Machine Learning Techniques using Key Performance Indicators for the Configuration Optimization of 4G Networks

Rui Miguel Maia Santos

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Doctor António José Castelo Branco Rodrigues
Doctor Pedro Manuel de Almeida Carvalho Vieira

Examination Committee

Chairperson: Doctor José Eduardo Charters Ribeiro da Cunha Sanguino
Supervisor: Doctor António José Castelo Branco Rodrigues
Member of the Committee: Doctor Pedro Joaquim Amaro Sebastião

November 2018

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

First and foremost, I would like express my very profound gratitude to my parents for all the support and encouragement they have given me over the years, as this accomplishment would not have been possible without them. I would also like to thank my sister for always being there when I needed.

I would also like to thank my supervisors Professor António Rodrigues and Professor Pedro Vieira, for all the support and insight they provided during the development of this thesis.

I would like to thank Instituto de Telecomunicações for providing me the required means for the completion of this dissertation.

I would like to thank Celfinet for giving me the opportunity develop this work in a company environment. A special acknowledgement to Eng. Marco Sousa and Eng. Rodrigo Veríssimo for providing valuable insight and support throughout the course of this project.

Last but not least, to all my friends that helped me through my time in Técnico by collaborating in projects and studying, or just being great friends overall.

Abstract

Mobile networks have been growing significantly in recent years, both in terms of number of subscribers and in the complexity of the network itself. As such, performing an efficient management of the network, to ensure that it maintains the desired performance, has become an increasingly difficult task. Moreover, due to the increase in the amount data available, namely Key Performance Indicators (KPI), it becomes unfeasible to do this management using the traditional methods.

This thesis focuses on the evaluation of the performance of a LTE network through the use of unsupervised learning techniques. The main objective is to detect groups of cells that show similar performances and, consequently, identify the groups that perform below the desired level. The advantage of this approach over the methods commonly used in performance evaluation is that it allows to scan multiple KPIs at once, not requiring, in a first instance, a manual analysis.

Furthermore, this thesis also aims to identify which cell configurations are associated with a better performance.

In order to fulfill the first objective, a methodology based on the application of clustering algorithms to features extracted from the original KPIs was developed. The following algorithms were tested: K-means, Expectation-Maximization using Gaussian Mixture Models, and Spectral Clustering. The selection of the optimal input parameters of each of the algorithms is based on a voting mechanism that used several internal clustering validity metrics.

Regarding the second objective, Fisher's exact test was used. This test evaluates the independence between the configuration values of the cells and the groups to which they belong.

Using this methodology it was verified that there is not a significant difference in the results obtained using the different algorithms. In the majority of the cases presented, only two groups of cells were identified: one group consisting essentially of the cells with the best performance and the other group containing the worst performing cells.

As far as the connection between configuration data and performance data is concerned, only one case, referring to a parameter associated with the subscription capacity of the cells, was detected.

Keywords: LTE, Mobile Networks, Machine Learning, Clustering, KPIs.

Resumo

As redes móveis têm verificado um crescimento acentuado nos últimos anos, tanto em número de utilizadores como na própria complexidade da rede. Como tal, realizar uma gestão eficiente da rede, de modo a garantir que esta tenha um desempenho desejado e consistente, tem-se tornado uma tarefa cada vez mais árdua. Mais ainda, face ao crescimento de quantidade de dados disponíveis, nomeadamente *Key Performance Indicators* (KPI), torna-se incomportável fazer essa gestão utilizando os métodos usuais.

Esta tese foca-se na avaliação do desempenho de células numa rede LTE através da utilização de técnicas de aprendizagem automática não supervisionada. O objetivo principal é detetar grupos de células que apresentem desempenhos semelhantes e, consequentemente, identificar os grupos que apresentam um desempenho abaixo do desejado. A vantagem desta abordagem relativamente aos métodos habitualmente utilizados na avaliação de desempenho é que permite fazer um varrimento sobre vários KPIs de uma só vez, não requerendo uma análise manual, numa primeira instância.

Após identificar os diferentes grupos de células, pretende-se identificar quais são as configurações das células que estão associadas a um melhor desempenho.

De modo a cumprir o primeiro objetivo, foi desenvolvida uma metodologia baseada na aplicação de algoritmos de *clustering* a dados extraídos a partir de KPIs. Foram testados os seguintes algoritmos: *K-means*, *Expectation-Maximization* utilizando *Gaussian Mixture Models*, e *Spectral Clustering*. A seleção dos parâmetros de entrada ótimos de cada um dos algoritmos é baseada num mecanismo de votação que utiliza diversas métricas de avaliação dos *clusters* de células obtidos.

Relativamente ao segundo objetivo, é utilizado o teste exato de Fisher, que avalia a independência entre os valores de configuração das células e os grupos a que estas pertencem.

Utilizando esta metodologia verificou-se que não existe uma diferença significativa nos resultados obtidos utilizando os diferentes algoritmos. Na maioria dos casos apresentados apenas foram identificados dois grupos de células: um grupo que é essencialmente constituído por células que apresentam o melhor desempenho e o outro grupo que é constituído pelas células com pior desempenho.

No que diz respeito à ligação entre dados de configuração e dados de desempenho, apenas um caso, referente a um parâmetro associado à capacidade de subscrição das células, foi detetado.

Palavras-chave: LTE, Redes Móveis, Aprendizagem Automática, Clustering, KPIs.

Contents

List of Tables

List of Figures

Acronyms

3GPP Third Generation Partnership Project

ARQ Automatic Repeat Request

AuC Authentication Centre

BCH Broadcast Channel

C-RNTI Cell Radio Network Temporary Identifier

CB Contention Based

CFI Control Format Indicator

CH Calinski-Harabasz

CLPC Closed Loop Power Control

CM Configuration Management

CQI Channel Quality Indicator

CS Circuit Switched

CVI Clustering Validity Index

DB Davies-Bouldin Index

DFT Discrete Fourier Transform

DL-SCH Downlink Shared Channel

DTW Dynamic Time Warping

E-RAB E-UTRAN Radio Access Bearer

E-SMLC Evolved Serving Mobile Location Centre

E-UTRA Evolved UMTS Terrestrial Radio Access

E-UTRAN Evolved UMTS Terrestrial Radio Access Network

ECDF Empirical Cumulative Distribution Function

EM Expectation-Maximization

eNodeB Evolved Node Base Station

EPC Evolved Packet Core

EPS Evolved Packet System

FDD Frequency Division Duplex

FDMA Frequency Division Multiple Access

FFT Fast Fourier Transform

FTP File Transfer Protocol

GERAN GSM/EDGE Radio Access Network

GMLC Gateway Mobile Location Centre

GMM Gaussian Mixture Models

GTP GPRS Tunneling Protocol

HARQ Hybrid Automatic Repeat Request

HSS Home Subscriber Server

IDFT Inverse Discrete Fourier Transform

IEEE Institute of Electrical and Electronics Engineers

IFFT Inverse Fast Fourier Transform

IP Internet Protocol

ITU International Telecommunication Union

KPI Key Performance Indicators

LCS LoCation Services

LTE Long Term Evolution

MAC Medium Access Control

MCH Multicast Channel

ML Machine Learning

MME Mobility Management Entity

NAS Non-Access Stratum

NE Network Element

NR Network Resource

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

P-GW PDN Gateway

PAR Peak-to-Average Ratio

PBCH Physical Broadcast Channel

PCH Paging Channel

PCRF Policy and Charging Rules Function

PDCCH Physical Downlink Control Channel

PDCP Packet Data Convergence Protocol

PDF Probability Density Function

PDN Packet Data Networks

PDSCH Physical Downlink Shared Channel

PDU Payload Data Unit

PLMN Public Land Mobile Network

PM Performance Management

PMCH Physical Multicast Channel

PRACH Physical Random Access Channel

PRB Physical Resource Block

PS Packet-Switched

PUCCH Physical Uplink Control Channel

PUSCH Physical Uplink Shared Channel

QoS Quality of Service

RA-RNTI Random Access Radio Network Temporary Identifier

RACH Random Access Channel

RAT Radio Access Technology

RBF Gaussian Radial Basis Function

RLC Radio Link Control

RRC Radio Resource Control

RRM Radio Resource Management

RSRP Reference Signal Received Power

RSRQ Reference Signal Received Quality

RSSI Received Signal Strength Indicator

S-GW Serving Gateway

SAE System Architecture Evolution

SAE GW SAE Gateway

SC Single Carrier

SC-FDMA Single Carrier Frequency Division Multiple Access

SINR Signal to Interference and Noise Ratio

SNE Stochastic Neighbour Embedding

SNMP Simple Network Management Protocol

SR Scheduling Request

t-SNE t-Distributed Stochastic Neighbour Embedding

TB Transport Blocks

TDMA Time Division Multiple Access

TMN Telecommunication Management Network

TTI Transmission Time Interval

UE User Equipment

UL-SCH Uplink Shared Channel

UTRAN Universal Terrestrial Radio Access Network

WCNC Wireless Communications and Networking Conference

Chapter 1

Introduction

This chapter presents the motivation that led to developed work, as well as the objectives to be accomplished. The thesis outline is also included in this chapter.

1.1 Motivation

As the complexity of mobile networks increases due to the increase of performance requirements and number of subscribers, it becomes harder for the mobile network operators not only to maintain but also to optimize the performance of those networks. As a result, mobile network operators are focusing more and more in creating tools and procedures that aim to not only assist radio engineers in the process of maintaining and optimizing the mobile networks, but also making the network itself more autonomous.

At the same time, Machine Learning (ML) and associated technologies are revolutionizing the way current systems work, by allowing machines to learn from available data and perform actions that would normally be taken by humans. Thus, mobile network operators can take advantage of these techniques to assist in the network management and automation process, increasing its efficiency and reducing costs.

The data used to evaluate the performance of a network is composed of Key Performance Indicators (KPI), which are quantifiable performance metrics. Some work related with the application of ML techniques to KPIs collected from mobile networks has already been developed, as in [1], where unsupervised learning techniques are used to automatically detect faults in a Long Term Evolution (LTE) network.

This thesis focuses on applying unsupervised learning techniques to a set of KPIs from a LTE network to assess its performance.

1.2 Objectives

The main objective of this thesis is to develop a model that is able to evaluate the performance of a LTE network, based on KPIs collected from the network. The system should apply unsupervised learning techniques in order to find groups of cells that present similar performances.

Furthermore, the system should classify those groups regarding the performance of the cells that constitute them. This classification depends on levels of performance specified by the mobile network operator.

Thus, the end purpose of this system is to find groups of cells that exhibit poor performance, serving as a tool to assist radio engineers when assessing the network performance. The key advantage when compared to traditional methods to assess the network performance is that this system should be capable of analyzing multiple KPIs at once.

Additionally, this thesis also aims to analyse the relationship between the configuration parameters and the obtained groups of cells performance.

1.3 Thesis Outline

This work is divided into six chapters. Chapter 2 gives a technical overview of LTE networks, necessary to understand the available Performance Management (PM) and Configuration Management (CM) data.

Chapter 3 presents the used PM and CM data, consisting of KPIs and configuration parameters, respectively. This chapter also presents the data preprocessing steps applied to remove any data artifact and null values.

Chapter 4 presents the methodology proposed to accomplish the outlined objectives of this work, as well as the technical ML background necessary to develop that methodology.

Chapter 5 includes the results obtained through the application of the proposed methodology in Chapter 4. It includes both the clustering results and the results regarding the connection between the configuration parameters and the overall cluster performance.

In Chapter 6, a summary of the work carried out in this thesis is presented and some conclusions are drawn. Lastly, future work is suggested.

1.4 Publications

The following scientific paper was written in the context of this work:

- “Unsupervised Learning Approach for Performance and Configuration Optimization of 4G Networks” written by R. Santos, M. Sousa, P. Vieira, M. P. Queluz and A. Rodrigues. This paper was submitted to the 2019 Institute of Electrical and Electronics Engineers (IEEE) Wireless Communications and Networking Conference (WCNC), Marrakech, Morocco, 15th-18th April 2019.

Chapter 2

LTE Background

This chapter aims to give an overview of LTE networks. In Section 2.1 a short introduction about LTE is presented. Section 2.2 provides an overview of the LTE network architecture. Section 2.3 explains the multiple access techniques used in LTE. Section 2.4 briefly describes the physical layer design. In Section 2.5 are presented the mobility aspects of LTE networks. Section 2.6 introduces concepts regarding PM and CM.

The information provided in this chapter was based on the following literature: [2, 3] in Section 2.1; [2–5] in Section 2.2; [2, 5] in Section 2.3; [2, 3, 5, 6] in Section 2.4; [2, 6, 7] in Section 2.5; [8–11] in Section 2.6.

2.1 Introduction

LTE corresponds to a set of mobile network standards developed by the Third Generation Partnership Project (3GPP). The driving forces for the development of LTE were:

- the evolution of wireline capabilities that increased data rates;
- the requirement for more wireless capacity;
- the need to increase data delivery efficiency;
- the competition of other wireless technologies.

These driving forces pushed LTE to meet, among others, the following performance targets: increase the peak user throughput by a factor of 10; decrease the latency by a factor of 2-3; increase the spectral efficiency by a factor of 2-4.

In order to achieve those goals, LTE used some technologies such as: Orthogonal Frequency Division Multiple Access (OFDMA) for the transmission in the downlink direction, Single Carrier Frequency Division Multiple Access (SC-FDMA) for the transmission in the uplink direction and Packet-Switched (PS) radio interface. These technologies are explained in detail in the following subsections.

In LTE, the transmission bandwidth can be selected from 1.4 MHz up to 20 MHz, depending on the available spectrum.

2.2 Network Architecture

LTE is designed to only support PS services in contrast to the Circuit Switched (CS) model of previous systems. The combination of the radio access component, LTE, and the non radio access component, System Architecture Evolution (SAE) which includes the Evolved Packet Core (EPC) network, form the Evolved Packet System (EPS). EPS is responsible for providing the user with Internet Protocol (IP) connectivity to a Packet Data Networks (PDN) for accessing the Internet as well as running services built on top of IP.

The overall network architecture is shown in Figure 2.1. Each Network Element (NE) has a different function and different elements are interconnected through standardized interfaces. The radio access component, also known as Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), is essentially composed by only one type of nodes, the Evolved Node Base Station (eNodeB), that connects to the User Equipment (UE)s, while the EPC consists of many logical nodes.

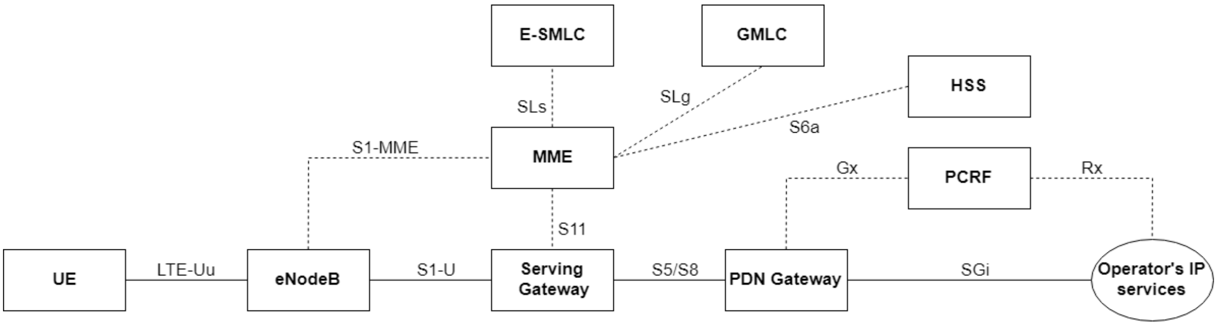


Figure 2.1: EPS network elements (adapted from [3]).

2.2.1 Core Network

The role of the core network is to control the UE and to establish the E-UTRAN Radio Access Bearer (E-RAB)s. E-RABs are used to route IP traffic, with a defined Quality of Service (QoS), between a gateway in the PDN and the UE. The PDN Gateway (P-GW), Serving Gateway (S-GW), Mobility Management Entity (MME) and Evolved Serving Mobile Location Centre (E-SMLC) are the main logical nodes of the core network. In addition to these main nodes there are other logical nodes and functions such as the Home Subscriber Server (HSS), the Gateway Mobile Location Centre (GMLC) and the Policy and Charging Rules Function (PCRF). All the nodes and functions listed are explained in more detail below:

- **P-GW** is the edge router that interconnects the EPS and external PDNs. It is responsible for traffic gating and filtering functions needed by the different services and is responsible for the allocation of an IP address to each UE. This IP allocation function can also be done by the external PDN, to which the UE is connected, and the P-GW tunnels all traffic to that network.

- **S-GW** is responsible for the management and switching of user plane data. Moreover, it serves as a mobility anchor so that the data transmission is continuous when the UE moves between different eNodeBs.

- **MME** is the main control entity in the EPC. Its main functions can be categorized into bearer management related functions and connection management related functions. The first one includes the establishment, maintenance and release of bearers while the second one is related to the connection establishment and the security of that connection.

- **E-SMLC** is responsible for managing the resources required to find the location of a UE.

- **HSS** is a database that stores user related information such as QoS, roaming restrictions and the PDNs to which the user can connect. It can also integrate the Authentication Centre (AuC) that, in turn, generates the vectors for both authentication and security keys.

- **PCRF** is the node responsible for policy control, which includes making the decisions on how each service should be handled in terms of QoS, and is also responsible for data charging functions.

- **GMLC** incorporates the required functionalities to support LoCation Services (LCS).

2.2.2 Radio Access Network

E-UTRAN is responsible for all the radio-related functions. As stated previously, this network component is only composed by eNodeBs which, in turn, are interconnected with each other through X2 interfaces and are also connected to the EPC via S1 interfaces, as shown in Figure 2.2.

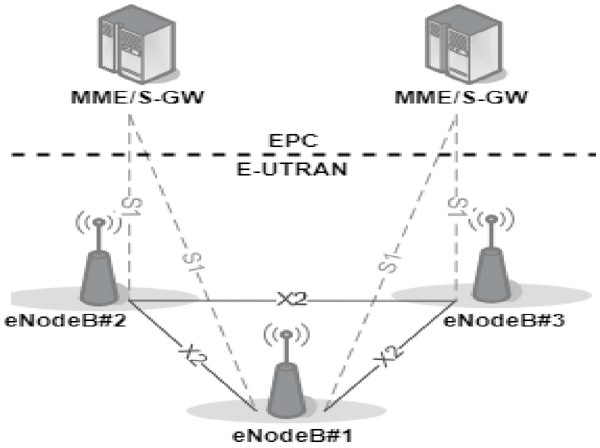


Figure 2.2: Overall E-UTRAN architecture (adapted from [3]).

The functions performed by the eNodeBs include, among others, the following:

- Radio Resource Management (RRM), i.e. managing radio bearers and radio link's resources;

- Compression and decompression of IP headers;

- Encryption of the data sent over the radio interface;

- Connectivity to the EPC;

- Handling handover between eNodeBs that are connected via X2 interfaces.

2.2.3 Radio Protocol Architecture

The protocol architecture of the radio access component in LTE can be divided into user plane architecture and control plane architecture. The role of these protocols is to set up, reconfigure and release the Radio Bearer that enables for transferring the EPS bearer.

The stack layer for both user and control plane in the radio access is shown in figure 2.3.

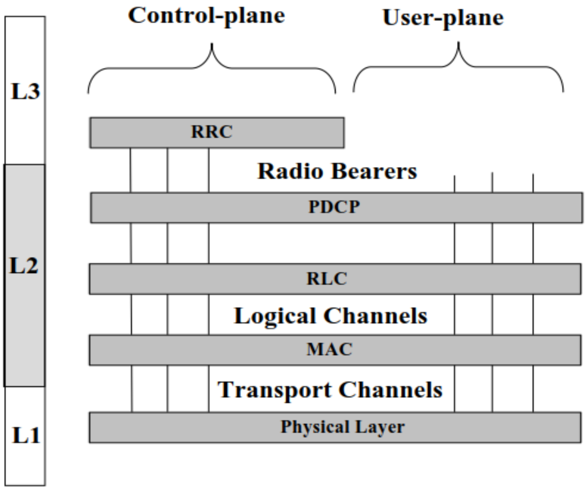


Figure 2.3: LTE Radio Protocol Stacks [2].

Both user and control plane include the Medium Access Control (MAC), Radio Link Control (RLC) and Packet Data Convergence Protocol (PDCP) layer, which are Layer 2 protocols, above the physical layer. Additionally, the control plane also includes the Radio Resource Control (RRC) protocol, which is a Layer 3 protocol. A brief description of these protocols is given below:

- **Physical layer** is responsible for carrying the information from the MAC transport channels over the air interface;

- **MAC** maps the logical channels to transport channels as presented in figure 2.3. Other functions in this layer include multiplexing/demultiplexing of RLC Payload Data Unit (PDU)s belonging to one or different radio bearers into/from Transport Blocks (TB) delivered to/from the physical layer on transport channels, perform error correction using Hybrid Automatic Repeat Request (HARQ) and handling priorities between logical nodes through dynamic scheduling;

- **RLC** is responsible for the segmentation and concatenation of the PDCP-PDUs to be transmitted and for the reassemble of RLC PDUs to reconstruct the PDCP PDUs and also performs error correction through the Automatic Repeat Request (ARQ) mechanism;

- **PDCP** main functions include IP header compression/decompression, integrity protection and ciphering/deciphering both the user plane data and most of the control plane data;

- **RRC** controls the radio resource usage. It is responsible for managing UE's signalling and data connections and also has functions related to handovers.

