Modelling and forecasting incidents in cellular networks

Daniel André Correia Almeida

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Luís Manuel de Jesus Sousa Correia

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Supervisor: Prof. Luís Manuel de Jesus Sousa Correia

Members of Committee: Prof. Paulo Luís Serras Lobato Correia

Eng. Jorge Seabra

November 2017

To my loved ones

Acknowledgements

The first appreciations go to my thesis supervisor, Prof. Luís M. Correia. It has been a remarkable opportunity to realise this work under his guidance, due to all knowledge and all the valued pieces of advice that he provided. Also, some special thanks to the opportunity of accomplishing this work in collaboration with NOS and for the experience of belonging to the GROW research group.

A special thanking to Eng. Jorge Seabra and Eng. João Duarte, for all the support and for being the point of contact with NOS. A sincere acknowledgement for the total availability to guide my work and for answering all my doubts about the course of the thesis, helping me to improve the quality of this thesis.

I would like to thank all GROW members, for their great friendship during this year, especially my master thesis' colleagues and friends Cristina Dias, Diogo Martins, Hugo Martins, João Cardoso, Miguel Ramos and Sónia Pedrinho. This experience would not have been the same without all of them.

To my friends and colleagues that accompanied me throughout IST: André Ramião, Diogo Gonçalves, Fábio Fernandes, Gonçalo Pais, Inês Costa, Joana Baleiras, João Pires, José Coelho, Marisa Gomes, Nuno Nico, Paulo Gonçalves, Pedro Mateus, Rodrigo Zenha, Rúben Pinto, Sílvia Figueira, Tiago Santos, Xavier Reis, and to all the remaining that were not mentioned, very special thanks for all the support during these years.

Special thanks for NEECIST for all the incredible people that I worked with, for making my college life a different and better experience.

To all my family, and the ones who are not legal family, but who are also very important: Adelino Gaspar, Alice Gaspar, Ana Paula Botas, Armando Botas, Bruno Botas and all the rest that were not mentioned, truthful thanks for all the good wishes and support during these last years.

To Cátia Botas, my love, and most important, my best friend, the sincerest thanks for all the patience and support during these years. It has been a pleasure to spend them besides you.

Abstract

This thesis addresses the study of the number of incidents in an operator's network, depending on meteorological factors. This study addresses the understanding of how weather is related to incidents, allowing the operator to better attend them, in order to maintain or increase the quality of service to its customers. Taking this into account, one developed a statistical model that correlates the number of incidents with weather factors in each region of Portugal. One also developed a forecasting model, 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 instead of the whole country of Portugal data. Regarding the forecasting model, one applied some methods to predict the number of incidents. The best results appear when applying the NARX Neural Network; however, in this case, the method has a mean square error of 3.6, hitting on average 24% of the peaks with 17% of false peaks predicted. This means that this approach cannot be applied in a real operation. Though, this is the first step into a study to be implemented in the real world.

Keywords

Alarms, Incidents, Faults, Correlation, Forecasting, Neural Networks.

Resumo

Esta tese aborda o estudo da quantidade de incidentes na rede de uma operadora, dependendo de fatores meteorológicos. Esta análise permite o entendimento de como o clima está relacionado com incidentes, permitindo que a operadora os resolva melhor, de modo a manter ou melhorar a qualidade de serviço dos seus clientes. Tendo isto em conta, foi desenvolvido um modelo estatístico que correlaciona o número de incidentes com os fatores meteorológicos em cada distrito de Portugal. Foi também desenvolvido um modelo de previsão, para ser possível prever da melhor forma o número de incidentes com um foco nos dias de pico. Foi concluído que cada distrito tem um comportamento diferente dependendo da variável meteorológica que está mais relacionada com o número de incidentes, levando a melhores resultados quando se usa dados de distritos em vez de dados de Portugal. Para o estudo de previsão, foram aplicados vários métodos para prever o número de incidentes. O melhor resultado aparece quando se utiliza a rede neuronal NARX. No entanto, para este caso, o método tem um erro quadrático médio de 3.6, atingindo em média 24% dos picos com uma previsão de 17% de falsos picos. Isto significa que este método não pode ser aplicado em operações reais, mas é o primeiro passo para um estudo a ser aplicado num cenário real.

Palavras-chave

Alarmes, Incidentes, Falhas, Correlação, Previsão, Redes Neuronais.

Table of Contents

List of Figures

List of Tables

List of Acronyms

2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
ANN	Artificial Neural Networks
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
BPSK	Binary Phase Shift Keying
BS	Base Station
BSC	Base Station Controller
BTS	Base Transceiver Station
CDMA	Code Division Multiple Access
CN	Core Network
CP	Cyclic Prefix
CS	Circuit Switch
DL	Downlink
DS-CDMA	Direct-Sequence Code Division Multiple Access
EDGE	Enhanced Data Rates for Global Evolution
EIR	Equipment Identity Register
eNodeB	Evolved Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FCC	Federal Communications Commission
FDD	Frequency Division Duplex
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway General Packet Radio Service Support Node
GMSC	Gateway Mobile Services Switching Centre
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HLR	Home Location Register
HSDPA	High-speed Downlink Packet Access
HSPA +	Evolved High Speed Packet Access
HSS	Home Subscription Server
HSUPA	High Speed Uplink Packet Access

IBM	International Business Machines
IMS	IP Multimedia Sub-system
IP	Internet Protocol
IPMA	*Instituto Português do Mar e Atmosfera*
ISDN	Integrated Service Digital Network
ISI	Inter Symbol Interference
JSON	JavaScript Object Notation
LTE	Long Term Evolution
M2M	Machine-to-Machine
MC	Multi-Carrier
ME	Mobile Equipment
MIMO	Multiple Input Multiple Output
MLP	Multi-layer Perceptron
MM	Mobility Management
MME	Mobility Management Entity
MS	Mobile Station
MSC	Mobile Services Switching Centre
MSE	Mean Squared Error
NARX	Nonlinear Autoregressive Network with Exogeneous Inputs
NE	Network Element
NN	Neural Network
NOC	Network Operation Centre
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PCC	Policy and Charging Control
PCRF	Policy and Charging Resource Function
P-GW	Packet Data Network Gateway
PLMN	Public Land Mobile Network
PS	Packet Switch
PSTN	Public Switched Telephone Network
QAM	Quadrature Amplitude Modulation
QoS	Quality of Services
QPSK	Quadrature Phase-Shift Keying
RB	Resource Block
RE	Resource Element
RMSE	Root Mean Squared Error
RNC	Radio Network Controller
RNN	Recurrent Neural Networks
RRM	Radio Resource Management
SAE	System Architecture Evolution

SAE-GW	System Architecture Evolution Gateway
SC-FDMA	Single Carrier Frequency Division Multiple Access
SGSN	Serving General Packet Radio Service Support Node
S-GW	Serving Gateway
SIP	Session Initiation Protocol
SIRESP	*Sistema Integrado de Redes de Emergência e Segurança de Portugal*
SLA	Service Level Agreement
SNOC	Services and Network Operation Centre
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machines
TCH	Traffic Channels
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
USIM	UMTS Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
WCDMA	Wideband Code Division Multiple Access

List of Symbols

α_0	Bias term
α_j	Connection Weights
β_{0j}	Bias term
β_{ij}	Connection Weights
ε_t	Random Shock
$\overline{\varepsilon^2}$	Mean Square Error
ξ	Slack variable
τ	Kendall's τ coefficient
ψ	Nonlinear function
A_{value}	Actual value of the weather variable
b	Bias term
b_0	Regression constant
c_{peaks}	Quantity of correct peaks
C	Regularisation constant
$d(X,V)$	Euclidean distance between two points
D	Number of discharges
e_{fpeaks}	Error between false peaks and total forecasted peaks
E	Number of concordant pairs
f_{peaks}	Quantity of false peaks
f_t	Forecasted value

G	Gust Speed
H	Humidity
I	Discharges Intensity
l	Size of sample
m_k	Regression coefficient
M_{value}	Maximum value of the weather variable
n	Number of ranks
n_{faults}	Number of faults from regression equation
n_u	Input order
n_y	Output order
P	Precipitation
q	Number of hidden nodes
Q	Number of discordant pairs
r	Pearson coefficient
r_s	Spearman coefficient
R^2	Coefficient of determination
R_{meteo}	Relative value presented
R_p	Rank of dataset in position p
s	Standard deviation
T	Temperature
u	Number of inputs
$u(t)$	Input of the network at time t
v	Number of vectors

w	Weight Vector
W	Wind Speed
x_k	Predictor
X	Generic dataset
$\bar{X}$	Mean of dataset X
X_i	Generic dataset in index i
X_j	Dataset in index j
X_n	Dataset in index n
X_v	Dataset in index v
$\widehat{X_i}$	Predicted value of X_i
Y	Generic dataset
y_{MLP}	Output of MLP
$y_{predictand}$	Multiple Liner Regression Predictand
$y_{surface}$	Regression with two variables
$y(t)$	Output of the network at time t
$\bar{Y}$	Mean of dataset Y
Y_i	Generic dataset in index i
Y_j	Dataset in index j
Y_p	Rank of dataset in position p
Y_t	Actual value
Y_{t-1}	Inputs of MLP
(Y_i, X_i)	Input-Output pair

List of Software

Atom	Python editor
IBM SPSS Statistics 24	Statistics Computing Environment
Matlab R2016a	Numerical Computing Environment
Mendeley	PDF repository and paper's research
Microsoft Excel 2016	Excel processor
Microsoft Paint	Image processing
Microsoft Visio 2013	Flowchart and diagrams software
Microsoft Word 2016	Word processor
Weka 3.8.1	Collection of machine learning algorithms

Chapter 1

Introduction

In this chapter, one provides an overview of mobile communications systems evolution, followed by an introduction of the importance of alarms in an operators' network. One also presents the thesis motivation and the content of the report.

1.1 Overview

The popularity of mobile devices has been increasing, as well as the demand for mobile communications technologies, due to the growth of mobile subscribers' number. According to [Cisc16], approximately 563 million of new mobile devices were added in 2015 to networks, accounting 7900 million global mobile devices in 2015. This growth can be explained by low-cost cell phones and improvements in network coverage and capacity.

More and more mobile phones are used to access mobile networks, contributing to growth in global mobile traffic. According to [Cisc16], the mobile data traffic is expected to reach 30.6 Exabytes, together with an increase of mobile-connected devices per capita, reaching 1.5 by 2020. Machine-to-Machine (M2M) communications are expected to grow 38%, being the most noticeable growth from 2015 until 2020. Another significant trend is the increase of data consumed by smartphones, reaching 81% of all consumed data, Figure 1.1.

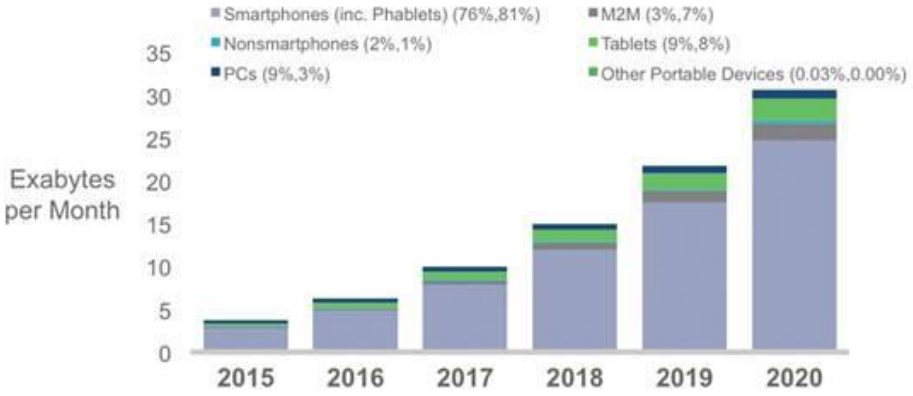


Figure 1.1. Forecast of Global Mobile Data Traffic Growth by Device Type between 2015-2020 (extracted from [Cisc16]).

As communications systems became more complex to deal with the growth of devices and data traffic, the task of identifying and correcting faults in a network has turned into a critical task of network management, concentrated in the Network Operation Centre (NOC). NOC is the centre where monitoring and alarm management are done, visualising all network connections, acting as the principal place for network troubleshooting, software updating and performance monitoring. Still, the concept of NOC is changing to a Services and Network Operation Centre (SNOC), where the quality of the overall service is also monitored, and, if necessary, actions are taken if the service has degradation or outages.

A fault that can interfere with the services provided by the operator is costly. The detection of these faults before users can suffer from service degradation is a needed requirement of a proper communications system, where these service issues may be detected by monitoring error rates and alarms. Since it is not possible to avoid faults in communications systems, their detection and correction are essential. The use of mechanisms for controlling user's service parameters is getting critical, by the detection of failures and then providing a notification to network managers.

In fault management, there are some basic concepts, but there is no standard in naming them. An Alarm (also known as an event) is the exceptional condition occurred in the operation of hardware or software in a managed network, e.g. an open-door notification or a severe problem in the network. A Root Cause indicates the origin of an abnormal condition in the system. An Incident (also used in terms fault and root cause) is a malfunction of the system that can trigger several alarms. Finally, a Ticket, which can also be referred to as an alarm, is the notification that the network manager receives information of a changed state in the network; it contains the information of alarms that originated the incident and its impact on the network.

With the increase in size and complexity in a communications network, an enormous amount of information that needs to be analysed is created. According to [Wall09], from an alarm database belonging to a large mobile operator with approximately 15 million alarms in 4 months, only about 3.5 million are real alarms and 90 000 are associated trouble tickets. One illustrates in Figure 1.2 this alarm database, representing some produced alarms. The managed alarms account for the number of alarms already correlated, the alarms with ticket represents the alarms handled by the network managers, and finally, the cleared alarms are the ones managed automatically by Network Elements (NE).

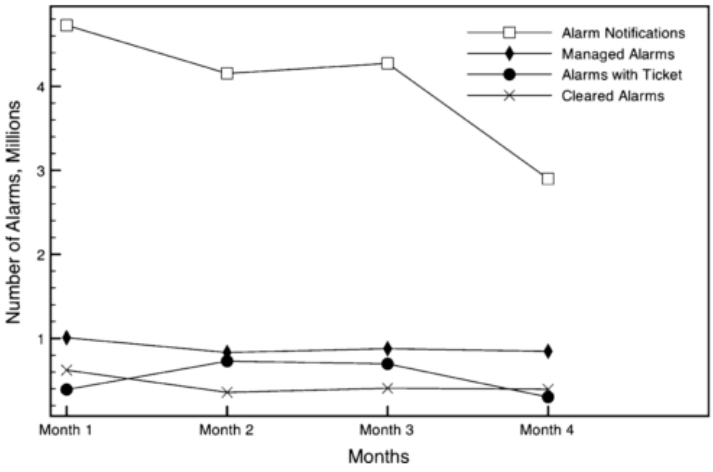


Figure 1.2. Number of alarms in a 4-month database (extracted from [Wall09]).

These failures have a major impact on both operator and customers. The former should provide an excellent service to its clients, for not losing them in the competitive market of telecommunications. The latter want to be able to call when wanted, and to make it properly. In order to achieve the satisfaction of both parts, prior work is necessary to be able to overcome these faults. The operator needs to organise its workforce to respond in a quick way to the incidents that may exist, providing a lower impact to their services, and then, a better service to customers. To complement this organisation, the forecast of the number of incidents, and to understand the severity of the failures that may exist, is a major information task to the network manager. This knowledge brings an advantage to the operator, not only by concerning organisation but also in an economical way. One can economise money more directly, by, for example, not putting people in prevention to these incidents, but also, in an indirect way, with customers pleased with the service, leading to more loyal clients.

However, in recent times, the problem of network failure is widely spoken in the Portuguese media, due

to the case of *Sistema Integrado de Redes de Emergência e Segurança de Portugal* (SIRESP), an operator dedicated to security and emergency forces. During the forest fires in Portugal, this network had several failures, leaving the emergency forces without a network to make calls. With the proportion that the media gave to this case, the population is now more alert to failures in a network, and if an operator does not provide an excellent service, this will be reflected in customers satisfaction. In addition, every year exists extreme meteorological conditions, such as extreme temperatures and strong wind speeds, occur, with the ability of reducing network capacity, this being a key factor to study.

1.2 Motivation and Contents

The demand for a more optimised network that consumes fewer resources with the same capacity to serve customers is a necessary challenge to be achieved by operators. Thus, the sites (localisation of the base stations) represents a significant location to be taken into consideration, due to its consumption of resources, such as electricity, the need of equipment maintenance and the displacement of a representative to repair it. Sometimes, many of these sites are placed in locations difficult to reach, requiring many work-hours just to arrive at the site. Also, in SNOC, the organisation of the workforce to manage all situations is necessary, which sometimes occurs when the necessary people to manage these situations is not available.

A prediction of the number of incidents, regarding some natural factors, such as temperature or humidity and the planned worked in the network, are becoming information even more relevant to operators, due to the importance of workforce allocation decisions and network maintenance planning. Besides, the study of the peaks of incidents is also important. These are days with an unusual quantity of incidents, which lead to a need for more workforce by the operator to deal with them.

The major relevance of this study relies on the fact that there is insufficient information about the importance of meteorological factors in the occurrence of incidents in a telecommunication network. There are many studies on the importance of weather factors in the area of health or electrical networks, however, in the field of telecommunications, this study is almost inexistent. Besides, this study provides a significant advantage to the network manager, by giving a sense of the number of incidents that the network may have, thus, enabling to organise the response team to these failures and to take measures so that clients do not suffer the consequences of these failures.

The goal of this master thesis is to develop a statistical model to relate meteorological variables with incidents, on both all days and peak days (the outliers), as well as to forecast the number of incidents. The first study is realised by measuring the importance of both meteorological variables and planned works in the occurrence of incidents, to understand the importance of each factor in the occurrence of these failures. This study implements an analysis of the regions of Portugal in separate and in Portugal as a whole. The second study relies in the forecasting on the number of faults. One aims at using some methods of forecasting but mainly focuses on Neural Networks, to develop a model that predicts the

number of faults regarding these factors.

This thesis was done in collaboration with NOS, a network operator in Portugal. The conclusions of this thesis are intended to give additional information to an operator, in order to be allow an understanding of how meteorological factors and planned works affect the number of incidents in the network. One also intends to provide a model to forecast the number of incidents, with a special analysis of days with an abnormal quantity of incidents.

Regarding contents, this thesis is divided into five chapters with a set of annexes with additional information and results to the main work. The present chapter makes a brief introduction and overview of the mobile communications evolution, as well an introductory presentation of incidents in these networks, providing the motivation behind this thesis.

Chapter 2 introduces some fundamental aspects regarding this work. One provides with a brief description of GSM, UMTS and LTE networks architectures, showing its main elements. Then, a definition of alarms and incidents, as well as these alarms are propagated, and correlated are presented. One also presents the approach of prediction of incidents, introducing regressions and neural networks, as well as the used performance measures.

Regarding Chapter 3, a description of the data, divided by the dataset and the meteorological description is given. Since the process of these data is necessary, one presents some flowcharts and description of how data was treated in order to extract the required information. Then, the statistical study is presented, regarding the number of incidents versus one, two and several weather variables. One also presents the study of forecasting, showing the application of regression, neural network and the Weka classification. Finally, one presents the assessment of these forecasting measures.

Chapter 4 presents the description of the scenario under study, as well as the analysis of the scenario. One introduces the statistical analysis of the number of incidents regarding meteorological variables, a study of the planned works, and finally the forecasting of the number of incidents on meteorological variables. The statistical study is presented in several steps, the first being a study in only a region. Afterwards, the study is performed in all regions in separate, and then it is applied to the whole data in Portugal. Regarding the forecasting approach, one first applies a linear regression and a neural network to predict the number of incidents using the data of Portugal. Then, the same study was implemented to the regions, and finally, for the Regions, another method to forecasting the number of incidents was applied, using the Weka software.

In Chapter 5, one presents the main conclusion of this thesis, together with an overall description of the study and suggestions for future work in this thematic.

A list of annexes with some additional information is given in the end. Annex A presents the list of meteorological stations used in Weather Underground. Annex B contains the confidential information, and Annex C covers additional information about the statistical data from the studied scenarios. Finally, Annex D presents the supplementary data about the Weka classification.

Chapter 2

Fundamental concepts and state of the art

In this chapter, an overview of the cellular systems is presented, showing the architectural design and radio interface of GSM, UMTS and LTE. A summary of alarms and incidents is also presented, followed by a brief description on alarms correlations. One finalises with the approach on the statistical and forecasting study of the number of faults and state of the art.

2.1 GSM and UMTS

In this section one introduces fundamental aspects about GSM and UMTS, focused on the network architecture presented jointly with its main elements and a description of the radio interface based on [HoTo06], [HoTo07] and [HaRM03].

2.1.1 Network Architecture

The need for using the same radio access network either in GSM and UMTS leads towards an architecture that can be efficiently integrated into a single UMTS multi-radio network. This architecture consists of some logical networks in which each one has a defined functionality, Figure 2.1. This architecture is composed of the GSM Enhanced Date Rates for GSM Evolution (EDGE) Radio Access Network (GERAN) and UMTS Terrestrial Radio Access Network (UTRAN), responsible for all radio-related functionalities at GSM and UMTS respectively. Another element in the architecture is the Core Network (CN), responsible for switching and routeing calls and also data connections to external networks. Finally, the User Equipment (UE), denomination for UMTS and the Mobile Station (MS), denomination for GSM, are the interfaces that connect the user to the network.

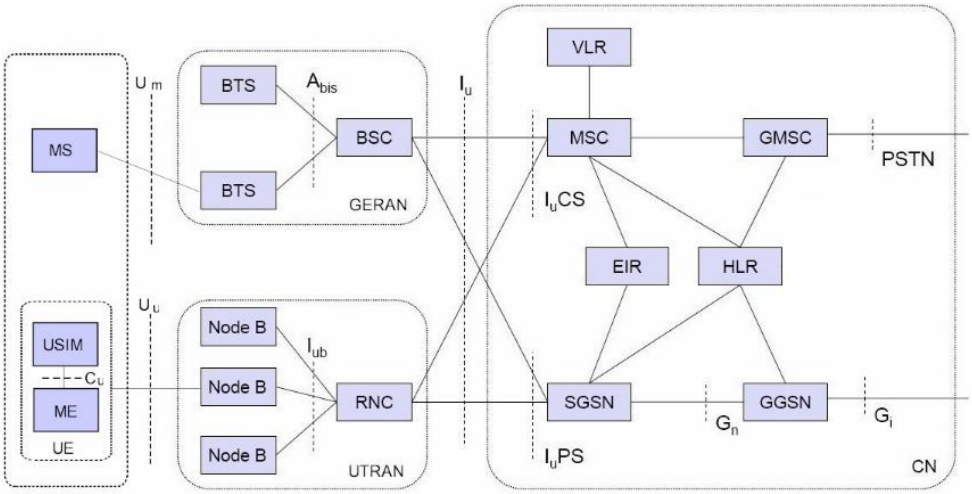


Figure 2.1. GSM and UMTS Network Architecture (adapted from [Netw16]).

The UMTS network can be divided into sub-networks (UE, UTRAN and CN), either on their own or together with other sub-networks, distinguished from each other with unique identities, allowing to call this sub-network as UMTS Public Land Mobile Network (PLMN). Typically, these sub-networks are operated by a single operator and are connected either to other PLMNs or to other types of external networks.

The UE is split in two parts: the Mobile Equipment (ME) and UMTS Subscriber Identity Module (USIM). The first is the radio terminal used for radio communication between the user and the UTRAN, and the second is the smartcard that identifies the subscriber, execute the authentication algorithms, stores authentication, encryption keys and some subscription information needed at the terminal.

Both UTRAN and GERAN are composed of two distinct elements: the Node B for UTRAN and the Base Transceiver Station (BTS) for GERAN, corresponding to the more generic term Base Station (BS), which provides the connection between the UE and Radio Network Controller (RNC) for UTRAN, and Base Station Controller (BSC) for GERAN, which control all radio resources. It can be used as the service access point for all services that UTRAN/GERAN provides to CN (e.g. management of connection from the UE for UMTS and the MS for GERAN).

CN has two domains, Circuit Switch (CS) providing circuit-switched connections for voice, and Packet Switch (PS) providing packet data connections. This division comes from the different requirements of data, depending on whether the domain is real-time (Circuit Switched) or non-real time (Packet Data). The CS domain has the following elements:

- Mobile Services Switching Centre/Visitor Location Register (MSC/VLR): the switch and database, correspondingly, that serves the UE/MS in its current location for CS services. The MSC objective is to switch CS transactions, and VLR's is to save the information of UE's location, as well as the visiting user's service profile.
- Gateway MSC (GMSC): the switch where PLMN is connected to external CS networks, with all the CS incoming and outgoing connections made through it.

The PS domain has the following elements:

- Serving General Packet Radio Service (GPRS) Support Node (SGSN): its functionality is similar to MSC/VLR, but it is typically used for PS services.
- Gateway GPRS Support Node (GGSN): the function is similar to GMSC, but for PS services.

In addition to the two domains, the network has some registers with valuable information:

- Home Location Register (HLR): a database that stores the master copy of the user's service profile, consisting of, for example, information on allowed services or forbidden roaming areas. A new entry is created when a new user subscribes to the system and remains stored as long this subscription is active.
- Equipment Identity Register (EIR): contains the information related to the terminal equipment, and can be used, for example, to block the access to the network of a specific terminal.

The connection of the network to external ones can be divided into two groups:

- CS Networks: connecting, like the existing telephony system, the Integrated Service Digital Network (ISDN) and the Public Switched Telephone Network (PSTN).
- PS Networks: providing connections for packet data services, Internet being the major example.

There are interfaces between logical elements, which are, as well, standardised. The Cu interface is the electrical interface between USIM smartcard and ME. The Uu and Um interfaces provide access of the UE to the fixed part of the system. The connection of UTRAN and GERAN to CN is done by the Iu interface for both CS and PS services; the connections inside UTRAN are done by Iub one, which connects the Node B to RNC, while for GERAN the connections between BTS and BSC are done by the Abis interface.

2.1.2 GSM and UMTS Radio Interface

GSM standard is based on a Multi-Carrier (MC), Time Division Multiple Access (TDMA) and Frequency Division Duplex (FDD) modes. A frame is subdivided into eight time-slots, data being transmitted in time-slots in bursts. Channels are described at two levels: Physical and Logical Channels. Logical Channels carry the information, being mapped onto Physical ones. Regarding the Logical Channels, one can be divide into two groups: Traffic (TCH) and Control Channels. The first one is used to carry user data, which can be either speech or data, and the second one is used for control and signalisation between the BS and the UE, e.g. synchronisation and information for a possible handover. In Portugal, [ANAC16], the three operators use the band of 900 MHz (GSM900) and 1800 MHz (GSM1800).

Wideband Code Division Multiple Access (WCDMA) is used as the radio interface of UMTS, and it is a wideband Direct-Sequence Code Division Multiple Access (DS-CDMA) system, using FDD. User information bits are spread over a wide bandwidth by multiplying this user information with chips (quasi-random bits derived from the Code Division Multiple Access (CDMA) spreading codes). With 3.84 Mcps as maximum chip rate and a carrier bandwidth of 5 MHz, WCDMA supports highly variable data rates. This data rate is kept constant during 2 ms, duration of a frame (one frame is capable of 38400 chips), on which data capacity can change from frame to frame. The network will typically control the allocation of this data capacity, to achieve optimum throughput for packet data services.

In order to support higher DL data rates, High-Speed Downlink Packet Access (HSDPA) was added to Release 5, mainly intended for non-real time traffic, on which the theoretical peak data rate is 14.4 Mbps. This improvement was achieved by an improvement in efficiency at modulation and coding, where Quadrature Phase-Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (QAM) and multicode operation with a spreading factor fixed in 16 are used. In Release 7, to improve data rate even more, 64QAM and Multiple Input Multiple Output (MIMO) were introduced. In Release 6, High-Speed Uplink Packet Access (HSUPA) was introduced, with the same objective of HSDPA, but for UL. In this case, the theoretical data rate is 5.8 Mbps, with the use of Binary Phase Shift Keying (BPSK). Table 2.1 summarises the fundamental properties of UMTS.

WCDMA mainstream band is 2100 MHz (Band 1), but in Europe and Asia Band 3, 1800 MHz, and Band 8, 900 MHz, are also used, [HoTo06]. Release 7, introducing Evolved High-Speed Packet Access (HSPA +), was the first step to LTE goals, with higher data rates achieved by improvements in modulation and with the use of MIMO.

Table 2.1. Fundamental properties in 3GPP Release 99, 5, 6 and 7 (adapted from [Vena14]).

	Release 99	**Release 5 (HSDPA)**	**Release 6 (HSUPA)**	**Release 7 (HSPA +)**
Spreading Factor	Variable	Fixed in 16	Variable	Variable
Modulation	Fixed (BPSK for UL, QPSK for DL)	Variable (16QAM, QPSK)	Fixed (BPSK)	QPSK(UL); 16QAM or 64QAM (UL/DL);
Maximum Data Rates [Mbit/s]	2	14.4	5.8	21.1 (UL); 42 (DL)

2.2 LTE

Section 2.2 introduces the fundamental concepts of LTE, mainly the aspects of network architecture and an explanation of the radio interface, based on [HoTo11], [CCox12] and [SeTB11].

