois, United States of America, 2013 (https://www.soa.org/Library/Newsletters/Forecasting-Futurism/2013/july/ffn-2013-iss7.pdf).

[20] L. Vaněk, Introduction into Bayesian Networks, Class Support, Faculty of Information Technology, Brno, Czech Republic, 2008 (http://www.fit.vutbr.cz/study/courses/VPD/public/0809VPD-Vanek.pdf).

[21] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, "When is "Nearest Neighbor" Meaningful?", Lecture Notes in Computer Science, Vol.1540, No. 1, Jan. 1999, pp. 217-235.

[22] J. Alonso, K-nearest neighbours, Class Support, Universitat Politècnica de Catalunya, Barcelona, Spain, 2012 (http://www.cs.upc.edu/~bejar/apren/docum/trans/03d-algind-knn-eng.pdf).

[23] H. Siegelmann, B. Horne and C. Giles, "Computational Capabilities of Recurrent NARX Neural Networks", IEEE Transactions on Systems, Man and Cybernetics, Vol.27, No. 2, April 1997, pp. 208-215.

[24] T. Lin, C. Giles, B. Horne and S. Kung, "A Delay Damage Model Selection Algorithm for NARX Neural Networks", IEEE Transactions on Signal Processing, Vol.45, No. 11, Nov. 1997, pp. 2719-2730.

[25] O. Renaud and M. Feser, "A robust coefficient of determination for regression", Journal of Statistical Planning and Inference, Vol.140, No. 7, July 2010, pp. 1852-1862.

[26] A. Bouloutas, S. Calo and A. Finkel, "Alarm Correlation and Fault Identification in Communication Networks", IEEE Transactions on Communications, Vol. 42, No. 2, Feb. 1994, pp. 523-533.

[27] D. Željko, M. Randić and G. Krčelić, "A Multivariate Approach to Predicting Quantity of Failures in Broadband Networks Based on a Recurrent Neural Network", Journal of Network and System Management, Electronics and Microelectronics, Vol.24, No. 1, Jan. 2016, pp. 189-221.

[28] H. Asgari, X. Chen, M. Morini, M. Pinelli, R. Sainudiin, P. Spina and M. Venturini, "NARX models for simulation of the start-up operation of a single-shaft gas turbine", Applied Thermal Engineering, Vol.93, No. 1, Jan. 2016, pp. 368-376.

[29] V. Barrera, J. Meléndez, S. Herraiz, A. Ferreira and A. Muñoz, "Analysis of the influence of weather factors on outages in Spanish distribution networks", in ISGT Europe 2011 – 2nd International Conference and Exhibition Innovative Smart Grid Technologies, Manchester, United Kingdom, Dec. 2011.

[30] J. Denissen, L. Butalid, L. Penke and M. Aken, "The Effects of Weather on Daily Mood: A Multilevel Approach", Emotion, Vol.8, No. 5, Jan. 2008, pp. 662-667.

[31] A. Yang, J. Fuh, N. Huang, B. Shia, C. Peng and S. Wang, "Temporal Associations between Weather and Headache: Analysis by Empirical Mode Decomposition", PLoS One, Vol.6, No. 1, Jan. 2011, pp. 1-6.

[32] NOS – Portuguese Operator, http://www.nos.pt/, Sep. 2017.

[33] Weather Underground – Meteorological Information, https://www.wunderground.com/, Feb. 2017.

[34] IPMA – Instituto Português do Mar e Atmosfera, http://www.ipma.pt/, Feb. 2017.

[35] ScyPy – Open Source Software, https://docs.scipy.org/doc/scipy-0.14.0/reference/stats.html, Mar. 2017.

[36] MathWorks – Estimate Multiple Linear Regression Coefficients, https://www.mathworks.com/help/stats/regress.html, May 2017.

[37] M. Beale, M. Hagan and H. Demuth, Neural Network Toolbox, User's guide, MathWorks, 2017 (https://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf).

[38] Weka – Collection of machine learning algorithms - http://www.cs.waikato.ac.nz/ml/weka/, July 2017.