Additionally to the protocols related to the LTE radio interface, there are protocols between the UE and the core network that are transparent to the radio layers and are generally referred to as Non-Access Stratum (NAS) signaling.

2.3 LTE Multiple Access

LTE uses OFDMA as the downlink multiple access scheme while for the uplink it uses SC-FDMA. The principle of Single Carrier (SC) transmission is to modulate the information to only one carrier. This can be done by adjusting the carrier, with respect to the information that needs to be transmitted, in phase or amplitude or both. This principle is shown in Figure 2.4.

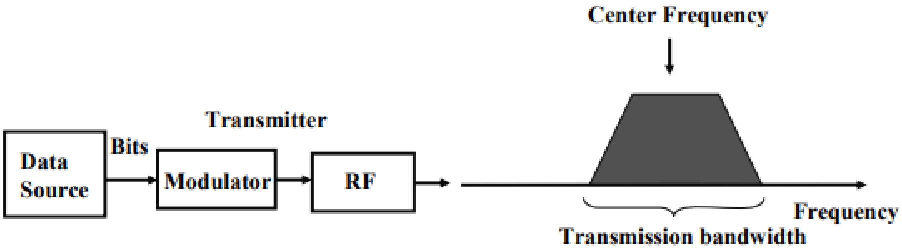


Figure 2.4: Single carrier transmitter [2].

In a Frequency Division Multiple Access (FDMA) system, different users use different carriers or sub-carriers in order to access it simultaneously. Consequently, interference between carriers may arise which leads to the use of guard bands in order to minimize that interference. However, these guard bands cannot be too extensive since that would lead to an inefficient use of the available bandwidth. The FDMA principle is presented in Figure 2.5.

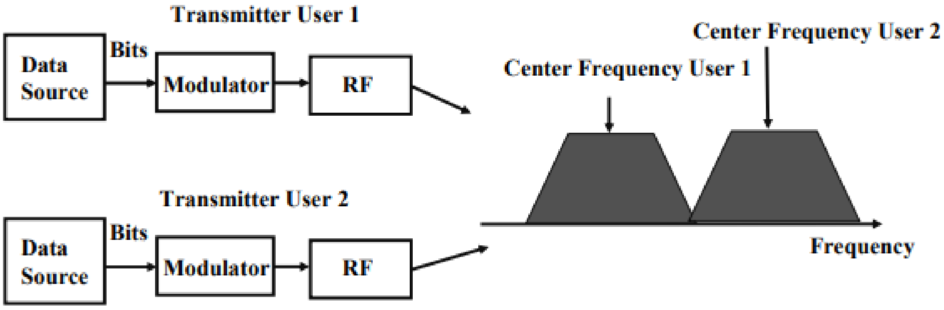


Figure 2.5: FDMA principle [2].

These techniques allow for a better use of the available bandwidth and also to reduce inter-symbol interference and fading. Thus, these multiple access schemes have a key role in achieving the performance targets defined in LTE.

2.3.1 OFDMA

In OFDMA the sub-carriers are mutually orthogonal. This means that at a sampling instant of a given sub-carrier all the others have zero value, as shown in figure 2.6. It is then intuitive to understand that such scheme allows for a better use of the available bandwidth since the sub-carriers overlap without interference. In the specific case of LTE, the sub-carrier spacing has been specified to be 15 kHz regardless of the available bandwidth.

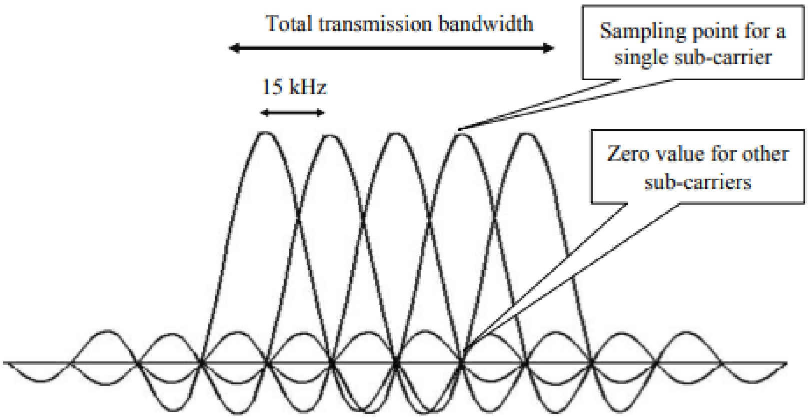

Total transmission bandwidth
Sampling point for a single sub-carrier
15 kHz
Zero value for other sub-carriers

Figure 2.6: Maintaining the sub-carriers' orthogonality [2].

The practical implementations of an OFDMA system use both the Fast Fourier Transform (FFT) and the Inverse Fast Fourier Transform (IFFT), where the first moves a time-domain signal into the frequency domain while the second does the inverse operation. These operations can be carried out back and forth as long as the sampling rate requirements of digital signal processing are fulfilled.

The transmitter of an OFDMA system has an IFFT block that transforms each sub-carrier from frequency to time domain. The IFFT block is fed by a serial-to-parallel block which has the data source as the input and the different sub-carriers as the output. Following the IFFT block it is added a cyclic extension so that inter-symbol interference is avoided. Figure 2.7 illustrates the architecture of an OFDMA system.

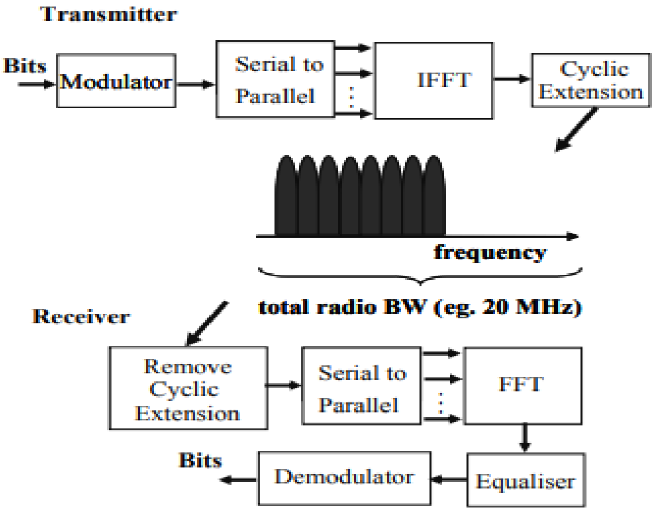


Figure 2.7: OFDMA transmitter and receiver [2].

The addition of a cyclic extension is preferable when compared to breaking the transmission (guard interval) since then the Orthogonal Frequency Division Multiplexing (OFDM) signal is periodic. The impact on the channel, due to the periodic nature of the signal, ends up corresponding to a multiplication by a scalar. Not only that, but the periodicity of the signal also allows for a discrete Fourier spectrum which, in turn, enables the use of the Discrete Fourier Transform (DFT) in the receiver and Inverse Discrete Fourier Transform (IDFT) in the transmitter.

On the receiver end, the inverse operations are applied and there is also an equalizer that reverts the channel impact for each sub-carrier.

A major aspect of using an OFDMA scheme in a base station transmitter is that it can allocate any of its sub-carriers in the frequency domain to its users. On one hand, the scheduler is able to benefit diversity in the frequency domain. On the other hand this implies a practical limitation since the signalling resolution caused by the resulting overhead prevents the allocation using single sub-carriers. Therefore, in LTE, the allocation is done using Physical Resource Block (PRB)s, each consisting of 12 sub-carriers which translates into a minimum allocated bandwidth of 180 kHz. The allocation in the time-domain is done in intervals of 1 ms, designated by Transmission Time Interval (TTI), even though each PRB lasts only 0.5 ms.

2.3.2 SC-FDMA

One major challenge of using OFDMA is the need for high linearity in the transmitter due to the large variations of the transmitted signal power that result in a high Peak-to-Average Ratio (PAR). Since SC-FDMA enables better power amplifier efficiency when compared to linear amplifiers that are required by OFDMA, then SC-FDMA is the selected multiple access scheme used in the uplink transmission because it allows to use cheaper transmitters when compared to the ones used in base stations.

SC-FDMA is similar to Time Division Multiple Access (TDMA) in the way that each symbol is sent one at a time. The generation of the signal in the frequency domain, which is shown in Figure 2.8, adds the OFDMA property of good waveform. Therefore, similarly to what happens with the OFDMA scheme used in the downlink, there is no need to use guard intervals between different users.

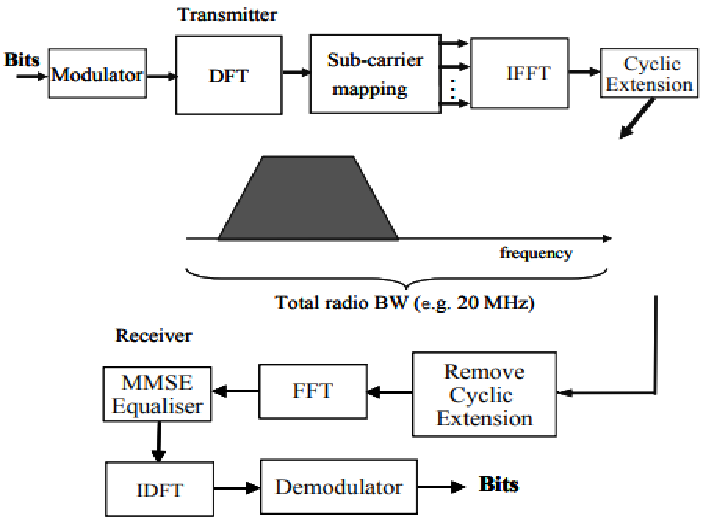


Figure 2.8: SC-FDMA transmitter and receiver with frequency domain signal generation [2].

As in the OFDMA scheme, a cyclic extension is added periodically to the signal in order to prevent inter-symbol interference and to simplify the receiver design. However, this cyclic extension is not added after each symbol because the symbol rate is faster in SC-FDMA than in OFDMA. Thus, the cyclic extension only prevents inter-symbol interference between blocks of symbols and inter-symbol interference between symbols of the same block still needs to be taken into account. In the receiver, each block of symbols is handled by the equalizer until reaching the cyclic prefix so that the further propagation of inter-symbol interference is prevented.

In LTE, the resolution allocation rate for the SC-FDMA system is 1 ms, which is the same value as for the OFDMA system, and the transmission occupies the whole part of the user's allocated spectrum. Moreover, the SC-FDMA system uses the same values for the sub-carrier spacing and for the resource blocks bandwidth, these values being 15 kHz and 180 kHz, respectively. It is worth noticing that even though the transmission is, by name, a single carrier, the signal generation phase uses a sub-carrier term.

Since only a single modulation symbol is transmitted at a time, then a low transmitter waveform is ensured and the modulation method used highly influences the waveform characteristics. Consequently, using SC-FDMA, a low PAR can be achieved and power amplifiers with low power consumption and good power conversion efficiency can be used.

The receiver located in the base station that is used in the SC-FDMA system is more complex than the receiver used in the OFDMA system due to the fact that in SC-FDMA the inter-symbol interference is terminated only after a block symbols while in OFDMA the inter-symbol interference is terminated after every symbol. However, this increased complexity in the receiver using SC-FDMA is outweighed by the benefits of the uplink range and better device battery life. Additionally, the dynamic resource usage with a 1 ms resolution guarantees that there is no base-band receiver per UE on standby and enables the base station to be used in a dynamic fashion by users that have data to transmit.

2.4 Physical Layer Design

The physical layer of LTE is designed in order to maximize the efficiency of packet-based transmission, meaning that the channels are shared, enabling dynamic resource utilization, instead of having dedicated resources reserved for each user. The deployment of the physical layer in a radio access system plays an essential role on the resulting system's capacity and, consequently, is a relevant point of comparison on the expected performance of different systems.

2.4.1 Transport Channels

Transport channels are responsible for connecting the physical layer to the MAC layer and they are briefly described in the following points:

- **Uplink Shared Channel (UL-SCH)** carries user data and control messages in the uplink direction;
- **Random Access Channel (RACH)** acts on the uplink direction and enables the mobile to contact the network without prior scheduling;

- **Downlink Shared Channel (DL-SCH)** carries user data and control messages in the downlink direction;

- **Paging Channel (PCH)** transports paging messages in the downlink direction;

- **Broadcast Channel (BCH)** broadcasts, in the downlink direction, information required for the devices to access the system and to identify the operator;

- **Multicast Channel (MCH)** is used in the downlink direction for carrying multicast service content to the UE.

The mapping between the transport channels described above and the corresponding physical channels is the following:

In the uplink direction, the UL-SCH and RACH are mapped to the Physical Uplink Shared Channel (PUSCH) and Physical Random Access Channel (PRACH), respectively.

In the downlink direction, both the PCH and the DL-SCH are mapped to the Physical Downlink Shared Channel (PDSCH) while the BCH and MCH are respectively mapped to the Physical Broadcast Channel (PBCH) and Physical Multicast Channel (PMCH).

2.4.2 Uplink User Data Transmission

The PUSCH carries the user data in the uplink direction. It has a 10 ms frame structure and is based on the allocation of frequency and time resources with 1 ms and 180 kHz resolution. The resource allocation is performed by a scheduler located in the eNodeB, as shown in Figure 2.9. Only random access resources may be used if there is no prior signaling from the eNodeB and since resource allocation is done in a dynamic fashion, there are no fixed resources for the devices. Thus, the UE needs to provide information to the eNodeB of both the its transmission requirements and its available transmission power resources.

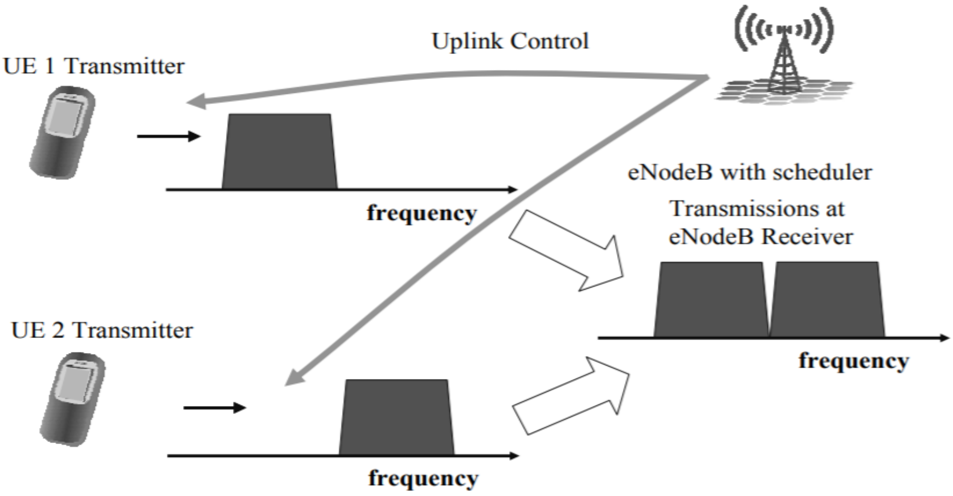


Figure 2.9: Uplink resource allocation controlled by eNodeB [2].

The 10 ms frame structure can be divided into subframes of 1 ms each, which constitutes the allocation period. Each subframe can be further divided into slots of 0.5 ms each. Within the 0.5 ms slot there are both user data symbols and reference symbols, in addition to the signalling. The 10 ms frame structure is illustrated in Figure 2.10.

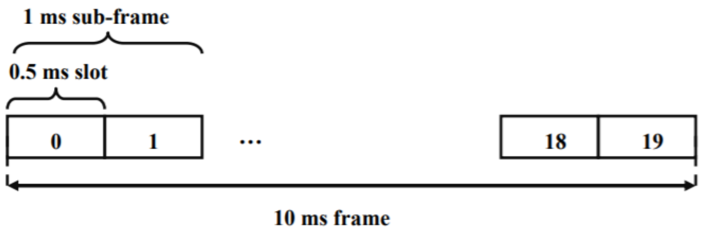


Figure 2.10: LTE Frequency Division Duplex (FDD) frame structure [2].

In the frequency domain, the bandwidth is allocated between 0 and 20 MHz in the steps of 180 kHz. Since the uplink transmission is FDMA modulated with only one symbol being transmitted at a time, then the allocation is continuous. The slot bandwidth adjustment between consecutive TTIs is shown in Figure 2.11, where it can be observed that doubling the data rate results in doubling the bandwidth being used. Moreover, it can be seen that a higher data rate results in a corresponding increase for the reference symbol data rate since the reference symbols occupy the same space in the time domain.

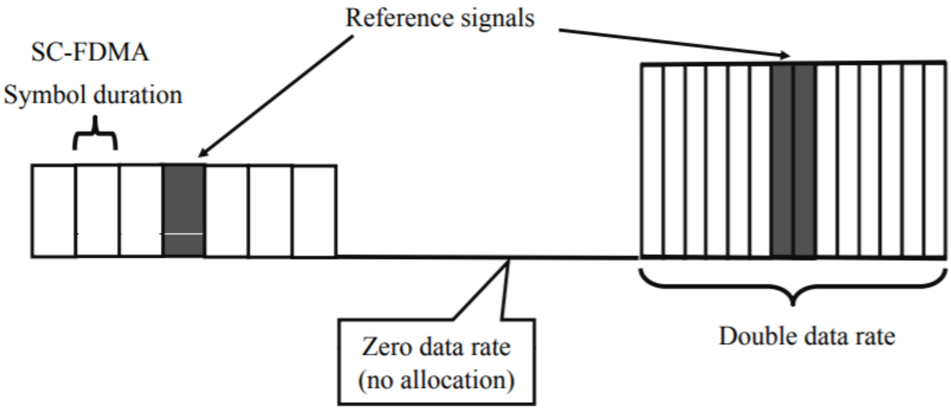


Figure 2.11: Data rate between TTIs in the uplink direction [2].

In the uplink, the cyclic prefix may take two values depending on whether a short or extended cyclic prefix is used. If an extended cyclic prefix is applied then the data payload is reduced and only 6 symbols are transmitted per slot, instead of the 7 symbols transmitted with the short cyclic prefix, so this is not frequently used because the benefits of having more payload are greater than possible degradation caused by inter-symbol interference when the channel delay spread is higher than the cyclic prefix.

2.4.3 Downlink User Data Transmission

The PDSCH carries the user data in the downlink direction. Similarly to what happens in the uplink direction, the resources are allocated, in the frequency domain, in blocks of 180 kHz. The user data rate is dependent on the number of allocated sub-carriers for a given user as the multiple scheme is OFDMA which translates into each sub-carrier being transmitted as parallel 15 kHz sub-carrier. The eNodeB allocates the resources based on the Channel Quality Indicator (CQI) specified by the UE. The resource allocation is done in both the time and the frequency domain, as in the uplink direction, and is illustrated in Figure 2.12.

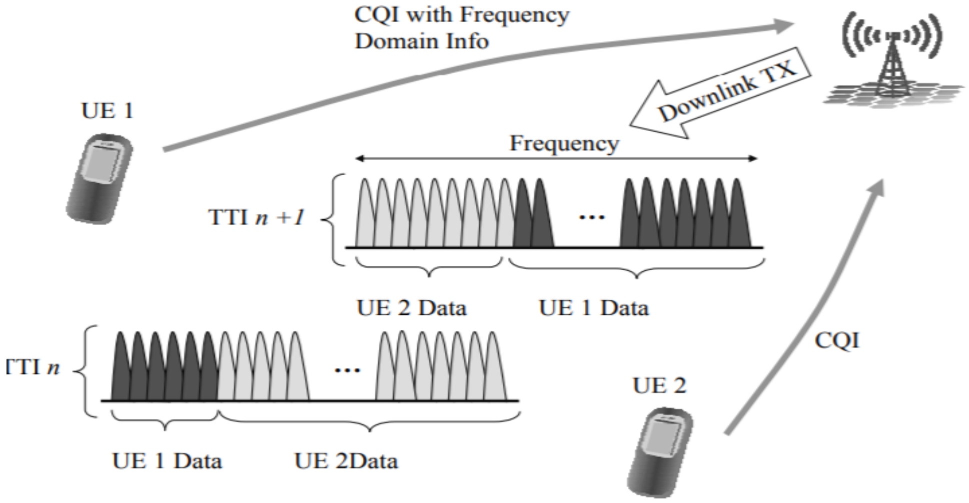


Figure 2.12: Downlink resource allocation at eNodeB [2].

Each UE is informed about the corresponding allocated resources through the Physical Downlink Control Channel (PDCCH) and, once again, the resource allocation is done dynamically with a 1 ms granularity. PDSCH data can occupy from 3 up to 6 symbols in each 0.5 ms slot, depending on the type of cyclic prefix used (short or extended) and also on the allocation for the PDCCH. Within each 1 ms subframe, only the first 0.5 ms slot contains control symbols PDCCH while the second 0.5 ms slot is used solely to transmit data symbols (PDSCH). A short cyclic prefix allows to fit 7 symbols in each 0.5 ms slots, while with an extended cyclic prefix each 0.5 ms accommodates 6 symbols. Figure 2.13 illustrates the downlink slot structure assuming that 3 symbols are used for PDCCH. Besides the control symbols, also the reference symbols, synchronization signals and broadcast reduce the space available for the user data.