2.2.1 Network Architecture

The need for PS services optimisation, and improvements in the user bit rates led the discussion for System Architecture Evolution (SAE), implemented in Release 8. However, some development already started in Release 7, where some 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 supports PS services.

Figure 2.2 describes the basic LTE architecture and elements configuration, with the logical nodes and connections. The architecture can be divided into four main domains: UE, Evolved UTRAN (E-UTRAN), Evolved Packet Core Network (EPC) and the Services.

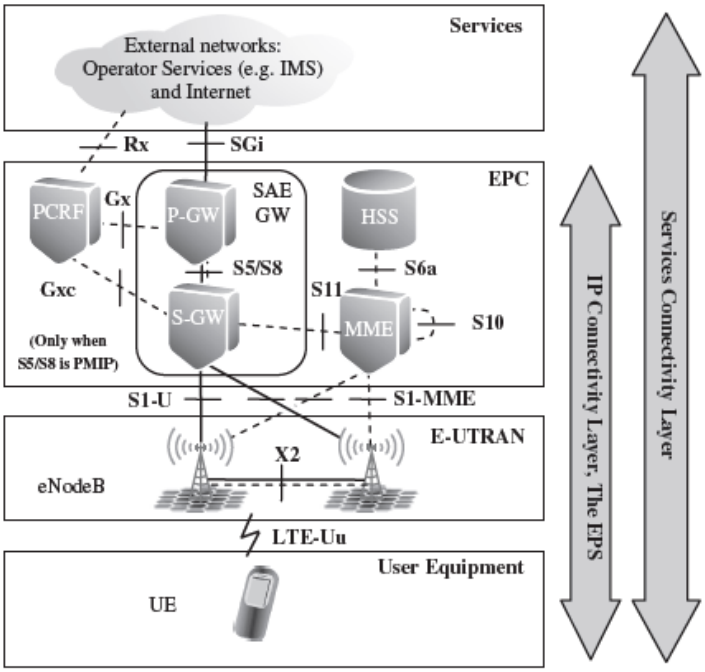


Figure 2.2. System architecture for E-UTRAN (extracted from [HoTo11]).

UE, E-UTRAN and EPC represent the Internet Protocol (IP) Connectivity Layer, also called as the Evolved Packet System (EPS), with the main function to provide IP-based connectivity. EPS together with Services is represented as the Service Connectivity Layer. UE is equally represented in the UMTS architecture and has the same functionality on LTE network, also composed of ME and USIM.

In the E-UTRAN domain, the only node is the Evolved Node B (eNodeB). These are radio base stations control all radio related functions in the fixed part of the network between UE and EPC. It is also responsible for Radio Resource Management (RRM), controlling the allocation of resources, prioritising

and scheduling traffic, conforming to the required Quality of Services (QoS). Another important role is Mobility Management (MM), where the eNodeB, to make decisions in UEs handovers, controls and analyses radio signal level measurements.

Regarding EPC, Mobility Management Entity (MME) is the primary control element, being also the primary control channel between UE and the network, and also taking care of user authentication and management of security. MME is permanently requesting information to the Home Subscription Server (HSS), a database that contains the authentication information, e.g. the UE profile and location and the permanent key used to calculate the authentication vectors.

EPC is also composed of the Serving Gateway (S-GW) and the Packet Data Network Gateway (P-GW), who are part of the System Architecture Evolution Gateway (SAE-GW). The high-level function of S-GW is the responsibility for tunnel management and switching, acting as the local mobility anchor during mobility between eNodeBs. Concerning P-GW, it is the edge router between EPS and external packet data networks, performing also traffic gating and filtering functions, usually acting as IP point of attachment for the UE.

The Policy and Charging Resource Function (PCRF) is the Network Element responsible for Policy and Charging Control (PCC), making the decision on how to handle the services regarding QoS, providing information for the P-GW and the S-GW, so that appropriate bearer and policing can be set up.

Services are external networks, which the operator does not provide directly, but which can be accessed by the network. An example is the IP Multimedia Subsystem (IMS), a service machinery used to provide services, using Session Initiation Protocol (SIP) or the common connection to a server on the internet.

2.2.2 Radio Interface

LTE DL multiple access is based on Orthogonal Frequency Division Multiple Access (OFDMA), and the UL on Single Carrier Frequency Division Multiple Access (SC-FDMA). This technique for radio transmission and reception is a powerful way to minimise the problems of fading and Inter-Symbol Interference (ISI). Orthogonal Frequency Division Multiplexing (OFDM), from where OFDMA is based, has fewer problems with the interference and bit errors at the receiver problem. Instead of sending the information as a single stream, an OFDM transmitter divides data into several parallel sub-streams, sending each sub-stream on a different frequency know as sub-carrier.

In OFDMA, the base station shares its resources by transmitting at different times and frequencies, sharing up to a maximum of 1200 sub-carriers at Release 8, with a fixed spacing of 15 kHz with a symbol duration of 66.7 µs. The overall motivation for OFDMA in LTE has been due to the excellent performance in frequency selective fading channels, its good spectral properties, handling of multiple bandwidths and the compatibility with advanced receiver and antenna technologies.

Regarding SC-FDMA, it is used in UL, because the power of the signal transmitted by the UE is subject to significant variations, which can cause problems related to the distortion of the waveform, leading to leaks into an adjacent frequency band that would cause interference to other receivers.

DL transmission resources in time-frequency are subdivided in a structure, where the largest unit of time is a 10 ms frame, divided into ten 1 ms subframes, each of which separated into two 0.5 ms slots. Each slot, in normal Cyclic Prefix (CP), has seven OFDM symbols or six symbols in the extended CP case. In frequency, resources are grouped into 12 sub-carriers, occupying a total of 180 kHz (12 × 15 kHz), which is called a Resource Block (RB). The smallest resource unit is the Resource Element (RE), which consists of one sub-carrier for a duration of one OFDM symbol.

LTE also uses FDD and TDD modes, where 22 bands are specified for FDD and 9 for TDD, but only six from this bands are used in Europe. The most relevant, [HoTo11], is Band 7, at 2600 MHz. Both Band 3 and Band 8, at 1800 MHz and 900 MHz respectively, are also used by GSM, are attractive from the coverage viewpoint, and finally Band 20 at 800 MHz.

In Portugal, [ANAC16], LTE FDD is used by the three major network operators in Band 20, Band 3 and Band 7. In Band 20, each of the operators has a total spectrum of 2 × 10 MHz of exclusive utilisation, and for Band 3 and 7 this value increases to 2 × 20 MHz.

LTE uses three modulation schemes altogether: QPSK, 16 QAM and 64 QAM. For DL, all modulation schemes are used, but for UL only QPSK and 16 QAM are used. Combining the utilisation of the new architecture, this type of modulation and the use of MIMO, the theoretical maximum bit rate increases to 300 Mbps in DL and 75 Mbps in UL, [3GPP16].

2.3 Modelling alarms and incidents

2.3.1 Alarm Definition

An alarm notification can be described, [Wall09], by a set of valuable information with which network managers are alerted, to prevent service outage or degradation, described as a set of five conditions. One represents an alarm by the Resource, where the abnormal condition appeared, the Alarm Type, representing the classification referring to the undesirable state, and Time, Severity and Information. Network managers use the last parameter to add extra information to be used in the future.

The Severity parameter is used to range malfunctions, from the most severe to the least one, as Critical, Major, Minor or Warning, and being described also as Cleared and Indeterminate, [ITUT92]:

- Critical: The service affecting condition has occurred, and an immediate corrective action is required, which can occur when, for example, an element becomes totally out of service.
- Major: The service affecting condition has developed and an urgent corrective action is required; it can be reported, for example, when there is a severe degradation in the element's capability.
- Minor: A non-service affecting the fault condition and corrective action should be taken to prevent a more serious fault; it can be reported, for example, when the alarm is not currently degrading the capacity of the element.
- Warning: The detection of a potential or impending service that can cause a fault, before any

significant effects have been felt; an action should be taken to diagnose further (if necessary) and correct the problem to prevent it from becoming a more serious service-affecting fault.

- Indeterminate: The Severity level cannot be determined.
- Cleared: The Severity level indicates the clearing of previously reported alarms.

A distribution of severities from 2nd (2G) and 3rd Generation (3G) networks, together with a normalised plot of the distribution of manually assigned priorities in the trouble ticket system is presented in Figure 2.3. This representation is complemented with the presentation of a recommend severity distribution from [HoHa06], analysed from an alarm database [Wall09]. From Figure 2.3, one can see that the maturity of 3G is not satisfactory, because the distribution is almost the opposite of the recommendation. On the other hand, 2G had reached the proposed maturity, with alarms distribution almost reaching the recommendation.

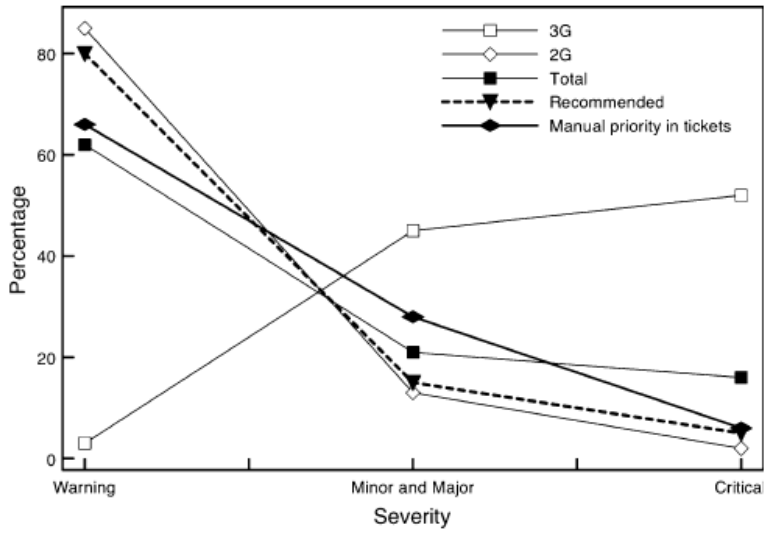


Figure 2.3. Severity distribution (extracted from [Wall09]).

One can generically categorise the origin of alarms into five categories, [ITUT92]:

- Communication alarm type: Associated with the procedures and/or processes required to carry information from one point to another.
- Quality of Service alarm type: Associated with a degradation in the quality of service.
- Processing Error alarm type: Associated with a software or processing fault.
- Equipment alarm type: Mainly associated with equipment fault.
- Environmental alarm type: Mainly related to the condition of an enclosure in which the equipment resides.

A combination of probable causes that could trigger alarms with the kind of alarm is described in Table 2.2. To categorise the alarms by their propagation in the network, [Wall09] uses a denomination of four levels, where each level denotes a different step in the alarm chain, Figure 2.4.

The first level, Level 0, also called The Phenomenon, is where the Resource state changes, interpreted as an alarm. A base station detects a problem considered to be an alarm, and an alarm software management is used to transfer it to the management system, over Level 1.

Table 2.2. Combination of probable causes and alarm types (adapted from [ITUT92]).

Category of an alarm type	Probable cause
Communication	• Loss of signal • Call establishment error • Communications protocol error
Quality of Service	• Response time excessive • Performance degraded • Congestion
Processing Error	• Software error • Out of memory • File error
Equipment	• Power problem • Receiver failure • Input device error
Environmental	• Temperature unacceptable • Fire detected • Leak detected

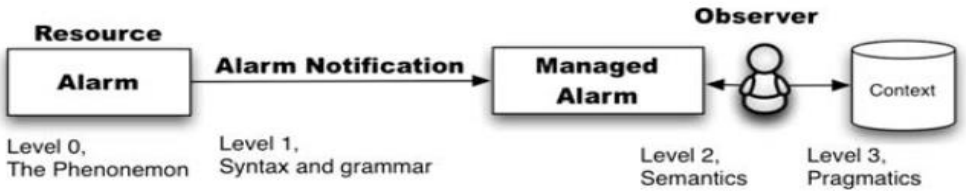


Figure 2.4. Alarm taxonomy (extracted from [Wall09]).

At Level 1, also called Syntax and grammar, the protocol and information modelling language defines the alarm interface, where the global interface is defined.

Regarding Level 2, called Semantics, it is the interpretation of the various alarm fields, such as severity, probable cause and the managed object (element). At this level, it is possible, for example, to check if a virtual connection was cut in the transport network between the RNC and Radio Base Station.

Finally, Level 3, also called Pragmatics, corresponds to the understanding of which cells and channels are affected by the alarm. An investigation on the Service Level Agreement (SLA) from service providers to the customer is also performed, as well as demographic data to understand the problem's real impact.

2.3.2 Faults and network faults propagation

According to [Kuhn97], the primary sources of failures are human errors and acts of nature. The former can be split into two types: caused by operator's employees and or by non-employees. In Table 2.3, one describes the most common causes of network faults, percentage of downtime (measured in customer minutes value, i.e., affected customers multiplied by the outage duration in minutes), and the number of outages from US Federal Communications Commission (FCC) reports from April 1992 to March 1994.

One can draw from Table 2.3 that the number and magnitude of outages differ significantly for each failure categories. For example, overloads only had 18 outages in two years, representing 44% of

downtime and human errors caused 150 outages in the same period, but “only” 28% of downtime, which can be explained by the maintenance working period, done overnight, with a less volume of traffic.

In the interest of diagnosing a fault, this process usually involves three steps, [StSe04]:

- Fault detection: a process of capturing indications of network disorder, provided by malfunctioning devices in the form of alarms.
- Fault localisation (also stated as fault isolation, event correlation and root cause analysis): a set of observed faults indications is analysed, to find an explanation for the alarms.
- Testing: a process that, given several possible hypotheses, determines the actual faults.

Table 2.3. Failure categories (adapted from [Kuhn97]).

Category	Source	Examples	Percentage of Downtime	No. Outages per year
Overloads	Service demand exceeds the system capacity	• Big events.	44%	9
Acts of nature	Large and minor natural events or Natural disasters	• Cable, power supply, or facility damaged; • Earthquakes, hurricanes, floods.	18%	16
Human error - company	Errors made by operator’s personnel	• Maintenance; • Mismatches in software versions.	14%	38
Human error - others	Errors made by non-operator personnel	• Cable cutting; • Accidents.	14%	37
Hardware failures	Hardware component failures	• Failures of cable components	7%	28
Software failures	Internal errors in the software	• Software errors.	2%	22
Vandalism	Sabotage or other internal damage	• Copper Stealing.	1%	2

Fault localisation techniques are based on many paradigms., which derive from different areas of computer science, including artificial intelligence, neural networks (NN) and information theory. Figure 2.5 shows some solutions, which include artificial intelligence or fault propagation models.

Artificial intelligence reflects actions of a human expert when solving problems in a particular domain, by imitating the knowledge of a person, which can be accomplished by the understanding of the system behaviour from its principles, or from experience. The system uses a rule-based representation of their knowledge-base, where, in each cycle, it chooses rules for execution, whose antecedents (conditions) match the content of the working memory.

Model traversing techniques use a formal representation of a communication system with clearly marked relationships among network entities. By exploring these relationships, the fault identification process can determine which alarms are correlated and locate faulty NEs. They are robust against frequent network configuration changes and are particularly attractive when automatic testing of a managed

object is done as a part of the fault localisation process. On the other hand, they are unable to model situations in which the failure of a device may depend on a logical combination of another device failure.

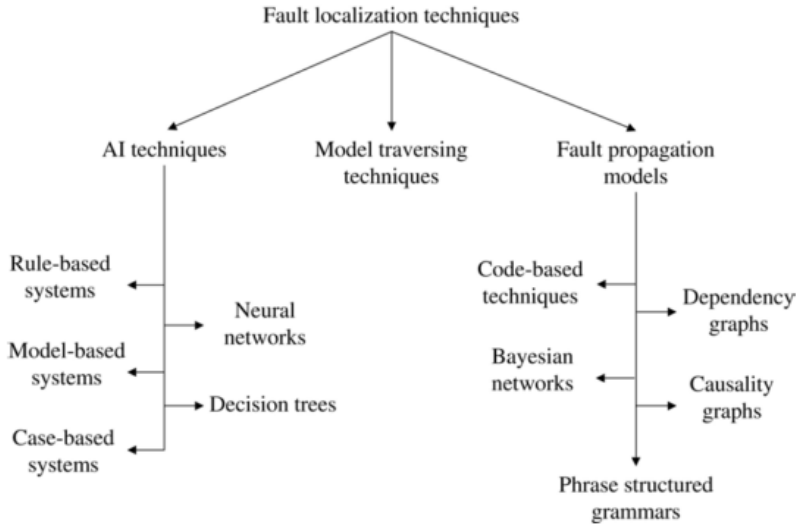
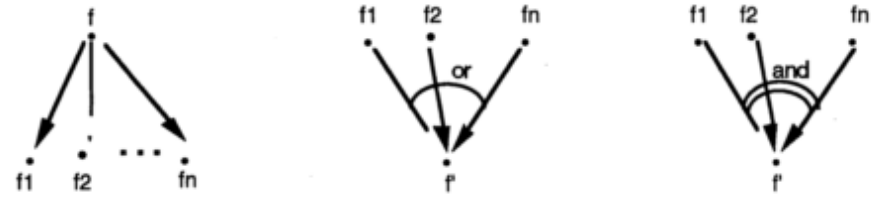


Figure 2.5. Classification of fault localisation techniques (extracted from [StSe04]).

A fault propagation model describes the alarms observed when a fault occurs, and includes the representation of all faults and alarms, requiring a prior specification of how a failure or alarm, in the components, are related to failures or alarms in other elements. According to [JaWe95], if faults can be linked, i.e. if they are not independent, causal relations between faults can be created, represented by faults propagations rules. One represents in Rule a) of Figure 2.6 a fault *f* as a root cause for multiple faults *f1, f2, …, fn*; in rule b), the fault *f'* can be due to any of the faults *f1, f2, …, fn*; finally, in rule c), all faults *f1, f2, …, fn* should be present to cause fault *f'*.



a) $f \rightarrow f1, f2, \ldots fn$ b) $f1 \vee f2 \vee \ldots \vee fn \rightarrow f'$ c) $f1 \wedge f2 \wedge \ldots \wedge f3 \rightarrow f'$

Figure 2.6. Fault propagation rules (extracted from [JaWe95]).

There are several reasons why a single fault in a network causes multiple alarms, being sent to SNOC, including, [HoCF95]:

- multiple alarms generated by the same device for a single fault (known as alarm streaming);
- the fault is intermittent and each re-occurrence results at the beginning of new alarm;
- the fault is reported each time a service provided by the failing component is invoked;
- multiple components detect the same condition;
- the fault propagates by causing dependent failures and resultant alarms.

By testing a real alarm stream data, [HoCF95] encountered a few practical problems on finding the root

failure. One of them was the hidden dependencies in the network, where sometimes the visibility of the network is often limited and not all the dependencies that point to the root cause of a set of symptoms are known, described in Figure 2.7.

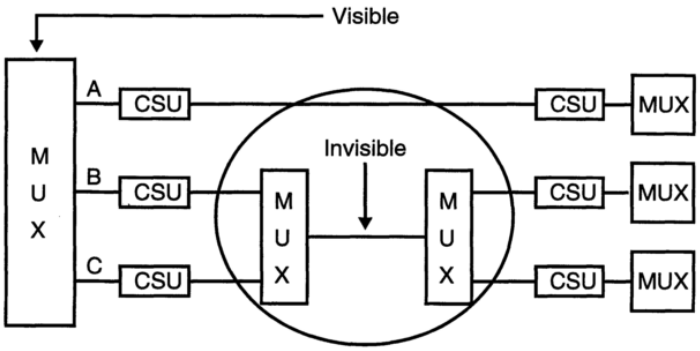


Figure 2.7. Example configuration with Hidden Dependencies (extracted from [HoCF95]).

The other problem is the complex dependencies of the network, in which, if one resource fails, all dependents fail with it. Since not all the dependencies are simple to understand, the root problem is hard to find. Sometimes some data is missing, thus complicating the correlation of alarms.

2.3.3 Alarms Correlation

The use of a management centre allows correlation of alarms as they are sent by NEs, Figure 2.8, by combining the fragmented information and interpreting the flow of alarms, reducing the amount of information presented to network managers. The correlation is achieved by removing redundant information, filtering low-priority alarms when higher-priority alarms are present, and replacing a set of alarms by some latest information.

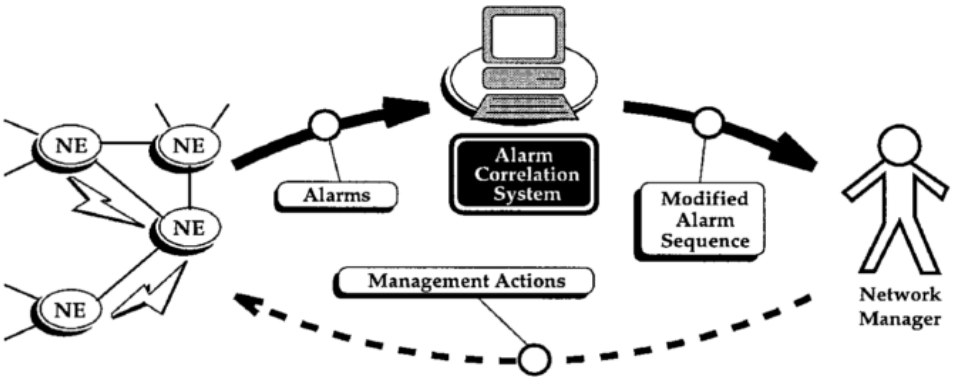


Figure 2.8. Flow of alarms with an alarm correlation system (extracted from [KlMT99]).

A situation can be recognised in an alarm sequence within a time window by a correlation pattern, which, typically, is an expression of the set of active alarms of, for example, the last five minutes. Associated with each correlation pattern is an action, which is executed when there is an occurrence of the corresponding pattern. One example is when, for instance, both alarms A and B co-occur, which can be followed by fatal problems by their sender, Figure 2.9.

Figure 2.9. Example of a correlation action (extracted from [KIMT99]).

Depending on the nature of alarms, one can consider the types of correlation described in Table 2.4. On alarm compression (2.1), one can reduce multiple occurrences of identical alarms into a single one representative of the whole event. Regarding alarm filtering (2.2), if some parameter $p(a)$ of alarm a, e.g. priority or location of the NE, does not fall into a set of predefined values of H, the alarm is discarded or sent to a log file; this decision is based on the characteristics of a. Concerning event suppression (2.3), it is a context-sensitive process, in which a is temporarily constrained, depending on the context C of the network management process; the presence of other alarms or other external requirements determine C. One can also make a correlation between counting and thresholding (2.4) the number of repeated arrivals of identical alarms. Alarm escalation (2.5) assigns a higher value to some parameter $p'(a)$ of a, e.g. severity, depending on the number of occurrences of the alarm. Alarm generalisation (2.6) is a correlation in which a is replaced by its superclass b, allowing the network manager to analyse situations from a higher-level perspective of network alarms. Regarding alarm specialisation (2.7), it is a different procedure from alarm generalisation, substituting an alarm with a more specific subclass of the event. Temporal relation (2.8) T between a and b allows them to be correlated to the order and the time of arrival. Finally, event clustering (2.9) allows the creation of complex correlation patterns, using logical operators $\wedge$ (and), $\vee$ (or) and $\neg$ (not); these patterns can be such as another correlation, network events or tests of network connectivity.

Table 2.4. Types of alarm correlation (adapted from [JaWe95]).

Correlation		
Compression	$[a, a, \dots, a] \Rightarrow a$	(2.1)
Filtering	$[a, p(a) < H] \Rightarrow \emptyset$	(2.2)
Suppression	$[a, C)] \Rightarrow \emptyset$	(2.3)
Count	$[n \times a] \Rightarrow b$	(2.4)
Escalation	$[n \times a, p(a)] \Rightarrow a, p'(a), p' > p$	(2.5)
Generalisation	$[a, a \subset b] \Rightarrow b$	(2.6)
Specialisation	$[a, a \supset b] \Rightarrow b$	(2.7)
Temporal Relation	$[a\ T\ b] \Rightarrow c$	(2.8)
Clustering	$[a, b, \dots T, \wedge, \vee, \neg] \Rightarrow c$	(2.9)

After the correlation of an alarm, an examination is required on if this alarm is important enough for a trouble ticket to the network manager, and then, if necessary, the creation of the ticket and of its priority, Figure 2.10.

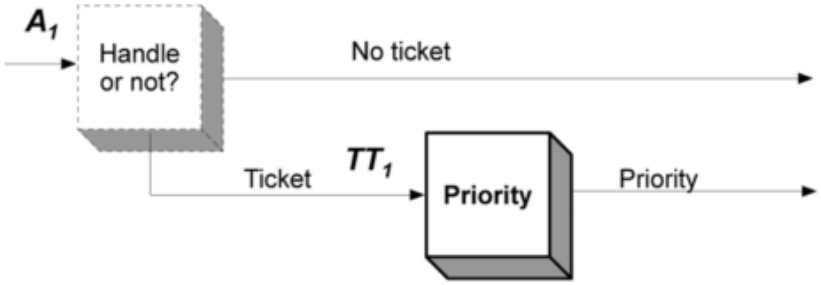


Figure 2.10. Two-step alarm management (extracted from [WaLL09]).

To prioritise an alarm, [WaLL09] suggested the integration of a neural network into the trouble ticket system, using manually assigned trouble ticket priorities and associated alarms as learning data. When the alarm is received, it questions the trained neural network, which outputs a priority to that alarm, indicating if it should be handled or not, and, if so, the priority. To train the network, A_1 data is used, to judge if a trouble ticket should be created, and TT_1 data is used to train the network to assign priorities.

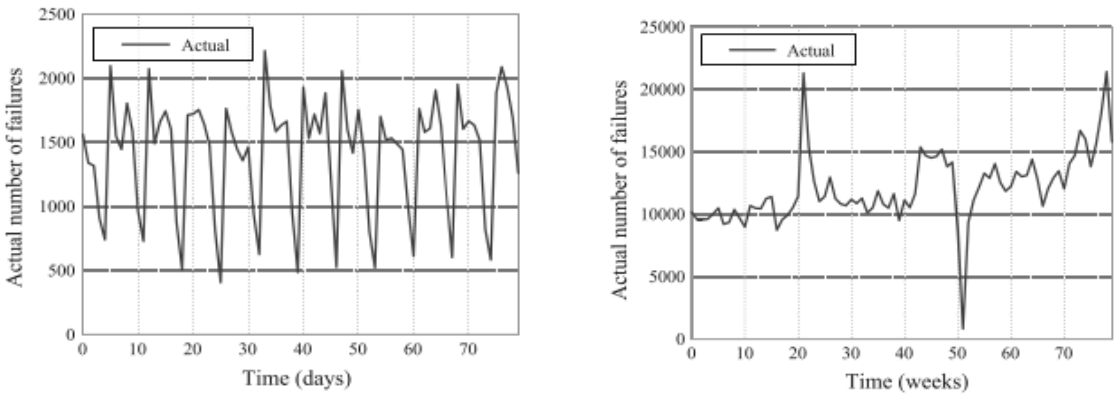
2.4 Failure prediction approach

2.4.1 Time Series

To make a proper system prediction, one needs to characterise data. As reported by [ŽeKu10], the arrival of faults can be considered as a statistical process in time, shown as an ordered time series, which can be univariate (one set of data for one period) or multivariate (multiple sets of data for one period) data collected over time. One can also describe them as a stochastic series, since future results can only be estimated and not calculated precisely. Knowing that are two relevant characteristics to describe a time-series, which are stationary and linearity/nonlinearity, [ŽeKS11a] outlines the time-series of faults as non-stationary, due to its high-level of daily fluctuations; concerning linearity, one should note that the variables can be both linear or nonlinear, meaning that specific causes have a nonlinear effect on the number of reported faults.

After the analysis of real data from an operator, [ŽeKS11b] identifies a certain periodicity in hourly and daily intervals in the number of faults, which can be explained by the daily routines that characterise the usage of these services. Furthermore, in series with weekly and monthly intervals, the seasonality is not notable due to the influence of random factors, such as extreme weather or unexpected breakdowns. One presents in Figure 2.11 the number of failures in daily and weekly intervals, where it is possible to observe the periodicity in Figure 2.11 a) and its absence in Figure 2.11 b).

Linear time series can be described by an Autoregressive Integrated Moving Average (ARIMA), while nonlinear ones are more adequately described by neural networks. According to [JaIH04], neural networks have been applied effectively in the identification and control of dynamic systems, being effective when applied to problems whose outputs require previous knowledge. The efficiency of neural networks depends strongly on inputs, hence, their importance on fault analysis.



a) Number of failures in a daily interval. b) Number of failures in a weekly interval.

Figure 2.11. Number of failures in a daily and weekly interval (extracted from [ŽeRK16]).

2.4.2 Data correlation

To better understand the association between the number of incidents and weather variables, one needs to understand the relationship between these variables. A possible method to quantify this relation is by calculating a correlation coefficient, which can be done following three methods: Pearson [RFGD08], Spearman [JZar05] and Kendall's τ [ChPo02]. These methods output a number between -1 and +1, expressing how closely the two variables are related to each other. The ±1 shows a perfect relationship and 0 indicated no connection. According to [Dumm17], one can also relate a value within ± 0.3 as a weak relationship, of ± 0.5 as a reasonable relationship, and finally larger than ±0.7 as a strong one. To exemplify the correlation method, one defines two generic datasets, X e Y. The X dataset can be generically defined as the target, and the Y one as the inputs that lead to the target.