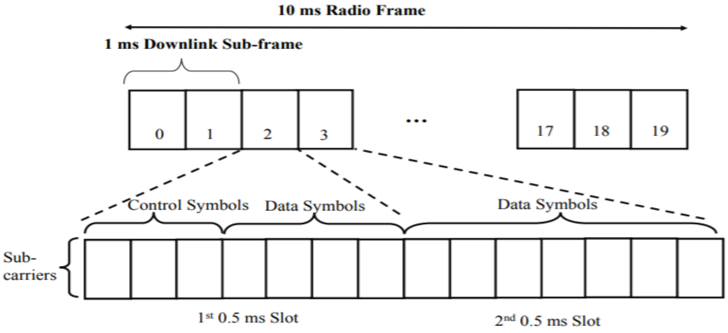


Figure 2.13: Downlink slot structure for bandwidths above 1.4 MHz [2].

2.5 Mobility

In this section is presented a brief overview of how mobility management is performed in LTE. Mobility is a key aspect in LTE, and in mobile networks in general, because in order to provide ubiquitous coverage to different users it is required that they are able to access and maintain the services as they move across the network coverage area. Such coverage comes with the cost of increased network complexity. Thus, LTE aims to provide seamless mobility but taking into account that complexity is an important factor and must be minimized.

There are two possible states that a UE can take: RRC_IDLE and RRC_CONNECTED. The mobility management procedures are then chosen with respect to the UE state. The RRC_IDLE state corresponds to the UE being switched on but not connected to the network while the RRC_CONNECTED state corresponds to when there is a connection between the network and the UE. In the former the mobility procedures are triggered by the UE and are related to cell re-selection according to parameters sent by the network; in the latter it is the E-UTRAN that decides, according to the reports sent by the UE, whether or not to perform an handover.

In order to provide user mobility in a LTE network, the following measurements are performed:

- **Reference Signal Received Power (RSRP)** corresponds, for a given cell, to the average power measured per resource element that contains cell-specific reference signals;
- **Reference Signal Received Quality (RSRQ)** is the ratio of the RSRP and the Evolved UMTS Terrestrial Radio Access (E-UTRA) Carrier Received Signal Strength Indicator (RSSI), for the reference signals;
- **RSSI** is the total received wideband power on a specific frequency.

2.5.1 Idle Mode Mobility

In Idle mode, the cell selection is done by the UE and is based on radio measurements that the UE itself performs. When a suitable cell is selected by the UE it is said that the UE is camped in that cell. In order for a cell to be considered a suitable candidate it must have good radio quality and not be blacklisted. More specifically, it is required that the cell fulfills the S-criterion:

$$S_{rxlevel} > 0, \tag{2.1}$$

where

$$S_{rxlevel} > Q_{rxlevelmeas} - (Q_{rxlevmin} - Q_{rxlevelminoffset}) \tag{2.2}$$

The $S_{rxlevel}$ is the Rx level value of the cell, the $Q_{rxlevelmeas}$ is the measured cell received level RSRP, $Q_{rxlevelmin}$ is the minimum required received level, measured in dBm, and the $Q_{rxlevelminoffset}$ is an offset used when searching for a higher priority Public Land Mobile Network (PLMN).

The UE continuously tries to find better cells as candidates for reselection, according to the reselection criteria, even after being camped on a cell. Some cells may be blacklisted by the network which means that those cells cannot be considered by the UE for the reselection process. In order to reduce the reselection measurements, the UE considers that if $S_{ServingCell}$, which corresponds to the Rx level value of the serving cell, is high enough then it does not need to carry out further intra-frequency, inter-frequency or inter-system measurements. When $S_{ServingCell} \leq S_{intrasearch}$ or $S_{ServingCell} \leq S_{nonintrasearch}$, the UE starts performing measurements for intra-frequency or inter-frequency reselection, respectively. The $S_{intrasearch}$ corresponds to the serving cell's Rx level threshold for the UE to start making intra-frequency measurements while the $S_{nonintrasearch}$ corresponds to the serving cell's Rx level threshold for the UE to start making inter-frequency measurements. Both intra-frequency reselection and equal priority E-UTRAN frequency reselection are based on the R-criterion which aims to find the best cell for the UE to camp on. The R-criterion comprehends both the serving cell ranking (R_s) and the neighbouring cell's ranking (R_n) where the former ranks the serving cell while the latter ranks the different neighbouring cells:

$$R_s = Q_{meas,s} + Q_{hyst} \tag{2.3}$$

$$R_n = Q_{meas,n} + Q_{offset}, \tag{2.4}$$

where Q_{meas} is the RSRP measurement, the Q_{hyst} is the power domain hysteresis to avoid ping-pong between cells and Q_{offset} is an offset that controls different frequency specific characteristics.

In order to restrain the amount of reselections that are performed it is used the parameter $T_{reselection}$. If the best ranked neighbor cell is better ranked than the serving cell for a period of time longer than $T_{reselection}$ then, the reselection occurs.

A method known as absolute priority based reselection is used in LTE, allowing the operators to control how UE prioritizes camping on different Radio Access Technology (RAT)s or frequencies of E-UTRAN. It is assigned a priority to each layer (different RAT/frequency) and the UE tries to camp on the highest priority layer, as long as it can provide decent service. This means that a threshold, $Thresh_{high}$, needs to be fulfilled for a period of time longer than $T_{reselection}$ before reselection is performed. Furthermore, a reselection to a lower priority is only performed if the higher priority layer drops below the threshold $Thresh_{high}$ and the lower priority layer rises above the $Thresh_{low}$.

2.5.2 Connected Mode Mobility

When the UE is in the RRC_CONNECTED state its mobility is controlled by the network, with the decision taken by the E-UTRAN to perform an handover being based on measurements carried out by the UE. Those measurements and their reporting are controlled by parameters given by the E-UTRAN. Since the handovers are targeted to be lossless, packet forwarding between the source and the target eNodeB is used. Once the handover is completed the core network S1 connection is updated. This is also known as Late path switch. The core network has no influence over the handovers. Figure 2.14 shows an overview of the intra-frequency handover procedure.

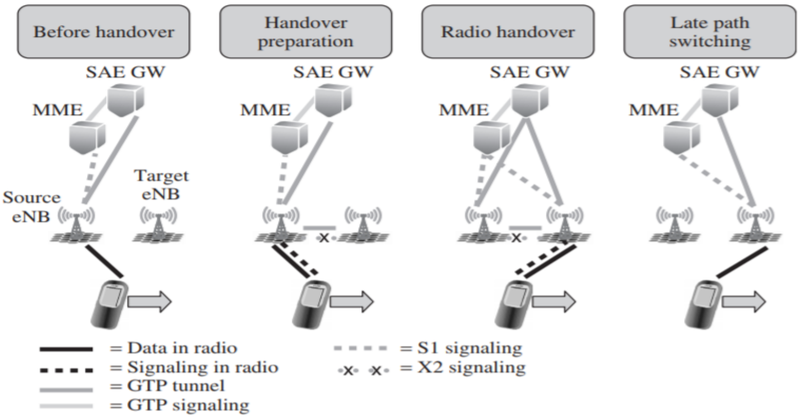


Figure 2.14: Intra-LTE handover procedure [7].

Firstly, it is illustrated that the UE is moving from left to right and has a user plane connection to the source eNodeB and further to the SAE Gateway (SAE GW). Moreover, the S1 connection exists between the source eNodeB and the MME. As the UE moves closer to the other eNodeB shown in the figure, the measurements relative to that target cell get closer to a reporting threshold, until they eventually fulfill that threshold. Once the threshold is fulfilled, the UE sends the measurement report to the source eNodeB. The source eNodeB then establishes both the signalling connection and the GPRS Tunneling Protocol (GTP) tunnel to the target cell. The source eNodeB sends the handover command to the UE when the target eNodeB has the resources available and then the UE can switch the radio connection to the target eNodeB. Once that connection is established, the core network connection is updated.

Regarding connected mode mobility, LTE also supports handovers to other RATs, namely Universal Terrestrial Radio Access Network (UTRAN), GSM/EDGE Radio Access Network (GERAN) and cdma2000® [2].

2.6 Performance Data Collection

The increasing complexity of telecommunication networks results in major challenges to telecommunication operators in both monitoring and managing the performance of those networks. In order to face those challenges, network operators use a set of methods that allows to collect data originated from the networks. These data provides insights about the network and thus can be used to monitor, plan and optimize the network.

2.6.1 Performance Management

International Telecommunication Union (ITU) developed Telecommunication Management Network (TMN) which is the framework used to manage telecommunications networks and services. In each layer of the TMN reference model, five different management functional areas are considered: Fault, Configuration, Accounting, Performance and Security. PM consists in evaluating and reporting the behavior and effectiveness of network elements by gathering statistical information, maintaining and examining historical logs, determining system performance and altering the system modes of operation. With that, the network use can be optimized, allowing increased Quality of Service (QoS) for the end users applications. PM has a key role for network operators because it allows them to detect the deteriorating trend in advance and thus solve potential threats and prevent faults.

The architecture of a PM system is divided in the following layers:

- **Data Collection and Parsing Layer:** where data is collected from the NEs using network specific protocols, like File Transfer Protocol (FTP) or Simple Network Management Protocol (SNMP);
- **Data Storage and Management Layer:** where the data coming from the Data Collection and Parsing Layer is stored into a database;
- **Application Layer:** is responsible for the processing of collected and stored data. It is also responsible for storing and sharing of generated KPIs and reports;
- **Presentation Layer:** provides a web-based user interface that shows the generated PM results, in the form of dashboards, charts and real-time graphs.

The main challenges faced in PM are related to the high volume of performance measurements data that is collected over time periods which rises difficulties in making an efficient administration, and to the fact that certain performance measurements don't have an unified structure or content thus creating difficulties in handling those measurements.

2.6.2 Key Performance Indicators

Generally speaking, KPIs are measures of the performance of essential operations and/or processes in an organization. In the context of telecommunications, KPIs are measures of the performance of a network. These KPIs are obtained through statistical calculations based on counters installed on the NEs, that register many indicators such as dropped calls, failed handovers or handover types. KPIs are fundamental in the context of PM since they provide valuable information about the network performance which can be used not only to identify performance gaps between current and desired performance but also to provide indication concerning the progress in closing those gaps.

Telecommunication specific KPIs can be divided into different categories, based on the measurement targets. Usually these categories are the following: Accessibility, Retainability, Integrity, Availability and Mobility . However, this is not standardized, therefore vendors may define additional categories such as Utilization [10]. A brief description and a KPI example for each category are provided in Table 2.1.

Categories	Description	Examples
Accessibility	Indicates if services requested by a user can be accessed within specified tolerances in the given operating conditions	Random Access Success Rate
Availability	Indicates the percentage of time that a cell is available	Cell Availability
Integrity	Indicates the E-UTRAN impacts on the service quality provided to the end-user	Downlink and Uplink Throughputs
Mobility	Evaluates the performance of E-UTRAN mobility	Intra-frequency Handover Out Success Rate
Retainability	Evaluates the network capability to retain a connection from its initiation until its disconnection by the user	Call Drop Rate
Utilization	Evaluates the network capability to meet the traffic demand	Resource Block Utilizing Rate

Table 2.1: KPI categories description.

2.6.3 Configuration Management

CM allows the operator to assure correct and effective operation of the network as it evolves. CM actions purpose is to control and monitor the active configuration of the NEs and Network Resource (NR)s which can be initiated by the operator or by functions in the Operation Systems or NEs. These actions may be carried out as part of implementation programmes, optimisation programmes and to maintain the overall QoS [11].

CM Service Components

When a network is installed and brought into service and following its installation, the network operator needs to enhance and adapt the network so that short and long term requirements are met and customer needs are satisfied. Thus, the network operator should be provided with a set of capabilities, such as initial system installation, system operation to adapt the system to short term requirements, system update to overcome software bugs or equipment faults and system upgrade to enhance or extend the network by features or equipment respectively. Such capabilities are provided by the CM system through its service components – system modification and system monitoring.

The system modification service component is used to adapt the system data to a new requirement due to optimization or new network configurations while the system monitoring service component allows the operator to receive reports, from managed NEs, on the configuration of the entire network, or parts of it.

CM Functions

Due to the requirements of CM and their usage, some basic system modification functions need to be defined for the network: creation of NEs and NRs; deletion of NEs and NRs; conditioning of NEs and NRs. To each of these functions, the following requirements apply:

- affected resources should be taken out of service only if needed, resulting in minimal network disturbance;
- physical modifications and the related logical modifications should be independent;
- in order to bring resources into service all the actions needed to satisfy a certain task should be completed correctly;
- data consistency checks should be performed.

Chapter 3

PM and CM Data

This chapter focuses on both the PM and CM data used in this thesis. This data is from a live network, deployed in a urban environment, and was provided by a mobile network operator.

For both datasets, a description of the most relevant features is provided and, in addition, each KPI in the PM dataset is classified into one of the classes considered in Section 2.6. Since the available data was originated from eNodeBs deployed with Ericsson equipment, both the PM and CM feature description and KPIs classification of each KPI are based on the available information provided in [12].

Each site may support from one up to three of the following frequency bands: L800 (800 MHz), L1800 (1800 MHz) and L2600 (2600 MHz).

This chapter presents the PM data in Section 3.1 and the CM data in Section 3.2.

3.1 PM Data

For each one of the analyzed cells, deployed in a site, the PM data was collected every 15 minutes, during 10 days, consisting of multiple KPIs and counters. Thus, for each cell, each PM data feature corresponds to a time-series.

The first step towards achieving the goals of this work is to understand what each feature in the dataset represents and, from there, select which KPIs are more relevant to analyse the network performance and classify each one of them into one of the classes defined in Section 2.6.2.

3.1.1 KPIs Selection and Description

The KPIs selected to evaluate the performance of each cell in the network are presented in Table 3.1, according to the KPI class to which they belong.

Accessibility	Integrity	Availability
CB_RACH_fail%	DL_Throughput_per_UE(Mbps)	CellAvail_perc
Added_E_RAB_Estab_fail%	UL_Throughput_per_UE(Mbps)	CellAvailAuto_perc
Init_E_RAB_Estab_fail%	DL_Pdcp_Cell_Tput(Mbps)	CellAvailMan_perc
RRC_Estab_fail%	DL_MAC_Cell_Tput(Mbps)	
S1_Estab_fail%	UL_Pdcp_Cell_Tput(Mbps)	
	UL_MAC_Cell_Tput(Mbps)	

Table 3.1: Selected KPIs.

Accessibility

The Init_ERAB_Estab_fail% indicates the E-RABs fail rate for end-user services that are carried by E-RABs included in the Initial UE Context setup procedure. Figure 3.1 illustrates the Initial UE Context setup procedure, where the attempts to establish the initial E-RAB are counted in point A while the failed established initial E-RABs are counted in point B.

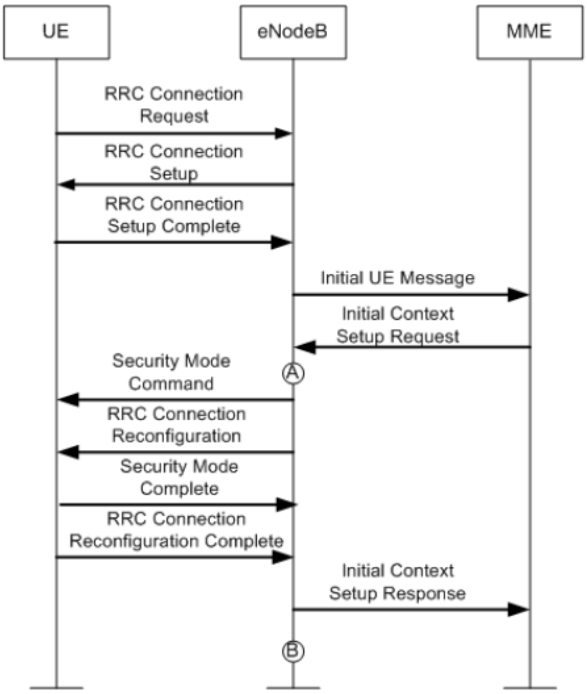

UE
eNodeB
MME
RRC Connection Request
RRC Connection Setup
RRC Connection Setup Complete
Initial UE Message
Initial Context Setup Request
Security Mode Command
A
RRC Connection Reconfiguration
Security Mode Complete
RRC Connection Reconfiguration Complete
Initial Context Setup Response
B

Figure 3.1: Initial UE context setup procedure [10].

When a UE requests a service that requires an improved level of QoS, the network decides if it establishes a new E-RAB (added E-RAB) for that service or if it modifies the QoS of an existing E-RAB. The Added_ERAB_Estab_fail% corresponds to the fail rate for end-user services that are carried by E-RABs included in the E-RAB setup procedure.

The RRC_estab_fail% measures the fail rate regarding the RRC connections establishment and the S1_Estab_fail% gives the fail rate for the establishment of signaling connections over the S1 interface.

Lastly, the CB_RACH_fail% indicates how often, in a Contention Based (CB) RACH procedure, a transmitted RaMsg2 does not result in a successfully received RaMsg3. The contention based RACH procedure can be observed in Figure 3.2 and it is performed by a UE in order to be synchronized with the network. Each cell has 64 preambles reserved, from which a fraction is reserved for the non-contention based procedure and the remaining are reserved for the contention based procedure. In the contention based procedure, a UE selects, randomly, a preamble sequence from the ones that are reserved for this procedure and transmits it on PRACH to the eNodeB. Then, a preamble response is sent to the UE on the DL-SCH resource assigned on PDCCH through the Random Access Radio Network Temporary Identifier (RA-RNTI), which is derived by the eNodeB through the time slot in which it receives the preamble. This response message carries, among others, information regarding the synchronization for uplink transmission and a temporary Cell Radio Network Temporary Identifier (C-RNTI) which identifies the UE [3].

Since the preamble sequence is selected randomly, a collision may happen, if two or more UEs select the same preamble and transmit it simultaneously. Thereupon, those UEs will receive the same C-RNTI and will also transmit the RaMsg3 on the same time-frequency resources, which may result in the eNodeB not being able to decode the message due to the resulting interference, meaning that the RaMsg3 is not successfully received.

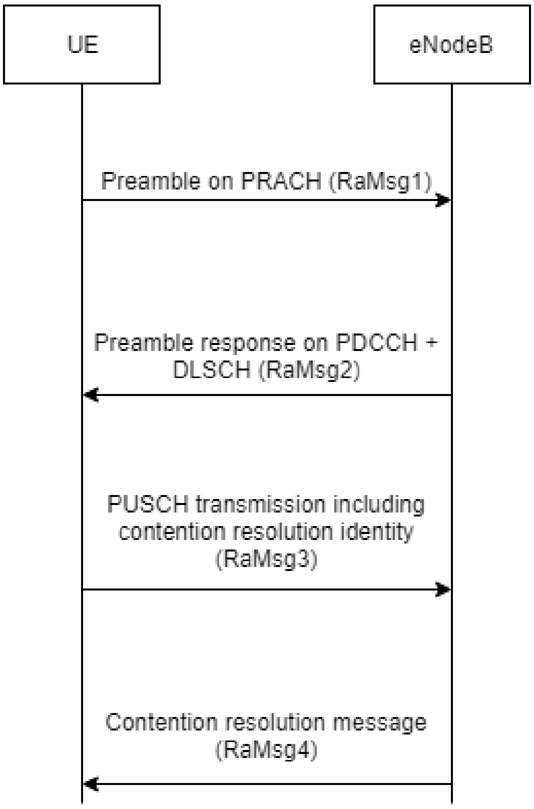

UE
eNodeB
Preamble on PRACH (RaMsg1)
Preamble response on PDCCH + DLSCH (RaMsg2)
PUSCH transmission including contention resolution identity (RaMsg3)
Contention resolution message (RaMsg4)

Figure 3.2: Contention based random access procedure (adapted from [2]).

Availability

The CellAvailAuto_perc provides the percentage of time that a given cell is available with respect to the time that has been disabled due to a fault while the CellAvailMan_perc provides the percentage of time that a given cell is available with respect to the time that has been disabled due to a reconfiguration request performed by the operator. The CellAvail_perc provides the overall percentage of time that a given cell is available.

Integrity

Regarding Integrity, DL_Throughput_per_UE and UL_Throughput_per_UE give, respectively, the average throughput measures per user in the downlink and uplink direction. The Pdcp_Cell_Tput and the MAC_Cell_Tput provide the average cell throughput with respect to the PDCP layer and the MAC layer, respectively. Once again, the DL and UL tags refer to downlink and uplink, respectively.

3.1.2 PM Data Preprocessing

Since the PM data was collected through measurements from multiple NEs in a live network, it is expected that it contains both missing values and noise. Consequently, the available data must go through a cleaning process before applying any clustering algorithm.

The preprocessing for the PM data consisted in the following steps:

1. Verify for missing values;
2. Verify for unexpected negative values;
3. Verify for unexpected big values.

Figure 3.3 shows that the dataset does not have missing data and, consequently, there is no need for preprocessing regarding the missing values.

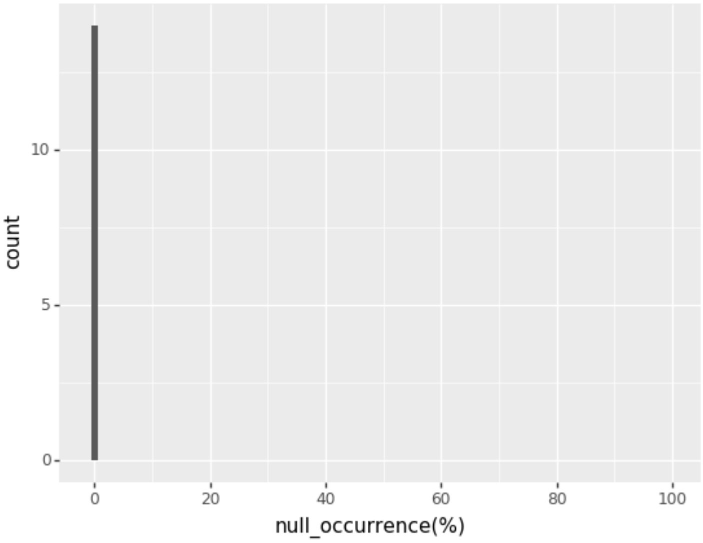


Figure 3.3: Frequency of KPI null occurrence percentage.

Given the KPIs presented in Section 3.1.1 it is quite straightforward to understand that their values should always be positive. Thus, it was verified if there were any KPIs that presented negative values.

Figure 3.4 illustrates the obtained percentages of negative values per KPI and it can be verified that there are three KPIs that have negative values, which are presented in Table 3.2.

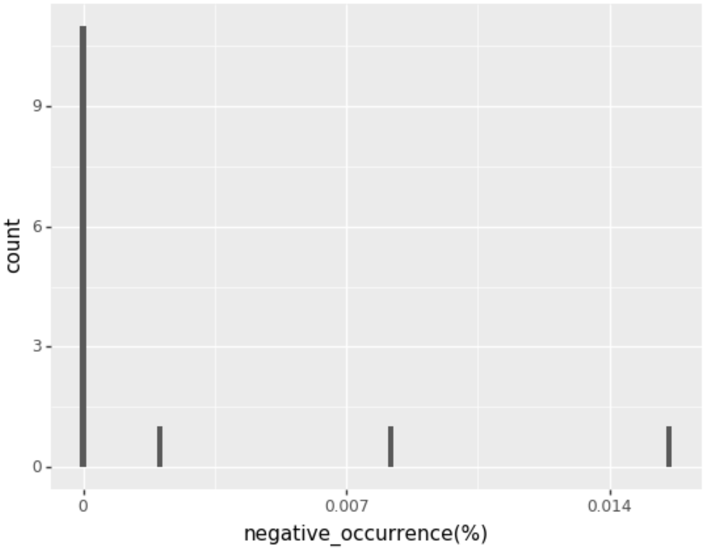


Figure 3.4: Frequency of KPI negative occurrence percentage.

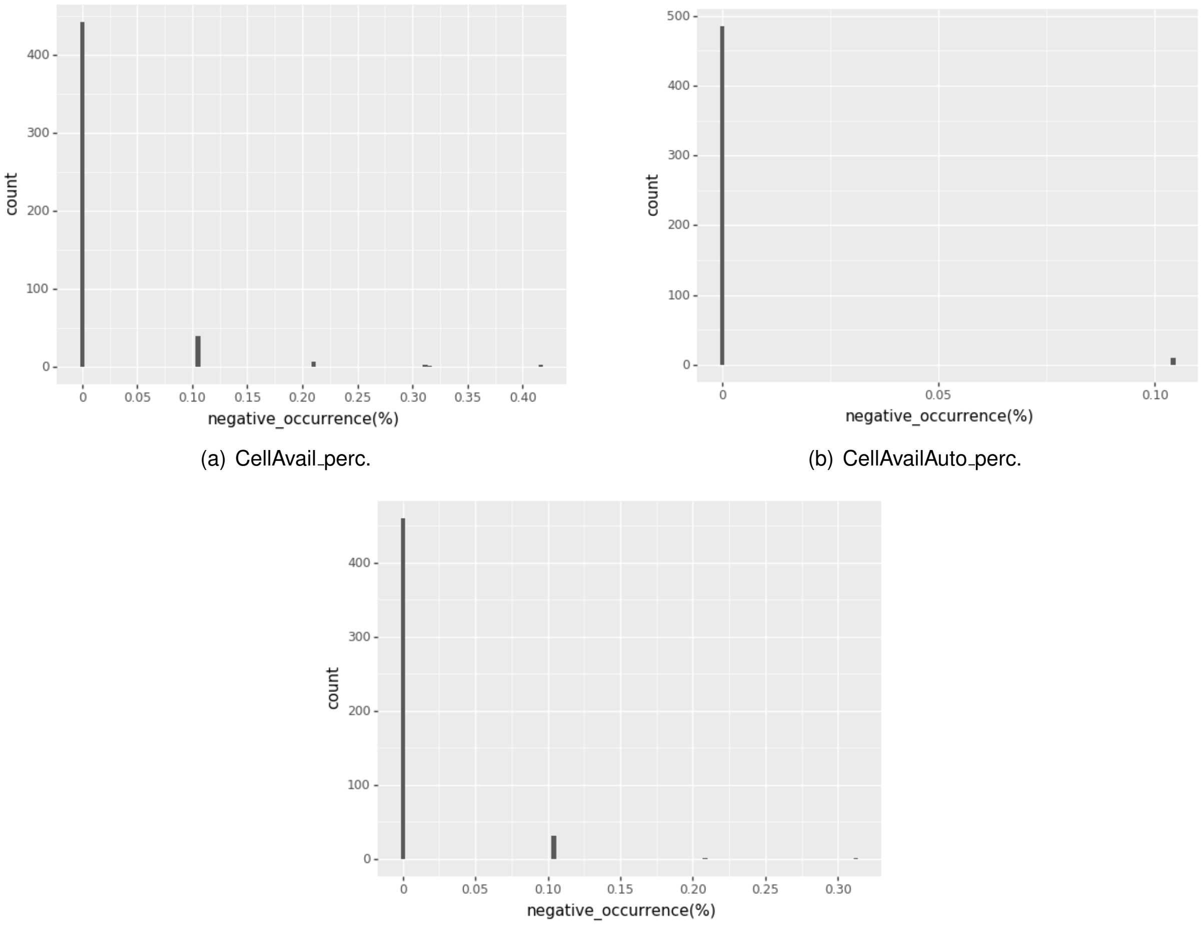
KPI	Negative occurrence (%)
CB_RACH_fail%	0.008210
CellAvail_perc	0.015578
CellAvailAuto_perc	0.002105

Table 3.2: Negative occurrence percentage per KPI.

Additionally, the distribution of the occurrence of negative values, in percentage, for each one of the three KPIs shown in Table 3.2, can be observed in Figure 3.5. Since, for each KPI, the observed values for the percentage of negative values per cell are rather low, with the maximum being 0.4167% for the CellAvail_perc feature in only one cell, it was decided to simply remove the rows that contained at least one negative value because the loss of information would not be significant.

Since all KPIs from the Availability and Accessibility classes are measured in terms of percentages, it was verified if there were any values above 100%. Since the maximum value observed for each one of these KPIs was 100%, further preprocessing was not needed.

Regarding the Integrity KPIs, the downlink and uplink throughputs were compared against the respective maximum theoretical downlink data rate (300 Mbps) and the maximum uplink data rate (75 Mbps) presented in [2]. It was verified that for all Integrity KPIs, with the exception of the DL_MAC_Cell_Tput, the values were coherent with the theoretical ones. The histogram of outliers regarding the Dl_MAC_Cell_Tput is illustrated in Figure 3.6. Similarly to the approach taken for the negative values, the decision regarding the outliers for the Dl_MAC_Cell_Tput KPI was to simply remove the rows of the dataset that contained them.



(a) CellAvail_perc.

(b) CellAvailAuto_perc.

(c) CB_RACH_fail%.

Figure 3.5: Frequency of negative occurrence percentage per KPI and cell.

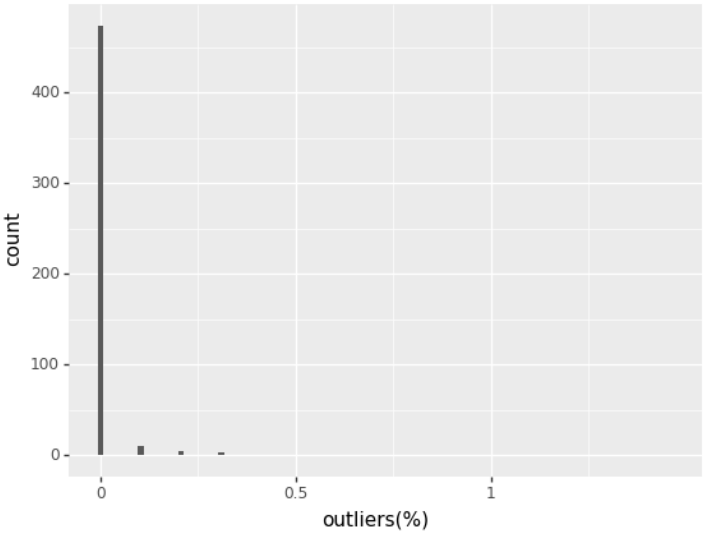


Figure 3.6: Frequency of outliers occurrence percentage per cell for Dl_MAC_Cell_Tput.

3.2 CM Data

For each cell in the PM dataset there is the correspondent configuration parameters in the CM dataset. The configuration parameters available are described in the following points [12]:

- **EARFCNDL and EARFCNUL** - channel number for the central downlink and uplink frequency, respectively;
- **DLCHANNELBANDWIDTH and ULCHANNELBANDWIDTH** - cell's downlink and uplink channel bandwidth, respectively;
- **NOOFPUCCHCQIUSERS** - number of CQI resources available on the Physical Uplink Control Channel (PUCCH);
- **NOOFPUCCHSRUSERS** - number of Scheduling Request (SR) resources available on the PUCCH;
- **ULINTERFERENCEMANAGEMENTACTIVE** - specifies if uplink interference management is enabled or disabled;
- **NOCONSECUTIVESUBFRAMES** - number of consecutive downlink sub-frames with positioning reference signals;
- **COVTRIGGERDBLINDHOALLOWED** - indicates whether a blind handover from this cell can be initiated when a UE reports bad coverage, or not;
- **MIXEDMODERADIO** - determines whether this SectorEquipmentFunction is shared with another node;
- **CELLSUBSCRIPTIONCAPACITY** - normalized subscription capacity of the cell. The value represents the total capacity of the cell used for traffic load balancing purposes;
- **PDCCHCFIMODE** - controls the Control Format Indicator (CFI) used for the control region;
- **THRESHSERVINGLOW** - specifies the threshold that the signal strength of the serving cell must be below for cell reselection towards a lower priority inter-frequency or inter-RAT frequency;
- **FREQBAND** - primary frequency band the cell belongs to according to its defined EARFCN;
- **LBTPNONQUALFRACTION** - fraction of non-qualified UEs at UE selection for throughput aware load balancing;
- **LBTPRANKTHRESHMIN** - minimum threshold for the relative gain at throughput aware load balancing;
- **ALLOCTHRPUCCHFORMAT1** - threshold in terms of number of remaining SR resources available for the cell. Below this threshold, allocTimerPucchFormat1 for allocation of an additional PUCCH format 1 PRB pair is triggered;

- **DEALLOCTHRPUCCHFORMAT1** - threshold in terms of number of remaining SR resources available for the cell. Above this threshold, deallocTimerPucchFormat1 for deallocation of a PUCCH format 1 PRB pair is triggered;

- **DEALLOCTIMERPUCCHFORMAT1** - defines a guard time. After this time, a PUCCH format 1 PRB pair is deallocated if threshold deallocThrPucchFormat1 is still passed;

- **ALLOCTIMERPUCCHFORMAT1** - defines a guard time. After this time, a PUCCH format 1 PRB pair is allocated if threshold allocThrPucchFormat1 is still passed;

- **TRANSMISSIONMODE** - defines the Transmission Mode that shall be used for the UEs that are connected to the cell;

- **INTERFERENCETHRESHOLDSINRCLPC** - Threshold value for measured noise plus interference level. If measured noise plus interference is higher than interferenceThresholdSinrClpc, then the Signal to Interference and Noise Ratio (SINR)-based UL Closed Loop Power Control (CLPC) can be considered;

- **RXSINRTARGETCLPC** - SINR target value for the PUSCH SINR-based CLPC;

- **OPERATIONALSTATE** - indicates the operational state of the cell.

The available CM data did not required any preprocessing since it did not contained neither missing values nor other kinds of artifacts.

Given that the OPERATIONALSTATE feature indicates the operational state of the cell, *i.e.* if the cell is active or not, only active cells were considered.

Chapter 4

PM Clustering

The PM data presented in section 3.1 is constituted by a special type of data, named multivariate time-series. This kind of data is characterized for having multiple features whose values change over time. The main goal of this thesis is to develop a mechanism that has the ability to learn from the PM data in order to evaluate the performance of an LTE network. This can be done using ML.

The purpose of ML is to provide to computers the ability to learn, without them being explicitly programmed. This is done through mathematical models that are built based on statistical and probability theory. The data that is used to train those models is known as training data and, depending on the available training data, there are two main categories into which the learning process may be classified: supervised or unsupervised.

Regarding supervised learning, the training data is composed by an input vector and a target vector. A supervised learning problem may be further divided into [13] a classification problem or a regression problem, the difference being that, for the former, the target data is categorical while for the latter the target data is continuous. In both cases, the goal is to use the model to make predictions in the future.

About unsupervised learning, the training data does not have have any target data to train the model, so the goal in such cases is to identify complex processes or patterns. Unsupervised learning problems may be further classified into [13]: clustering, dimensionality reduction and density estimation. Regarding clustering, the goal is to divide the input vector into groups with similar characteristics. In dimensionality reduction the objective is to map the input vector from a higher dimension to a lower dimension space. Lastly, in density estimation the objective is to find the distribution of the input vector.

Since the PM data is not labelled, meaning that there is no target data to train the model, the focus will be set on developing an unsupervised learning based system that, given the PM data, is able to not only find groups of cells that present similar performance but also classify the performance of each group.

To develop such system, it is necessary to understand the different approaches that can be taken when clustering multivariate time-series, the clustering algorithms that can be used and how to choose the number of groups in which the cells are going to be categorized.

This chapter is organized as follows: Section 4.1 briefly explains the approaches usually taken when clustering time-series; Section 4.2 gives the theoretical background behind the clustering algorithms used in this work; Section 4.3 presents the clustering validation metrics considered; Section 4.4 introduces a technique for dimensionality reduction; Section 4.5 details the methodology used to accomplish the objectives of this work.

4.1 Clustering Time-Series

The following challenges are faced when clustering time-series [14]:

- Time-series data is naturally high dimensional and large in data size which results in an exponential decrease of the clustering process speed;
- Computing the similarity between time-series in order to group them into clusters due to the data itself which is prone to be noisy, contain outliers and shifts but also to its length that can vary.

When clustering time-series, one of the following approaches is usually taken [14]:

- **Shape-based** - the raw time-series are clustered based on their shape, using conventional clustering algorithms and an appropriate similarity measure for time-series, such as Dynamic Time Warping (DTW) [15];
- **Feature-based** - each raw time-series is converted into a lower dimension feature vector and the extracted feature vectors are then clustered;
- **Model-based** - for each raw time-series, it is assumed that it was generated by a known model [16]. Using a suitable distance measure to compute the similarity, a conventional clustering algorithm is applied over the model parameters.

In this work it was taken a feature-based approach, in which the raw time-series are converted into single values that are clustered to find groups of cells with similar performance. Since we are now clustering a vector for each cell, instead of a multivariate time-series, the complexity of the clustering process is reduced significantly.

This process is explained in more detail in Section 4.5.

4.2 Clustering Algorithms

There are several clustering algorithms defined in the literature that depending on the starting point and criteria to cluster the data, can be classified into one of the following categories [17, 18]:

- **Representative-based clustering** - the goal of this type of clustering is to partition the data, containing *n* objects, into *k* desired clusters, each one having at least one object. Moreover, there is a representative point for each cluster that summarizes it, a common choice being the mean (*centroid*);

- **Hierarchical clustering** - data objects are grouped in a sequence of partitions, either starting with each data object being a partition (bottom-up approach) or with a partition including all data objects (top-down approach). For the former, the most similar pair of clusters are successively merged, until all objects are grouped into only one cluster. For the latter, the opposite process is performed, with each cluster being successively split until each object belongs to a different cluster.

- **Density-based clustering** - aims to find regions with a high density of points that are separated from one another by sparse or empty regions;

- **Graph clustering** - objects are clustered over a graph that models the similarity between each pair of objects.

Since different clustering algorithms have distinct underlying principles and assumptions, it is unclear which one fits the available data the best. Consequently, different algorithms were considered: K-means, Expectation-Maximization (EM) using Gaussian Mixture Models (GMM) and Spectral Clustering.

4.2.1 K-means

K-means is a representative-based clustering that, given a cluster set $\mathcal{C} = \{C_1, ..., C_k\}$, aims to minimize the following sum of squared errors:

$$SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2 \tag{4.1}$$

where $\mathbf{x}_j$ and $\boldsymbol{\mu}_i$ are the j^{th} point and centroid of cluster C_i, respectively. Formally, this is given by:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \{SSE(\mathcal{C})\} \tag{4.2}$$

The pseudo-code for K-means is shown in Algorithm 1, where it can be seen that the algorithm is iterative, with each iteration consisting of two steps: cluster assignment and centroid updates. The inputs D, k and ϵ, where $\epsilon > 0$, correspond to the dataset, number of desired clusters and convergence threshold, respectively.

Algorithm 1: K-means algorithm (adapted from [17]).

```
Input : $\mathbf{D}, k$ and $\epsilon$
$t \leftarrow 0$
Randomly initialize the centroids $\boldsymbol{\mu}_1^t$,..., $\boldsymbol{\mu}_k^t$
repeat
    $t \leftarrow t + 1$
    foreach $\mathbf{x}_j \in \mathbf{D}$ do
        $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^t \right\|^2 \right\}$
        $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
    end
    foreach $i = 1$ to $k$ do
        $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
    end
until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\| \leq \epsilon$;
```

It is important to notice that K-means is particularly suited for convex shaped clusters, namely spherical shaped clusters, or circular in two dimensions. Furthermore, K-means is a clustering algorithm used for general purpose due to its simplicity and speed [19].

4.2.2 Expectation-Maximization Clustering using Gaussian Mixture Models

The EM using GMM is also a representative-based clustering algorithm.

Let $\mathbf{D} = \{\mathbf{x}_j\}_{j=1}^{n}$ be a dataset of n points in $\mathbb{R}^d$ and $\mathbf{X} = (X_1, ..., X_d)$ be the vector random variable for the d-attributes. In a GMM each cluster C_i is represented by a multivariate Gaussian distribution, so, the probability density at $\mathbf{x}$ relatively to cluster C_i is given by:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{ - \frac{(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)}{2} \right\} \tag{4.3}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ are the cluster mean and covariance, respectively.

The probability density function of each point $\mathbf{x}$ over all the k specified clusters is a GMM defined by:

$$f(\mathbf{x}) = \sum_{i=1}^{k} f_i(\mathbf{x}) P(C_i) \tag{4.4}$$

where $P(C_i)$ are the prior probabilities which must satisfy $\sum_{i=1}^{k} P(C_i) = 1$.

Therefore, the set of parameters $\boldsymbol{\theta}$ that characterize the GMM are the mean $\boldsymbol{\mu}_i$, the covariance $\boldsymbol{\Sigma}_i$ and the probability $P(C_i)$ of each one of the k clusters:

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(C_1), ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(C_k)\}$$

The goal is to estimate $\boldsymbol{\theta}$ so that the conditional probability of the data $\mathbf{D}$ given the model parameters

$\boldsymbol{\theta}$ is maximized. Formally, this is given by:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}}\{P(\mathbf{D}|\boldsymbol{\theta})\} \tag{4.5}$$

where, considering that each point $\mathbf{x}_j$ is independent and identically distributed as $\mathbf{X}$, $P(\mathbf{D}|\boldsymbol{\theta})$ is defined as

$$P(\mathbf{D}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\mathbf{x}_j) \tag{4.6}$$

In order to estimate $\boldsymbol{\theta}$ that maximizes Equation 4.6, the Expectation-Maximization algorithm is used, as shown in Algorithm 2. The Expectation-Maximization algorithm consists in two steps, the expectation step and the maximization step, that are iteratively performed until a convergence condition is satisfied.

Algorithm 2: Expectation-Maximization algorithm (adapted from [17]).

```
Input : $\mathbf{D}, k$ and $\epsilon$
$t \leftarrow 0$;
Randomly initialize $\boldsymbol{\mu}_1^t$,..., $\boldsymbol{\mu}_k^t$
$\boldsymbol{\Sigma}_i^t \leftarrow \mathbf{I}, \forall i = 1, ..., k$
$P^t(C_i) \leftarrow \frac{1}{k}, \forall i = 1, ..., k$
repeat
    $t \leftarrow t + 1$
    // Expectation Step
    for $i = 1, ..., k$ and $j = 1, ..., n$ do
        $w_{ij}^t \leftarrow \frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)P(C_i)}{\sum_{a=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)P(C_a)}$ // where $w_{ij}^t = P^t(C_i|\mathbf{x}_j)$
    end
    // Maximization Step
    for $i = 1, ..., k$ do
        $\boldsymbol{\mu}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$
        $\boldsymbol{\Sigma}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij}(\mathbf{x}_j-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$
        $P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$
    end
until $\sum_{i=1}^{k}\left\|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1}\right\| \leq \epsilon$;
```

Clustering using GMMs is a generalized version of the K-means where each point can now belong to multiple clusters instead of only one cluster. This is also known as “soft” assignment. Since each cluster is described by two parameters, mean and variance, GMM is more flexible when compared to K-means, being able to find elliptical shaped clusters.