Pearson's correlation is [RFGD08] ideal if data follow a bivariate Normal Distribution, being a method vulnerable to data deviation of any kind, thus, data transformation to approach the required bivariate Normal Distribution is necessary. Pearson's coefficient r is calculated by [RoNi88],

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \ \sum(Y_i - \bar{Y})^2}} \tag{2.10}$$

where:

- X_i: Dataset X in index i;
- Y_i: Dataset Y in index i;
- $\bar{X}$: Mean of dataset X;
- $\bar{Y}$: Mean of dataset Y.

Spearman method provides a nonparametric (distribution-free) measure of correlation between two variables, requiring only an either increasing or decreasing monotonically relationship. The ranks of the sorted values determine the result, not the actual data values, thus, data is first sorted, and then, the Spearman correlation of the ranks is computed. One of the significant advantages of Spearman's

correlation, [RFGD08], is that results are the same for the original data or any linear transformation, as the transformation does not disrupt the order of data values. According to [JZar05], one obtains:

$$r_s = 1 - \frac{6\ \sum\left(R_p - T_p\right)^2}{n^3 - n} \tag{2.11}$$

where:

- R_p: Rank of dataset X in position p;
- T_p: Rank of dataset Y in position p;
- n: Number of ranks.

Kendall's τ method, quite similar to Spearman's, also measures the range of increasing or decreasing relationships between pairs of variables monotonically. This method is relatively robust against data deviation, since if the sign of the slope does not change, the result will stay the same. Thus, it is widely independent of the actual data values, and a linear transformation does not change the estimated correlation coefficient. One can concordant define pairs by, [ChPo02]:

$$Y_i < Y_j \text{ if } X_i < X_j \ \vee\ Y_i > Y_j \text{ if } X_i > X_j \ \vee\ (X_i - X_j)(Y_i - Y_j) > 0 \tag{2.12}$$

where:

- X_j: Dataset X in index $j \neq i$;
- Y_j: Dataset Y in index $j \neq i$.

and discordant ones by,

$$Y_i < Y_j \text{ if } X_i > X_j \ \vee\ Y_i > Y_j \text{ if } X_i < X_j \ \vee\ (X_i - X_j)(Y_i - Y_j) < 0 \tag{2.13}$$

To calculate the rank correlation, one uses,

$$\tau = \frac{2\ (E - Q)}{l(l-1)} \tag{2.14}$$

where:

- E: Number of concordant pairs;
- Q: Number of discordant pairs;
- l: Actual size of the sample.

2.4.3 Forecasting methods

One of the objectives is the forecasting of the number of incidents, and the use of linear regression is one of the simplest ways to accomplish this. Since one needs to predict the number of incidents regarding several variables, the multiple linear regression method is used, employing a single predictand, the number of incidents, with more than one predictor, i.e., several variables. The prediction equation is defined by, [Wilk06],

$$y_{predictand} = b_0 + m_1 x_1 + m_2 x_2 + \cdots + m_k x_k \tag{2.15}$$

where:

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

When representing the number of incidents versus two variables, one uses a variation of the regression equation, a factor representing the interaction between variables being included,

$$y_{surface} = b_0 + m_1 x_1 + m_2 x_2 + m_3 x_1 x_2 \quad (2.16)$$

where:

- $y_{surface}$: Regression surface;
- m_3: Interaction between variables.

However, in more complex cases, other forecasting models are used. For example, in data similar to the one studied in this work, Artificial Neural Networks (ANN) present reliable results. The most widely used neural network is the multi-layer perceptron (MLP) one, which is the simplest topology for time series prediction with time-delayed inputs. The essential objective of an ANN is [AdAg13] to build a model to resemble the function of a human brain, a recognition of input data patterns and experiential learning being done to model the network, and then provide outputs based on the previous knowledge; this model is characterised by a network of three layers: input, hidden (could be more than one) and output, connected by acyclic links. A three-layer feedforward architecture is described in Figure 2.12.

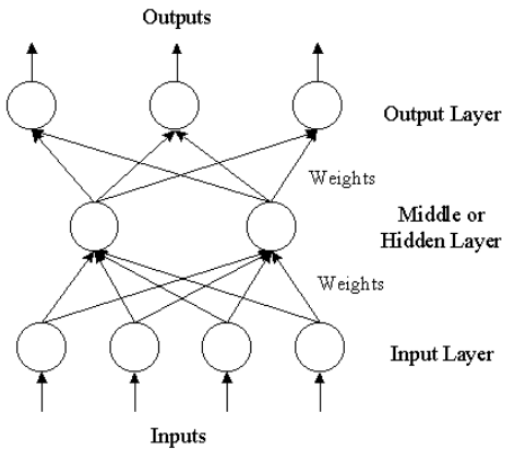


Figure 2.12. Three-layer feeds forward neural network architecture (extracted from [AdAg13]).

Neural networks have the advantage of not needing to specify a particular model form or any assumption about the statistical distribution, being based on the features presented from data, which is useful for many practical situations. Since the neural network is nonlinear, it is more accurate to model complex data patterns, [AdAg13]. Concerning the disadvantages, [JaTu96] refers the difficulties in identifying possible causal relationships, being likely to overfit. One can define the output of MPL by [AdAg13],

$$y_{MLP} = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{u} \beta_{ij} Y_{t-1}\right) + \varepsilon_t, \forall t \quad (2.17)$$

where:

- 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 a method to describe time series, initially developed to solve pattern classification problems, such as text classification and face identification, but new applications, such as regression estimation and time series prediction problems, are being studied. The main objective is to find a decision rule capable of selecting some particular subsets of training data, [AdAg13]. One can define SVM in two cases: when data is linearly separable, and when it is not (i.e., nonlinearly). The former is defined by, [AdAg13],

$$\left.\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|w\|^2 \\ \text{Subject to} & Y_i(w^t X_i + b) \geq 1;\ \forall i = 1,2,\dots,v \end{array}\right\} \tag{2.18}$$

where:

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

While the latter, e.g. XOR classification, is, [AdAg13],

$$\left.\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{N}\xi_i\right) \\ \text{Subject to} & Y_i(w^t X_i + b) \geq 1 - \xi;\ \forall i = 1,2,\dots,v \wedge \xi_i \geq 0 \end{array}\right\} \tag{2.19}$$

where:

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

An advantage of SVM is that the solution is always unique and globally optimal, but on the other hand, it has the disadvantage of, when having a large training size, taking a considerable amount of computation time.

Regarding Bayesian Networks, [Heat13] refers them 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 input data, typically used for predictive modelling and pattern recognition, being defined by, [Vaně08],

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{v=1}^{n} P(X_v = x_v | X_{v+1} = x_{v+1}, \dots, X_n = x_n) \tag{2.20}$$

where:

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

The advantages of using Bayesian Network relies on being correctly in handling missed values with the possibility of doing queries. A problem, as stated in [Vaně08], is that to calculate the probability of any branch of the network it is necessary to calculate all branches, leading to computational difficulties.

Another method, Nearest Neighbours, determines a point in a dataset that is the nearest to a query one, [BGRU99], which is accomplished by examining the distribution of the distance between query and data points. The identification of the distance is made by an evaluation of the number of points that are longer than a factor of the distance to the query point; one typically uses the Euclidean distance, [Alon12],

$$d(X,Y) = \sqrt{\sum_i (X_i - Y_i)^2} \tag{2.21}$$

Besides, [Alon12] refers that Nearest Neighbours have the advantages of having a lower learning process cost, and that the complex concept can be learned by local approximation using simple procedures. Regarding the disadvantages, it is computationally expensive to find the nearest neighbours when the dataset is vast, and performance depends on the number of dimensions.

After a comparison of the different prediction methods for the forecasting of faults, [ŽeKS11a] concluded that the best results were not achieved by using traditional methods, but either by applying dynamic models. One, Nonlinear Autoregressive Network with Exogenous Inputs (NARX), uses a back connection from the output layer as a component of the input, accentuating the output values sequences where the output data is preserved in a delayed memory line, Figure 2.13.

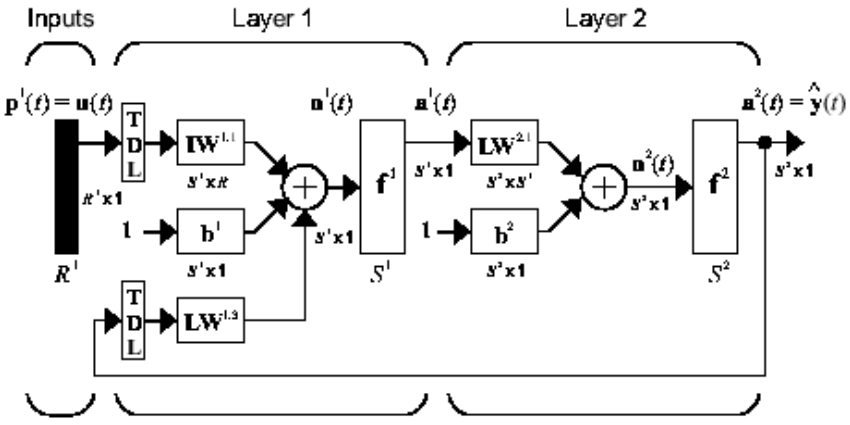


Figure 2.13. NARX scheme (extracted from [Matl16]).

One of the advantages of using NARX, [XiTL09], is its effectiveness in the gradient-descent learning rather than other architectures. Despite this, one of the drawbacks is that its feasibility as a nonlinear tool for time series modelling and prediction has not been fully explored yet. It is defined by, [SiHG97],

$$y(t) = \psi(u(t-n_u), \dots, u(t-1), u(t), y(t-n_y), \dots, y(t-1)) \tag{2.22}$$

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.

When $\psi(.)$ is approximated by a Multilayer Perceptron neural network, the resulting system is a NARX neural network, [LGHK97], which is well suited for modelling several nonlinear systems, such as heat exchangers, time series and various artificial nonlinear ones. The authors of [LGHK97] refer that NARX networks perform better on long-term dependencies problems, i.e., when the desired output of a system at a time T depends on inputs presented at times $t << T$.

2.4.4 Performance measures

To evaluate models' accuracy, one needs to measure and compare their performance, various performance measures, [AdAg13], being proposed in the literature to estimate forecast accuracy and to compare different models. The Mean Squared Error (MSE) is the most common performance measure,

$$\overline{\varepsilon^2} = \frac{1}{l}\sum_{t=1}^{l}(Y_t - f_t)^2 \tag{2.23}$$

where:

- Y_t: Actual value;
- f_t: Forecasted value.

One of the MSE proprieties is the possibility of showing if significant individual errors affect the total forecast error, [AdAg13], being also sensitive to the change of scale and data transformations, and penalising extreme errors while forecasting. Two attractive features of MSE, [WaBo09], are its simplicity, due to its inexpensive computation, and being widely used for optimising and assessing a variety of signal processing applications.

Regarding the performance of a linear regression, the coefficient of determination, R^2, gives one simple fit indicator, [ReFe10], providing a reasonable and rapid model fit indication,

$$R^2 = 1 - \frac{\sum_{i=1}^{l}\left(X_i - \widehat{X}_i\right)^2}{\sum_{i=1}^{l}(X_i - \bar{X})^2} \tag{2.24}$$

where:

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

Another method to evaluate performance is the standard deviation, [Wilk06], which defines the square root of the average squared difference between data points and their sample mean,

$$s = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}(X_i - \bar{X})^2} \tag{2.25}$$

2.5 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 even more alarms. The minimisation of failures with proper design and preventive maintenance is becoming more important. For operators, it is beneficial to predict and solve these failures to increase the quality of service offered to customers, to accomplish the SLAs and some rules from regulatory agencies. The opportunity of acting preventively and proactively enhances the motivation of using failure prediction to forecast possible failures in the network, enabling a quick reaction for solving these problems by operators. GSM, UMTS and LTE have a similar approach regarding the number of incidents, being more important the manner how the base station is protected from meteorological factors.

Some methods in cellular networks for correlating alarms aimed at the identification of faults have been studied. In [BoCF94], the authors proposed a framework to analyse the information given by alarms, suggesting possible hypotheses of faults, which can be used to perform alarm correlation to reduce the number of alarms presented to network managers. This framework consists of the representation of the systems and devices as nodes in a graph, and then when there is a failure, by traversing the graph one reaches the node that caused the fault; Figure 2.14 shows a possible graph with the dependence of alarms.

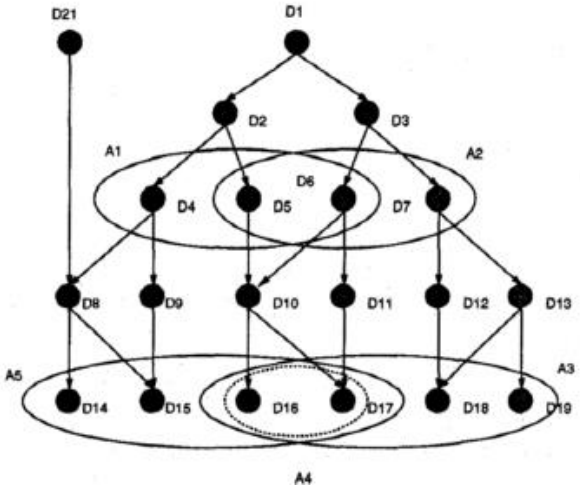


Figure 2.14. Graph demonstrating faults in dependent devices (extracted from [BoCF94]).

In [ŽeRK16], concerning the need to enable proactive action, a prediction of the expected number of failures in the network is achieved, and thus, with this information, giving the operator more time, and allowing the anticipation for future occurrences. A modelling of the number of failures as a time series is done, influenced by many factors, leading to a complex and nonlinear time series. By applying a statistical method, some elements such as outages, lightning, rainfall and announced work on the networks, are identified as the most significant predictable causes of faults, a temporal analysis of meteorological factors and failures being done. Using the results from [ŽeKS11a], the use of NARX is proposed, as the most likely network for predicting quantities of reported failures in complex systems.

By a correlation analysis, calculating Spearman's rank correlation coefficient, [ŽeRK16] determines the significance of the input factors to ensure a more accurate prediction of faults. The configuration of the NARX network is presented in Figure 2.15. One can conclude that the accuracy of prediction declines when the prediction period exceeds 3 or 4 days, this being the reasonable horizon to provide a forecast with considerable precision. This period is sufficient for operators to allocate their workforce and for network maintenance planning.

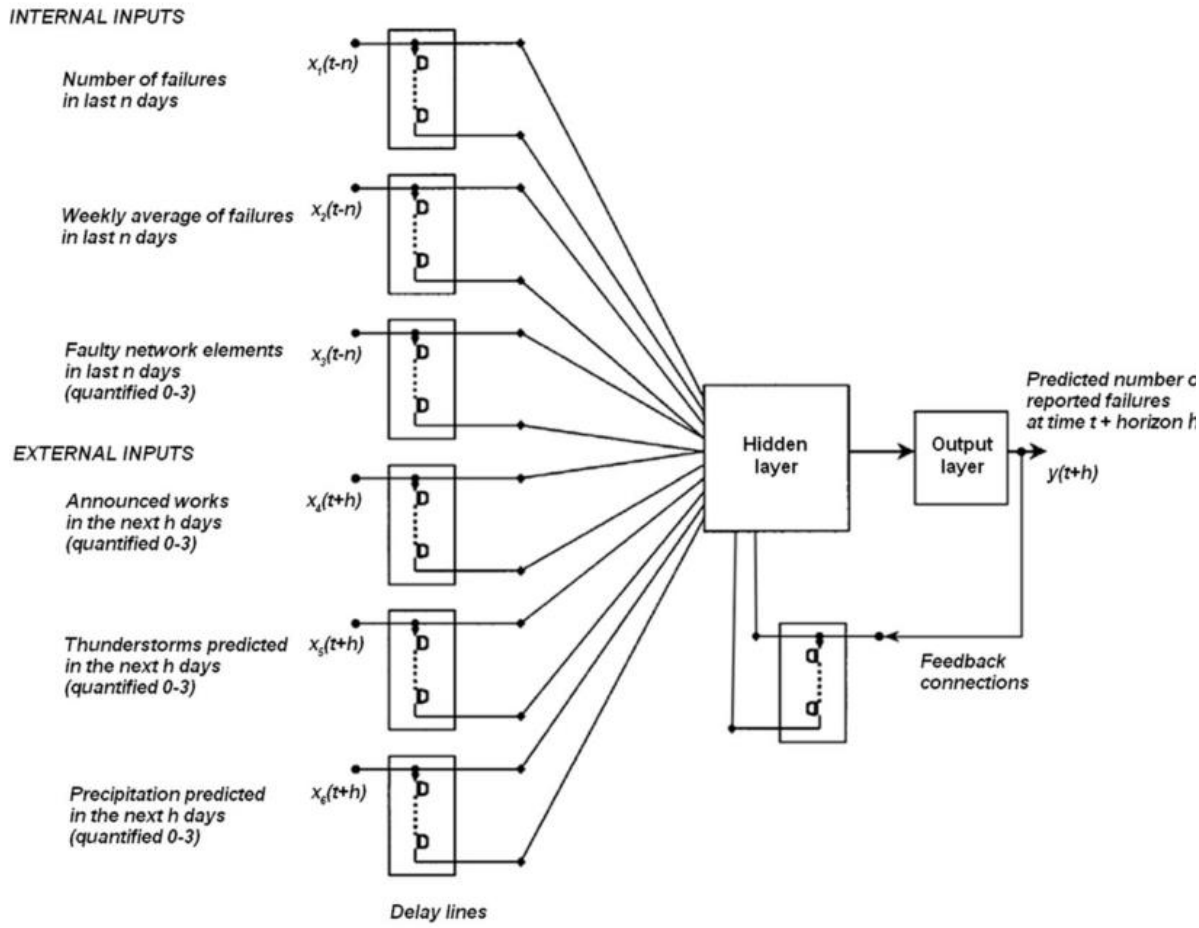


Figure 2.15. NARX network configuration (extracted from [ŽeRK16]).

Another use of NARX networks is the study of turbines, more specifically in the start-up operation of a gas turbine, [ACMP16]. The use of this kind of network was taken due to the capability of capturing the dynamics of complicated systems, as in the case of gas turbines. However, another Neural Network to study faults in turbines is typically used, e.g. [AGLA03] uses an ANN for fault diagnosis of a single-shaft industrial gas turbine and [OgSP02] applied an ANN for multi-sensor fault diagnosis of a stationary twin-shaft gas turbine.

Weather influence is extensively studied in other areas, like the effect of weather in power distribution networks; [BMHF11] presents the study of network's outages in electric facilities, mainly related to weather factors. One of the leading causes of incidents in telecommunications is electricity failures. In this study, the authors present a database with more than 40000 events in the distribution system, and

weather variables such as the quantity of lightning, wind speed, gust speed and rainfall level are considered, in the north western of Spain. These factors were the main cause for the more significant problems in the network, such as floods, storms or melted fuse, and a relationship between them is shown.

Another area of study on the influence of weather on the people's lives is health. [DBPA08] presents the study of how weather affects the mood of citizens, conducting a work about the effect of six weather parameters on three states of mood (positive affect, negative affect and tiredness). With a correlation study, it is concluded that the effects of those parameters on people's mood exist, being possible to deduce the consequences of weather on people's health.

Another example is presented in [YFHS11], similar to [DBPA08], with a correlation study between weather variables and patients' headache. In this case, with the help of 52 patients and five weather parameters, a correlation study on weather parameters together with headache incidence is conducted. A linear regression prediction of daily headache incidence using two methods is made, by using single and multiple weather variables. The authors conclude that with raw data no association between weather variables and headache incidence is found, but by using the empirical mode decomposition method in weather time-series parameters, an association between them is obtained.

Chapter 3

Dataset and Implementation Description

This chapter contains a description of the dataset used, explaining the variables under study. One also describes the statistical and forecasting study developed in this thesis, finalising with the assessment of the model.

3.1 Data Description

3.1.1 Dataset description

The data used in this thesis can be divided into two groups: Incidents and Meteorological. The former, provided by NOS [NOSP17], consisted of automatic reports extracted from the network management centre, with information sent automatically from base stations with some notes from the network manager. One presents in Table 3.1 the type of information, source and some examples possible to observe in this file.

Table 3.1. NOS incidents file information description with examples.

Type of information	Source	Examples
Origin	Group of the fault.	• Equipment; • Planned work.
Impact	If the service has suffered the impact.	• No impact; • Service Interruption;
Group	The team which deals with the fault.	• Core and Services.
Kind of equipment	Faulty equipment	• NodeB; • BTS.
Equipment	Kind of equipment with the fault.	• Antenna; • Cables.
Equipment fault origin	The source equipment that causes the fault.	• Base Station code; • Cell phone number.
Technology source	The technology which causes the fault.	• GSM; • LTE.
Problem beginning	Date and hour of problem start.	• Hour and date.
End of the problem	End date and time of problem	• Hour and date.
Description	A brief description of the network manager.	• Problem description.

A file with the location of the base stations was also provided by NOS, Table 3.2.

Table 3.2. NOS Base station localisation file description with examples.

Information	Source	Example
Base Station Name	Name of the base station.	LONDON_123L4
Site Code	Code that identifies the base station.	123L4
Technology	Technology used in the base station.	2G
Region (*Distrito*)	Region where the base station is placed.	Lisboa
Municipality (*Concelho*)	Municipality where the base station is placed.	Sintra
Parish (*Freguesia*)	Parish where the base station is placed.	Odivelas

The second set of data contains the information received from Weather Underground [Weat17], Table 3.3, and *Instituto Português do Mar e Atmosfera* (IPMA) [IPMA17], Table 3.4. The files from the former contain the meteorological information, except for electric discharges, each file being associated

with one month at a location. Each region is represented by a weather station, shown in Annex A, using, whenever possible, the airport meteorological station; for Lisbon, one did not use this weather station, since it lacks data from precipitation. There are two types of Weather Underground files: one, more detailed, presenting information in smaller intervals, e.g. 5 or 30 minutes intervals, depending on the weather station, and other with maximum values of each variable in a 24-hour interval.

Table 3.3. Weather Information provided by Weather Underground API.

Type of information	Information received	Units
Time	Time of the measure	h:m
Temperature	Temperature at the measure	°C
Humidity	Humidity at the measure	%
Precipitation	Precipitation at the measure	mm
Wind Speed	Wind Speed at the measure	km/h
Gust Speed	Gust Speed at the measure	km/h

Regarding the files received from IPMA, they were divided into several ones, each being associated with one month at one of three meteorological stations that IPMA has to collect electric discharges. The Reliability Parameter, Maximum axis and Minimum axis factors refer to the reliability of the location of the discharge. IPMA ensures that the provided discharges are under certain validation condition, i.e., an error of 50 km on which the discharge has 50% of location probability, distance to the detector of under 625 km and a reliability parameter under 10.

Table 3.4. Weather Information provided by IPMA regarding electrical discharges.

Type of information	Information received	Units
Id	IPMA identification of discharge	-
Date	Date of the discharge	Hour, minutes and seconds
Latitude	Latitude of discharge	GPS Coordinates
Longitude	Longitude of discharge	GPS Coordinates
Amperage	Intensity of discharge	kA
Reliability	Reliability parameter of discharge	-
Maximum axis	Localisation Error (Largest axis of an ellipse)	km
Minimum axis	Localisation Error (Smallest axis of an ellipse)	km

A positive correlation between some of the meteorological aspects, such as temperature or rain, in the occurrence of faults, is expected. Some of these relationships factors were studied in [ŽeRK16] to understand the relationship between them and the number of faults. [ŽeRK16] concluded that rain or electrical discharges have a positive correlation with the occurrence of faults, but on the other hand, fog and snow have a negative one.

The information received by Weather Underground about these weather variables is not complete. It misses, for example, data about snow or fog. Since these data are not possible to be obtained from the Weather Underground website, it was not studied. One examines the correlations between the weather variables available presented in Table 3.3 and electric discharges, with the number of incidents.

3.1.2 Meteorological Data

As referred to before, one needed to obtain historical meteorological data for Portugal, the Weather Underground website having been selected, due to the opportunity of getting these data based on an Application Programming Interface (API) for developers, provided by the website.

One developed an application based on JavaScript Object Notation (JSON) to receive the information from the website. A Python application was incorporated, to save data in an Excel file to be used in further applications. One presents the application flowchart in Figure 3.1, showing an explanation of each call requesting the detailed information at the weather station interval and the maximum values of the day. The use of the free version of this API meant that one could perform only 10 calls per minute and a maximum of 500 calls per day. Regarding electrical discharges, the data provided by IPMA is from the 3 meteorological stations that cover Portugal and part of Spain.

To understand the behaviour of the weather variables, one studied the maximum values recorded at each meteorological station. Regarding discharges, the highest value of the day is shown, if a discharge occurred. One presents in Table 3.5 the values of the mean, standard deviation (represented by Std. Dev.) and maximum (represented by Max) values of each weather variable: T describes Temperature, H describes Humidity, P defines Precipitation, W expresses Wind Speed, G defines Gust Speed, and I express Electrical Discharges Intensity.

Table 3.5. Weather variables, regarding mean, standard deviation and maximum, in Portugal.

	T [ºC]	**H [%]**	**P [mm]**	**W [km/h]**	**G [km/h]**	**I [kA]**
Mean	20.94	88.12	2.64	20.17	26.63	43.33
Std. Dev.	7.80	13.98	11.85	11.99	15.68	50.07
Max.	54.6	100	498.3	238	238	374.1

One should note that, due to problems in some meteorological stations, an approximation regarding Gust Speed was done: some stations consider zero when there was no Gust, but others take the maximum value of Wind Speed as the Gust Speed of that day, although there was no Gust. To normalise data, one considered the second case, that is, the maximum value of Wind Speed is considered the Gust Speed when the value is zero.

One presents in Figure 3.2 the histograms of all meteorological variables, to show data distribution. In this representation, one does not show the outliers of each variable, to obtain a better representation of the statistics of each weather variable. These outliers are one of the focus of this study, since they are related to the increase of incidents in the operator's network, however, they do not represent the usual behaviour of a variable.

One presents another example of the meteorological information used in [BLMD02], wherein a Mediterranean climate, a study about the influence of weather on a milk production is accomplished in two periods, Spring and Summer, and the values are presented in

Table 3.6. One can see this climate does not have an elevated mean Temperature. One can assume

that one of the probable causes of incidents peaks is the maximum values of weather variables, but still, one has analysed all data to make a relationship with weather variables. There is a lack of information regarding variables in this climate, most of the studies addressing extreme variables and not the distribution of weather and its statistics.