4.2.3 Spectral Clustering

Spectral clustering [20] is a form of graph clustering. Let $\mathbf{D} = \{\mathbf{x}_j\}_{j=1}^{n}$ be a set of n points in $\mathbb{R}^d$. The algorithm starts by computing the pairwise similarities through the Gaussian Radial Basis Function (RBF)

[21], resulting the similarity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, whose diagonal elements are A_{ii} = 0.

Then, the normalized symmetric Laplacian matrix is computed by [17]:

$$\mathbf{L}^s = \mathbf{\Delta}^{-1/2}\mathbf{L}\mathbf{\Delta}^{-1/2} \tag{4.7}$$

where $\mathbf{L} = \mathbf{\Delta} - \mathbf{A}$ is the Laplacian matrix and $\mathbf{\Delta}$ is the $n \times n$ diagonal matrix that represents the degree of each vertex in the graph. The degree of a vertex x_i is defined as:

$$d_i = \sum_{j=1}^{n} A_{ij} \tag{4.8}$$

After that, the algorithm finds the k eigenvectors, $\mathbf{u}_1, \ldots, \mathbf{u}_k$, corresponding to the k smallest eigenvalues of $\mathbf{L}^s$ [22]. These vectors are then used as columns to construct the matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$. Further, the matrix $\mathbf{U}$ is normalized so the rows have unit norm, resulting the matrix $\mathbf{Y} \in \mathbb{R}^{n \times k}$.

Each row of $\mathbf{Y}$ is then treated as point in $\mathbb{R}^k$ and clustered into k clusters using K-means. The pseudo-code for spectral clustering is presented in Algorithm 3.

Algorithm 3: Spectral clustering algorithm (adapted from [20]).

```
Input: $\mathbf{D}, k$
Compute the similarity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$
Compute the diagonal matrix $\mathbf{\Delta} \in \mathbb{R}^{n \times n}$
Construct the matrix $\mathbf{L}^s \in \mathbb{R}^{n \times n}$
Find $\mathbf{u}_1, \ldots, \mathbf{u}_k$, the $k$ eigenvectors corresponding to the $k$ smallest eigenvalues of $\mathbf{L}^s$
Form the matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns
Form the matrix $\mathbf{Y}$ from $\mathbf{U}$ by normalizing each row of $\mathbf{U}$ to have unit norm: $Y_{ij} \leftarrow \frac{U_{ij}}{(\sum_j U_{ij}^2)^{1/2}}$
Treating each row of $\mathbf{Y}$ as point in $\mathbb{R}^{n \times k}$, cluster them into $k$ clusters using K-means
Assign each original point $x_i$ to cluster $j$ if row $i$ of the matrix $\mathbf{Y}$ was assigned to cluster $j$
```

Compared with the K-means and the EM using GMM, the Spectral Clustering algorithm is able to find non-convex clusters, as it is neither tied to any ideal cluster spherical shape, as K-means, nor to elliptical shapes, as the GMM.

4.3 Clustering Validation

The information provided in this section was mainly based in [23].

Clustering validation is the process of evaluating the goodness of partitions after clustering and it can be categorized, depending if external information is used for clustering validation or not, into two main classes: internal clustering validation and external clustering validation.

External Clustering Validity Index (CVI)s are used when the ground truth of the data is available and they are indicators for choosing an optimal clustering algorithm in a specific data set.

Additionally, internal CVIs are used when the ground truth is not available and they are indicators of the optimal number of clusters. Taking into account that the goal of clustering is to make objects in a cluster similar and objects in different clusters distinct, internal CVIs are usually based on one or both of the following criteria:

Compactness: measures how closely related the objects in a cluster are.

Separation: measures how well-separated or distinct a cluster is from other clusters.

Given that, in the context of this work, the true cluster labels are not available, only internal CVIs are presented. Furthermore, only the CVIs that contemplate both evaluation criteria, compactness and separation, were considered.

Calinski-Harabasz Index

The Calinski-Harabasz (CH) index is a ratio-type index that measures the compactness and the separation through the average between- and within-cluster sum of squares, respectively. It is computed using the following expression:

$$CH = \frac{\sum_i n_i d^2(c_i, c)/(NC-1)}{\sum_i \sum_{x \in C_i} d^2(x, c_i)/(n-NC)} \tag{4.9}$$

where i is the i^{th} cluster, n is the number of points in the data set, n_i is number of points in the cluster C_i, c_i is the i^{th} cluster centroid, c is the data set centroid, NC is the total number of clusters, x is a data point and d is the Euclidean distance function.

I Index

The I index evaluates the separation through the maximum distance between cluster centroids and the compactness by the sum of distances between points and their respective cluster centroid. The value of this index is given by:

$$I = \left(\frac{1}{NC} \cdot \frac{\sum_{x \in D} d(x, c)}{\sum_i \sum_{x \in C_i} d(x, c_i)} \cdot max_{i,j} d(c_i, c_j) \right)^p \tag{4.10}$$

where p is the number of attributes of the data set and the remaining variables have the same definition as in Equation 4.9.

Dunn's Index

Dunn's index evaluates the cluster separation through the minimum pairwise distance between points in different clusters and the cluster compactness through the maximum diameter between all clusters. It is computed as follows:

$$D = min_i \ \{ \ min_j \left(\frac{min_{x \in C_i, y \in C_j} d(x, y)}{max_k \ \{ \ max_{x,y \in C_k} d(x, y) \ \}} \right) \} \tag{4.11}$$

where y is also a data point and the remaining variables have the same definition as previous CVIs.

Silhouette Index

The Silhouette index is a summation-type index that validates the clustering performance through the pairwise difference of between-cluster distances (separation) and the pairwise difference of within-cluster distances (compactness). It is given by:

$$\frac{1}{NC}\sum_i \left\{ \frac{1}{n_i} \sum_{x\in C_i} \frac{b(x)-a(x)}{max\left[b(x),a(x)\right]} \right\} \tag{4.12}$$

where $a(x)$ and $b(x)$ are defined in 4.13 and 4.14, respectively. The remaining variables were already defined.

$$a(x) = \frac{1}{n_i - 1} \sum_{y\in C_i, y\neq x} d(x,y) \tag{4.13}$$

$$b(x) = min_{j,j\neq i} \left[\frac{1}{n_j} \sum_{y\in C_j} d(x,y) \right] \tag{4.14}$$

where x and y are data points and the remaining variable were already defined.

Davies-Bouldin Index

The first step to determine the Davies-Bouldin Index (DB) is to compute, for each cluster C, the similarities to all the other clusters and assign to that cluster C the highest similarity value obtained. The DB is then obtained by averaging all the cluster similarities:

$$DB = \frac{1}{NC}\sum_i max_{j,j\neq i} \left\{ \frac{\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)}{d(c_i,c_j)} \right\} \tag{4.15}$$

where all variables were previously defined.

S_Dbw

The S_Dbw index [24] introduces the notions of scattering and density to measure the compactness of the clusters and the separation between clusters, respectively. It is computed as follows:

$$S\_Dbw = Scat(NC) + Dens\_bw(NC) \tag{4.16}$$

where $Scat(NC)$ corresponds to the average cluster scattering and is given by:

$$Scat(NC) = \frac{1}{NC}\frac{\sum_i ||\sigma(C_i)||}{||\sigma(D)||} \tag{4.17}$$

where NC is the number of clusters, $\sigma(C_i)$ is the ith cluster variance vector and $\sigma(D)$ is the whole data set variance vector. The variance of the data set is $\sigma(D)$ and its p$^{\text{th}}$ dimension value is given by:

$$\sigma_D^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x}^p\right)^2 \tag{4.18}$$

where $\overline{x}^p$ is the p$^{\text{th}}$ dimension of:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in D \tag{4.19}$$

The p$^{\text{th}}$ dimension of the cluster C_i variance is defined as:

$$\sigma_{C_i}^p = \frac{1}{n_i}\sum_{k=1}^{n_i}\left(x_k^p - c_i^p\right)^2 \tag{4.20}$$

where n_i is the number of objects belonging to cluster C_i.

$Dens\_bw$ evaluates the average density in the region among clusters in relation with the density of the clusters and is computed by:

$$Dens\_bw(NC) = \frac{1}{NC(NC-1)}\sum_{i=1}^{NC}\left[\sum_{j=1,j\neq i}^{NC}\frac{density(u_{ij})}{max\,\{\ density(c_i), density(c_j)\ \}}\right] \tag{4.21}$$

where NC is total number of clusters, c_i and c_j are the centers of the clusters C_i and C_j, and u_{ij} is the middle point of the line segment defined by c_i and c_j. The density function $density()$ is defined as follows:

$$density(u) = \sum_{l=1}^{n_{ij}} f(x_l, u) \tag{4.22}$$

where n_{ij} is the number of tuples that belong to the clusters C_i and C_j, i.e., $x_l \in C_i \cup C_j \subseteq D$, representing the points in the neighborhood of u. A point belongs to the neighborhood of u if its distance from u is smaller than the average standard deviation of clusters:

$$f(x,u) = \begin{cases} 0, & \text{if } d(x,u) > stdev \\ 1, & \text{otherwise} \end{cases} \tag{4.23}$$

where the average standard deviation of clusters is given by:

$$stdev = \frac{1}{NC}\sqrt{\sum_{i=1}^{NC}\left\|\sigma(C_i)\right\|} \tag{4.24}$$

It is worth noticing that, even though all the CVIs presented above evaluate both the compactness and the separation of the obtained partitioning, the optimal number of clusters can be given by the lower or the higher value of the CVIs. For the Calinski-Harabasz, I, Dunn's and Silhouette indices, a higher value indicates a better partitioning whereas for the Davies-Bouldin and S_Dbw indices the opposite happens, with a lower value indicating a better partitioning of the data.

4.4 Clustering Visualization with t-SNE

t-Distributed Stochastic Neighbour Embedding (t-SNE) [25] is an algorithm for dimensionality reduction. Given a high-dimensional dataset $X = \{x_1, x_2, ..., x_n\}$, where each x_i represents a data point in the high-dimensional space, t-SNE converts it into a two or three-dimensional dataset $Y = \{y_1, ..., y_n\}$, where each data point y_i is the low-dimensional representation of the correspondent x_i data point. The low-dimensional data Y and the low-dimensional representation y_i of individual data points are also referred to as a map and map points, respectively. The dimensionality reduction aims to preserve as much of the significant structure of the original data as possible in the low-dimensional map and allows for visualization of high-dimensional data since the map points can be displayed in a scatterplot.

This algorithm corresponds to a variation of the Stochastic Neighbour Embedding (SNE) [26].

t-SNE starts by converting the high-dimensional Euclidean distances between each pair of data points into conditional probabilities that represent similarities. More specifically, each conditional probability, $p_{j|i}$, is the probability that the data point x_i would pick x_j as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian centered at x_i and it is computed using the following expression:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{4.25}$$

where σ_i is the variance of the Gaussian centered at x_i. For any particular value of σ_i a probabilistic distribution, P_i, is induced over all data points. The value of σ_i is obtained through a binary search, performed by t-SNE, that produces a P_i with a fixed perplexity which is a parameter specified by the user and can be interpreted as a smooth measure of the effective number of neighbours.

The joint probabilities in the high-dimensional space, p_{ij}, are then determined by:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{4.26}$$

where n is total number of high-dimensional datapoints.

It is also possible to compute a joint probability, denoted by q_{ij}, for the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j. However, in order to compute q_{ij}, a heavy-tailed distribution (Student-t distribution) is used to convert the Euclidean distance into probabilities instead of a Gaussian distribution. This is a direct consequence of the "crowding problem": when a high-dimensional dataset is modeled into two or three dimensions, it is difficult to segregate the nearby data points from moderately distant data points and gaps can not form between natural clusters. Hence, using a heavy-tailed distribution allows a moderate distance in high-dimensional space to be modeled by a larger distance in the low-dimensional space when compared to using a Gaussian distribution. The joint probability q_{ij} is given by:

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}} \tag{4.27}$$

If the similarity between the high-dimensional data points x_i and x_j is correctly modeled by the map points y_i and y_j, the joint probabilities p_{ij} and q_{ij} will be equal. Thus, the goal is to find a low-dimensional map that minimizes the mismatch between p_{ij} and q_{ij}. The faithfulness with which q_{ij} models p_{ij} can be measured by the Kullback-Leibler divergence. The cost function is given by:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4.28}$$

where P and Q correspond to the joint probability distributions in the high- and low-dimensional space, respectively. Both p_{ii} and q_{ii} are set to zero since we are only interested in modeling pairwise similarities.

A gradient descent method is performed in order to minimize the cost function presented in (4.28).

4.5 Performance Evaluation Methodology

Developing an automatic procedure to evaluate the performance of LTE networks, or any other cellular networks for that matter, is rather challenging. The existence of databases with labelled cases of faults or performance evaluations is scarce, making it unfeasible to use supervised learning techniques.

Hence, the developed methodology consists on the application of clustering techniques in order to find groups of cells with similar performances. Between the identified groups, it is expected to distinguish groups of cells with good performance while others should correspond to cells where the performance is below the desired level.

The proposed method consists in a feature-based approach, where new features are extracted through the comparison of each time-series against desired targets for each KPI. Even though this approach results in the loss of information regarding the cell behavior, in the time domain, it allows to evaluate each cell overall performance level and group the cells accordingly. The methodology flowchart is illustrated in Figure 4.1.

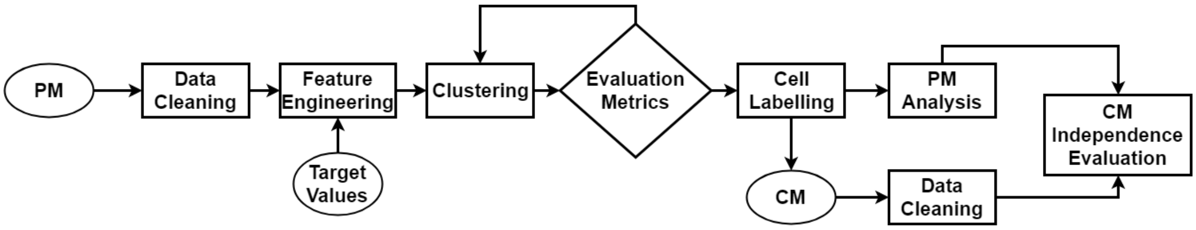


Figure 4.1: Methodology flowchart.

As figure 4.1 shows, the first stage, which was already described in detail in Section 3.1.2, corresponds to the data cleaning step, where the PM data goes through the process of data preprocessing to ensure the data integrity. This step is also applied to the CM data with the same purpose.

4.5.1 Feature Engineering

The feature engineering step corresponds to the process of extracting new features from the original data. The process of feature extraction consists in defining a set of target values $T = [T_1, T_2, ..., T_M]$, where each T_p corresponds to the target value for the p^{th} KPI, from the M considered KPIs. Then, for each cell, c, and KPI, p, each measure, x_{cp_t}, acquired in the instant t of the time-series of size N, is compared against the defined target, T_p, for that KPI.

If the KPI is from the Accessibility group then the target is said to be satisfied if x_{cp_t} is lower than the target, since each Accessibility KPI in the PM dataset corresponds to a fail rate. The value of x_{cp_t} is then changed to 0 or 1 accordingly:

$$x_{cp\_t} = \begin{cases} 1, & \text{if } x_{cp\_t} \leq T_p \\ 0, & \text{otherwise} \end{cases} \tag{4.29}$$

Contrarily, for an Integrity or Availability KPI, it is desired that x_{cp_t} is greater than the target value. Thus, the new value for x_{cp_t} is given by:

$$x_{cp\_t} = \begin{cases} 1, & \text{if } x_{cp\_t} \geq T_p \\ 0, & \text{otherwise} \end{cases} \tag{4.30}$$

Thereupon, for each cell, c, and KPI, p, the ratio between the number of times that KPI satisfies its defined target and the total number of times the KPI was measured is computed:

$$feature_{cp} = \frac{1}{N} \sum_{t=1}^{N} x_{cp\_t} \tag{4.31}$$

A visual interpretation of the feature engineering process is shown in Figure 4.2. Considering one cell, the KPI RRC_Estab_fail% is plotted over the time that was measured, against the target value for that KPI. The new feature generated from this KPI, and all the remaining extracted features for that matter, can be interpreted as the time period over the total considered time that the defined target for the original KPI was satisfied. Since the KPI shown in Figure 4.2 corresponds to a fail percentage then satisfying the target means that the measured value for the KPI is below its target.

Intuitively, every value obtained for the generated features is comprised between 0 and 1, with 1 representing a cell that satisfied the target set for a KPI during the total period of time in which the time series were obtained.

As a result, for each pair, cell and KPI, the correspondent time series is converted into a single value, thus the resulting dataset is composed by only one row per cell. This is exemplified for one KPI in Figure 4.3. It can be observed the Init_E_RAB_Estab_fail% KPI as a time-series, for a cell univocally identified by the _MECONTEXT and _VSDATAEUTRANCELLFDD columns, prior to the feature engineering stage and the extracted feature for that same cell.

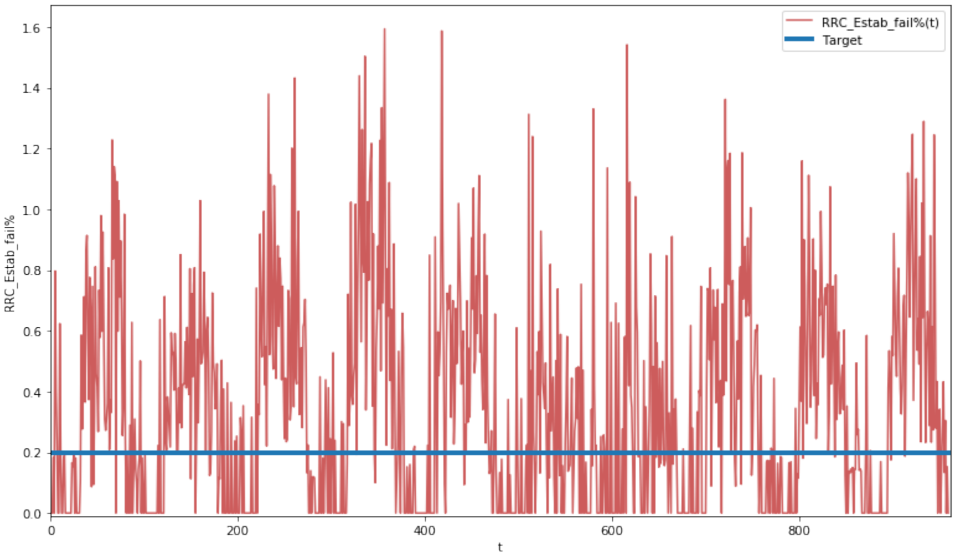
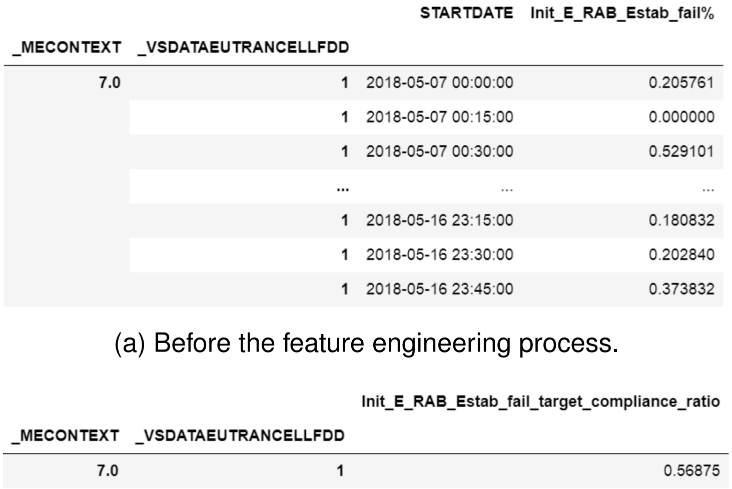


Figure 4.2: Feature engineering example.



		STARTDATE	Init_E_RAB_Estab_fail%
_MECONTEXT	_VSDATAEUTRANCELLFDD		
7.0	1	2018-05-07 00:00:00	0.205761
	1	2018-05-07 00:15:00	0.000000
	1	2018-05-07 00:30:00	0.529101
	...	...	...
	1	2018-05-16 23:15:00	0.180832
	1	2018-05-16 23:30:00	0.202840
	1	2018-05-16 23:45:00	0.373832

(a) Before the feature engineering process.

		Init_E_RAB_Estab_fail_target_compliance_ratio
_MECONTEXT	_VSDATAEUTRANCELLFDD	
7.0	1	0.56875

(b) After the feature engineering process.

Figure 4.3: Example of PM data before and after the feature engineering process.

4.5.2 Clustering and Labelling Stages

The next step consists in applying a clustering algorithm to the dataset generated in the previous stage in order to find groups of cells that present similar behaviours. This is an iterative process:

1. Choose the clustering algorithm;
2. Run the clustering algorithm with different input parameters and evaluate the clustering result using the Kolmogorov-Smirnov test;
3. Select the set of input parameters that give the best clustering results using the metrics presented in Section 4.3.

Regarding step 2 of the above process, each pair of clusters must have a distinct statistical behaviour for at least one feature in order to consider the clustering result relevant [1]. This is verified using the two sample Kolmogorov-Smirnov test [27], to test if the observed values for a feature of two different clusters are generated by the same distribution (null hypothesis). The null hypothesis is rejected when the resulting *p*-value is lower than the chosen significance level, thus a lower *p*-value indicates a more distinct statistical behaviour between the pair of clusters and feature being evaluated. If there is at least one pair of clusters that present a similar statistical behaviour for every feature, meaning that the null hypothesis is never rejected, then the number of clusters that originated that partition is automatically discarded. The significance level used to test the null hypothesis was 0.01.

Each one of the clustering algorithms presented in Section 4.2 requires the number of clusters to be specified beforehand. In this work, it was chosen to run the chosen clustering algorithm for set of number of clusters that ranges from 2 to 8, to ensure that a wide enough range of possible partitions is analyzed by the system.

In step 3, the selection of the set of input parameters, which includes the optimal number of clusters, k, that give the best partitioning of the data is attained using multiple internal validation metrics. This approach was taken because not only it is extremely hard to, intuitively, understand which metric is the most appropriate, for this data, but also considering that even though all proposed metrics evaluate the partitioning through the clusters separation and compactness, they do not always return the same optimal number of clusters [23]. Therefore, the optimal number of clusters is chosen through an election mechanism where all metrics have the same importance, meaning that the optimal number of clusters is the one that obtains more votes, with each metric contributing with one vote. The validation metrics used were the ones presented in Section 4.3.

Once the set of input parameters that give the best partitioning result, according to the election mechanism, is identified, the algorithm is performed again with the optimal input in order to assign a label to each cell, identifying to which cluster that cell belongs. Thus, this is called the labelling stage.

4.5.3 PM Analysis

After each cell has been labelled according to the cluster to which it belongs, PM analysis is performed. This stage includes the following actions:

- **Data visualization** - t-SNE is used for dimensionality reduction and the resulting points are plotted in a two-dimensional space for visual inspection;
- **Cluster scoring and classification** - a score for each cluster is computed. A higher score indicates a better performance. The clusters are then classified based on the obtained score;
- **Feature distribution analysis** - the distribution of each feature per cluster is plotted allowing the user to gain insight about the performance of each cluster regarding each KPI.

The cluster score is a weighted average of the scores computed for each feature considered, and is given by:

$$score_{cluster} = \frac{1}{M}\sum_{p=1}^{M}\alpha_p score_{feature\_p} \tag{4.32}$$

where α_p corresponds to a weight given to $feature_p$ and $\sum_{p=0}^{M} \alpha_p = 1$ and $score_{feature\_p}$ is the score for the p^{th} feature of the considered cluster and is given by:

$$\begin{aligned} score_{feature} &= P(feature \geq time_{threshold}) \\ &= 1 - P(feature < time_{threshold}) \end{aligned} \tag{4.33}$$

where $time_{threshold}$ is a parameter that can take values between 0 and 1 and is specified by the user depending on the level of exigency desired. This score can be interpreted as the probability of a cell in the cluster being compliant with the target set for the feature being evaluated, for a period of time above the $time_{threshold}$. This is illustrated in Figure 4.4, where the $time_{threshold}$ was set at 0.8. The intersection of the vertical red line, representing the threshold, with the Empirical Cumulative Distribution Function (ECDF) of a cluster gives the value of $P(feature < time_{threshold})$. It is straightforward to see that the probability of a cell belonging to cluster 1 being compliant with the target defined for the RRC_Estab_fail% for more than 80% of the total period of time of the original time-series, is greater than for a cell belonging to cluster 0. This probability gives the $score_{feature}$ for each cluster.

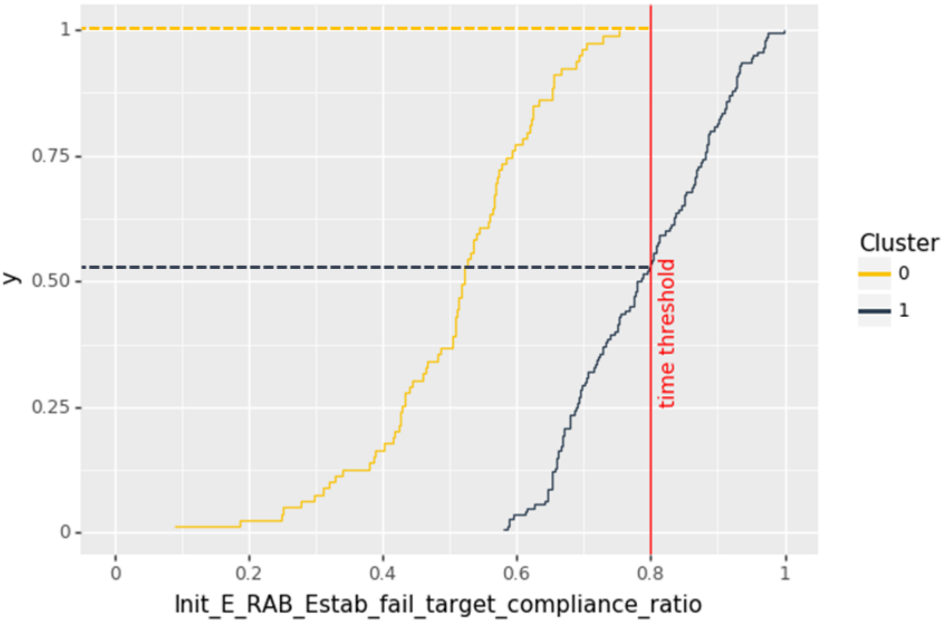


Figure 4.4: Visual interpretation of the cluster feature score.

So, setting $time_{threshold}$ to 1 means that the user wants to obtain the probability of a cell in the cluster to be compliant with the defined target during the total period of time during which the original time series were obtained. Furthermore, by multiplying the obtained $score_{feature}$ with the number of cells in the cluster being evaluated one gets the number of cells that are compliant with the target for the specified $time_{threshold}$.

Through the combination of the cluster scoring and the feature distribution analysis, one is able to immediately tell which clusters have better performance and what are the most distinct performance features.

4.5.4 CM Independence Evaluation

Regarding the CM features, the goal is to find the set of configuration parameters that are most distinct between clusters and evaluate if, for the class of KPIs being analyzed, there is a correlation between them and the performance of the cluster.

The process of identifying the most distinct configuration parameters is based on the Fisher's exact test of independence [28] while the process of evaluating if there is a correlation between the configuration parameters and the performance relies, at this stage, on the expertise of radio network engineers.

Chapter 5

Network Performance Analysis

In this chapter, the results related with the clustering process and subsequent analysis of the clusters performance are presented. As described in 4.5, these results include the optimal number of clusters, the visualization for the optimal partitioning, the comparison of the Probability Density Function (PDF) and ECDF plots for each feature in different clusters and the respective cluster scoring.

It was already mentioned in Section 3.1 that the network under analysis contains sites from three distinct frequency bands (L800, L1800 and L2600). Moreover, each frequency band in this network is associated with a specific bandwidth; for L800 the bandwidth is 10 MHz while for both L1800 and L2600 the bandwidth is 20 MHz. Since different frequency bands may serve different purposes and the bandwidth affects the performance of a telecommunications system, it was decided to divide the available PM dataset into three smaller datasets, each one corresponding to a different frequency band.

Moreover, the PM data is further grouped according to the classes of KPIs presented in Section 3.1.1, so the performance for each frequency band deployed is evaluated for each KPI class individually.

Both the clustering algorithms (K-means, EM using GMM and Spectral Clustering) and the t-SNE algorithm used in this chapter are from the Scikit-learn library [29].

Additionally, each feature in the dataset being clustered, was standardized by removing the mean and scaling to unit variance.

The figures where the cells locations are shown were obtained using the software in [30].

The scores obtained for each cluster using Equation 4.32 were computed with $time_{threshold} = 0.8$, unless explicitly stated otherwise.

This chapter is organized as follows: Section 5.1 gives a brief introduction regarding the targets used in the feature engineering step; in Section 5.2 the clustering results, using the K-means, are presented; Section 5.3 presents a comparison between the results obtained with K-means, Spectral Clustering and EM using GMMs; lastly, Section 5.4 details the method used to link the configuration parameters with the performance parameters and presents a use case related with the cells subscription capacity.

5.1 Targets

A reasonable choice regarding the KPIs targets is perhaps the most influential step when applying the methodology presented in Section 4.5, as the dataset resulting from the feature engineering step depends on that choice. On one hand, a set of targets that is too demanding, regarding the KPI class being evaluated, may result in all cells showing a similar poor performance. On the other hand, if the targets are too loose it may result that all cells show great performance. Therefore, the targets should be defined taking into account the knowledge of experts and the performance level that the operator wants to provide.

In the context of this work, it was necessary to define targets for the three identified KPI classes: Availability, Accessibility and Integrity. For both the Accessibility and Availability KPIs it was considered that their targets should be independent of the frequency band. However, the bandwidth is intrinsically related to the provided QoS, which in turn is evaluated through the Integrity KPIs, so the targets defined for this class vary with respect to the used bandwidth. Thus, two sets of targets for the Integrity KPIs are specified, one for the L800 frequency band and the other for both the L1800 and L2600 frequency band.

The target values for both the Availability and Accessibility KPIs were specified with the help of engineers from Celfinet. Since all available Integrity KPIs are different measures of throughputs, which depend not only on the established service(s) for the UEs connected to the cell but also on other factors such as the modulation used or the channel coding rate [2], it is rather difficult and somewhat naive to define targets for this category. The approach taken for this case was to compute, for each Integrity KPI, the 25^{th} percentile with respect to the 10 MHz and 20 MHz bandwidths. This percentile should be sufficiently low that different performing groups of cells are found but not too low so all cells appear to have an excellent performance.

5.2 Clustering using K-means

K-means was the first algorithm used to cluster the PM data due to its simplicity. The only input parameter that was changed, when using K-means during the clustering stage, was the number of clusters, k. For the initialization of the k centroids , K-means++ [31] was used.

The clustering results regarding the Availability KPIs will only be presented for the L1800 frequency band, as most cells were always available, during the time window for which the KPIs were collected.

5.2.1 L800

This section presents the clustering results obtained for the cells operating in the 800 MHz frequency band. The PM dataset for this frequency band is constituted by 219 cells.

Accessibility

Following the methodology presented in Section 4.5, and since the preprocessing of the PM data was already explained in Section 3.1, the targets for the Accessibility KPIs need to be defined in order to perform the feature engineering step. The targets for each Accessibility KPI are presented in Table 5.1. These targets are the same for the L1800 and L2600 frequency bands.

KPI	**Target [%]**
CB_RACH_fail%	5
Added_E_RAB_Estab_fail%	0.01
Init_E_RAB_Estab_fail%	0.05
RRC_Estab_fail%	0.25
S1_Estab_fail%	0.25

Table 5.1: Targets for Accessibility KPIs.

Executing the K-means for values of k ranging from 2 to 8, the optimal number of clusters for each CVI, presented in Section 4.3, is shown in Figure 5.1.

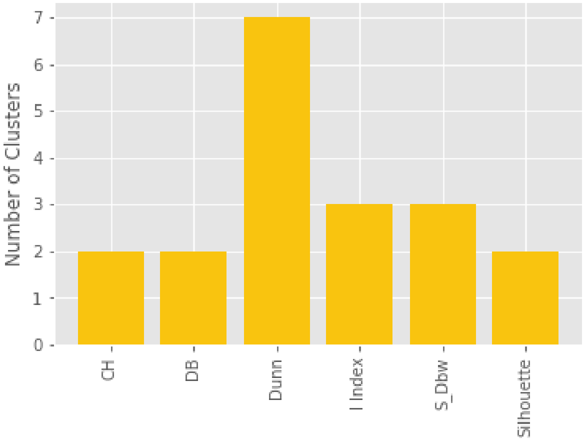


Figure 5.1: CVI results for K-means in L800 (Accessibility).

It can be seen that the most voted optimal number of clusters was 2. With the exception of the Dunn Index, all metrics return an optimal k of 2 or 3. Using the t-SNE for dimensionality reduction, the obtained clusters for $k = 2$ can be visualized in Figure 5.2.

There are 79 cells belonging to cluster 0 while cluster 1 contains 140 cells. It can be visualized that there is not a clear separation between the two clusters. Even though the Figure 5.2 points correspond to a mapping of the original data in two dimensions, where some information was lost, it can be seen that the points have a somewhat homogeneous distribution.

Applying the procedure detailed in Section 4.5.3, the scores for each cluster were obtained. Furthermore, the operator may choose to qualitatively classify the overall performance cluster based on the obtained score. In this work, to ease the results analysis, the following qualitative levels of performance were considered: *unsatisfactory*, for a cluster score below 0.25; *below average*, for a score between 0.25 and 0.5; *average*, for a cluster score between 0.5 and 0.75; and *above average* for a cluster score greater than 0.75.

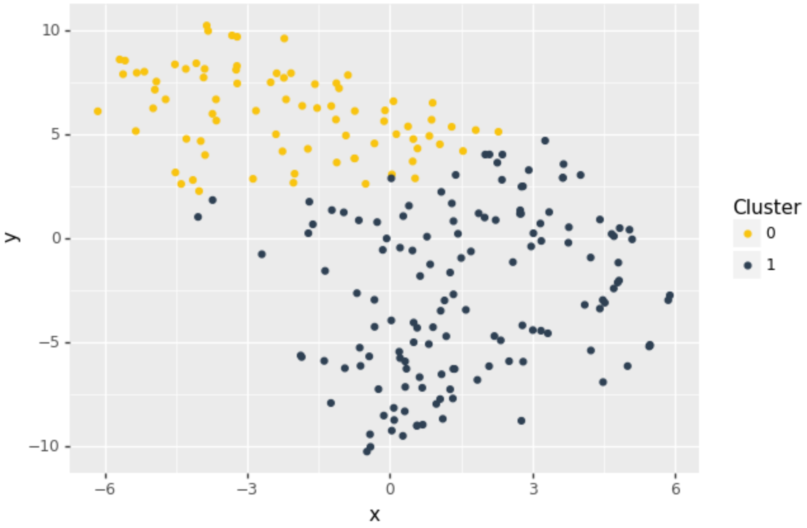


Figure 5.2: Clustering visualization for K-means in L800 (Accessibility).

Both the quantitative and qualitative scores obtained for the clusters are presented in Table 5.2.

Cluster	Number of Cells	Score	Classification
0	79	0.32	Below average
1	140	0.57	Average

Table 5.2: Cluster classification for K-means in L800 (Accessibility).

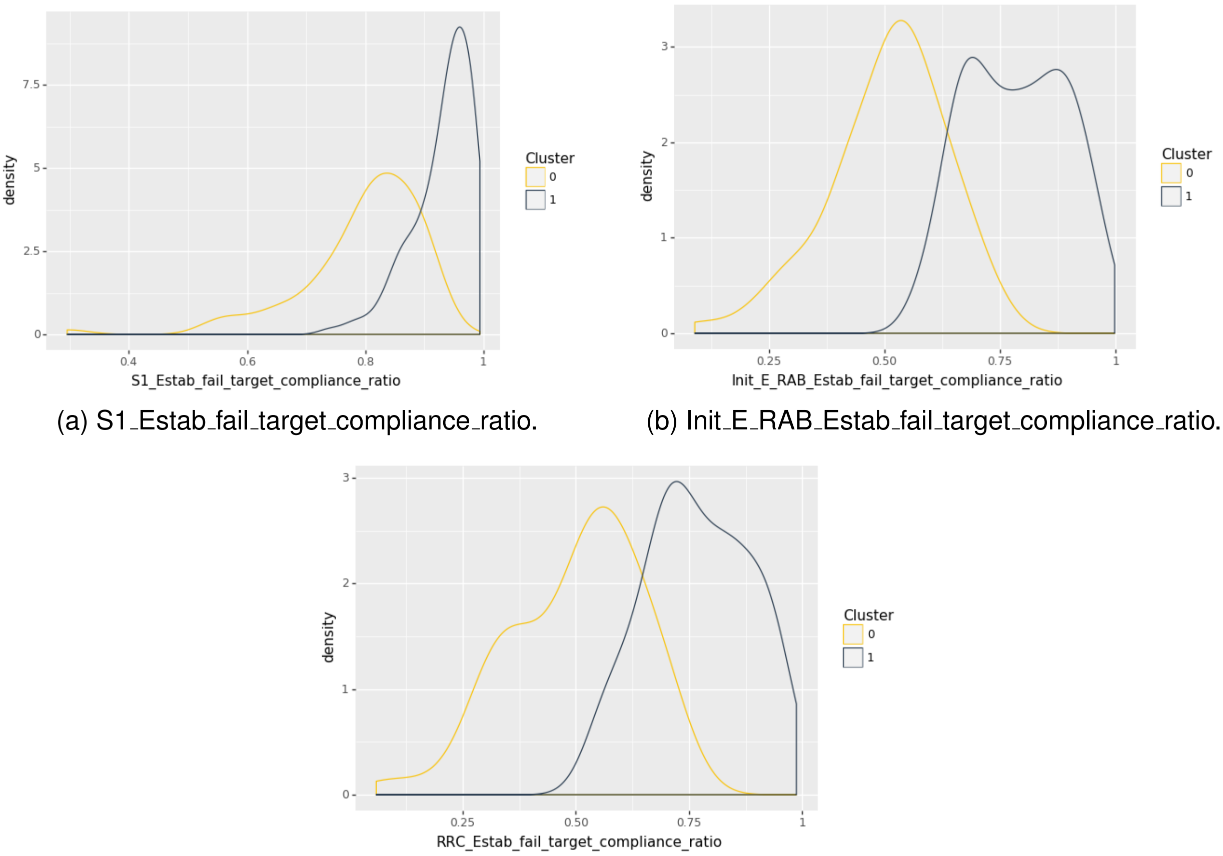
Cluster 1 contains approximately 64% of the cells being evaluated and presents an average overall performance, whereas, cluster 0, containing the remaining cells, presents a performance that is below average. From an operational perspective, the cluster scores would allow to radio engineers to focus their network optimization efforts towards the cluster 0.

Even though the main goal of the Kolmogorov-Smirnov test is to guarantee that the obtained clusters are statistically significant, it is also possible to gain knowledge regarding the features that contribute the most for the attained clustering partitions. Following the line of though explained in Section 4.5.2, regarding the obtained *p*-value through the Kolmogorov-Smirnov test, the features that have a lower *p*-value for a pair of clusters are the ones that better explain the difference between those clusters. The results from the Kolmogorov-Smirnov test, applied to clusters 0 and 1, are presented in Table 5.3, where the column *Feature* contains the names of the features extracted from the respective KPIs.

Feature	*p*-value
CB_RACH_fail_target_compliance_ratio	6×10^{-15}
Added_E_RAB_Estab_fail_target_compliance_ratio	2×10^{-11}
Init_E_RAB_Estab_fail_target_compliance_ratio	7×10^{-30}
RRC_Estab_fail_target_compliance_ratio	1×10^{-23}
S1_Estab_fail_target_compliance_ratio	4×10^{-25}

Table 5.3: Kolmogorov-Smirnov test for K-means in L800 (Accessibility).

The resulting *p*-value for each feature is very low when compared to the significance level of 0.01 that was considered. Thus, clusters 0 and 1 are statistically significant, for a 99% confidence interval, in all features. Furthermore, the visual comparison of the histograms of each feature allows to intuitively understand the features that are more distinct between clusters. Figure 5.3 shows the histograms of each cluster for the features that presented the lowest *p*-values in Table 5.3.



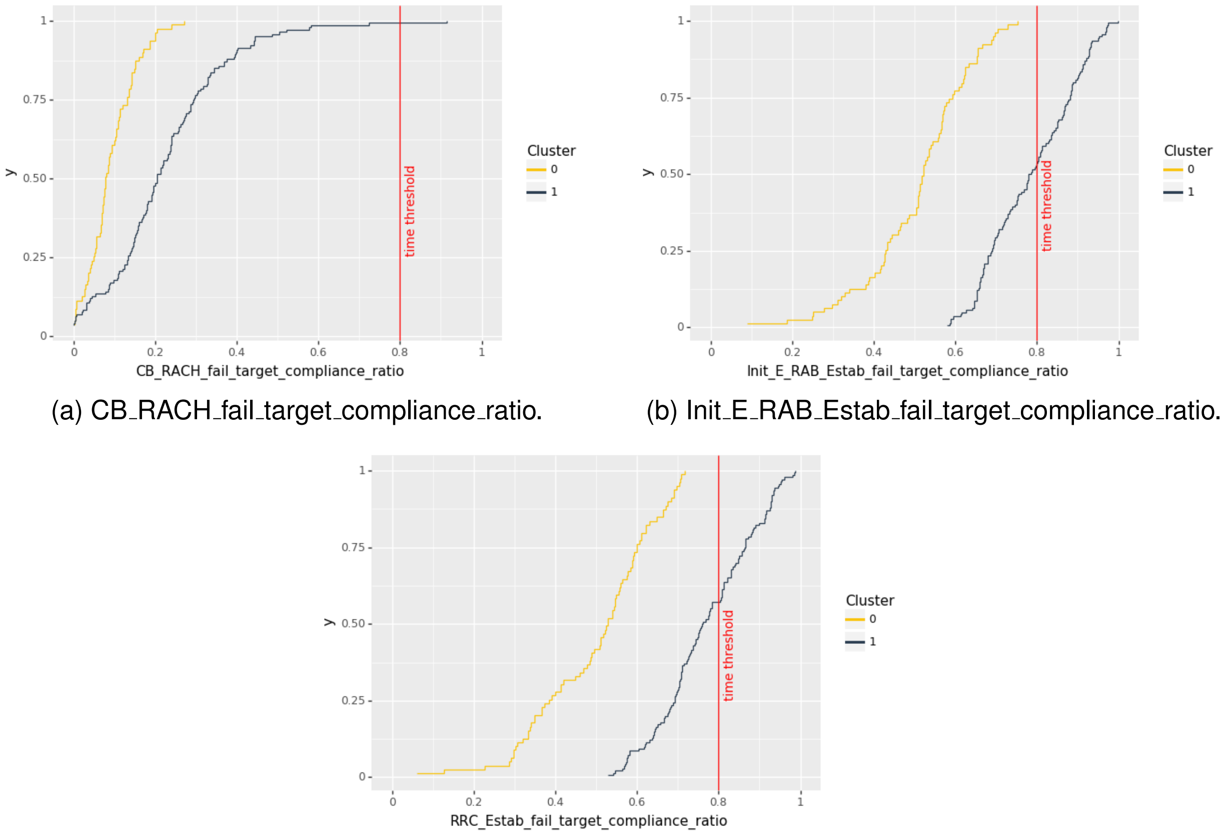
(a) S1_Estab_fail_target_compliance_ratio.