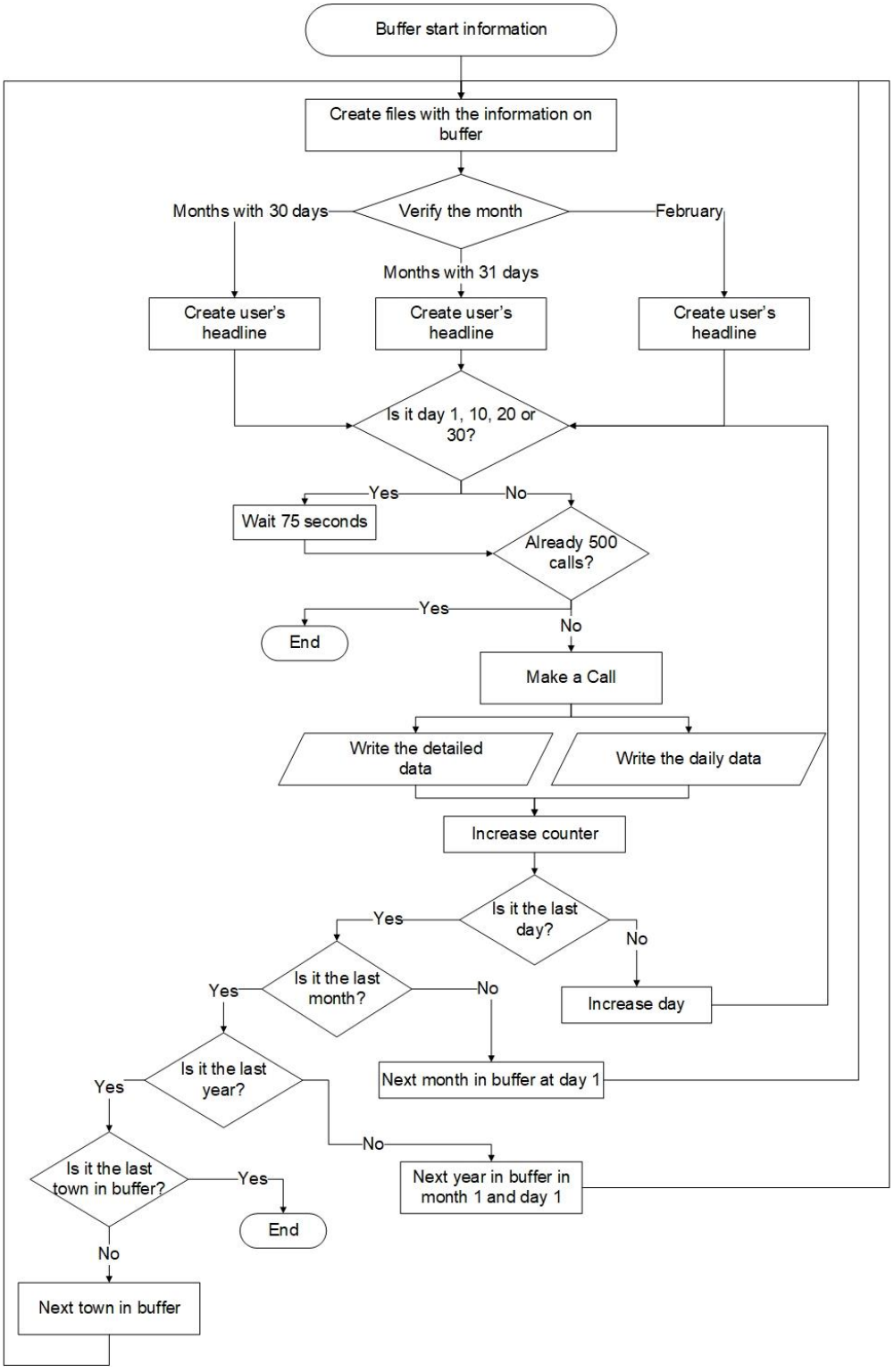
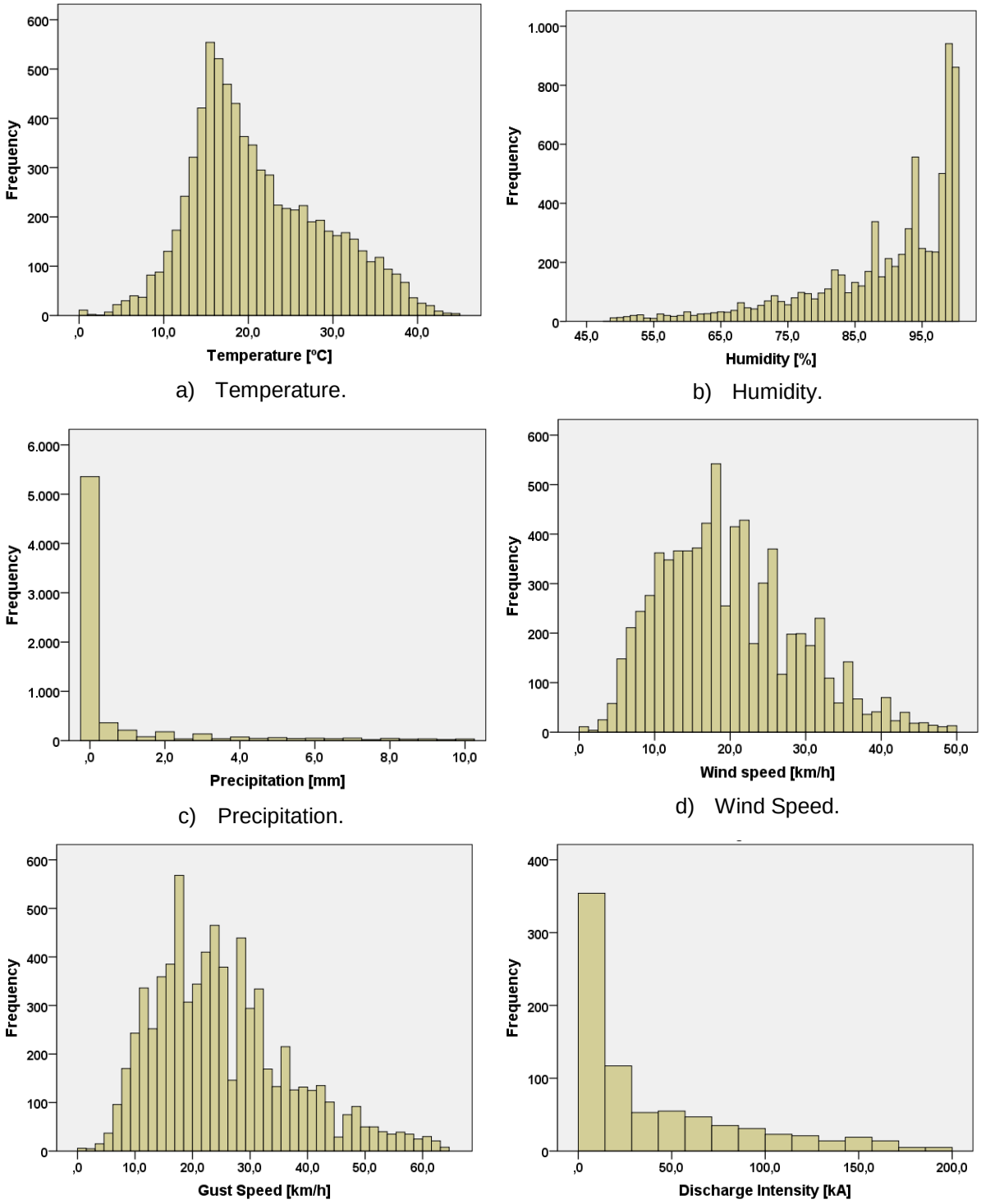


Figure 3.1. Flowchart about Weather Underground API application.



a) Temperature. b) Humidity.

c) Precipitation. d) Wind Speed.

e) Gust Speed. f) Discharge intensity.

Figure 3.2. Representation of the study about Portugal weather variables.

Table 3.6. Weather variables study on a Mediterranean climate (adapted from [BLMD02]).

	Mean T [ºC]	T Std. Dev. [ºC]	Mean H [%]	H Std. Dev. [%]
Spring	21.6	2.69	55.7	0.07
Summer	29.8	2.5	45.0	0.06

3.2 Data processing

Since one uses data from multiple sources, a structure of the several steps to organise and process data from each entity to obtain the final file is needed. This processing comprises several procedures, including the removal of information not needed for this study. Thus, one merged and process the information necessary, Figure 3.3 showing the procedures to get this file.

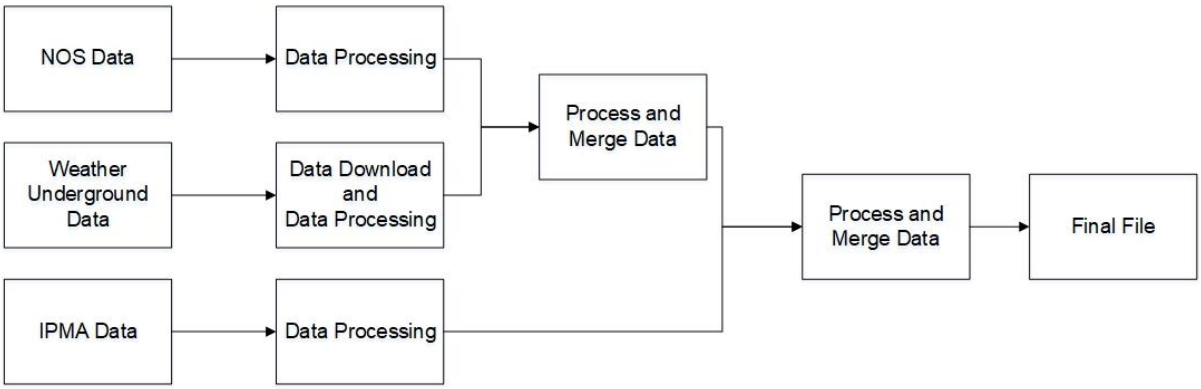


Figure 3.3. Steps for processing NOS, Weather Underground and IPMA data to obtain a single file.

Two major objectives compose this thesis, a statistical study being the first one, and forecasting of the number of incidents being the second, Figure 3.4. For the former, one used two schedules, 24 and 12-hour, to understand the importance of time intervals. One also includes a study regarding the influence of the meteorological variables in the occurrence of incidents, using one, two, and several meteorological factors to understand how they are related to the events.

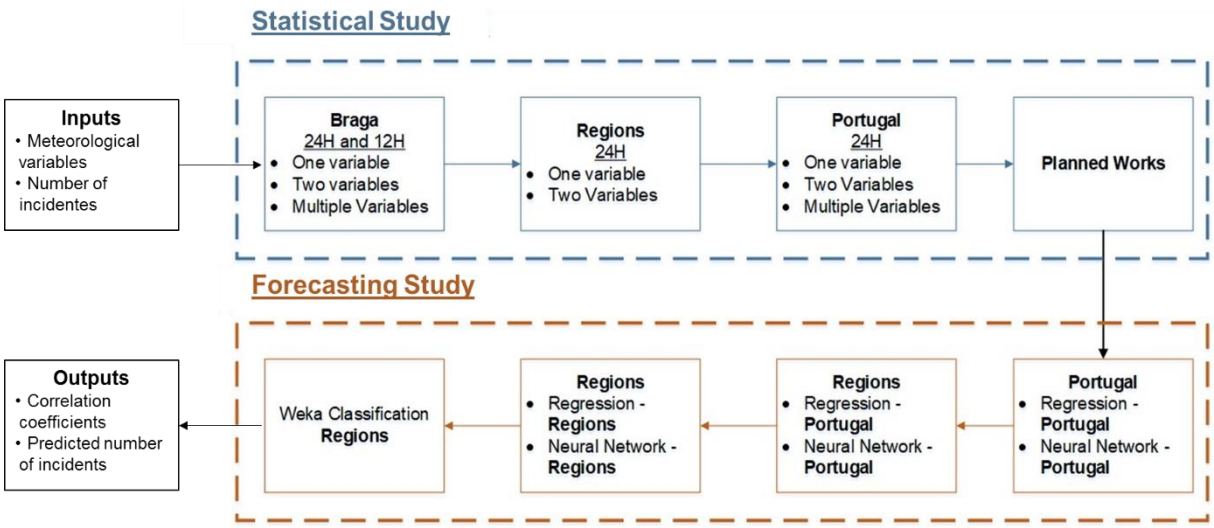


Figure 3.4. Schematisation of the statistical and forecasting studies, pointing each step of work.

For the second objective, one accomplished this study by the use of the information of Portugal as a whole and for each region. The first forecasting is by the utilisation of Portugal data to train both the Regression equation and the Neural Network, and then simulated with the same data. The second step uses the Regression equation and the Neural Network trained with the information of Portugal, but

simulating with the data of each region. The third forecasting is by training and forecasting both methods with the information of the Regions. Finally, one uses the Weka [Weka17] software, to establish the classification of each Region into three different classes, regarding the number of incidents.

To accomplish the objectives of the statistical study, some steps needed to be achieved. The analysis of the correlation between meteorological factors and the number of faults was the first goal, due to the importance of getting the importance of weather variables in incidents occurrence. To complete this objective, weather information and incidents data were merged by location and time. One presents in Figure 3.5 the process to calculate the correlation between meteorological data and number of faults.

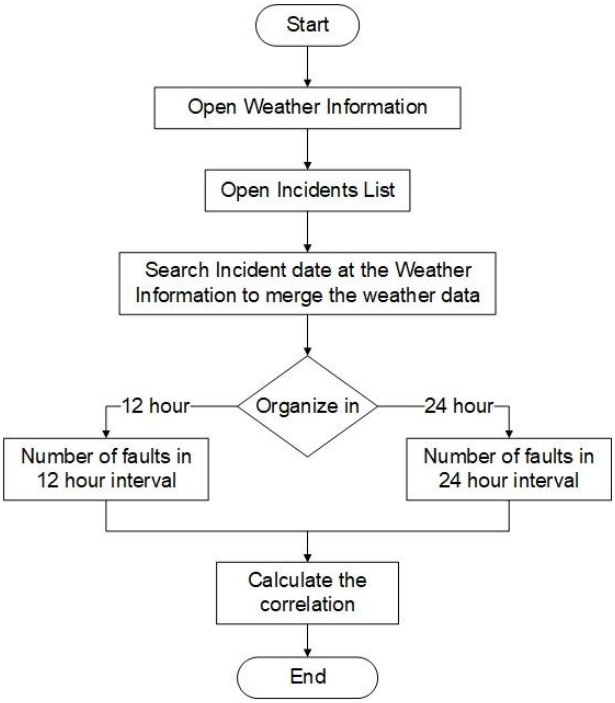


Figure 3.5. Correlation methodology.

However, there are several intermediate steps that need to be explained and tested. The first phase for the application of the model is the understanding of where the incident occurred. In Figure 3.6, one can see how the information about the city is assembled by combining two separate files, one containing information about incidents, and the other about the location of each base station. The combined file has information for each region with the base stations information.

Knowing where the incident is located, one needs to know the time of the incident in order to relate it with meteorological data, meteorological data was organised in two intervals: 12-hour and 24-hour. The former has data in 12-hour intervals with the maximum values of each variable inside that interval, for which one needs to analyse each 12-hour interval, and get the maximum value of the variable; Figure 3.7 shows the procedure for organising the meteorological file into a 12-hour interval. For the 24-hour interval, Weather Underground provides directly the maximum value of each day.

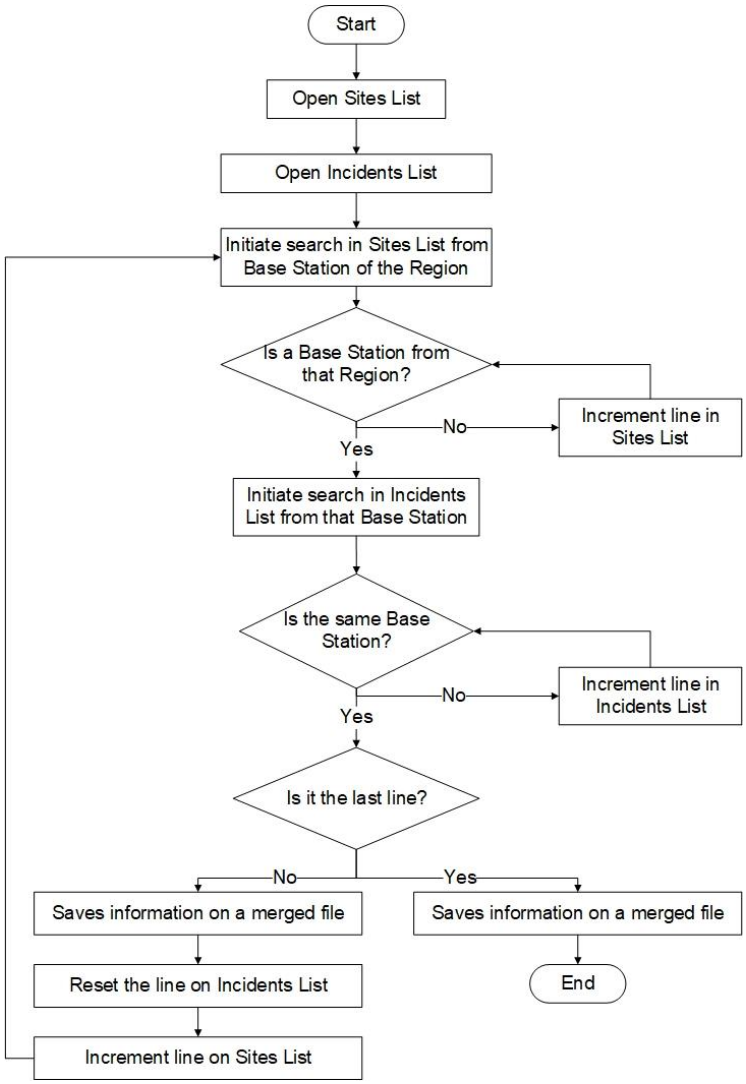


Figure 3.6. Combination to determine the city of each base station.

Having both files organised, one containing the location of incidents, and the other with the meteorological data, only the information about electrical discharges is missing, which was obtained spread into several files, each of which containing the data of three meteorological stations, organised by month. The first step was the conversion of all these single files, into just one; a simple script, which opened each single file and copies the information to the final one, was developed for this purpose. The next step was to locate the region of each electrical discharge, to be able to establish the relationship with the remaining data, by using an API of Google Maps, using Global Positioning System (GPS) coordinates to discover the region where the electrical discharge occurred, Figure 3.8. One should mention that the free version of this API was used, and due to this, it was only possible to perform 2 500 calls each time.

With the information of the discharges location, one needs to merge this information with the remaining one, for which the sum of the electrical discharges during the time intervals studied and the maximum discharge intensity were calculated, in order to have all data standardised.

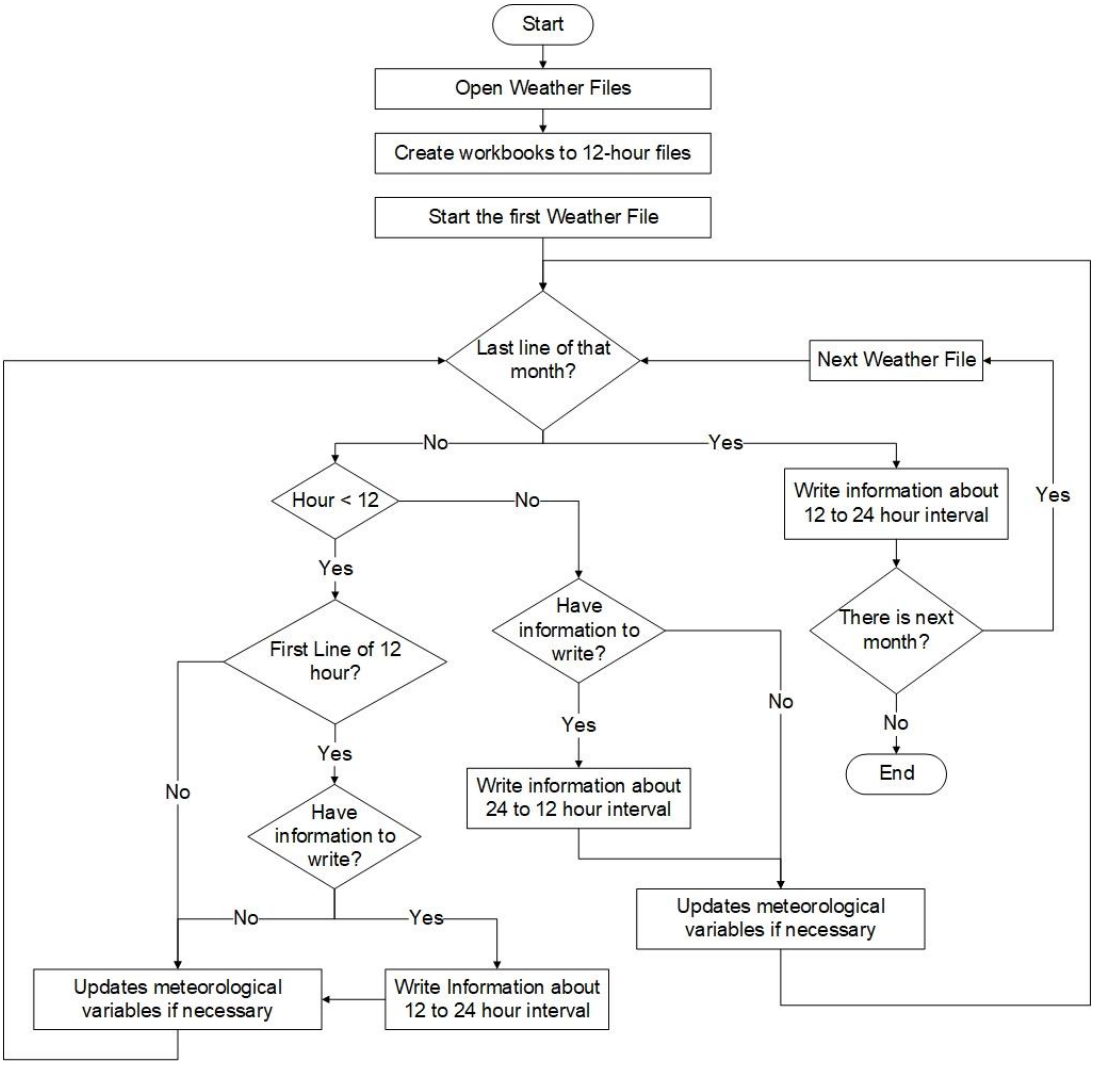


Figure 3.7. Method to organise data in 12-hour intervals.

Having all the information about weather variables processed, one needed to finalise the processing in the incidents file. The first step was the calculation of the number of incidents by day and merging with the weather data, which was done by an application that collects the information of weather variables in each region and incidents, afterwards producing the count of incidents at each period, excluding the planned work, as shown in Figure 3.9 for the 24-hour interval; the only difference between this case and 12-hour one is that the count for the latter is made at a 12-hour interval instead of a daily count. Regarding the file with the planned works, it has the same flow of Figure 3.9, with the only difference of skipping the count of incidents if a planned worked caused that incident.

The last step necessary was to relate the information between electric discharges and incidents by the time that each one occurred, in both files, for which one needed a verification if the day and the region are the same in both files, to enable the processing of information, Figure 3.10. This being the last step, a file with all the information needed to accomplish the statistical and the forecasting study is achieved. The previous procedures have been performed for each region, in order to have a file per region; regarding the data of Portugal, and to complete the same statistical and forecasting study, one compiled all the regions files into a single file, this being the Portugal data.

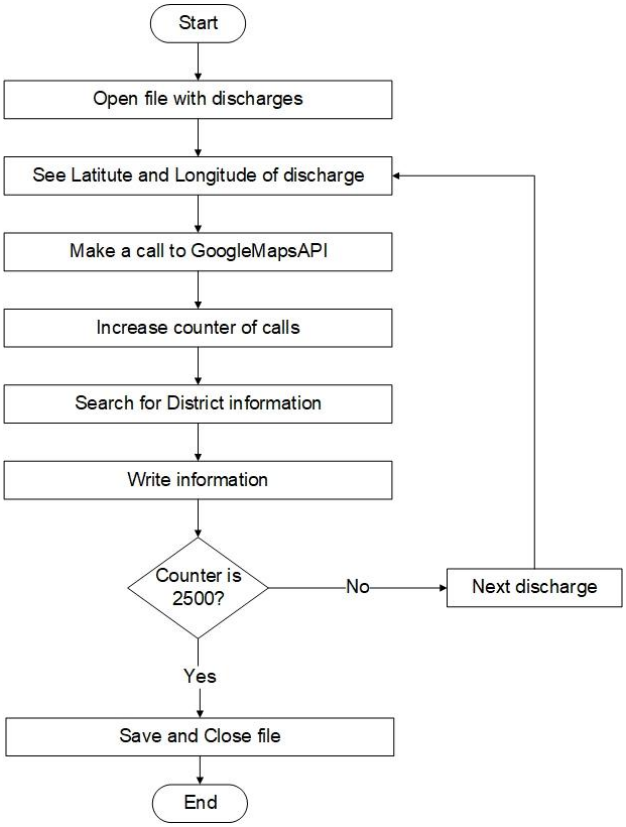


Figure 3.8. Use of Google Maps API to search for regional information.

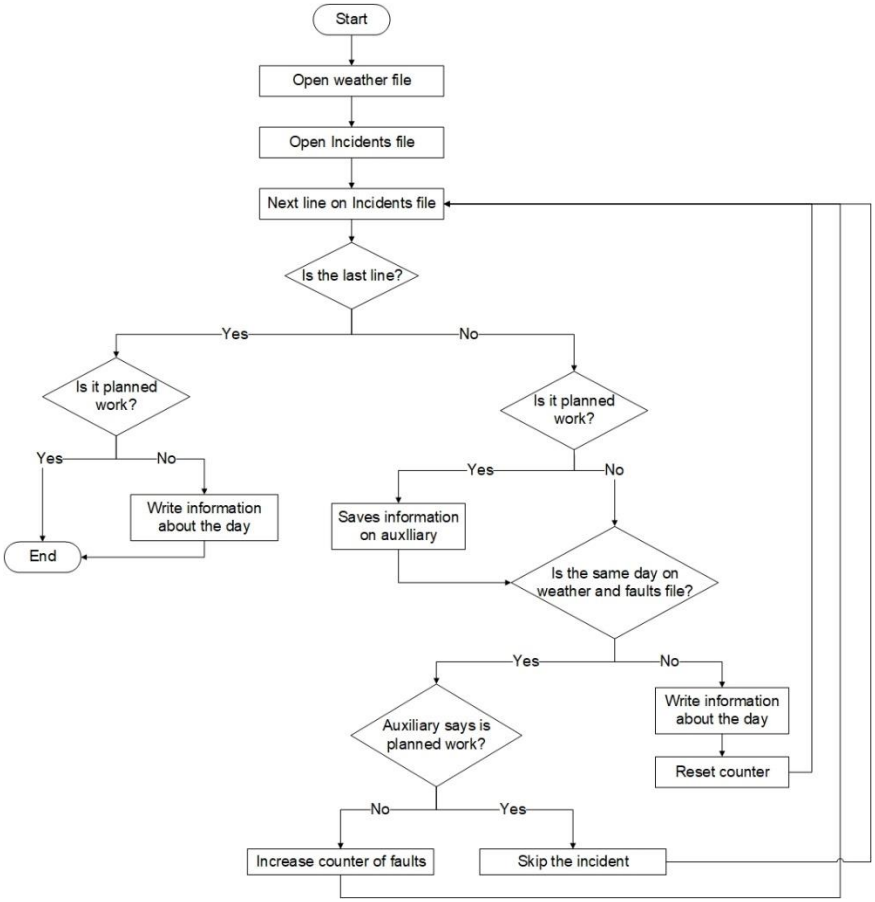


Figure 3.9. Count of the incidents number and relation with weather information in 24 hour-intervals.

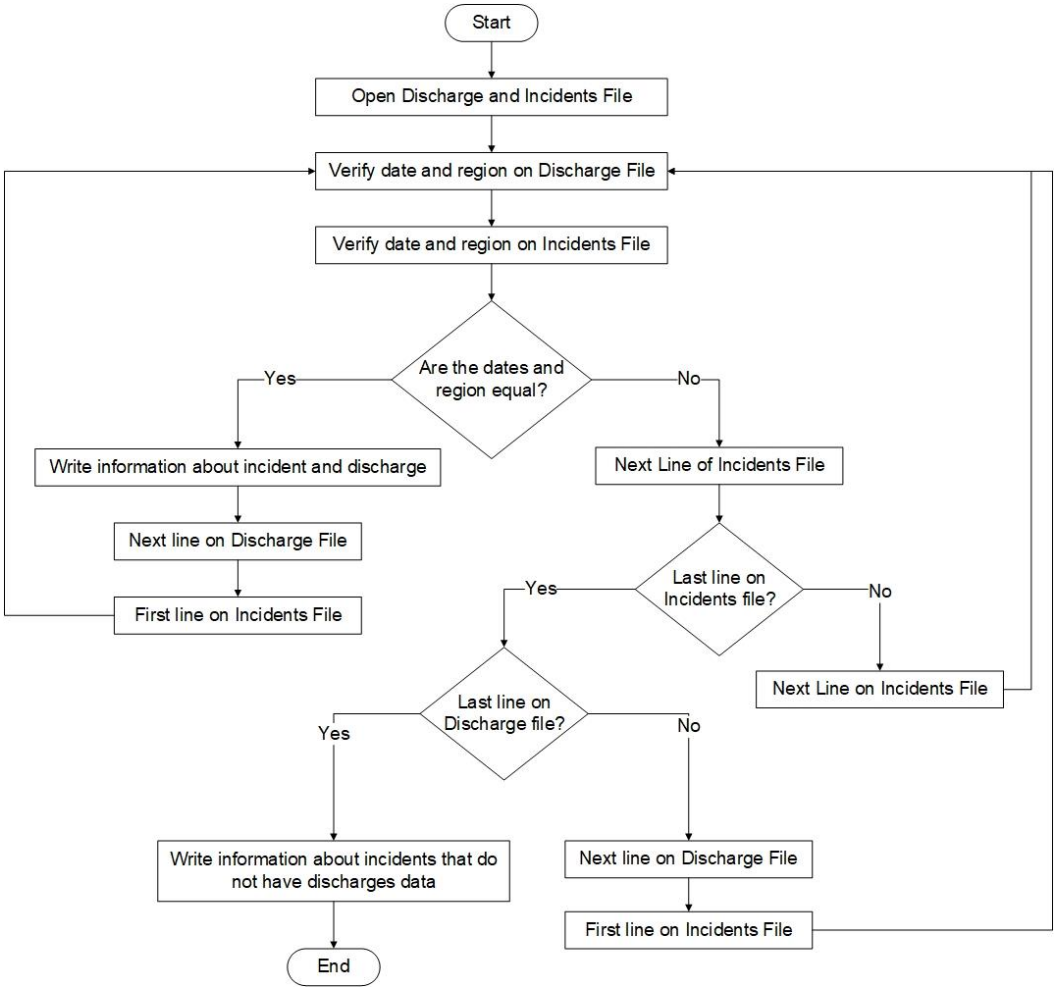


Figure 3.10. Method to relate the information between electrical discharges and incidents.

3.3 Statistical Study

The first step for the statistical tests on the number of incidents was their relationship versus one weather variable, which was used as relative values,

$$R_{meteo} = \frac{A_{value}}{M_{value}} \times 100 \tag{3.1}$$

where:

- A_{value}: Actual value of the weather variable;
- M_{value}: Maximum value of the weather variable.

One studied this relationship in two parameters: Correlation and Regression. The former is achieved by using the three methods presented in Section 2.4.2, leading to a coefficient that enables a proper analysis. Regarding the latter, the linear regression presented in Section 2.4.3 was used, enabling the

knowledge on the important parameter of the slope. One also used the regression study to understand the importance of the quantity of weather variables concerning the number of incidents, performing the study about the number of incidents versus one, two and multiple weather variables.

The calculation of the correlation coefficient used a file with the number of incidents and all weather variables, using the approaches of Pearson, Spearman and Kendall's τ correlation coefficients, following SciPy [ScyP17] libraries on Python.

Regarding the regression method, one used the same file with the number of incidents and all weather variables, and via the Excel chart tool, a scatterplot of the number of incidents regarding one weather variable can be obtained, calculating both the linear regression equation by (2.15), and the determination coefficient by (2.24). Again, relative values were used for the weather variables.

After the study of the number of incidents versus one variable, one could observe that some weather conditions are more critical on the occurrence of incidents, and that more weather variables should be included in the study. Thus, a study on the number of incidents versus two weather variables was performed, using Matlab, [Math17], which estimates and shows the information of multiple linear regression coefficients from a 3D scattergram. One used the linear regression with two dependent variables, adding a third variable that represents the interaction between two factors, defined by (2.16).

The last statistic study was the number of incidents versus more than two weather variables, which was done using the actual values collected from the weather variables, instead of the relative values used before. This study enables to analyse how several weather variables are related to the number of incidents, leading to a better understating of reality. Using (2.15), one established a linear equation of several variables, using 3 different weather variables, and increasing the number of variables until reaching all of them. The calculation of the coefficients of each variable, and thus, the regression equation, was done by the software of International Business Machines (IBM), named Statistical Package for the Social Sciences (SPSS), where one can subdivide the calculation of each coefficient in the linear regression into three steps: first, opening the Excel file containing all the information about incidents, then, the second phase is the configuration of the several variables to calculate, i.e., setting the dependent variable and the several independent ones, and finally, the last step is the calculation of each coefficient and the setting of the regression equation.

3.4 Forecasting Study

3.4.1 Regression

The simplest procedure to implement a prediction is using a regression equation. One describes in Section 2.4.3 the method to calculate the regression equation with the several variables studied in this work, which enables to understand which of the variables are more relevant to perform the prediction of incidents. To calculate each regression coefficient, one used SPSS, and then compiled each regression

coefficient into a final regression equation.

To simulate each day, one used the weather information in the regression equation, leading then to the number of incidents. An Excel file was used to compile all the information needed to reach this prediction. The information depends on the equation that is studied, since each equation uses different variables.

Finally, one calculated the predicted number of incidents in each case and compared with the real one. Using the performance measure presented in Section 2.4.4, one can perceive which of the equations achieve a better result in forecasting the number of incidents, Figure 3.11.

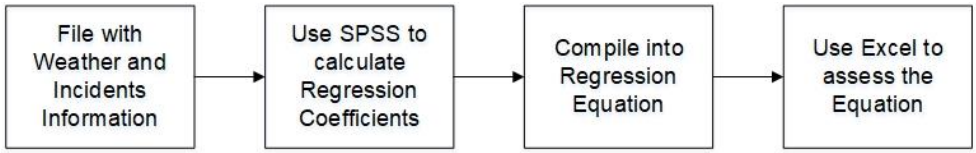


Figure 3.11. Process of the calculation the regression equation.

3.4.2 NARX Neural Network

To properly identify the system and to predict time series, one used Recurrent Neural Networks (RNN), [LGHK97]. NARX is a subclass of these recurrent networks, using embedded memory as connections, which provide shorter paths for propagating gradient information more efficiently, reducing the network's sensitivity to long-term dependencies problem, being computational powerful in theory.

The 14 months of data were divided into three subsets, [BeHD17], for the training, validation and test for the forecast method. The training part, which represented 70% of data, was used for computing the gradient and update network weights and biases. The validation subset, representing 15% of data, was when validation errors were being monitored during the training part. Finally, the test subset, also representing 15% of data, used to compare different models, is useful to plot the errors during the training process. These 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 was trained using a second-order algorithm, the Levenberg-Marquardt algorithm, due to the significantly increased training speed compared to a first-order algorithm, such as the Error Backpropagation algorithm, [WiYu10].

The implementation of the NARX network in the Matlab was done by the Neural Network Toolbox. The first step was the organisation of data to train the Network, by dividing information into two files: the output, referring to the objective of the network, which is the number of incidents; the input, relating to the several weather variables.

The next step was to set the size of the network, where the number of neurons and the delay change is possible. To achieve the best network, one trained the network with several numbers of neurons and delays: the different use sizes rely on the fact that the utilisation of few neurons can cause an underfitting of the network, and the use of too many neurons can contribute to overfitting, [BeHD17]; regarding delay, it could happen that the current state could depend on previous ones at various times. One presents in

Figure 3.12 the representation of the network in training, where the subsets of training as inputs to the network are shown, and *x* represents the number of variables present in training, the delay being represented by *d* and the number of neurons by *n*.

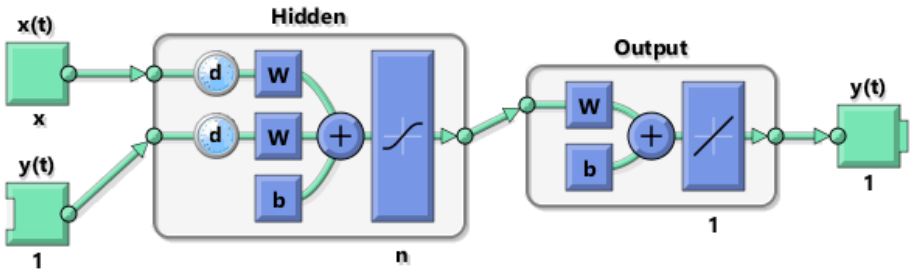


Figure 3.12. Representation of the neural network in training.

After each training, one considers the MSE of the three subsets previously determined. Since in each training, values are randomly chosen, the network is always different at each simulation; due to this, several pieces of training are necessary until reaching the optimal network.

3.4.3 Weka classification

Weka [Weka17] is a software that collects a set of machine learning algorithms for predictive and classification tasks, developed in the University of Waikato, New Zealand, by the Machine Learning Group. One presents in Figure 3.13 the workspace of Weka, where one can understand the potential of this software: however, in this thesis, only the Classify option is used.

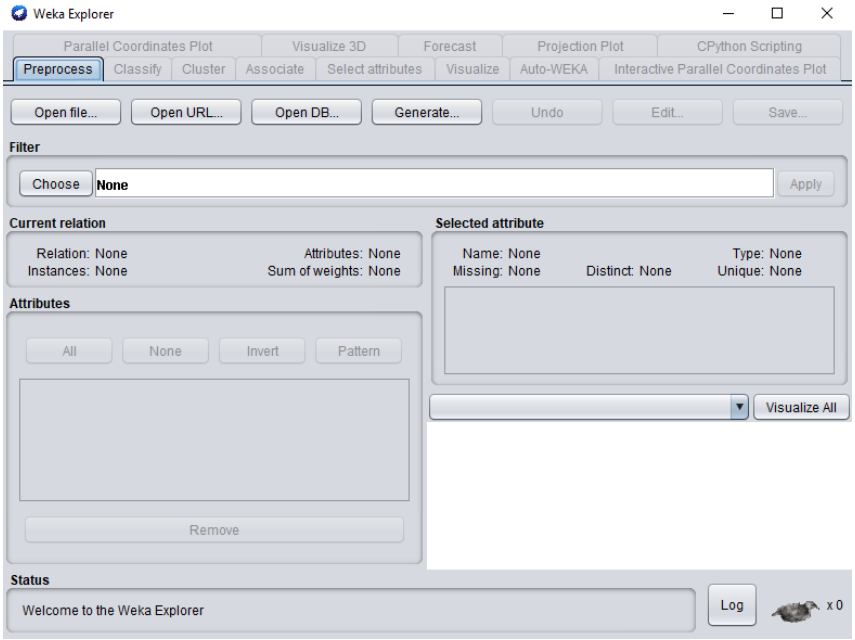


Figure 3.13. Weka workspace.

The first step for classification is the division per classes of each file, according to the number of incidents, each class being the target of the classification method. One divided the file into three categories, A, B and C: the A class refers to the days with a low quantity of incidents, B to the

intermediate days regarding how many incidents occurred, and C represents the days with an abnormal number of incidents. This classification was added to the file, replacing data on the number of incidents.

To accomplish the training, one used the option of Weka to cross-validation the data ten times, meaning that data is divided into ten equal parts, and then, uses nine of these parts to realise the training and the last one to test. This was done ten times, using every time a different part to test. The performance of each method is the mean of all simulations.

One uses four methods: Bayes Network, MLP, SVM and Nearest Neighbour, described in Section 2.4.3. This classification outputs the class that each method calculates for each case. Then, one can compare it with the remaining methods to evaluate the performance of forecasting the number of incidents regarding the several variables, Figure 3.14.

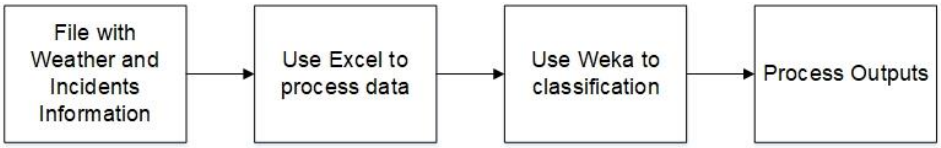


Figure 3.14. Process of the use the Weka software.

3.5 Forecasting Assessment

To assess the several forecasting methods, and for Regression and Neural Networks, one used two approaches. The first used the outputs to study the results that these methods calculate. The second studied only the peaks of incidents, to recognise how these methods behave when the severity of incidents increases. Regarding datasets, one used two different datasets in the forecasting, to perceive which produces better results: the first used the data of Portugal as one single file, and the second each region in separate.

Regarding the Weka classification, one used only the data from Regions to classify each class. Three different studies were done: the first was the percentage of correct classification, in a general way: the second was the number of correct and false peaks classified; and last one was the percentage of correct classification for each of the three classes.

Concerning the study of Neural Networks and of the Regression equation, one calculated the MSE to verify the quality of each method. This approach gave an optimal first perception of the total behaviour of each case under study. Regarding the study of the peaks of incidents, one calculated the mean and the maximum error together with the peaks that the method hits.

One divided the assessment and the process of forecasting of Regression and Neural Network into three steps. The first phase was the creation of the regression equation and the training of the network with the data from Portugal. Then, using the same data, one calculated which were the best regression equation and neural network using MSE. Then, using the best results of each case, one obtained the

study of the peaks of incidents, to understand the behaviour of each method to this specific instance.

The second step, using the best regression equation and the best network, one performed the same study but using the data from each region. Regarding the regression equation, one used the same coefficients for each parameter as used in Portugal. From the neural network, one used the network trained with the Portugal data, but the input parameters were from each region, Figure 3.15. To evaluate this case, one calculated the mean error occurred in the calculation of the number of incidents, as well as the maximum error in the prediction of incidents.

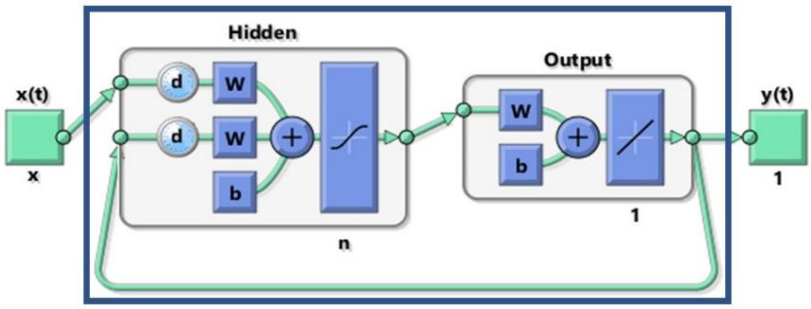


Figure 3.15. Example of using a trained neural network with the inputs other inputs.

The third step was to create both a regression equation and a trained network with the data from each region, and then, to assess both methods, one used the data from each region to calculate the performance parameters. The study on the peaks of incidents enabled to evaluate performance in this specific case.

One of the forecasting evaluation approaches uses the error of false peaks, considered between the number of false peaks and the number of correct forecasted peaks, providing a mechanism to understand the percentage of false peaks regarding the total peaks forecasted,

$$e_{fpeaks\,[\%]} = \frac{f_{peaks}}{c_{peaks} + f_{peaks}} \times 100 \tag{3.2}$$

where:

- f_{peaks}: Quantity of false peaks;
- c_{peaks}: Quantity of correct peaks.

Chapter 4

Results Analysis

This chapter presents the considered scenarios and the analysis of results. Both statistical and forecasting study results are presented in this chapter.

4.1 Scenario Description

The scenario is composed of data from 1st January 2016 to 28th February 2017, i.e., 14 months of data in Continental Portugal, which was given by NOS [NOSP17], with all the incidents occurred in this period. One used firstly a pilot city, Braga, to get familiar with the data, and then, the same study was applied to all other regions, and finally to Portugal as a whole. The remaining description of the scenario is presented in Annex B; most of the data is shown normalised, usually to its maximum, in order to keep its confidentiality, Annex B containing the values that allow to obtain the actual absolute values, and the corresponding denormalization.

The dataset of incidents, is shown in Figure 4.1, includes all of them except planned works, where the observation of two peaks of unusual events is visible. The first on 14th February 2016 with a normalised value of 0.86 incidents and the second on 3rd February of 2017 with the maximum value of incidents. Both were caused due to severe weather conditions in Continental Portugal, with elevated levels of precipitation and strong wind gusts.

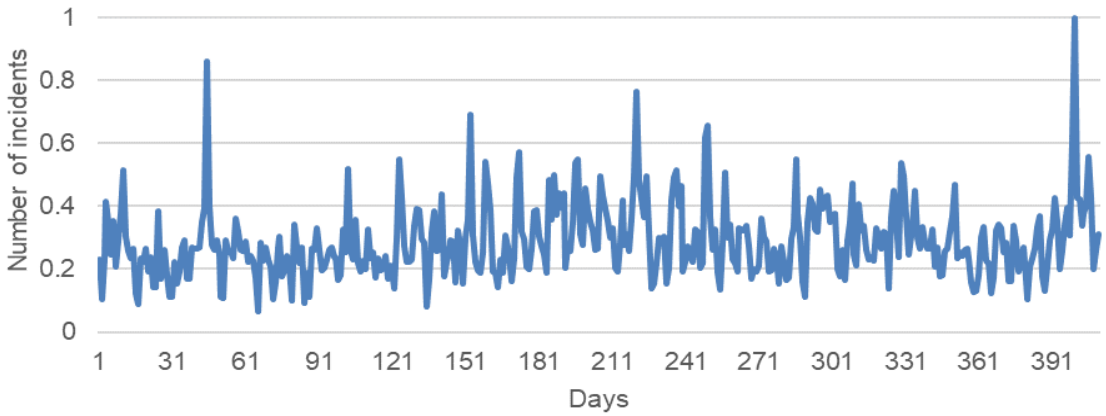


Figure 4.1. The number of incidents from January 2016 until February 2017.

The number of incidents occurred in each region from January 2016 to February 2017 is presented in Figure 4.2, without the planned works.

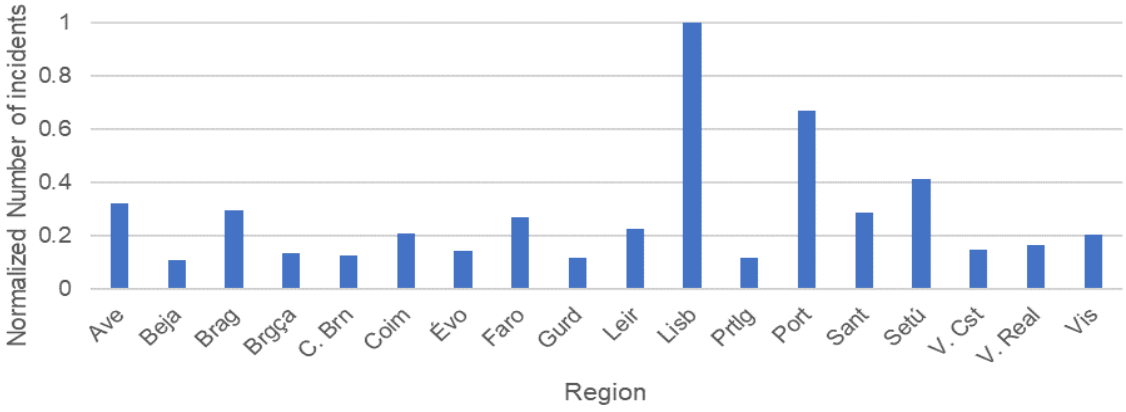


Figure 4.2. Quantity of incidents per region.

The number of base stations in each region is also an important parameter to take into account, being shown in Figure 4.3, per region.

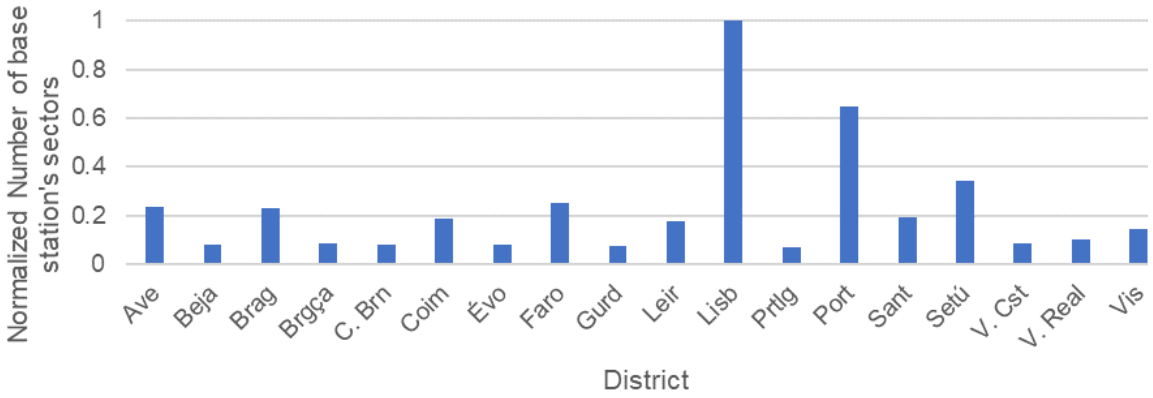


Figure 4.3. Quantity of base station's sectors per region.

The number of incidents is directly related to the number of base station in each region, but it does not mean that with the increase of base stations, the number of incidents also increases. To better understand this behaviour, one conducted a study to relate the ratio of incidents per base station and other metrics, shown in Annex B. From this study, one can observe that some regions with more incidents and base stations are the ones with fewer incidents per base station. Another conclusion is that the number of incidents per region size, in some regions, is significantly higher than in the rest of the regions, and the same applies to the number of incidents versus population.

These faults can occur in two major situations, either when critical conditions occur or in a typical normal situation. In the former, it is easy to understand that faults can happen, but it is very difficult to forecast the severity of the problem, while in the latter, there is usually no extreme seriousness of the situation, but there is also no significant interest in forecasting it. A first processing of incidents data, to better understand some of the cause-effect problems, is presented in Table 4.1.

Table 4.1. Cause-effect brief analysis.

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

Regarding the scenarios that are analysed, the study addresses the following problems:

- Number of incidents regarding one weather variable, including correlation;
- Number of incidents regarding two weather variables, including the interdependence of weather variables;

- Number of incidents regarding several weather variables, including the error associated with the use of regression with multiple variables;
- Forecasting the number of incidents using several weather variables.

4.2 Scenarios Processing

To test the procedures presented in Section 3.2, and the statistical study from Section 3.3, one selected a pilot city, Braga, due to the vast number of incidents during the period. In order to understand the importance of meteorological factors in this city, one investigated how these factors are correlated to the number of faults, presenting the number of faults regarding the meteorological variables divided into two intervals: 24-hour and 12-hour interval. The division into these intervals is important to understand if the variance of the time interval is important, and also to understand the best granularity in the organisation of the files to achieve the best results.

As mentioned in Section 3.3, meteorological data are presented in relative values, normalised to the following maxima: 40.3 ºC for Temperature, 67.3 mm for Precipitation, 45.7 km/h for Wind Speed, 64.4 km/h for Gust Speed, 34 for the electrical discharges in one day, and 191.5 kA for the discharge.

The representation of the number of incidents regarding one weather variable in the 24-hour interval is given in Annex B.1.2. In Table 4.2, one presents the linear equations regarding one weather variable, using (2.15). The previous designation is used, adding D for the number of electrical discharges.

Table 4.2. Linear equation variables for the 24-hour interval study in Braga.

	T	**H**	**P**	**W**	**G**	**D**	**I**
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

From Table 4.2, one concludes that the number of faults is positively related to all variables, except humidity, drawn by the values of the slope. The determination coefficient, R^2, is small for all variables, except those from discharges. The lack of a good adjustment between the linear regression and the observed values can easily explain this fact, and since this information is not very useful, one does not present it; on the other hand, for the electrical discharges, R^2 is not that significant, because it has only a few values. The study of the three methods for correlation introduced in Section 2.4.2 is presented in Table 4.3. A colour scheme by correlation method is also presented, where the green cells represent the maximum correlation in each method, the red cells represent the coefficients near zero and the yellow cells the mean ones. The basic conclusion is equal to the previous one, where the number of faults is positively related to all variables, except humidity.

Table 4.3. Correlation results for the 24-hour interval study in Braga.

	T	H	P	W	G	D	I
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

From the values presented in Table 4.3, one can see that the Pearson approach leads to a stronger correlation between number of incidents and the quantity and intensity of discharges, but using the other two more robust methods, Kendall Tau's and Spearman's, these differences are lower, with temperature presenting the lower correlation. With this information, one concludes that the quantity and the intensity of electrical discharges are the meteorological variables more related to the number of incidents, followed by Precipitation, Wind and Gust Speed as the factors most related to the number of incidents. One presents the plots of this study in Annex B.

Regarding the 12-hour interval study, one approach is the same as for the 24-hour one. The maximum values used to calculate the relative meteorological values are the same, except the quantity of discharges that is 30. Concerning the relationship between the number of incidents and one weather variable, Table 4.4 presents the linear regression equation parameters.

Table 4.4. Linear equation variables for the 12-hour interval study in Braga.

	T	H	P	W	G	D	I
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

The conclusions of the determination coefficient for the 12-hour interval are the same as for the 24-hour one, Table 4.5 presenting the correlation results using the three methods. From both Table 4.4 and Table 4.5 one can take conclusions similar to the previous case. Although having more accurate data for the 12-hour interval, the results are nearly the same, leading to very similar correlation coefficients. Both cases have the number of electrical discharges as the most severe variable in the occurrence of incidents, the remaining being similar in severity. One presents the plots of this study in Annex B.

Table 4.5. Correlations results from 12-hour interval study in Braga.

	T	H	P	W	G	D	I
Pearson	0.10	-0.16	0.08	0.31	0.32	0.51	0.17
Spearman	0.10	-0.14	0.12	0.21	0.21	0.30	0.15
Kendall Tau	0.08	-0.11	0.10	0.16	0.17	0.23	0.13

After the analysis of the number of incidents regarding one variable, the examination regarding two weather variables followed, as presented in Section 2.4.3. The importance of pairs of variables was checked by using (2.16). One has not analysed all possible cases, but only those that are the most important ones. For example, Wind Speed and Gust Speed are directly related, so it is not interesting

to see a direct relationship between these two. One presents in Annex B the representations as well as the regression equation of this study in a 24-hour interval. However, to better get the conclusions and to obtain a comparison method, one calculated the maximum value that each regression equation has, to obtain the pairs of variables that are more severe in the occurrence of incidents.

One presents in Table 4.6 the results of the number of incidents versus two weather variables, where the green cells represent the maximum value, the yellow cells the mean ones, and finally the red cells represent the lowest one. The pair of variables that include discharges is not represented in this colour scheme, as one intended to understand the gravity of other weather variables rather than this one, since the importance of discharges is already known. Regarding the arrows, one indicates the behaviour that the surface has in the simulation: if both arrows are up and green, it means that the peak of the surface occurs when both variables increase: if the first arrow is up and green and the second is down and red, it means that the peak takes place when the first variable increases and the second decreases.

Table 4.6. Pair of variables regression in Braga in a 24-hour interval.

	W P	W T	P T	G P	T H	W H	P I	T D
Peak Surface	**8**	**8.45**	**9.46**	**10.5**	**7.3**	**15.1**	17.1	39.2
Behaviour	↑ ↑	↑ ↓	↑ ↓	↑ ↑	↑ ↓	↑ ↓	↓ ↑	↓ ↑

For each regression equation, shown in Annex B, one can see that the last coefficient gives the interaction between both variables, an objective for this study. However, since the coefficient is always small relative to the remaining ones, one cannot take very reliable information about the interdependence of the two variables. Still, from Table 4.6, one can conclude on the severity that each pair has on the occurrence of incidents: (Wind Speed, Humidity) is the most severe pair, while (Temperature, Humidity) is the least one.

As the analysis of the time interval is important, one conducted the same survey for the 12-hour interval, presenting the complete results in Annex B. Table 4.7 shows the results in a similar way.

Table 4.7. Pair of variables regression in Braga in a 12-hour interval.

	W P	W T	P T	G P	T H	W H	P I	T D
Peak Surface	4.95	5.59	5.48	5.36	5.89	6.16	13.9	31.8
Behaviour	↑ ↑	↑ ↓	↑ ↓	↑ ↑	↑ ↓	↑ ↓	↓ ↑	↓ ↑

One can conclude from Table 4.6 and Table 4.7 that the results using 24-hour or 12-hour intervals do not differ much: (Wind Speed, Humidity) remains the most severe case, and also the behaviour. As expected, the surface peak decreases, because in a 12-hour interval fewer incidents occurs than in a 24-hour one. Given this result, from this point on, the remainder of study uses only the 24-hour interval.

Another cause of faults are the planned works in the network. In Braga, there are 0.03 normalised incidents originated by this cause, which compares quite low with 0.3 from the other causes. There are

also differences in the number of maximum of faults, where the normalised peak due to planned works is 0.09, comparing with 1 for the other cases. Another difference is the number of days that incidents took place, which can be easily explained since planned works are previously scheduled.

4.3 Regions analysis

The next step was the analysis of each region. The first phase of this analysis was the study of the independent weather variable in the occurrence of incidents in each region. The linear regression equations regarding the number of incidents with one variable are presented in Annex C, and the correlation coefficient in Table 4.8, for the Spearman approach, the remaining ones being presented in Annex C. One displays the values around zero in red, the maximum value in green, and the intermediate one in yellow; again the values for electrical discharges are not represented in the colour scheme, but the numbers are shown, to see the importance of this variable in the occurrence of incidents.

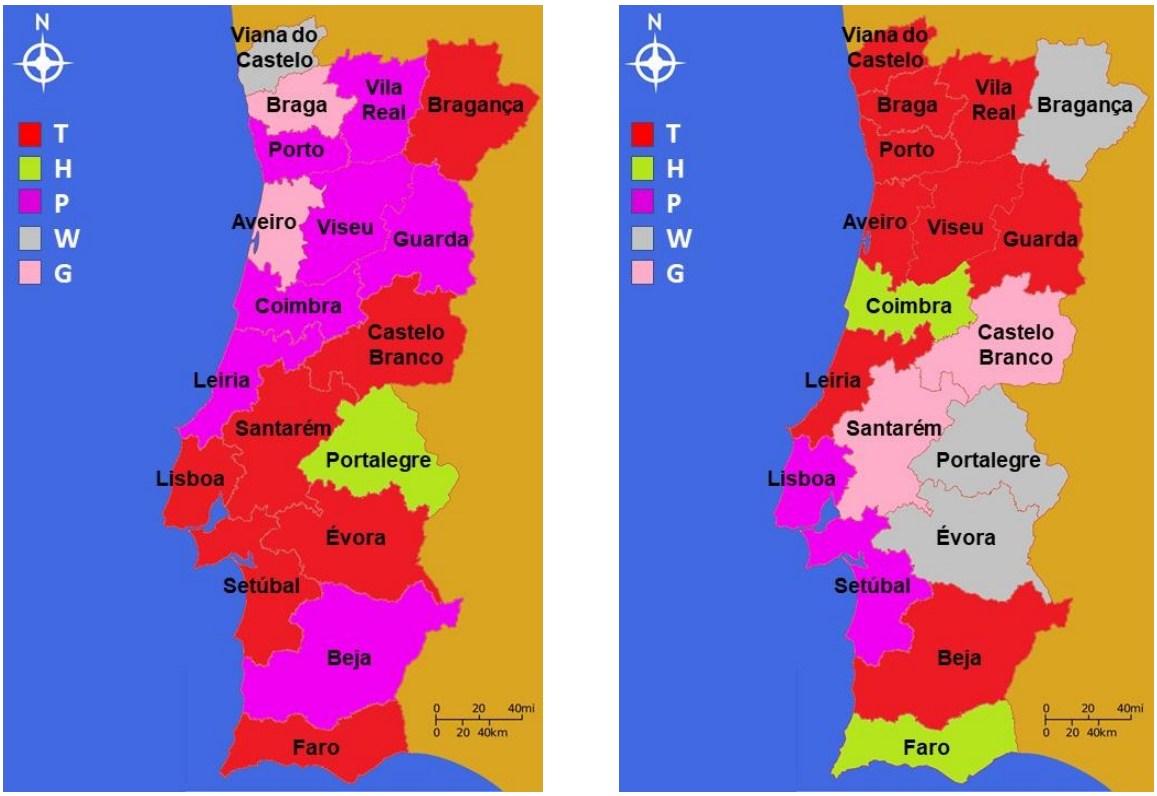
Table 4.8. Spearman correlation coefficient in each region.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
Aveiro	0.06	-0.07	0.13	0.19	0.20	0.15	0.20
Beja	0.05	-0.12	0.12	0.07	0.07	0.20	0.10
Braga	0.06	-0.08	0.14	0.19	0.20	0.30	0.26
Bragança	0.09	-0.05	-0.01	0	0.02	0.50	0.40
C. Branco	0.14	-0.05	0.03	0.01	0.01	0.17	0.26
Coimbra	0.07	0.06	0.11	0.11	0.08	0.17	0.12
Évora	-0.10	-0.08	0.08	0.04	0.04	0.02	0.15
Faro	0.19	0.04	0.16	0.07	0.10	0.41	0.24
Guarda	-0.02	0.05	0.22	0.17	0.21	0.20	0.38
Leiria	0.02	0.08	0.17	0.09	0.09	-0.42	-0.19
Lisbon	0.36	-0.21	-0.07	0.08	0.07	0.09	-0.07
Portalegre	-0.06	-0.11	0.03	0.01	0.05	0	0.12
Porto	0.01	0.04	0.18	0.13	0.15	0.69	0.60
Santarém	0.19	-0.08	-0.05	0.02	-0.01	0.28	0.20
Setúbal	0.14	0.10	0.01	0.06	0.04	0	0.04
V. Castelo	0.02	-0.06	0.16	0.22	0.19	0.69	0.55
V. Real	-0.03	-0.11	0.25	0.24	0.23	0.36	0.40
Viseu	-0.04	-0.05	0.19	0.10	0.06	0.57	0.71

One concludes from Table 4.8 that there is no single weather variable most related to the occurrence of faults: excluding discharges, Temperature and Precipitation are the most related ones, and Humidity is the least. Regarding the discharges, as expected from the literature and from network management

experience, the correlation is the highest in most of the cases.