(b) Init_E_RAB_Estab_fail_target_compliance_ratio.

(c) RRC_Estab_fail_target_compliance_ratio.

Figure 5.3: Histograms of Accessibility features for K-means in L800.

Regarding the histograms analysis, the cluster 0 histogram is systematically worse than cluster 1, in terms of performance. Even though Figure 5.3 presents the three most distinct features between the two clusters, these features are not necessarily the ones that better explain the obtained scores for the clusters. The visualization of the ECDFs of the features for each cluster allows one to understand which are the features that most contribute to lower the score of the clusters. In Figure 5.4 are presented the ECDFs of the features that better explain the scores presented in Table 5.2.



(a) CB_RACH_fail_target_compliance_ratio.

(b) Init_E_RAB_Estab_fail_target_compliance_ratio.

(c) RRC_Estab_fail_target_compliance_ratio.

Figure 5.4: ECDFs of Accessibility features for K-means in L800.

It can be verified that the cells of both clusters have a poor performance regarding the CB_RACH_fail% KPI. For the other two features it can be seen that approximately half of the cells belonging to cluster 1 are compliant with the $time_{threshold}$ while none of the cells belonging to cluster 0 are compliant with that same value.

Since the behaviour that each cell presents is heavily influenced by its location, it can be interesting to observe if there are geographical areas with a high density of similarly performing cells. Figure 5.5 shows the cells location, with each cell being identified with the color of the cluster to which it belongs.

It is possible to identify a few areas where the concentration of cells belonging to one of the clusters, is predominant. Moreover, there is a specific area, highlighted in the figure with a blue circle, which is mainly populated with cells from cluster 0, thus being a zone with accessibility issues for the 800 MHz frequency band.

The joint analysis of the obtained cluster scores and their geographical distribution, presents valuable information from the network optimization perspective. It can be used to prioritize network optimization actions, that would have the most impact in network performance and respective network users' QoS.

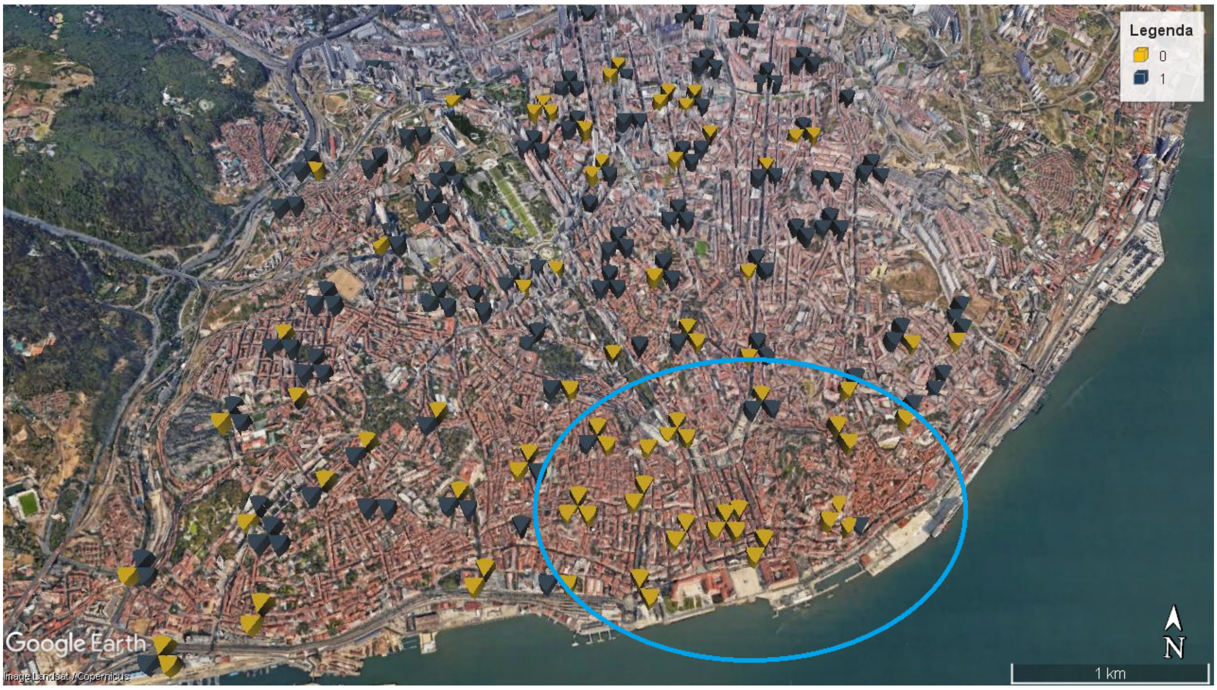


Figure 5.5: Clusters geographical distribution for Accessibility KPIs (L800).

Integrity

The first step before applying the methodology presented in Section 4.5 for the Integrity KPIs is to define the respective targets. Following the idea described in Section 5.1, and since all cells operating in this frequency band have a 10 MHz bandwidth, the set of targets for the Integrity KPIs with respect to this frequency band was obtained and is presented in Table 5.4.

KPI	Target [Mbps]
DL_Tput_per_UE(Mbps)	12.6
DL_Pdcp_Cell_Tput(Mbps)	7.2
DL_MAC_Cell_Tput(Mbps)	8.1
UL_Tput_per_UE(Mbps)	0.4
UL_Pdcp_Cell_Tput(Mbps)	0.59
UL_MAC_Cell_Tput(Mbps)	0.75

Table 5.4: Targets for Integrity KPIs (10 MHz bandwidth).

The optimal number of clusters obtained for each CVI, after executing the K-means for the different values of k, is shown in Figure 5.6.

Since each CVI has the same weight in the election system, the resulting optimal number of clusters is $k = 2$. The attained clusters can be visualized in Figure 5.7.

Cluster 0 and cluster 1 contain 130 and 89 cells, respectively. The quantitative score and respective classification of the clusters are presented in Table 5.5.

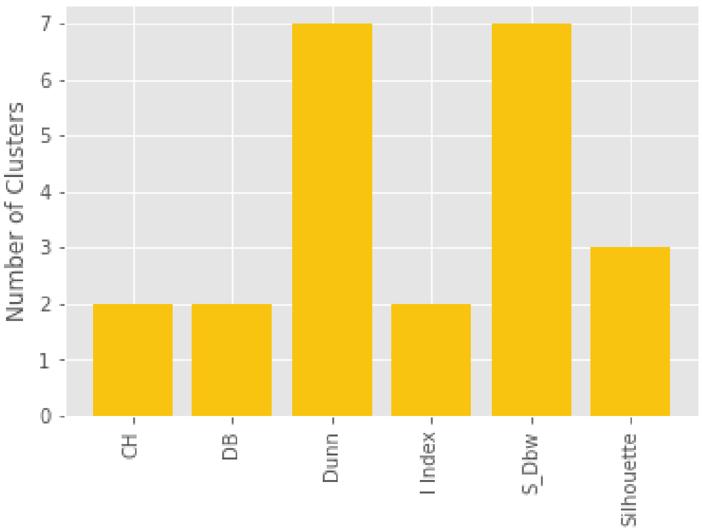


Figure 5.6: CVI results for K-means in L800 (Integrity).

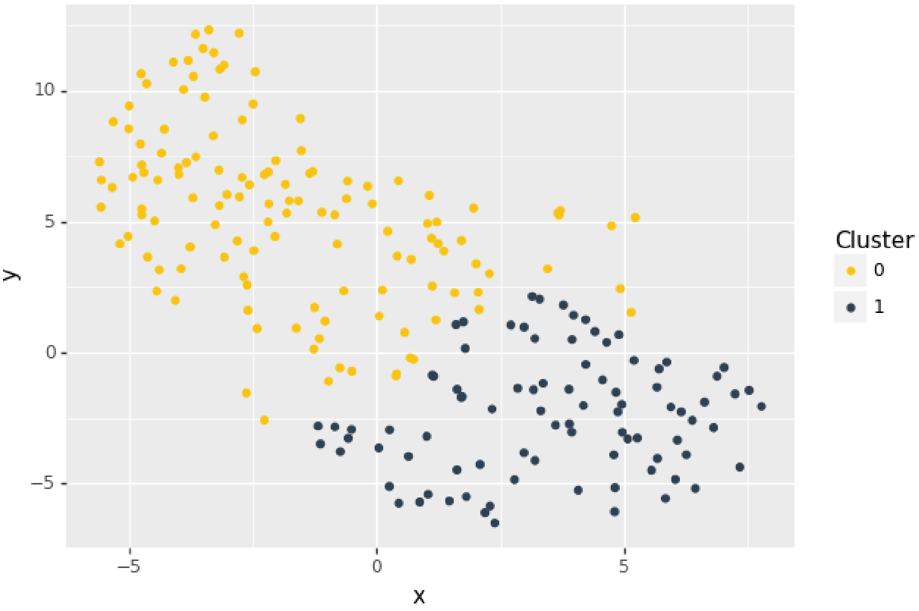


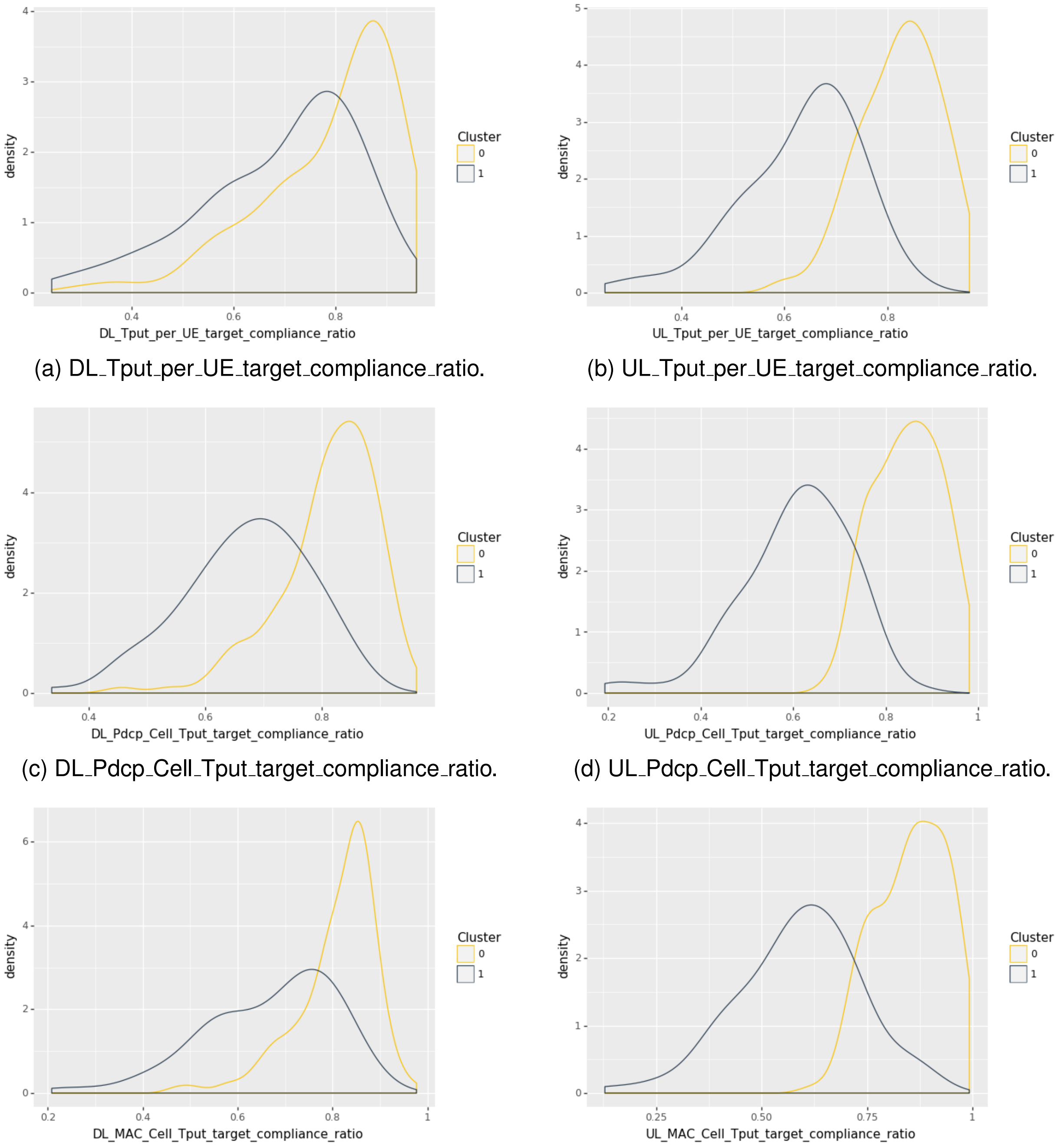
Figure 5.7: Clustering visualization for K-means in L800 (Integrity).

Cluster	Number of Cells	Score	Classification
0	130	0.67	Average
1	89	0.11	Unsatisfactory

Table 5.5: Cluster classification for K-means in L800 (Integrity).

It can be verified that the two obtained clusters have very distinct scores, with cluster 0 being categorized as average while cluster 1 is categorized as unsatisfactory. As mentioned in Section 5.1, the throughputs depend on the established services. Consequently, when a cell measures low throughputs, whether in uplink or in downlink, it does not necessarily imply that that cell is underperforming. Nonetheless, given the low targets that were shown in Table 5.4 when compared to the maximum theoretical throughputs [2], the score attained for cluster 1 may very well be an indicator that the cells belonging to that cluster do not present the desired behaviour.

Through the inspection of the histograms of the two clusters (Figure 5.8), it can be verified that they have very distinct behaviours, for each one of the features, as one would expect given the scores presented in Table 5.5.



(a) DL_Tput_per_UE_target_compliance_ratio.

(b) UL_Tput_per_UE_target_compliance_ratio.

(c) DL_Pdcp_Cell_Tput_target_compliance_ratio.

(d) UL_Pdcp_Cell_Tput_target_compliance_ratio.

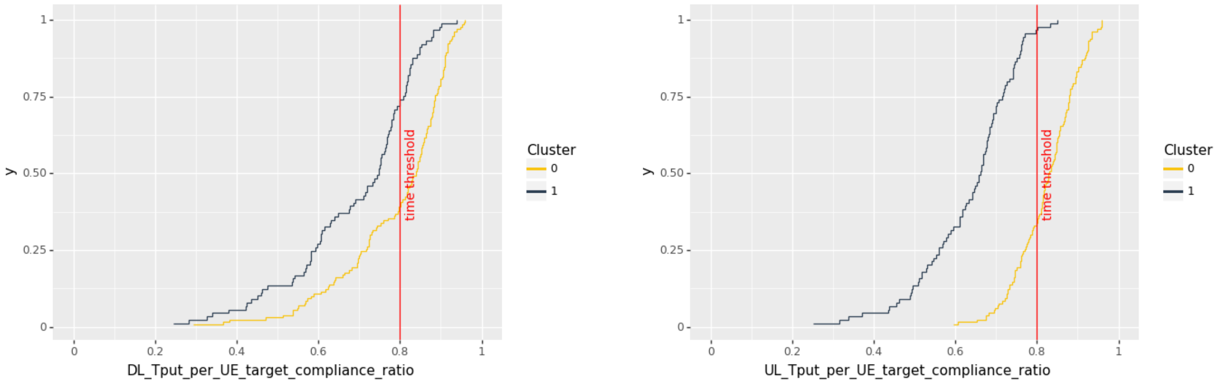
(e) DL_MAC_Cell_Tput_target_compliance_ratio.

(f) UL_MAC_Cell_Tput_target_compliance_ratio.

Figure 5.8: Histograms of Integrity features for K-means in L800.

The difference in the scores of the two clusters can be visually understood from Figure 5.9, which shows the ECDFs for the DL_Tput_per_UE_target_compliance_ratio and the UL_Tput_per_UE_target_compliance_ratio. The scores obtained for the

DL_Tput_per_UE_target_compliance_ratio and the UL_Tput_per_UE_target_compliance_ratio are representative of the scores obtained for the remaining downlink and uplink features, respectively.



(a) DL_Tput_per_UE_target_compliance_ratio.

(b) UL_Tput_per_UE_target_compliance_ratio.

Figure 5.9: ECDFs of Integrity features for K-means in L800.

Similarly to the Accessibility KPIs, the Integrity KPIs of a cell are also impacted by its location. Figure 5.10 illustrates the locations of the different cells belonging to clusters 0 and 1.

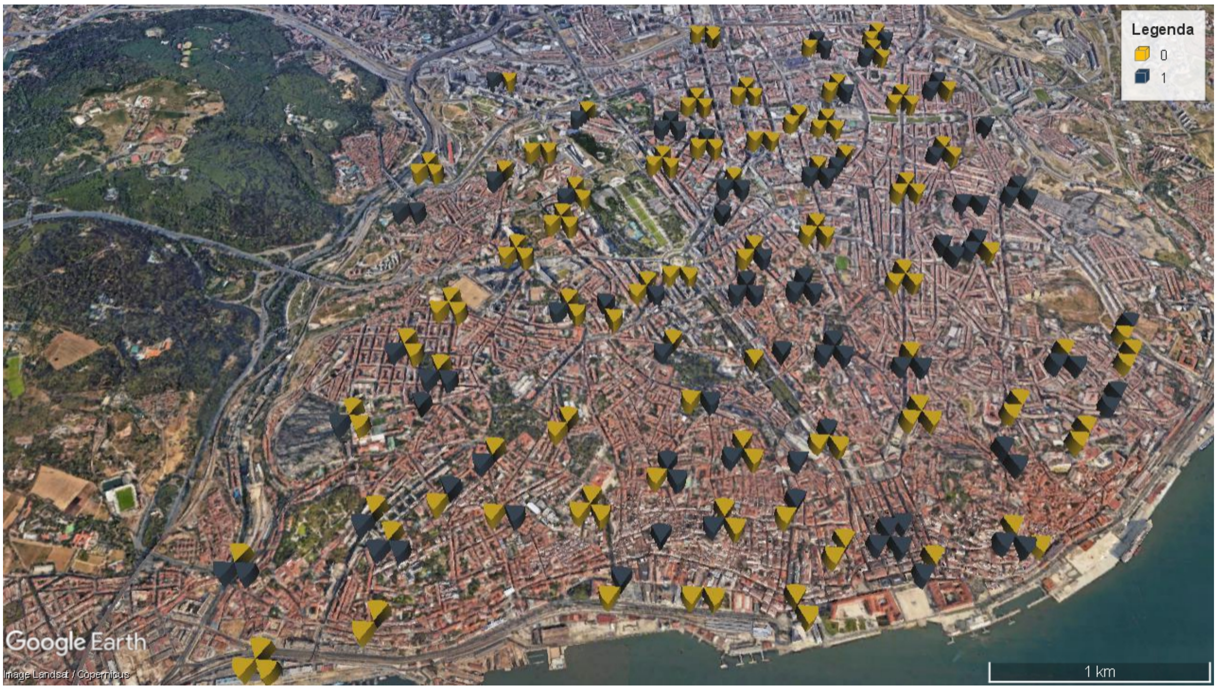


Figure 5.10: Clusters geographical distribution for Integrity KPIs (L800).

Through the visualization of Figure 5.10, it is possible to identify a few areas where cells predominantly belong to only one cluster. In this case, the goal is to find the areas that are mainly populated by cells belonging to cluster 1 since, given the overall performance of cluster 1, those areas should be the most problematic.

Cells from cluster 1 located in areas of high population density might be struggling with capacity issues. Also, the mobility strategy could be revised to optimize the network in this areas.

5.2.2 L1800

The clustering results, regarding Accessibility and Integrity, for cells operating in the L1800 frequency band are presented in this section. This frequency band contains 69 cells.

Accessibility

The feature engineering step is carried out using the targets defined in Table 5.1, and K-means is applied on the extracted features for the different values of k, as in Section 5.2.1. The acquired results regarding the optimal number of clusters for each CVI are presented in Figure 5.11.