To better perceive the geographical distribution of the importance of each weather variable, excluding discharges, one presents in Figure 4.4 a) the most severe variables and in Figure 4.4 b) the least ones. In the cases where the Spearman coefficient is equal between two variables, the remaining methods were used to take a decision.



a) Most severe variable. b) Least severe variable.

Figure 4.4. The most and least severe variable in each region.

For a more detailed analysis of the regions, continental Portugal was divided into three zones, North, Centre and South, according to the following regions:

- North: Viana do Castelo, Braga, Porto, Aveiro, Vila Real, Viseu, Bragança and Guarda;
- Centre: Coimbra, Leiria, Lisboa, Santarém, Castelo Branco and Portalegre;
- South: Setúbal, Évora, Beja and Faro.

One can conclude from Figure 4.4 that Precipitation is the variable with the highest correlation in the North, but Wind and Gust Speed also have meaningful results in Aveiro, Braga and Viana do Castelo, while in Bragança Temperature is the key variable. The typical weather at these regions explains these conclusions, since in all of them, except Bragança, there is a high level of precipitation and wind speed during the year, leading to the occurrence of incidents; regarding Bragança, the importance of Temperature can be explained due to the large thermal amplitude during the year, and probably most of the base stations not being prepared for such intervals of temperature.

Concerning the least severe variables in the North, Temperature is the dominant variable, explained by the low temperature registered there, while Bragança shows Wind Speed as the least severe variable.

One cannot draw significant conclusions from the most important variables in the Centre; since the weather is milder here, there is no predominant factor for the occurrence of incidents in here. However, Precipitation and Temperature are, again, the variables most present in the occurrence of incidents, similar conclusions apply to the least severe variables. However, in the more interior regions, Wind is the one with fewer relationships with incidents, explained again by the typical weather in these areas.

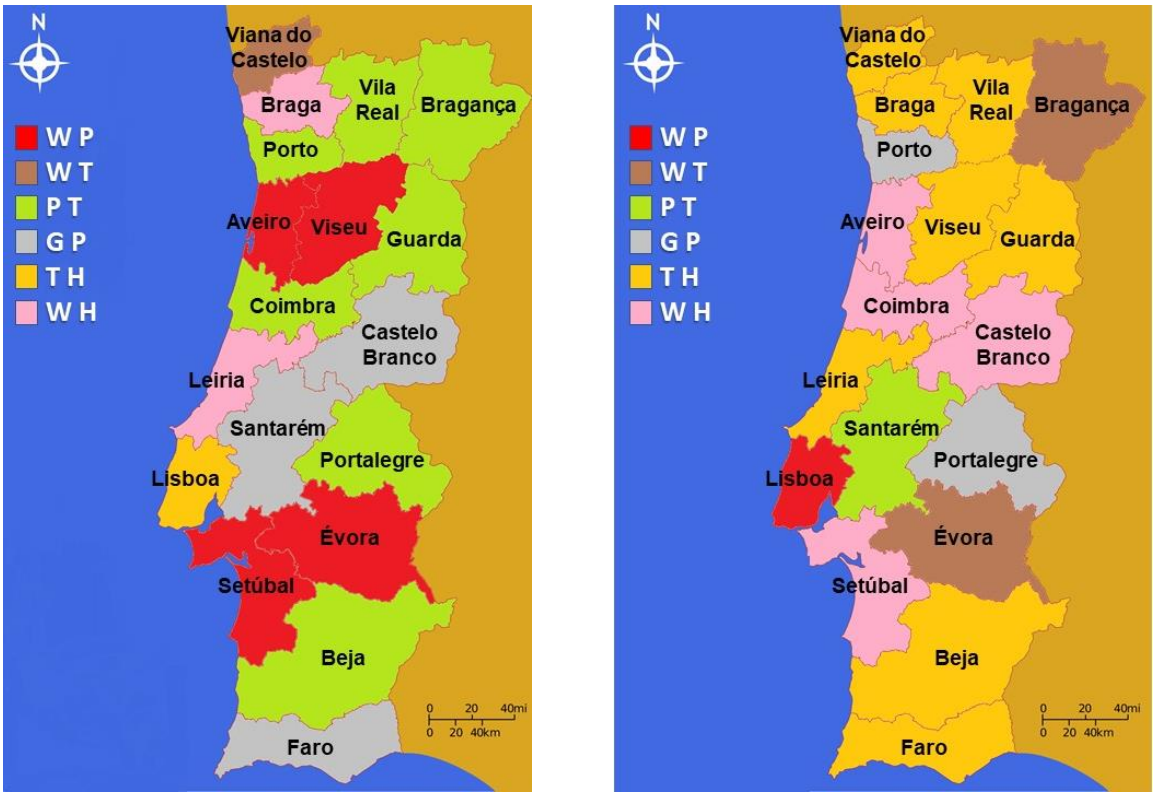
Finally, for the South, as expected, Temperature is the most critical variable, due to high temperatures that occur in this zone. Still, in Beja, Precipitation is the most severe variable, which can be, at first sight, surprising; an explanation for this fact is that Beja is known for the high temperatures, hence, precaution being taken by using air conditioning in base stations installations. Regarding the lower severity variables, there is, again, no major factor, but it is curious to see that in Beja it is the Temperature, which reinforce the explanation previously presented.

The next step for this analysis concerns the pairs of weather variables, Table 4.9; normalisations have been done to the maximum of each region. The highest value is in green, the values near zero are in red, and the intermediate ones in yellow.

Table 4.9. Pair or variables regression in each region.

	W P	W T	P T	G P	T H	W H
Aveiro	1	0.45	0.35	0.91	0.38	0.31
Beja	0.79	0.42	1	0.57	0.28	0.38
Braga	0.53	0.56	0.63	0.69	0.49	1
Bragança	0.96	0.47	1	0.93	0.71	0.48
C. Branco	0.75	0.36	0.52	1	0.29	0.26
Coimbra	0.85	0.43	1	0.87	0.63	0.3
Évora	1	0.36	0.93	0.88	0.37	0.38
Faro	0.94	0.43	0.69	1	0.18	0.2
Guarda	0.78	0.39	1	0.87	0.31	0.37
Leiria	0.16	0.38	0.27	0.16	0.12	1
Lisbon	0.52	0.84	0.76	0.63	1	0.81
Portalegre	0.90	0.61	1	0.16	0.23	0.29
Porto	0.38	0.51	1	0.35	0.38	0.38
Santarém	0.57	0.11	0.09	1	0.14	0.10
Setúbal	1	0.70	0.95	0.46	0.75	0.35
V. Castelo	0.64	1	0.43	0.57	0.33	0.83
Vila Real	0.63	0.22	1	0.57	0.11	0.21
Viseu	1	0.38	0.82	0.74	0.35	0.93

Again, one presents the most and least severe pairs of variables, Figure 4.5. In the North, one can see that (Precipitation, Temperature) is the pair most related to incidents, (Wind Speed, Precipitation) being the second; Precipitation is again the most important variable. Regarding the least severe variables, the pair (Temperature, Humidity) shows up, Temperature being the least severe variable.



a) Most severe pair of variables. b) Least severe pair of variables.

Figure 4.5. The most and least severe pair of variables in each region.

Regarding the Centre, one cannot take any conclusion about the pairs, in a situation similar to the one variable case. However, the two most severe pairs, (Precipitation, Temperature) and (Gust Speed, Precipitation) do have a common variable. The least severe pairs also do not show any predominance.

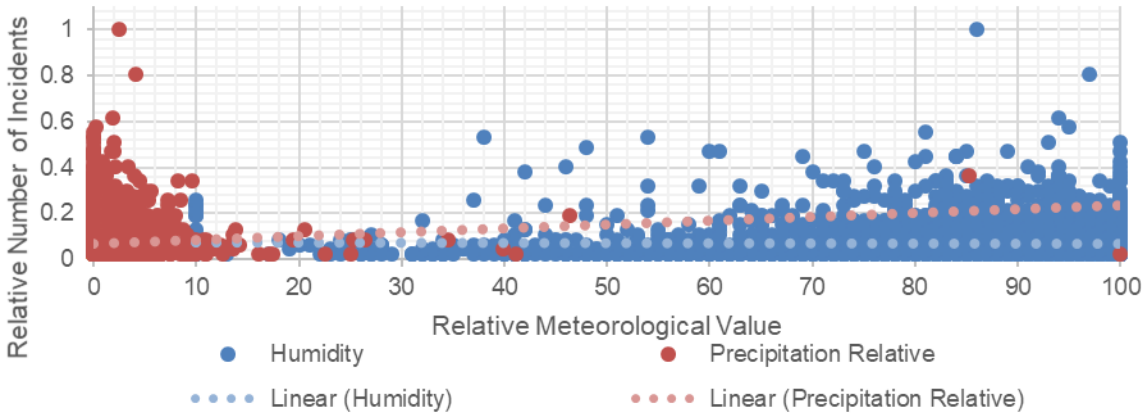
Finally, for the South, (Wind Speed, Precipitation) is the most severe in two regions, but in the remaining ones, (Precipitation, Temperature) and (Gust Speed, Precipitation) again play this role. Regarding the least severe pair, two regions have (Temperature, Humidity), while the remaining have (Wind Speed, Temperature) and (Wind Speed, Humidity).

4.4 Portugal analysis

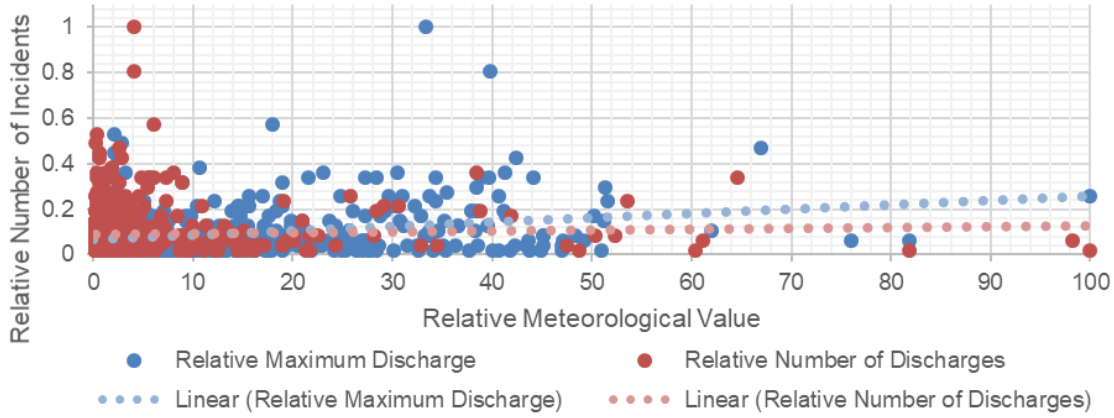
4.4.1 Statistical study

With the analysis done for each region, the analysis of the whole Portugal follows. First, one addresses the one variable case, Figure 4.6 showing the quantity of incidents against the various variables,

Table 4.10 the parameters of the linear equations, and Table 4.11 the correlation coefficients using the three methods presented previously. As before, the values of the determination coefficient are also small in this case.



a) Incidents vs. relative value of Humidity and Precipitation.



b) Incidents vs. Relative Value of Number of Discharges and Maximum Discharge.

Figure 4.6. Quantity of incidents vs. relative values of the weather variables in Portugal.

Table 4.10. Linear equation variables for Portugal.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.01	- 0.01	0.08	0.09	0.06	0.02	0.09
b	2.57	3.50	3.08	2.36	2.56	4.29	3.26

Table 4.11. Correlations results for Portugal.

	T	H	P	W	G	D	I
Pearson	0.06	-0.02	0.07	0.14	0.16	0.03	0.26
Spearman	0.04	<0.01	0.09	0.19	0.14	0.07	0.21
Kendall Tau	0.03	<0.01	0.08	0.14	0.10	0.05	0.15

One can observe that, again, the maximum discharge intensity is the variable with the highest correlation coefficient, but the number of discharges is not highly correlated; Wind variables are moderately correlated, but Temperature and Humidity show very low correlation.

The next step is the statistical study of two variables regarding the number of incidents, some equations being presented in Table 4.12, and the remaining equations being in Annex B.

Table 4.12. Some equations of the surface of Portugal.

Variables	Equation	
W P	$R = 0.0514 + 0.0017 \times W - 0.0003 \times P + 0.0001 \times W \times P$	(4.1)
P T	$R = 0.0482 + 0.0091 \times P + 0.0004 \times T - 0.0002 \times P \times T$	(4.2)
W H	$R = 0.0789 - 0.0004 \times W - 0.003 \times H + 0 \times W \times H$	(4.3)

One presents in Figure 4.7 the graphical distribution for each equation in Table 4.12. One observes the distribution of the meteorological variable in the form of the scatterplot, with the representation of the regression equation presented in Section 2.4.3. The remaining graphical distributions are presented in Annex B.

Equation (2.16) gives a coefficient that presents the interaction between the two variables. From the studied cases, this value is many times very low, and one cannot take reliable conclusions from it.

Table 4.13 shows the results for pairs of variables, with the same colour scheme in previous cases, as well as the arrows, to understand the behaviour for each meteorological variable.

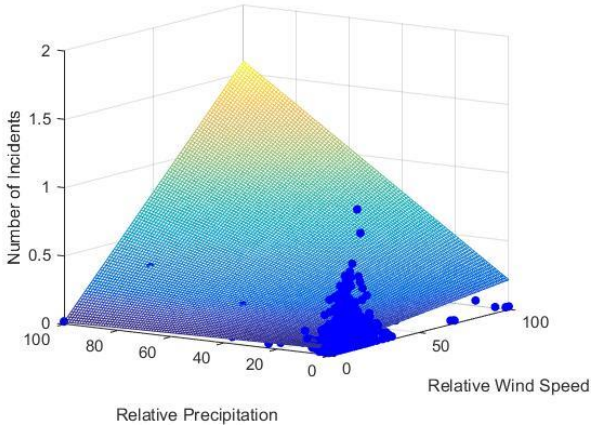
Table 4.13. Pair or variables regression in Portugal.

	W P	W T	P T	G P	T H	W H	P I	T D
Peak Surface	**74.5**	**18.6**	**44.5**	**44.8**	**7.26**	**12.3**	81.7	15.9
Behaviour	↑↑	↑↓	↑↓	↑↑	↑↓	↑↑	↑↑	↓↑

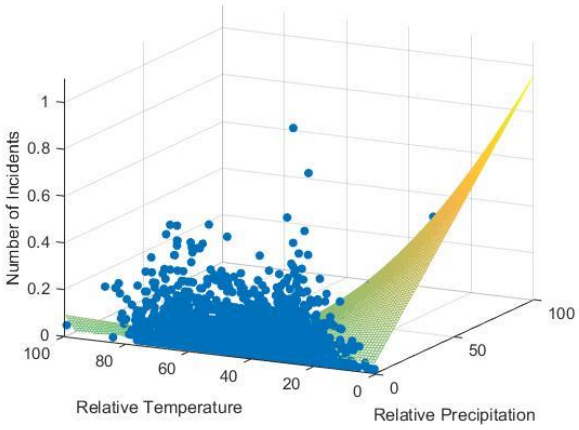
The most severe pair of variables is (Wind Speed, Precipitation), and the least one is (Temperature, Humidity). Regarding the ones with electrical discharge, one observes that (Precipitation, Intensity of discharge) are very severe in the occurrence of incidents.

For further analysis, one needs to understand the behaviour of multiple variables regarding the occurrence of incidents. Using (2.15), one uses as least 3 weather variables to obtain the regression equation for the number of faults, n_{faults}, Table 4.14 presenting the regression equation results, which were obtained as described in Section 3.4.1.

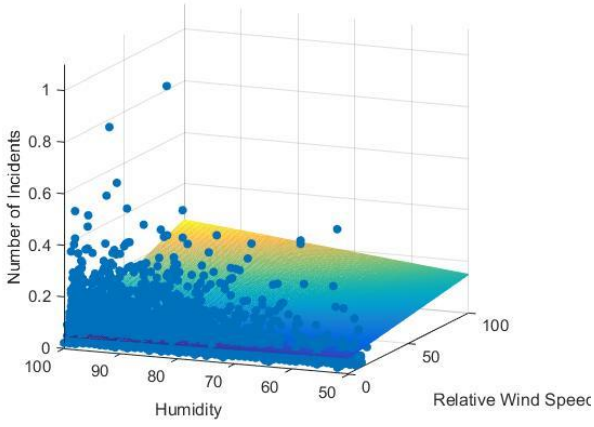
Using the equations from Table 4.14, one is able to have the first approximation for the prediction of the number of incidents, analysed further in this thesis.



a) Incidents vs. Precipitation and Wind Speed.



c) Incidents vs. Temperature and Precipitation.



c) Incidents vs. Humidity and Wind Speed.

Figure 4.7. Incidents vs. Weather variables on 24-hour-interval in Portugal.

Table 4.14. Portugal multiple variables regression equation.

Variables	Equation	
T H P D	$n_{faults} = 2.37 + 0.03 \times T_{[ºC]} - 0 \times H_{[\%]} + 0.018 \times P_{[\mathrm{mm}]} + 0.1 \times D$	(4.4)
T G P I	$n_{faults} = 1.63 + 0.039 \times T_{[ºC]} + 0.024 \times G_{[\mathrm{km/h}]} + 0.009 \times P_{[\mathrm{mm}]} + 0.023 \times M_{[\mathrm{kA}]}$	(4.5)
H G P D	$n_{faults} = 2.96 - 0.005 \times H_{[\%]} + 0.025 \times G_{[\mathrm{km/h}]} + 0.01 \times P_{[\mathrm{mm}]} + 0.01 \times D$	(4.6)
T H W G I	$n_{faults} = 1.1 + 0.04 \times T_{[ºC]} + 0.002 \times H_{[\%]} + 0.023 \times W_{[\mathrm{km/h}]} + 0.017 \times G_{[\mathrm{km/h}]} + 0.023 \times M_{[\mathrm{kA}]}$	(4.7)
T H W P D	$n_{faults} = 1.41 + 0.039 \times T_{[ºC]} - 0 \times H_{[\%]} + 0.038 \times W_{[\mathrm{km/h}]} + 0.015 \times P_{[\mathrm{mm}]} + 0.009 \times D$	(4.8)
T W G P I	$n_{faults} = 1.23 + 0.043 \times T_{[ºC]} + 0.023 \times W_{[\mathrm{km/h}]} + 0.017 \times G_{[\mathrm{km/h}]} + 0.008 \times P_{[m\mathrm{m}]} + 0.023 \times M_{[\mathrm{kA}]}$	(4.9)
H W G P D I	$n_{faults} = 2.9 - 0.007 \times H_{[\%]} + 0.02 \times W_{[\mathrm{km/h}]} + 0.016 \times G_{[\mathrm{km/h}]} + 0.005 \times P_{[\mathrm{mm}]} - 0.003 \times D + 0.023 \times M_{[\mathrm{kA}]}$	(4.10)
T H W G P D I	$n_{faults} = 1.08 + 0.045 \times T_{[ºC]} + 0.001 \times H_{[\%]} + 0.023 \times W_{[\mathrm{km/h}]} + 0.017 \times G_{[\mathrm{km/h}]} + 0.008 \times P_{[\mathrm{mm}]} - 0.004 \times D + 0.024 \times M_{[\mathrm{kA}]}$	(4.11)

4.4.2 Planned works analysis

The last statistical study to be addressed concerns the number of planned works, per each region, Table 4.15.

Table 4.15. Analysis of the incidents caused by planned works regarding other causes.

Region	Incidents planned works	Incidents other causes	Max planned works incidents in one day	Max other causes incidents in one day
Aveiro	0.02	0.32	0.06	0.81
Beja	0.01	0.11	0.04	0.17
Braga	0.03	0.30	0.09	1.00
Bragança	0.01	0.13	0.11	0.26
C. Branco	0.01	0.13	0.09	0.17
Coimbra	0.03	0.21	0.06	0.21
Évora	0.00	0.14	0.02	0.19
Faro	0.02	0.27	0.09	0.36
Guarda	0.01	0.12	0.04	0.34
Leiria	0.02	0.23	0.09	0.28
Lisbon	0.06	1.00	0.17	0.55
Portalegre	0.01	0.12	0.04	0.17
Porto	0.04	0.67	0.09	0.62
Santarém	0.02	0.29	0.06	0.32
Setúbal	0.02	0.41	0.06	0.45
V. Castelo	0.01	0.15	0.06	0.57
Vila Real	0.01	0.17	0.06	0.36
Viseu	0.02	0.21	0.09	0.34

An important note is that each time client services are affected a ticket is reported to control and register service unavailability. A comparison between the quantity of planned works and the number of incidents caused by other reasons is provided. One also presents the number maximum of incidents in both cases, in order to allow a comparison of the maximum number of incidents that occurred in one day. Due to the confidentiality of data, one presents the column for the quantity of incidents by planned works and other causes normalised to the maximum value of the latter. For the columns of the maximum incidents in one day, both columns are normalised to the maximum value of the other causes column. Annex B contains the complete results without the normalisation.

From the analysis of Table 4.15, one sees that the incidents caused by planned works are in a much smaller number than those caused by other reasons, the same being applied to the maximum number of incidents. This is due to the fact that planned works are scheduled in advance, hence, their severity and the number of incidents being controlled. Since incidents from number of planned works are not significant compared to those caused by other reasons, this parameter was not addressed in the forecasting methods.

4.5 Forecasting

4.5.1 Multiple Linear Regression

The simplest way of forecasting is by using a regression equation. Using the equations from Table 4.14, one has calculated the number of incidents for each variable, after which one calculated the MSE between the predicted number of faults that each of these equations gives and the actual number of incidents, Table 4.16.

Table 4.16. Error comparison among multiple variables regression, in Portugal.

	THPD	**TGPI**	**HGPD**	**THWGI**	**THWPD**	**TWGPI**	**HWGPDI**	**THWGPDI**
MSE	9.78	9.31	9.62	9.48	9.56	9.25	9.34	9.24

The best regression equation is the one using all variables, and the second-best is the one using all variables except Humidity. One can easily explain this situation due to the lack of correlation that Humidity has in the occurrence of incidents when dealing with the Portugal data. Still, one should note that the values of MSE are very similar, not enabling to establish a very strong conclusion.

By using the best regression equation, one can establish the first forecasting, by calculating the predicted number of incidents and having the real number of incidents, Figure 4.8, in relative terms, the figure with absolute data being provided in Annex B. The data is organised by day, from 1st January 2016 until 28th February 2017.

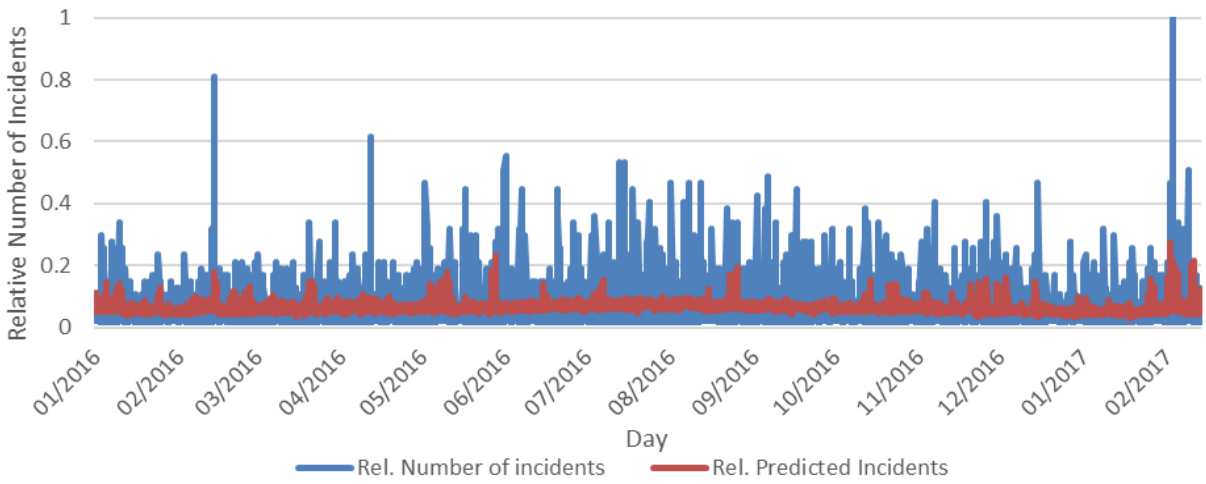


Figure 4.8. Real vs. predicted number of incidents in Portugal using the regression equation.

This regression equation can represent some of the incident's behaviour during the study period. However, most of these peaks are not reflected in this regression equation, as expected due to the complexity of the problem. To better perceive the behaviour of the regression equation in the peaks of incidents, one defines peak days as those one with top 5% of incidents, hence, being possible to calculate which quantity of incidents is considered a peak; data for this definition is given in Annex B.

Table 4.17 contains the MSE and the percentage of correct and false peaks that the forecasting method hits, as presented in Section 3.5., full results being shown in Annex B.

Table 4.17. Results for the forecasting study using the regression equation.

MSE	Mean error	Correct peaks [%]	False Peaks [%]
9.24	11.02	1.3	40

4.5.2 NARX Results

Using a NARX Neural Network to forecast the number of incidents requires first the training of the network, and then, with the trained network, the forecasting. Using the approach explained in detail in Section 3.4.2, one presents in Table 4.18 twelve networks with varied sizes and delays, only for the best simulation, since every time that the network is trained, the value of MSE also changes. The study was done only for Humidity, since the correlation of this variable with the occurrence of incidents is very low.

Table 4.18. NARX Neural Network study regarding the neurons, delay and error, for Humidity.

	Net 1	Net 2	Net 3	Net 4	Net 5	Net 6	Net 7	Net 8	Net 9	Net 10	Net 11	Net 12
Neurons	5	15	25	40	5	15	25	40	5	15	25	40
Delay	1	1	1	1	2	2	2	2	3	3	3	3
MSE no H	9.34	9.28	9.09	9.10	9.37	9.36	9.14	9.11	9.04	9.21	9.20	8.85
MSE with H	9.45	9.30	9.09	9.38	9.33	9.26	9.28	9.26	9.32	9.35	9.08	9.05

The best network is the one with 40 Neurons and 3 as Delay. However, the best network is when Humidity is not used as a variable, as shown in Figure 4.9 concerning the behaviour of the network, with the predicted and real number of incidents (the non-normalised figure is provided in Annex B).

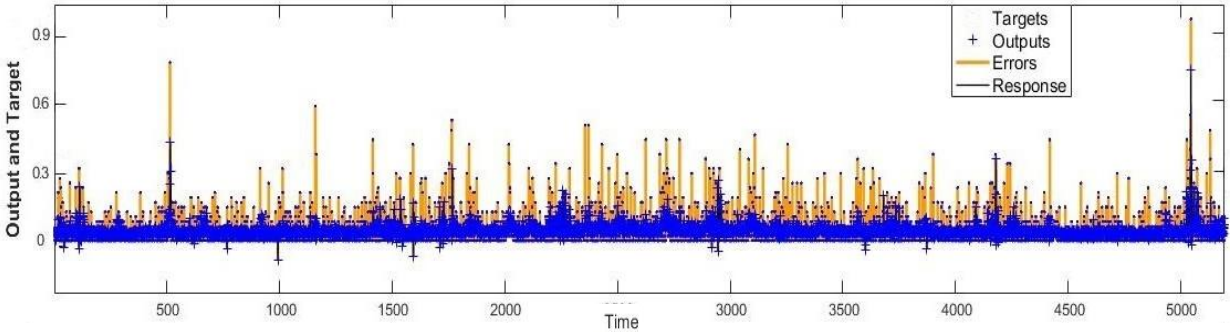


Figure 4.9. Real vs. predicted number of incidents in Portugal using the NARX neural network.

The study of how the neural network behaves when peaks of incidents occur is important, results being shown in Table 4.19, and full being presented in Annex B.

Table 4.19. Results for the forecasting study using the NARX neural network.

MSE	Mean error	Correct peaks [%]	False Peaks [%]
8.85	9.35	5.2	43

To enable a comparison between both methods, Table 4.20 shows the results from the previous studies, a colour scheme being used, where green represents the best case and red worst case.

Table 4.20. Comparison of NARX neural network and regression in forecasting study in Portugal.

	MSE	Mean error	Correct Peaks [%]	False Peaks [%]
Regression	9.24	11.02	1.3	40
NN	8.85	9.35	5.2	43

One concludes that regarding MSE and the Mean error, the NN has a better behaviour, but the values are not significantly better. Still, the best method to forecast peaks comes from Neural Networks, but it also increases the number of false peaks forecasted. Despite this conclusion, one can say that the percentage of correct peaks from both methods is low, leading to low confidence in the results.

4.5.3 Region forecasting

For the forecast in each region, one has used both the trained network and the regression equation with the information of Portugal. Table 4.21 shows both results from using the Neural Network and the Regression (represented as Reg), and a comparison between the MSE for both methods, where a red cell is positioned when the MSE is higher and a green one when it is lower; the absolute values are presented in Annex B. The values from MSE and the maximum error, from Neural Network and

Regression, are close, but Regression presents better results regarding MSE. Regarding the mean error, the value is approximately the same in each region and in both forecasting methods.

Table 4.21. Forecasting study by NARX neural network and regression.

	Relative Max incidents	MSE NN	Mean Error NN	Relative Max error NN	MSE Reg	Mean Error Reg	Relative Max error Reg
Aveiro	0.81	10.51	1.85	0.89	7.30	1.78	0.74
Beja	0.17	3.81	1.75	0.13	3.16	1.56	0.13
Braga	1.00	11.55	1.82	1.00	10.02	1.69	1.00
Bragança	0.26	4.84	1.80	0.23	3.58	1.56	0.19
C. Branco	0.17	4.58	1.61	0.30	2.63	1.42	0.10
Coimbra	0.21	3.47	1.54	0.17	3.03	1.42	0.16
Évora	0.19	6.25	1.76	0.49	3.25	1.53	0.15
Faro	0.36	6.07	1.74	0.43	4.31	1.55	0.30
Guarda	0.34	4.58	1.64	0.29	4.59	1.83	0.27
Leiria	0.28	3.55	1.49	0.23	3.69	1.49	0.31
Lisbon	0.55	38.22	4.66	0.56	44.07	5.05	0.56
Portalegre	0.17	6.00	1.72	0.43	4.51	1.84	0.19
Porto	0.62	19.80	2.93	0.68	19.24	2.97	0.63
Santarém	0.32	4.82	1.56	0.34	4.90	1.69	0.28
Setúbal	0.45	8.47	2.05	0.45	8.13	2.02	0.43
V. Castelo	0.57	9.98	1.98	0.58	5.55	1.49	0.57
Vila Real	0.36	9.44	1.92	0.75	4.88	1.79	0.25
Viseu	0.34	5.21	1.89	0.25	3.87	1.60	0.22

Figure 4.10 shows the MSE for both Neural Networks and Regression, together with the maximum incidents that occurred in each region. MSE in very related to the maximum incidents that occur in each region, due to huge difficulty from both methods to predict the peaks, which leads to an error increase.

Figure 4.11 shows the mean and maximum errors together with the maximum faults in each region. One can conclude that the maximum error follows the same behaviour as the number of maximum events. In some cases, the maximum error occurred on the day of the maximum incident. Regarding the mean value, the same value is usually achieved, except for Lisbon and Porto, where more incidents occur than for the other regions.

Another important study is the one realised on the peaks of incidents, presented in Table 4.22. One introduced the number of peaks in each region (represented as N. Peaks), the correct number of peaks in each method (represented as Correct), as well as the false peaks detected (represented as False). The full results are presented in Annex B.

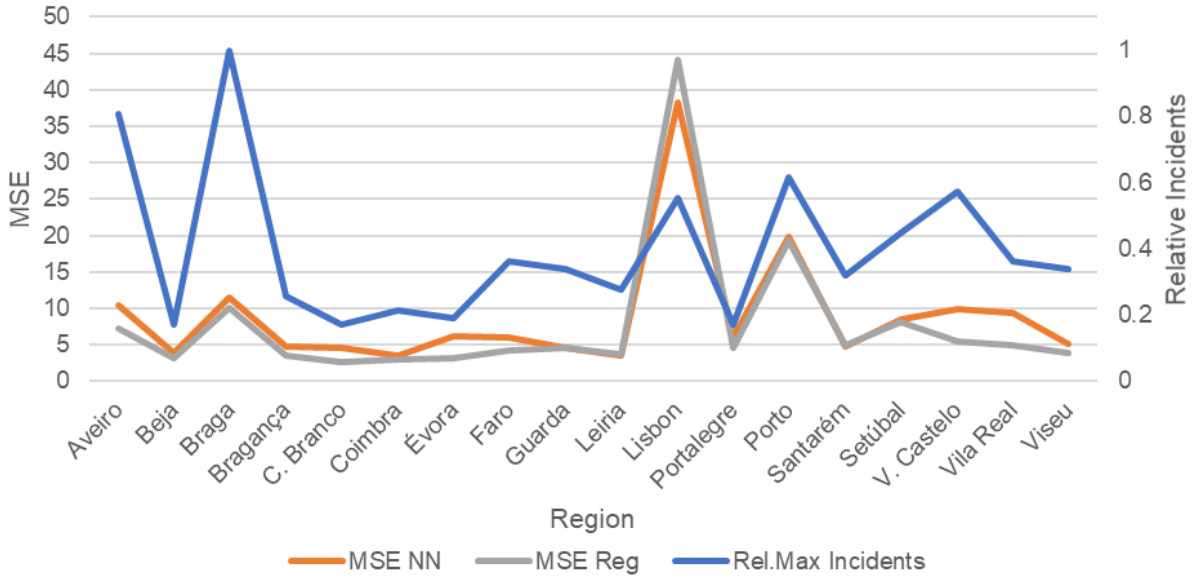


Figure 4.10. MSE from Neural Network and Regression vs. Maximum incidents.

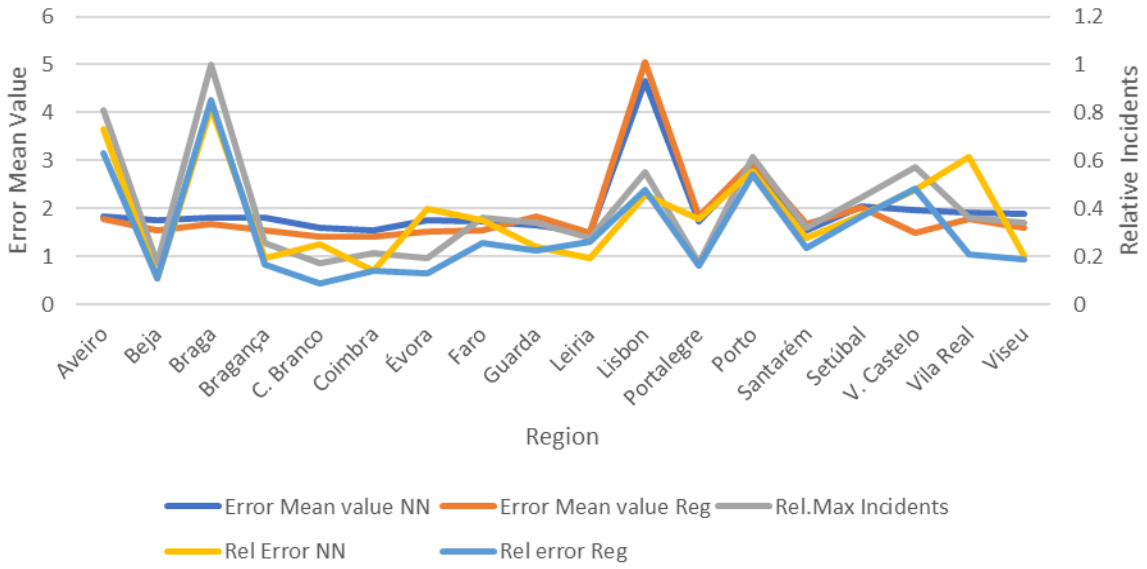


Figure 4.11. Mean and Maximum Error vs. Maximum incidents.

One presents in Table 4.23 the comparison between the Regression Equation and the NARX using the trained method with the data from Portugal, and forecasting using the Regions data. The full results are presented in Annex B.

The results for the forecast of the peaks are not good, since, for example, in the neural network the best case only hits 20% of the peaks, and in regression, it hits 36%. On average, these values are lower, and only 5% of peaks are detected in the Neural Networks, and 7% in the Regression. Another significant result is the false peaks error, where nearly 90% of peaks detected in the Neural Network and nearly 70% in the regression are false.

Table 4.22. Results of the forecasting by regression and NARX neural networks in each region.

	Correct NN [%]	Correct Reg [%]	False NN [%]	False Reg [%]
Aveiro	11	5	60	83
Beja	0	7	100	88
Braga	12	0	60	0
Bragança	0	25	100	93
C. Branco	10	0	88	0
Coimbra	8	0	80	0
Évora	0	10	100	91
Faro	7	7	75	88
Guarda	7	14	80	88
Leiria	0	6	100	86
Lisbon	0	0	0	0
Portalegre	13	19	90	86
Porto	0	0	100	0
Santarém	0	6	0	86
Setúbal	0	0	0	0
V. Castelo	20	0	50	0
Vila Real	9	36	75	90
Viseu	7	14	75	88

Table 4.23. Comparison of NARX neural network and regression in forecasting.

	Error Mean Value	Mean MSE	Correct Peaks [%]	Error False Peaks [%]
Regression	1.9	7.8	7.2	67
NN	2	9	5.3	86

To achieve better results, one conducted another experiment, creating a regression equation for each region, using the equation that uses all weather variables. One presents in Table 4.24 the coefficients for the variables, which all combined create the regression equation for each region, together with the MSE calculated in each region.

One can observe from Table 4.24 that the values from MSE follow the results previously presented in Figure 4.10. The MSE is always higher in Lisbon and Porto, which is expectable since the complexity of the network is higher. On the other hand, Beja and Portalegre have low MSE, which can be related to the low quantity of incidents that these regions have. However, one can conclude that in every case, the MSE is low than from the previous studies, as expected.

One completes the study about peaks of incidents, presenting the percentage of correct and false peaks. The visualisation of the mean and maximum error occurred during simulation is also possible. One describes in Table 4.25 the results for this study, with the full results presented in Annex B.

Table 4.24. Coefficients for the regression equation per region and its MSE.

	Constant	T	H	P	W	G	D	I	MSE
Aveiro	-0.91	0.110	-0.010	0.0470	0.081	0.014	0.135	0.030	6.43
Beja	2.77	-0.004	-0.015	0.004	0.011	0.007	0.001	0.009	1.05
Braga	3.03	0.054	-0.037	-0.006	-0.092	0.155	0.528	-0.006	8.22
Bragança	2.42	0.007	-0.012	0.024	0.085	-0.06	0.028	0.015	2.25
C. Branco	1.44	0.028	-0.002	0.023	0.017	-0.022	-0.002	0.016	1.30
Coimbra	1.02	0.046	-0.002	0.015	0.048	-0.025	0.007	0	2.14
Évora	2.55	-0.004	-0.007	-0.001	0.012	-0.003	-0.002	0.007	1.61
Faro	-1.36	0.093	0.023	0.049	-0.008	0.016	0.029	0.004	3.28
Guarda	0.95	0.003	0.002	0.039	-0.008	0.019	0.035	0	1.67
Leiria	0.73	0.041	0.005	0.085	0	0.008	-0.067	0.020	2.52
Lisbon	8.12	0.304	-0.073	0.224	0.011	-0.033	0.090	0.015	16.77
Portalegre	2.83	-0.010	-0.011	-0.001	-0.001	-0.001	0.002	0	1.05
Porto	3.43	0.153	-0.023	0.084	-0.046	0.071	0.250	0.036	12.67
Santarém	1.52	0.064	-0.009	0.011	-0.018	0.017	0.033	0.004	4.34
Setúbal	0.14	0.136	0.003	0.040	0.013	-0.009	0.013	0.005	7.08
V. Castelo	0.04	0.022	0.002	-0.005	0.083	0.025	0.401	0.015	2.82
Vila Real	0.80	-0.001	0.004	0.022	-0.092	0.101	0.009	0.017	3.03
Viseu	5.40	-0.020	-0.031	0.017	-0.103	0.090	-0.021	0.064	2.66

Table 4.25. Forecasting study of the regression per region.

	Mean error	Relative Max error	Correct [%]	False [%]
Aveiro	5.82	0.79	11	60
Beja	2.66	0.18	0	0
Braga	8.14	1.00	24	0
Bragança	5.05	0.29	13	0
C. Branco	3.30	0.14	0	0
Coimbra	4.34	0.23	0	0
Évora	7.00	0.13	0	0
Faro	4.04	0.25	21	25
Guarda	3.23	0.14	7	50
Leiria	4.22	0.34	0	0
Lisbon	7.95	0.51	0	0
Portalegre	2.68	0.19	0	0
Porto	8.61	0.74	25	29
Santarém	5.84	0.40	6	0
Setúbal	7.51	0.52	0	0
V. Castelo	4.90	0.34	70	0
Vila Real	4.33	0.34	27	0
Viseu	4.42	0.30	21	0

After the survey of the regression method to forecast the number of incidents using the data from each region, one can observe a good decrease in the false peaks detection. The next phase is the simulation using the Neural Network. The first step is the training of a network with the information of each region, and then the forecast, using the same data. The size and delay of the best network, as well as the MSE of the prediction, are presented in Table 4.26. One also presents the MSE of the regression method, to allow a comparison between both approaches. In green one presents the method that has low MSE and in red the one with higher.

Table 4.26. Region Neural Network neurons and delay, together with MSE of regression and NN.

	Neurons	Delay	MSE NN	MSE Reg
Aveiro	40	2	3.90	6.43
Beja	5	1	1.09	1.05
Braga	15	1	4.44	8.22
Bragança	25	2	1.73	2.25
C. Branco	15	1	1.20	1.30
Coimbra	40	1	1.75	2.14
Évora	40	1	1.30	1.61
Faro	15	3	2.89	3.28
Guarda	5	2	1.27	1.67
Leiria	15	3	1.88	2.52
Lisbon	40	2	14.98	16.77
Portalegre	5	2	1.08	1.05
Porto	15	3	11.43	12.67
Santarém	25	2	3.49	4.34
Setúbal	40	1	5.09	7.08
V. Castelo	15	2	2.28	2.82
Vila Real	25	1	2.30	3.03
Viseu	5	3	2.50	2.66

Given the training of a Neural Network for each region, the MSE using Neural Network is lower than the ones using the Regression equation. A different result from using the data from Portugal is obtained. These better results are due to the fact that one uses more accurate data for each region.

To complete the study, one addresses the study regarding the peaks of incidents using Neural Networks, described in Table 4.27. To have a comparison between both methods regarding the peaks of faults, one shows the final results in Table 4.28, drawn from Table 4.25 and Table 4.27, together with the percentage between false and total forecasted peaks. The full results are presented in Annex B.

One can conclude that when using Neural Networks, the number of correct forecasted peaks is higher than when using Regression. Despite the increase of false peaks when using Neural Networks, the false and correct peaks from both methods are approximately the same. Nevertheless, both approaches cannot achieve great performances in the forecasting of these peaks, despite the better results from Neural Networks and using Region data.

Table 4.27. Forecasting study of the NARX neural network per region.

	Mean error	Relative Max error	Correct [%]	False [%]
Aveiro	2.70	0.56	42	11
Beja	2.91	0.27	0	0
Braga	5.52	0.66	24	20
Bragança	2.67	0.34	50	0
C. Branco	2.84	0.20	10	0
Coimbra	3.34	0.30	8	0
Évora	2.93	0.26	0	100
Faro	2.97	0.52	29	20
Guarda	2.17	0.25	14	0
Leiria	2.67	0.46	18	0
Lisbon	6.39	0.64	20	20
Portalegre	2.60	0.30	0	0
Porto	8.25	1.00	25	17
Santarém	3.25	0.69	29	0
Setúbal	4.39	0.58	40	27
V. Castelo	1.39	0.48	60	25
Vila Real	2.75	0.50	55	0
Viseu	3.42	0.38	21	25

Table 4.28. Mean value of NARX neural network and regression in forecasting study per region.

	Error Mean Value	Mean MSE	Correct Peaks [%]	Error False Peaks [%]
Regression	5.2	4.5	11	19
NN	3.5	3.6	24	17

4.6 Weka Forecasting

The first step for the Weka [Weka17] classification is the processing of the data, which is independent of the study, meaning that each person can categorise data as considered to be more important. One decided to divide the file into three classes: A, B and C. A is the class with few incidents, B the intermediate one, and finally C is the one with higher incidents. The C class can be considered as a peak day, since the same number of incidents as the previous studies is used. Annex B contains the detailed description of the classification into the three classes for each Region.

As referred to in Section 3.4.3, one used four different methods, for the data of each region to forecast the number of incidents. One presents in Figure 4.12 the results from the Bayes Network, a method described in Section 2.4.3.

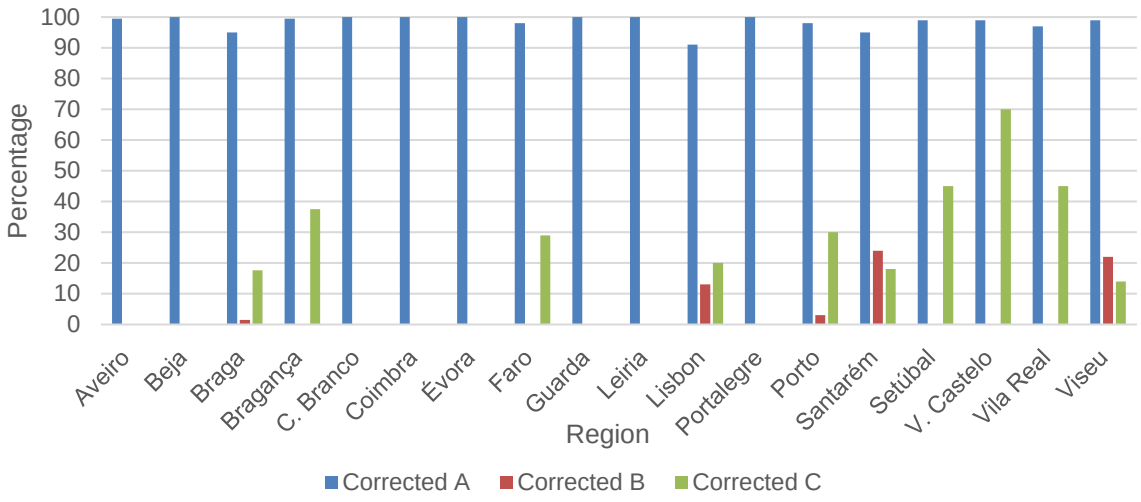


Figure 4.12. Bayes Network results per region from Weka.

The Bayes Network has a high accuracy in classifying the A class, but regarding the classification of classes B and C, the results are almost the opposite, the majority of Regions having 0% of accuracy regarding these two categories. The best results occur for the region of Viana do Castelo, with 100% of class A identified, 70% of class C and 0% of class B.

The results for MLP, described in Section 2.4.3, are presented in Figure 4.13.

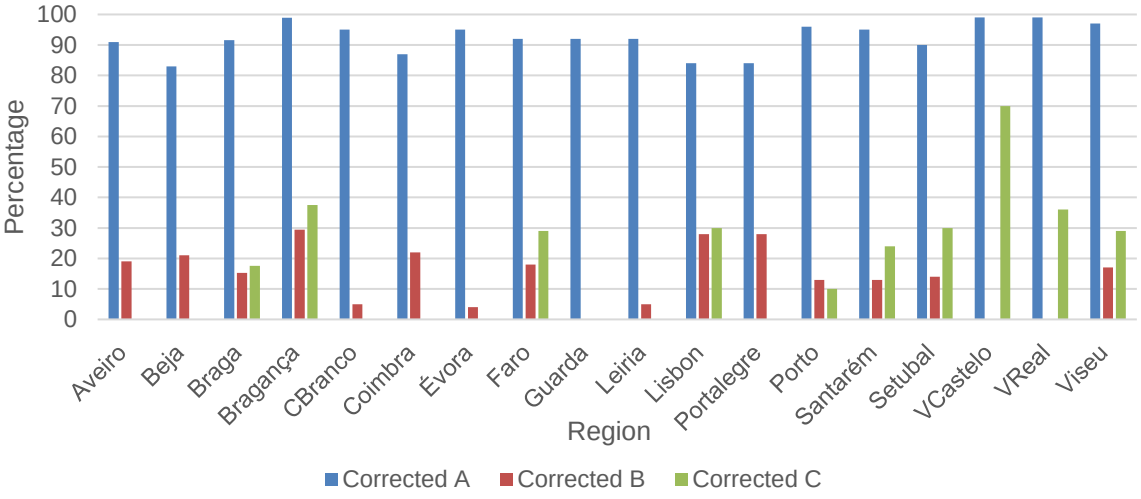


Figure 4.13. MLP results per region from Weka.

MLP has a high accuracy regarding the A class, but the accuracy is lower than the one observed from the Bayes Network. Regarding classes B and C, MLP does not achieve good results, but they are better than the ones from the Bayes Network. There is no predominant better region, however, Viana do Castelo appears from the A and C classes as the best case. Still, none of the Regions presents results that can be used in real scenarios.

Nearest Neighbours results are shown in Figure 4.14.

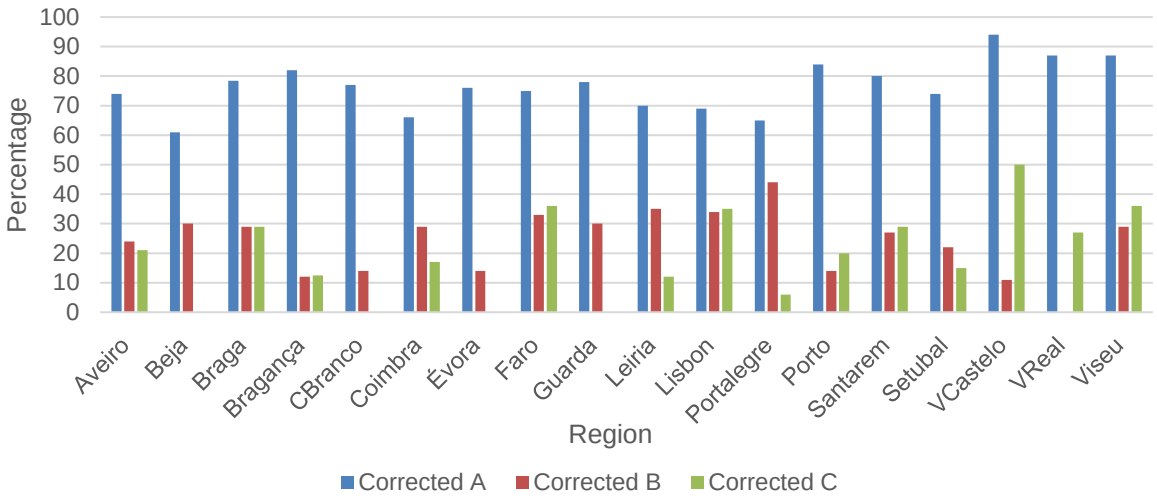


Figure 4.14. Nearest Neighbours results per region from Weka.

The accuracy of the Nearest Neighbours regarding the A class is the lowest so far, but still with a significant performance compared to the remaining classes. On the other hand, the accuracy of the remaining methods increases. This method presents the best results for class B. In this case, Faro and Lisbon present the more balanced results regarding the classification of the three categories.

The last method, SVM, is presented in Figure 4.15. The accuracy of class A is quite better to classes B and C. Regarding the correct classification, this is the weakest method.

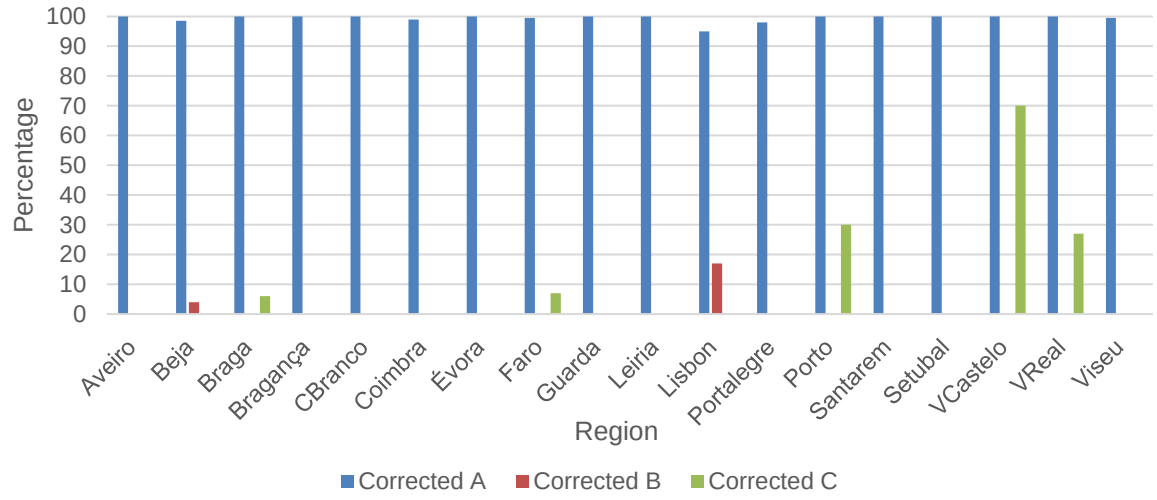


Figure 4.15. SVM results per region from Weka.

Table 4.29 shows the means of the accuracy of the three categories, for a comparison among the four methods. The percentage of the number of false and total forecasted peaks (represented by Error False P.) is also presented, calculated by (3.2). A colour scheme by column is also shown, where red represents the worst case, green the best case and yellow the mean ones. The complete results are available in Annex D.

Table 4.29. Mean values of the comparison among the four classification methods used in Weka.

	Mean [%]	Mean A [%]	Mean B [%]	Mean C [%]	Error False P. [%]
Bayes Network	73	98	4	18	56
MLP	72	92	14	17	62
Nearest Neigbour	63	77	24	19	78
SVM	73	99	1	8	22

Bayes Network, SVM and MLP have a very close accuracy, leading to be the best methods in this study. Regarding the accuracy in the three classes, one concludes that SVM is the best method to classify class A, with an accuracy of 99%. However, it has a low performance regarding the increase of incidents. Nearest Neighbour is the best method to forecast classes B and C, however, 78% of forecasted peaks are false, the highest number in the study: this method has also the most unsatisfactory result for the forecast of class A. Finally, for the Bayes Network and MLP, both have approximately the same results in the classification of classes A and C, and a lead to the error of false peaks, the only difference being that in the classification of class B, MLP has an advantage.

Chapter 5

Conclusions

In this chapter, conclusions are presented, finalising this work. One summarises the major findings of the study, as well as some aspects to be developed in future work.

The primary goal of this thesis was the study of the number of incidents regarding two main variables: Meteorological and Planned Works. After a brief analysis, one concluded that the incidents caused by planned works are much lower compared to the ones caused by weather. Due to this fact, only the study of meteorological variables was addressed. With the possibility of perceiving how the number of incidents is related to the various meteorological factors, the paradigm of the network managers can end up changing, since once getting a sense of the number of incidents and how this is related to weather, a new door opens to the operators in the organisation of their teams and network.

Five chapters compose this thesis, Introduction being the first. This chapter contains a summary of mobile wireless communications evolution to introduce the importance of the study of incidents for these types of network, as well as a brief introduction on incidents. One also presents the motivation for and contents in this work.

In Chapter 2, one provides a brief description of GSM, UMTS and LTE network architectures as well as the radio interface. Next, one describes the definition of alarms and incidents together with a brief description of how these faults occur and are propagated in the network, and then correlated until reaching the network manager. One also presents the approach for failure prediction, showing the possibilities to correlate incidents information with the meteorological variables. An introduction to forecasting methods is presented, showing the several possibilities to predict the number of incidents, as well as performance measures to assess these methods. Finally, one presents state of the art.

Regarding Chapter 3, one provides the description of the dataset used in this work, classifying data from NOS, IPMA and Weather Underground. Since each of these entities has the information organised differently, one first needed to process each dataset into a single file, in order to perform the different studies. One presents the organisation of the statistical and forecasting studies, as well as the several procedures to organise and relate the several datasets. Then, a detailed explanation is given on the statistical study presenting the steps to implement it, followed by the same detailed explanation for the forecasting study. Finally, the forecasting assessment is presented.

Concerning Chapter 4, it starts with the description of the scenario under study, providing a brief analysis of it, presenting some statistical data on the number of incidents and base stations sectors. Then, one addresses the ratio between the number of incidents and some variables, e.g. region size and base station sectors. With these results, one concludes that Portugal does not have a homogenous country in relation to these metrics, where each region differs from each other. This can be easily explained due to the vast diversity in the number of inhabitants and infrastructures that each region has in Portugal. Despite having a good amount of data, there are no parameters that link incidents with their cause, but it is possible to indirectly link the reasons for the faults with the network alarm, as shown in this chapter.

A pilot city, Braga, is used to understand how the meteorological variables are related to the number of incidents. The study was done using two intervals, 24-hour and 12-hour, enabling to understand the importance of data organisation by time intervals. The conclusion is that results are very similar, hence, one only uses the 24-hour interval in the rest of the work. The variable most related to the number of incidents is the quantity of electrical discharges, and the least one is Temperature. The importance of pairs of variables regarding the number of incidents is also addressed, and one concludes that the pair

of variables most related to the number of incidents is (Wind Speed, Humidity), but no pair shows to be the least related, although (Temperature, Humidity) and (Wind Speed, Precipitation) are the lower ones.

The regions in Portugal were taken for a study similar to the one for Braga. Every region behaves differently from the other, meaning that results of a national study do not reflect the behaviour of each region. However, dividing the country into three parts, North, Centre and South, results can sometimes be approximated. Precipitation is the variable with higher correlation in the North, but Wind and Gust Speed are also important in three regions, which can be explained due to the typical weather in those regions, with high levels of precipitation and wind speed during the year. For the least severe variables, Temperature is the one that occurs more often, due to the low temperatures registered during the year.

Regarding the Centre, one cannot draw a significant conclusion from the most and least severe variables, since there is no predominant factor in the occurrence of incidents, which can be explained by the mild weather, and the diversity of terrain and environment of each region. However, Precipitation and Temperature are the variables that appear more often and the most related to the number of incidents. On the other hand, Wind appears as the least severe in the interior, but in the coastal zone, there is no main conclusion.

Finally, for the South, Temperature is the most critical variable in the occurrence of incidents, explained by the high temperatures that occur in this area; however, in Beja, a surprising result appears, with Precipitation being the most severe variable. One can explain this situation since Beja is known by its high temperatures all over the year, hence, measures are taken to minimise this problem. Another explanation can be the vast differences that this region has in terms of terrain, since it ranges from the coast to the dense interior. For the least severe variable, there is no major factor related to the number of incidents, but, again Beja presents the result that Temperature shows up, which reinforces the previous possible explanation for the results.

Using the same division as before, pairs of variables have been studied for each region. For the North, (Precipitation, Temperature) appears more often, and (Wind Speed, Precipitation) appears as the second one. It is interested to observe that Precipitation appears in these two most influential pairs. Concerning the least severe pair, (Temperature, Humidity) shows up. For the Centre, one cannot observe any conclusions about the pair most and least related to the number of incidents. However, one sees that the two most severe pairs are (Precipitation, Temperature) and (Gust Speed, Precipitation). The pair most relevant for the South is (Wind Speed, Precipitation), appearing in two areas, while for the least severe pair (Temperature, Humidity) appears in two areas.

Then, one analysed the data from Portugal as a whole. For the one variable study, the maximum discharge intensity reveals to be the one most related to the number of incidents, and Humidity as the least one. For the pair of variables, and excluding the ones with electrical discharges, since it has few data, the most related one is (Wind Speed, Precipitation), and (Temperature, Humidity) the least. One also presents the study of planned works, concluding that the incidents with this origin are much less than the ones resulting from other causes. Due to this fact, one has not considered this variable for the forecasting study.

The forecasting study has been divided into 4 different procedures: the first uses the best regression equation and neural network using Portugal data; the second starts by using the regression equation and the neural network trained with Portugal data, but then using the data from each region; the third uses data from each region to train the neural network and to obtain the regression equation; finally, the fourth uses Weka for the classification of regions' data using the Bayes Network, SVM, MLP and Nearest Neighbours algorithms. A study about incidents peaks is done, which are the top 5% of incidents in each case. This is an important metric to network operators as it gives information about the chaotic days, giving data to organise their teams.

For the first study, one concludes that the best regression equation is when using all the meteorological variables available. This equation has an MSE of 9.24, hitting 1.3% of peaks and reaching 40% of false peaks detected. The value for the false peaks means a percentage of false peaks detected regarding all peaks detected. For the neural network, one concludes that the best network has 40 Neurons with a Delay of 3. The best result is reached when not using the Humidity variable, reaching an MSE of 8.85, hitting 5.2% of peaks and 43% of false peaks. The neural network has better results, but they are unsatisfactory.

The next study, using the trained equation and network of Portugal but with the data from each region, shows that the regression has on average an MSE of 7.8, hitting 7.2% of peaks and reaching 67% of false peaks. On the other hand, the neural network has on average an MSE of 9, hitting 5.3% of peaks and reaching 86% of false peaks. Regarding this case, the regression reaches better results, but these remain unsatisfactory, where nearly 70% of the peaks detected are false.

For the following study, training each equation and network with the region data and using the same data to predict, the regression has on average an MSE of 4.5, hitting 11% of peaks with 19% of false peaks. For the neural network, it has on average an MSE of 3.6, hitting 24% of peaks with 17% of false peaks. These are the best results as expected. One can conclude that for the regular days, the forecasting method manages to work, the problem being the prediction of peaks, where the best result only hits 24%.

One used the Weka software, which allows the use of four more algorithms of classification. In this case, a division of the dataset is necessary, class A being the day with lowest incidents, class B the days with common incidents, and class C the peaks days. The best method to predict peaks is the Nearest Neighbours with 19% of correct classification, but the problem is that this approach reaches 78% of false peaks. On average, the best methods are the Bayes Network and SVM, but the former only hits 4% of class B and has 56% of false peaks, while in the latter, the classification of classes B and C are poor.

One can globally conclude that the best results come from training the NARX neural network with the data from each region and use the same data to obtain the results, leading to the best overall results, as well as the best results in the peak study. However, regarding the peak study, this result is not satisfactory to use in a real scenario, where it hits only 24% of peaks, which is explained by the type of data that was available: one only uses one weather station per region, and regions are very vast to allow a precise correlation.

Regarding future work, a more profound analysis could be made, since this is a new field of study where there is very few information about the relationship between incidents and meteorological variables. This thesis is part of a new study about this area, and one of the main aims is 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 used. The second improvement is on weather data, since data from personal weather stations was used, which sometimes is not accurate. Then, the next improvement is the use of weather stations nearby the base stations, in order to link in a better mode, the weather which occurs in each base station; to improve the quality of the results, the base stations near to weather stations should be used, reaching a better correlation between the number of incidents and weather variables. One also misses the opportunity of considering the seasonality of weather data, a topic important to be taken in consideration. In this thesis one only studies the number of incidents, but it could be important to define the severity of each incident. Finally, one only addresses some of the machine learning algorithms, but there are many more algorithms that can be used, so the use of other algorithm or different algorithm configurations should be explored, one example being the use of the survival analysis, considering an incident as a death.

Annex A

Meteorological Stations

In this Annex the Weather Underground ID is presented for the weather station used.

A.1 List of Meteorological Stations

One presents in Table A.1 the Station ID of the meteorological stations used in Weather Underground. Whenever possible, airport meteorological is used. However, for Lisbon, precipitation information was not available, and therefore another meteorological station is used.

Table A.1. Meteorological Stations used in Weather Underground.

Region	Station ID
Aveiro	IAVEIROG2
Beja	LPBJ
Braga	IBRAGABR7
Bragança	IPORTUGA50
C. Branco	ICASTELO11
Coimbra	ICOIMBRA17
Évora	IVORAEVO2
Faro	LPFR
Guarda	IGUARDAG7
Leiria	LPMR
Lisbon	ILISBOAL20
Portalegre	IPORTALE14
Porto	LPPR
Santarém	IPORTUGA75
Setúbal	ISETUBAL4
V. Castelo	IVIANADO17
Vila Real	IVILAREA7
Viseu	IVISEUI6

Annex B

Confidential Information

One presents in this Annex the confidential information in this thesis.

Annex C

Statistical Study

In this Annex. the statistical study completed is presented.

C.1 Regions Study

C.1.1 Number of incidents vs. one variable

Table C.1. Equation from Aveiro.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.0093	-0.05	0.04	0.07	0.10	0.12	0.09
b	2.59	7.99	2.79	0.64	1.40	2.85	2.82

Table C.2. Equation from Beja.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0	-0.007	0.01	0.01	0.008	0.02	0.018
b	1.68	2.30	1.63	1.14	1.62	1.66	1.61

Table C.3. Equation from Braga.

	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	0.2	0.11
b	2.14	6.83	2.76	0.98	0.85	2.33	1.88

Table C.4. Equation from Bragança.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.01	-0.02	0.02	0.01	0.008	0.11	0.05
b	1.38	3.42	1.85	1.44	1.60	1.80	1.70

Table C.5. Equation from Castelo Branco.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.01	-0.005	0.008	0	0	0.02	0.02
b	1.40	2.31	1.87	1.93	1.96	1.90	1.84

Table C.6. Equation from Coimbra.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.02	0.003	0.01	0.01	0.01	0.02	0.008
b	1.38	1.94	2.20	1.80	1.93	2.21	2.20

Table C.7. Equation from Évora.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	-0.001	-0.008	-0.005	0.006	0.006	0.0003	0.006
b	2.08	2.68	2.01	1.83	1.83	2.01	1.99

Table C.8. Equation from Faro.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.02	0.01	0.07	-0.004	0.01	0.11	0.07
b	1.90	1.66	2.56	2.87	2.64	2.65	2.62

Table C.9. Equation from Guarda.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.002	0.004	0.02	0.01	0.02	0.11	0.05
b	1.87	1.55	1.83	1.52	1.32	1.77	1.79

Table C.10. Equation from Leiria.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.006	0.003	0.03	0.02	0.02	-0.008	0.01
b	2.05	2.05	2.20	2.16	2.16	2.35	2.32

Table C.11. Equation from Lisbon.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.13	-0.15	0.009	0.01	0.006	0.07	0.06
b	1.33	20.9	7.98	7.42	7.69	7.92	7.92

Table C.12. Equation from Portalegre.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.002	-0.009	-0.003	-0.005	-0.003	0.005	0.002
b	1.61	2.38	1.71	1.77	1.75	1.70	1.70

Table C.13. Equation from Porto.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.03	-0.02	0.06	0.09	0.07	0.18	0.17
b	4.21	6.93	5.11	3.28	3.13	5.21	5.16

Table C.14. Equation from Santarém.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.03	-0.01	-0.004	0.007	0.009	0.05	0.02
b	0.91	3.71	2.74	2.63	2.54	2.67	2.70

Table C.15. Equation from Setúbal.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	0.05	0.005	0.006	0.0003	-0.006	0.02	0.009
b	0.84	3.23	3.64	3.65	3.79	3.62	3.63

Table C.16. Equation from Viana do Castelo.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	-0.004	0.01	0.02	0.07	0.06	0.21	0.11
b	2.36	1.35	1.96	0.21	0.15	1.87	1.86

Table C.17. Equation from Vila Real.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	-0.007	0.001	0.14	0.04	0.05	0.08	0.10
b	2.66	2.21	2.10	0.62	0.47	2.17	2.10

Table C.18. Equation from Viseu.

	T [ºC]	H [%]	P [mm]	W [km/h]	G [km/h]	D	I [kA]
m	-0.007	-0.001	0.05	0.02	0.03	0.06	0.10
b	2.76	2.46	2.20	1.88	1.71	2.25	2.06

C.1.2 Correlation Coefficient

Table C.19. Correlation Coefficient in Aveiro.

	T	H	P	W	G	D	I
Pearson	0.05	-0.09	0.23	0.34	0.31	0.21	0.44
Spearman	0.06	-0.07	0.13	0.19	0.20	0.15	0.20
Kendall Tau	0.05	-0.06	0.11	0.13	0.14	0.11	0.15

Table C.20. Correlation Coefficient in Beja.

	T	H	P	W	G	D	I
Pearson	0	-0.08	0.14	0.16	0.20	0.18	0.05
Spearman	0.05	-0.12	0.12	0.07	0.07	0.20	0.10
Kendall Tau	0.04	-0.11	0.11	0.06	0.06	0.19	0.10

Table C.21. Correlation Coefficient in Braga.

	T	H	P	W	G	D	I
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

Table C.22. Correlation Coefficient in Bragança.

	T	H	P	W	G	D	I
Pearson	0.14	-0.16	0.11	0.12	0.08	0.63	0.60
Spearman	0.09	-0.05	-0.01	0	0.02	0.50	0.40
Kendall Tau	0.07	-0.04	-0.01	0	0.01	0.40	0.29

Table C.23. Correlation Coefficient in Castelo Branco.

	T	H	P	W	G	D	I
Pearson	0.20	-0.10	0.11	0	0	0.12	0.38
Spearman	0.14	-0.05	0.03	0.01	0.01	0.17	0.26
Kendall Tau	0.11	-0.03	0.03	0.01	0	0.13	0.20

Table C.24. Correlation Coefficient in Coimbra.

	T	H	P	W	G	D	I
Pearson	0.18	0.02	0.08	0.10	0.07	0.14	0.05
Spearman	0.07	0.06	0.11	0.11	0.08	0.17	0.12
Kendall Tau	0.05	0.04	0.10	0.08	0.06	0.13	0.09

Table C.25. Correlation Coefficient in Évora.

	T	H	P	W	G	D	I
Pearson	-0.02	-0.08	-0.03	0.10	0.10	-0.10	0.06
Spearman	-0.1	-0.08	0.08	0.04	0.04	0.02	0.15
Kendall Tau	-0.08	-0.06	0.07	0.03	0.03	0.04	0.11

Table C.26. Correlation Coefficient in Faro.

	T	H	P	W	G	D	I
Pearson	0.12	0.06	0.41	-0.02	0.05	0.55	0.21
Spearman	0.19	0.04	0.16	0.07	0.10	0.41	0.24
Kendall Tau	0.14	0.03	0.13	0.05	0.08	0.30	0.18

Table C.27. Correlation Coefficient in Guarda.

	T	H	P	W	G	D	I
Pearson	0.02	0.04	0.17	0.11	0.16	0.77	0.43
Spearman	-0.02	0.05	0.22	0.17	0.21	0.20	0.38
Kendall Tau	-0.02	0.04	0.19	0.01	0.16	0.15	0.30

Table C.28. Correlation Coefficient in Leiria.

	T	H	P	W	G	D	I
Pearson	0.05	0.02	0.20	0.08	0.08	-0.40	-0.19
Spearman	0.02	0.08	0.17	0.09	0.09	-0.42	-0.19
Kendall Tau	0.02	0.06	0.14	0.07	0.07	-0.32	-0.14

Table C.29. Correlation Coefficient in Lisbon.

	T	H	P	W	G	D	I
Pearson	0.44	-0.32	0.02	0.04	0.02	0.10	0.03
Spearman	0.36	-0.21	-0.07	0.08	0.07	0.09	-0.07
Kendall Tau	0.26	-0.15	-0.05	0.06	0.05	0.07	-0.04

Table C.30. Correlation Coefficient in Portalegre.

	T	H	P	W	G	D	I
Pearson	0.04	-0.15	-0.03	-0.04	-0.03	0.12	0.04
Spearman	-0.06	-0.11	0.03	0.01	0.05	0	0.12
Kendall Tau	-0.05	-0.08	0.03	0.01	0.04	0	0.10

Table C.31. Correlation Coefficient in Porto.

	T	H	P	W	G	D	I
Pearson	0.09	-0.02	0.20	0.20	0.27	0.71	0.76
Spearman	0.01	0.04	0.18	0.13	0.15	0.69	0.60
Kendall Tau	0.01	0.03	0.14	0.10	0.11	0.52	0.45

Table C.32. Correlation Coefficient in Santarém.

	T	H	P	W	G	D	I
Pearson	0.26	-0.15	-0.01	0.03	0.03	0.28	0.03
Spearman	0.19	-0.08	-0.05	0.02	-0.01	0.28	0.20
Kendall Tau	0.14	-0.07	-0.05	0.02	0	0.21	0.14

Table C.33. Correlation Coefficient in Setúbal.

	T	H	P	W	G	D	I
Pearson	0.28	0.02	0.02	0	-0.02	0.15	0.02
Spearman	0.14	0.10	0.01	0.06	0.04	0	0.04
Kendall Tau	0.10	0.07	0.01	0.04	0.03	0	0.03

Table C.34. Correlation Coefficient in Viana do Castelo.

	T	H	P	W	G	D	I
Pearson	-0.03	0.04	0.15	0.39	0.38	0.70	0.24
Spearman	0.02	-0.06	0.16	0.22	0.19	0.69	0.55
Kendall Tau	0.02	-0.04	0.13	0.17	0.15	0.53	0.34

Table C.35. Correlation Coefficient in Vila Real.

	T	H	P	W	G	D	I
Pearson	-0.07	0.01	0.51	0.35	0.39	0.31	0.45
Spearman	-0.03	-0.11	0.25	0.24	0.23	0.36	0.40
Kendall Tau	-0.02	-0.08	0.21	0.19	0.18	0.28	0.29

Table C.36. Correlation Coefficient in Viseu.

	T	H	P	W	G	D	I
Pearson	-0.08	-0.01	0.22	0.14	0.17	0.25	0.84
Spearman	-0.04	-0.05	0.19	0.10	0.06	0.57	0.71
Kendall Tau	-0.03	-0.04	0.16	0.07	0.05	0.43	0.59

Annex D

Weka Classification

One described in this Annex the full results for the Weka Classification.

D.1 Weka Results

One represents in the following tables the results for the Weka classification, using the Bayes Network (Represented by Bayes Net.), MLP, SVM and Nearest Neighbours (represented by Nearest N.)

Table D.1. Weka Results in Aveiro.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	69.5	1	99.5	0	0
MLP	68	6	91	19	0
SVM	69.8	0	100	0	0
Nearest N.	58.7	10	74	24	21

Table D.2. Weka Results in Beja.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	60	0	100	0	0
MLP	57	3	83	21	0
SVM	61	0	98.5	4	0
Nearest N.	46.3	13	61	30	0

Table D.3. Weka Results in Braga.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	73	8	95	1.5	17.6
MLP	73	4	91.6	15.3	17.6
SVM	75.6	0	100	0	6
Nearest N.	66.3	10	78.4	29	29

Table D.4. Weka Results in Bragança.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	82.5	2	99.5	0	37.5
MLP	82.5	2	98.9	29.4	37.5
SVM	81.6	0	100	0	0
Nearest N.	69.4	7	82	12	12.5

Table D.5. Weka Results in Castelo Branco.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	75	0	100	0	0
MLP	72	2	95	5	0
SVM	75	0	100	0	0
Nearest N.	60	13	77	14	0

Table D.6. Weka Results in Coimbra.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	66	0	100	0	0
MLP	64	0	87	22	0
SVM	65	0	99	0	0
Nearest N.	53	10	66	29	17

Table D.7. Weka Results in Évora.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	75	0	100	0	0
MLP	72	2	95	4	0
SVM	75	0	100	0	0
Nearest N.	60	11	76	14	0

Table D.8. Weka Results in Faro.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	73	8	98	0	29
MLP	72	5	92	18	29
SVM	73	2	99.5	0	7
Nearest N.	63	10	75	33	36

Table D.9. Weka Results in Guarda.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	76	0	100	0	0
MLP	70	1	92	0	0
SVM	76	0	100	0	0
Nearest N.	65	7	78	30	0

Table D.10. Weka Results in Leiria.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	65	0	100	0	0
MLP	61	4	92	5	0
SVM	65	0	100	0	0
Nearest N.	56	12	70	35	12

Table D.11. Weka Results in Lisbon.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	62	11	91	13	20
MLP	62	10	84	28	30
SVM	64	0	95	17	0
Nearest N.	56	18	69	34	35

Table D.12. Weka Results in Portalegre.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	57	0	100	0	0
MLP	58	2	84	28	0
SVM	56	0	98	0	0
Nearest N.	53	14	65	44	6

Table D.13. Weka Results in Porto.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	78	4	98	3	30
MLP	78	2	96	13	10
SVM	79	2	100	0	30
Nearest N.	69	11	84	14	20

Table D.14. Weka Results in Santarém.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	77	2	95	24	18
MLP	75	4	95	13	24
SVM	76	0	100	0	0
Nearest N.	67	10	80	27	29

Table D.15. Weka Results in Setúbal.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	74	9	99	0	45
MLP	70	11	90	14	30
SVM	73	0	100	0	0
Nearest N.	59	8	74	22	15

Table D.16. Weka Results in Viana do Castelo.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	90	4	99	0	70
MLP	89	2	99	0	70
SVM	91	1	100	0	70
Nearest N.	86	2	94	11	50

Table D.17. Weka Results in Vila Real.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	85	10	97	0	45
MLP	86	4	99	0	36
SVM	86	0	100	0	27
Nearest N.	75	7	87	0	27

Table D.18. Weka Results in Viseu.

	Correct class [%]	False Peaks	Correct A [%]	Correct B [%]	Correct C [%]
Bayes Net.	84	0	99	22	14
MLP	82	4	97	17	29
SVM	80	0	99.5	0	0
Nearest N.	77	8	87	29	36

References

[3GPP16] 3GPP, http://www.3gpp.org/technologies/ keywords-acronyms/98-lte, Oct. 2016.

[ACMP16] 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.

[AdAg13] R. Adhikari and R. Agrawal, *An Introductory Study on Time series Modeling and Forecasting,* Lambert Academic Publishing, Saarbrücken, Germany, 2013.

[AGLA03] J. Arriagada, M. Genrup, A. Loberg and M. Assadi, "Fault Diagnosis System for an Industrial Gas Turbine by Means of Neural Networks", in *Proc. IGTC'03 – 8th International Gas Turbine Congress*, Tokyo, Japan, Nov. 2003.

[Alon12] 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).

[ANAC16] ANACOM – *Autoridade Nacional de Comunicações*, http://www.anacom.pt/render.jsp?categoryId=382989#.WAd0yGYrLIU, Oct. 2016.

[BeHD17] 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).

[BGRU99] 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.

[BLMD02] R. Bouraoui, M. Lahmar, A. Majdoub, M. Djemali and R. Belyea, "The relationship of temperature-humidity index with milk production of dairy cows in a Mediterranean climate", *Animal Research*, Vol.51, No. 6, Nov. 2002, pp. 479-491.

[BMHF11] 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.

[BoCF94] 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.

[CCox12] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*,

Wiley, Chichester, United Kingdom, 2012.

[ChPo02] P. Chen and P. Popovich, *Correlation – Parametric and Nonparametric Measures,* Sage Publications, Thousand Oaks, California, United States of America, 2002.

[Cisc16] Cisco, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, Nov. 2016.

[DBPA08] 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.

[Dumm17] Dummies – How to interpret a correlation coefficient - http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/, Sep. 2017.

[HaRM03] T. Halonen, J. Romero and J. Melero, *GSM, GPRS and EDGE Performance – Evolution Towards 3G/UMTS*, Wiley, Chippenham, United Kingdom, 2003.

[Heat13] 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).

[HoCF95] K. Houck, S. Calo and A. Finkel, "Towards a Practical Alarm Correlation System" in A.S. Sethi et al. (eds.), *Integrated Network Management IV*, Chapman and Hall, London, United Kingdom, 1995.

[HoHa06] B. Hollifield and E. Habibi, *The Alarm Management Handbook*, PAS, Houston, Texas, United States of America, 2006.

[HoTo06] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS,* Wiley, Chippenham, United Kingdom, 2006.

[HoTo07] H. Holma and A. Toskala, *WCDMA for UMTS – HSPA Evolution and LTE*, Wiley, Chippenham, United Kingdom, 2007.

[HoTo11] H. Holma and A. Toskala, *LTE for UMTS – Evolution to LTE-Advanced*, Wiley, Chippenham, United Kingdom, 2011.

[IPMA17] IPMA – Instituto Português do Mar e Atmosfera, http://www.ipma.pt, Feb. 2017.

[ITUT92] 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).

[JaIH04] M. Jaudet, N. Iqbal and A. Hussain, "Neural Networks for Fault-prediction in a Telecommunications Network", in *Proc. of INMIC 2004 – 8th International Multitopic Conference,* Punjab, Pakistan, Dec. 2004.

[JaTu96] J. Tu, “Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes”, *Journal of Clinical Epidemiology*, Vol.49, No.1, Dec. 1996, pp. 1225-1231.

[JaWe95] 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.

[JZar05] J. Zar, “Spearman Rank Correlation” in John Wiley & Sons, Ltd, *Encyclopedia of Biostatistics*, Wiley, Chippenham, United Kingdom, 2005.

[KlMT99] M. Klemettinen, H. Mannila and H. Toivonen, “Rule Discovery in Telecommunication Alarm Data”, *Journal of Network and System Management*, Vol.7, No. 4, Dec. 1999, pp. 395-423.

[Kuhn97] R. Kuhn, “Sources of Failure in the Public Switched Telephone Network”, *IEEE Computer*, Vol. 30, No. 4, Apr. 1997, pp. 31-36.

[LGHK97] 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.

[LHTG96] T. Lin, B. Horne, P. Tine and C. Giles, “Learning Long-Term Dependencies in NARX Recurrent Neural Networks”, *IEEE Transactions on Neural Networks*, Vol.7, No. 6, Nov. 1996, pp. 1329-1338.

[Math17] MathWorks – Estimate Multiple Linear Regression Coefficients, https://www.mathworks.com/help/stats/regress.html, May 2017.

[Matl16] Matlab Documentation - Design Time Series NARX Feedback Neural Networks, https://www.mathworks.com/help/nnet/ug/design-time-series-narx-feedback-neural-networks.html, Dec. 2016.

[Netw16] Networking — Something Good to Know, Oct. 2016 (https://conningtech.files.wordpress.com/2010/07/umtsnetworkdomains1.jpg).

[NOSP17] NOS – Portuguese Operator, http://www.nos.pt/, Sep. 2017.

[OgSP02] S. Ogaji, R. Singh and S. Probert, “Multiple-sensor fault-diagnosis for a 2-shaft-stationary gas turbine”, *Applied Energy*, Vol.71, No. 4, April 2002, pp. 321-339.

[ReFe10] 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.

[RFGD08] 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.

[RoNi88] 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.

[ScyP17] ScyPy – Open Source Software, https://docs.scipy.org/doc/scipy-0.14.0/reference/stats.html, Mar. 2017.

[SeTB11] S. Sesia, I. Toufik and M. Baker, *LTE – The UMTS Long Term Evolution,* Wiley, Chippenham, United Kingdom, 2011.

[SiHG97] 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.

[StSe04] M. Steinder and A. Sethi, “A survey of fault localisation techniques in computer networks”, *Science of Computer Programming*, Vol. 53, No. 2, Nov. 2004, pp. 165-194.

[Vaně08] 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).

[Vena14] P. Venâncio, *Analysis of Network Quality Using Non-Intrusive Methods from an End-User Perspective in LTE*, M.Sc Thesis, Técnico Lisbon/University of Lisbon, Lisbon, Portugal, 2014.

[WaBo09] Z. Wang and A. Bovik, “Mean Squared Error: Love It or Leave It?”, *IEEE Signal Processing Magazine*, Vol. 26, No. 1, Jan. 2009, pp. 98-117.

[Wall09] S. Wallin, “Chasing a Definition of “Alarm””, *Journal of Network and System Management*, Vol.17, No. 4, Dec. 2009, pp. 457-481.

[WaLL09] S. Wallin, V. Leijon and L. Landén, “Statistical analysis and prioritisation of alarms in mobile networks”, *International Journal of Business Intelligence and Data Mining*, Vol. 4, No. 1, May 2009, pp. 4-21.

[Weat17] Weather Underground – Meteorological Information, https://www.wunderground.com/, Feb. 2017.

[Weka17] Weka – Collection of machine learning algorithms - http://www.cs.waikato.ac.nz/ml/weka/, July 2017.

[Wiki17a] Wikipedia – Portugal Region Size, https://pt.wikipedia.org/wiki/Lista_de_distritos_portugueses_ordenados_por_%C3%A1rea , July 2017.

[Wiki17b] Wikipedia – Portugal Population Size, https://pt.wikipedia.org/wiki/Lista_de_distritos_portugueses_ordenados_por_popula%C3%A7%C3%A3o, July 2017.

[Wilk06] D. Wilks, *Statistical Methods in the Atmospheric Sciences,* Elsevier, London, United Kingdom, 2006.

[WiYu10] B. Wilamowski and H. Yu, “Improved Computation for Levenberg-Marquardt Training”, *IEEC Transactions on Neural Networks*, Vol.21, No. 6, June 2010, pp. 930-937.

[Worl17] World O Meters, http://www.worldometers.info/world-population/portugal-population/, Aug. 2017.

[XiTL09] H. Xie, H. Tang and Y. Liao, “Time Series Prediction based on NARX Neural Networks: An Advanced Approach”, in *Proc. of 8th International Conference on Machine Learning and Cybernetics,* Baoding, China, July 2009.

[YFHS11] 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.

[ŽeKS11a] 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.

[ŽeKS11b] D. Željko, M. Kunstic and B. Spahija, “A Comparison of Traditional Forecasting Methods for Short-term and Long-term Prediction of Faults in the Broadband Networks”, in *Proc. MIPRO 2011 – 34th International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia, May 2011.

[ŽeKu10] 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.

[ŽeRK16] 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.