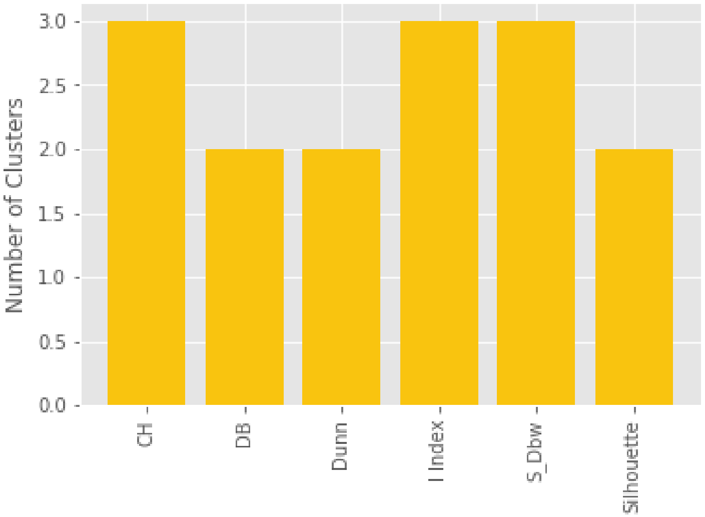


Figure 5.11: CVI results for K-means in L1800 (Accessibility).

From Figure 5.11, it can be observed that there was a tie between $k = 2$ and $k = 3$. In scenarios where a tie between two or more different configurations occurs, the configuration corresponding to higher number of clusters is selected. The rationale behind this decision is to have more granularity regarding the performance of the cells, since more groups with distinct behaviours are identified. Therefore, in this case the selected number of clusters was 3.

The obtained clusters with $k = 3$ can be visualized in Figure 5.12.

There are only 4 cells belonging to cluster 0, while clusters 1 and 2 have 39 and 26 cells, respectively. The scores and classification of the clusters are presented in Table 5.6

Cluster	Number of Cells	Score	Classification
0	4	0.7	Average
1	39	0.85	Above average
2	26	0.8	Above average

Table 5.6: Cluster classification for K-means in L1800 (Accessibility).

It can be verified that the performance level of cluster 0 is average while both clusters 1 and 2 exhibit above average performance. Figure 5.13 shows, through the histograms, which features present the

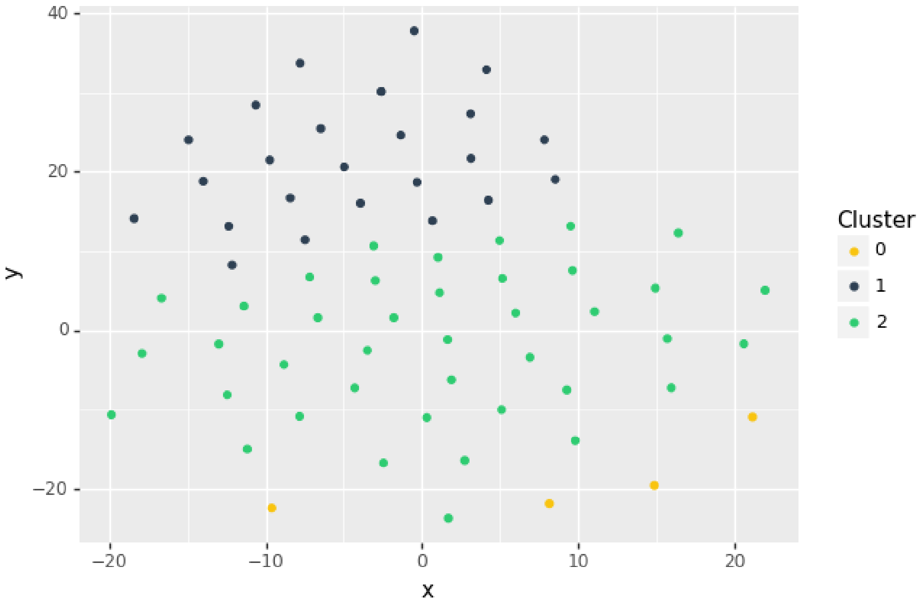


Figure 5.12: Clustering visualization for K-means in L1800 (Accessibility).

most distinct behaviour between clusters.

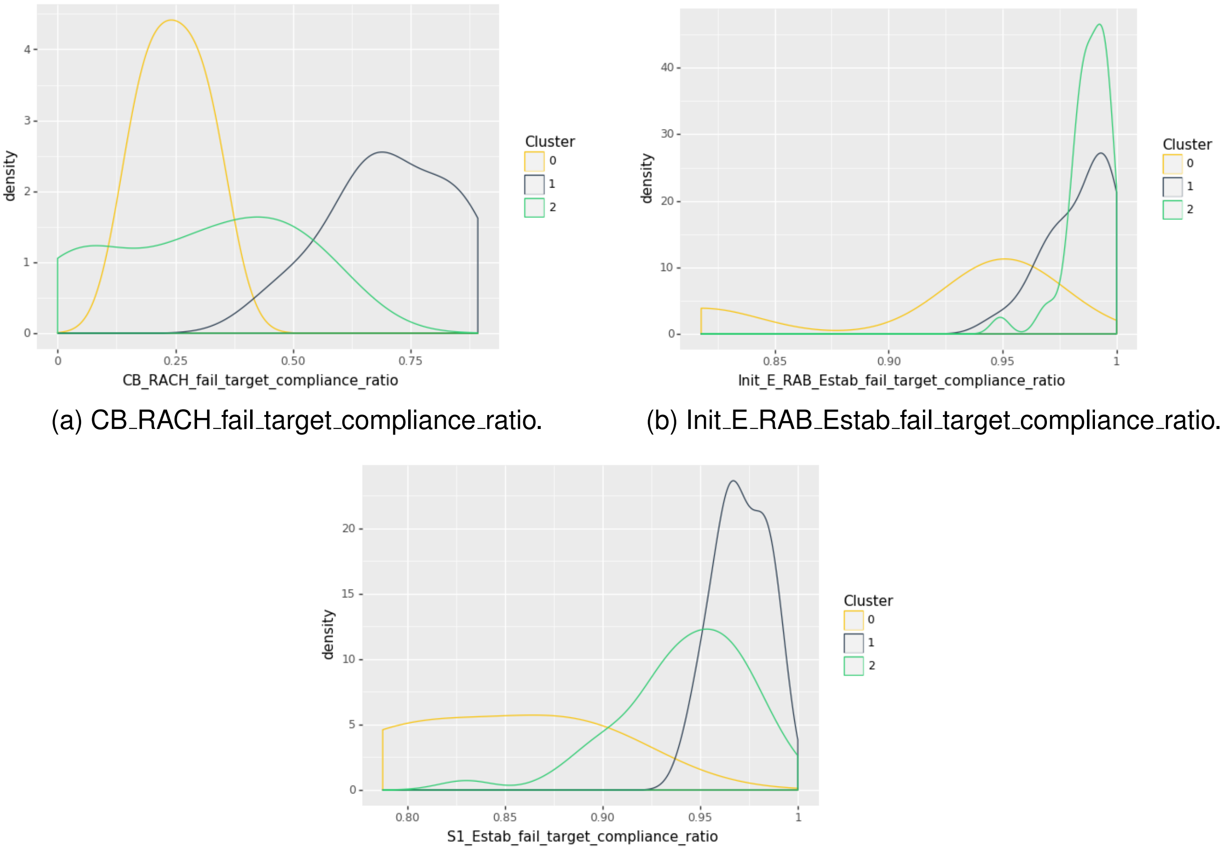
It can be verified that for the same features, their cluster respective histogram can be very different, indicating that, indeed, the cluster partition identified the main performance patterns in the data. Even though clusters 1 and 2 have very similar scores, which translates into having the same classification, it can be observed in Figure 5.13 that, for the CB_RACH_fail_target_compliance_ratio feature, they present very distinct behaviours.

Furthermore, the ECDFs of the features that explain the differences in the scores of the clusters are shown in Figure 5.14.

It can be verified that all clusters exhibit problems regarding the CB_RACH_fail_target_compliance_ratio, with only cluster 1 having a score above 0 for the corresponding extracted feature. Moreover, the lower score of cluster 0 can be explained through the score obtained obtained for the RRC_Estab_fail_target_compliance_ratio, which is due to the fact that this cluster only contains 4 cells and one of them is not compliant with the target defined for the RRC_Estab_fail% KPI for a period of time above $time_{threshold}$.

The location of each cell is illustrated in Figure 5.15.

The number of cells operating in this frequency band is much lower when compared with the number of cells operating in the L800 frequency band. As a result the locations of the cells are much more disperse. In this case, given that cluster 0 only contains 4 cells and clusters 1 and 2 have a similar score, with both having above average performance, it is the harder to find a relation between the locations of the cells and their performance. Nonetheless, it is still possible to identify a few areas where the cells predominantly belong to cluster 1 or cluster 2.

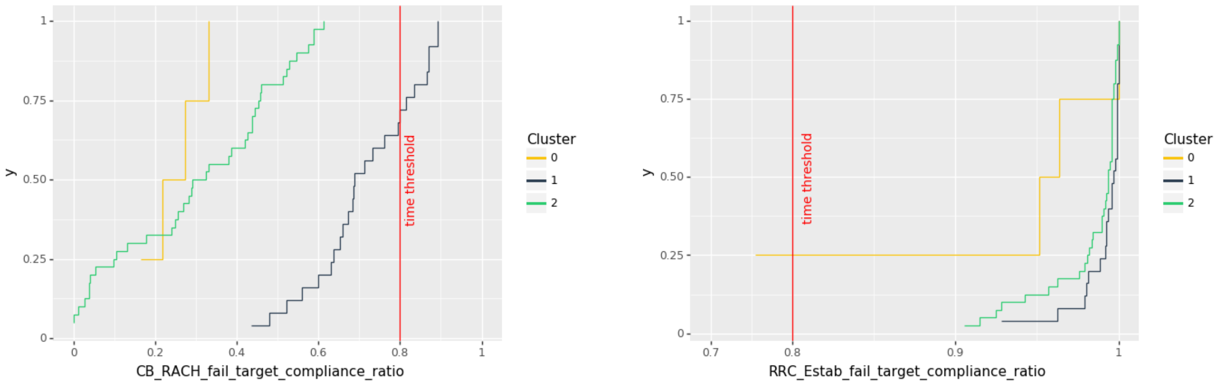


(a) CB_RACH_fail_target_compliance_ratio.

(b) Init_E_RAB_Estab_fail_target_compliance_ratio.

(c) S1_Estab_fail_target_compliance_ratio.

Figure 5.13: Histograms of Accessibility features for K-means in L1800.



(a) CB_RACH_fail_target_compliance_ratio.

(b) RRC_Estab_fail_target_compliance_ratio.

Figure 5.14: ECDFs of Accessibility features for K-means in L1800.

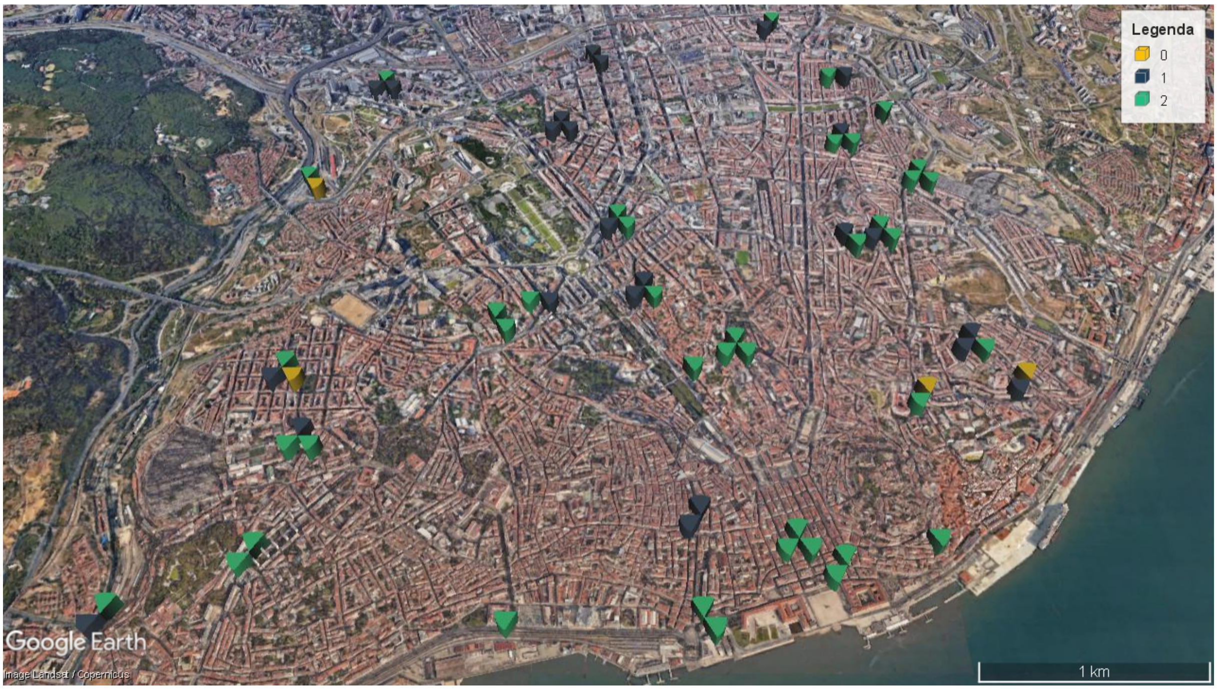


Figure 5.15: Clusters geographical distribution for Accessibility KPIs (L1800).

Integrity

All the cells operating on this frequency band have a 20 MHz bandwidth. For this bandwidth, the set of targets for the Integrity KPIs, obtained as described in Section 5.1, is presented in Table 5.7.

KPI	Target [Mbps]
DL_Tput_per_UE(Mbps)	15.8
DL_Pdcp_Cell_Tput(Mbps)	8.6
DL_MAC_Cell_Tput(Mbps)	9.5
UL_Tput_per_UE(Mbps)	0.5
UL_Pdcp_Cell_Tput(Mbps)	0.68
UL_MAC_Cell_Tput(Mbps)	0.99

Table 5.7: Targets for Integrity KPIs (20 MHz bandwidth).

The feature engineering step is then applied to the Integrity KPIs of the cells operating in the 1800 MHz, using the targets of Table 5.7. The optimal number of clusters, with respect to the CVIs, is shown in Figure 5.16.

From the examination of Figure 5.16, it results that $k = 2$ provides the best partitioning of the dataset being clustered. The resulting clusters can be visualized in Figure 5.17.

Cluster 1 only contains 9 cells while cluster 0 contains 60 cells, thus being the most representative cluster of the dataset. Table 5.8 shows the score for each cluster as well as its overall performance classification, regarding the Integrity class.

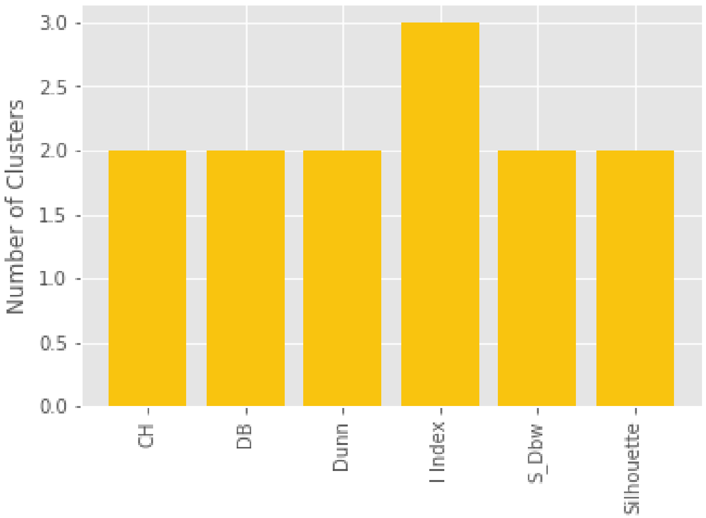


Figure 5.16: CVI results for K-means in L1800 (Integrity).

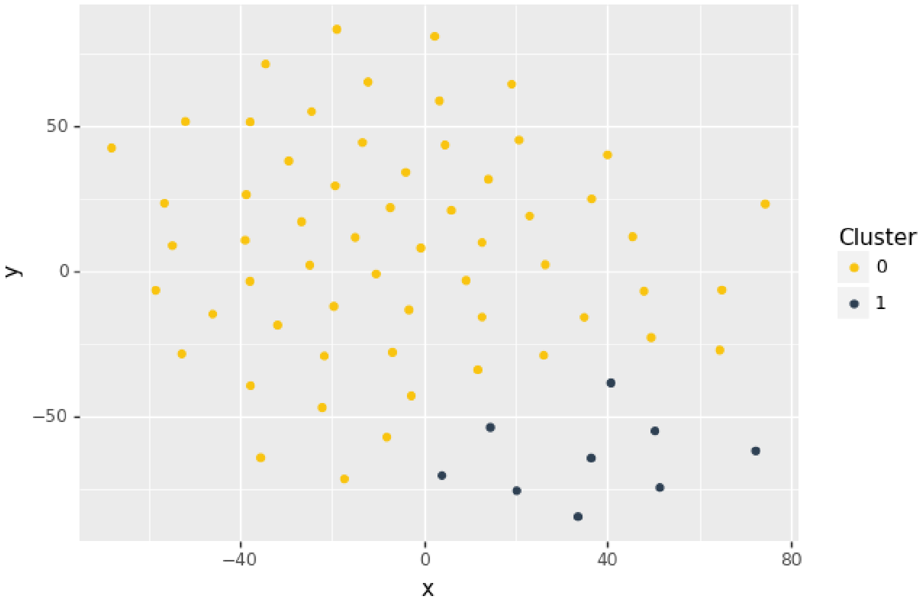


Figure 5.17: Clustering visualization for K-means in L1800 (Integrity).

Cluster	Number of Cells	Score	Classification
0	60	0.99	Above average
1	9	0.72	Average

Table 5.8: Cluster classification for K-means in L1800 (Integrity) and $time_{threshold} = 0.8$.

When comparing the histograms of the features of each cluster, it was verified an identical situation to the one observed in Section 5.2.1, where, for each feature, the histogram relative to the cluster with the lower score is shifted to the left and the histogram of the cluster with the higher score is shifted to the right.

It is interesting to see how the classification of a cluster changes by changing the value of $time_{threshold}$. Let us consider a more demanding cluster evaluation by setting $time_{threshold} = 0.9$

instead of $time_{threshold} = 0.8$, which was the value used thus far. The scores and classification obtained in this case are presented in Table 5.9.

Cluster	Number of Cells	Score	Classification
0	60	0.83	Above average
1	9	0.25	Unsatisfactory

Table 5.9: Cluster classification for K-means in L1800 (Integrity) and $time_{threshold} = 0.9$.

Comparing Tables 5.8 and 5.9, it can be verified that cluster 0 still has an above average performance. However, the score of cluster 1 drops drastically after changing the value of $time_{threshold}$, being classified as unsatisfactory.

Figure 5.18 exemplifies the difference in the score obtained for the feature UL_Tput_per_UE_target_compliance_ratio with $time_{threshold}$ set to 0.8 and 0.9.

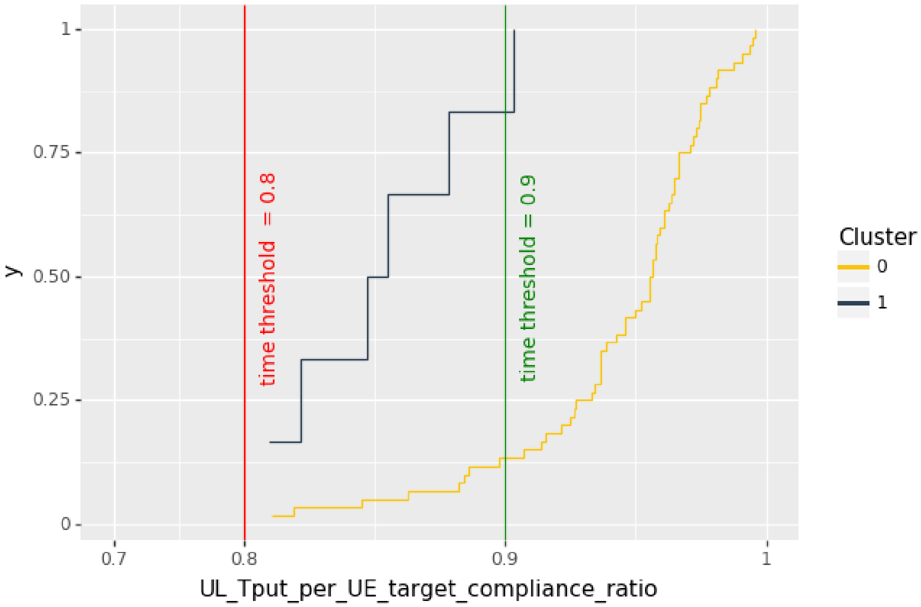


Figure 5.18: Score comparison for different $time_{threshold}$ values.

It can be verified that the ECDF of cluster 1 is shifted to the right of the red line, representing $time_{threshold} = 0.8$, thus the score obtained with respect to that value is 1. Contrarily, for $time_{threshold} = 0.9$, identified with the green line, the score for cluster 0 is nearly 0, with only one cell being compliant with the target for a period of time above the $time_{threshold}$. This demonstrates the importance of setting the value of $time_{threshold}$ to a value that is aligned with the expectations of the network operator, regarding the performance of the network.

The geographic location of the cells is shown in Figure 5.19. Cluster 0 contains about 86% of the cells of the dataset, therefore it is only natural that there are areas that are mainly populated with cells of cluster 0. Regarding the cells from cluster 1, it were not identified any particular small areas with a large concentration of cells, since they have disperse locations.

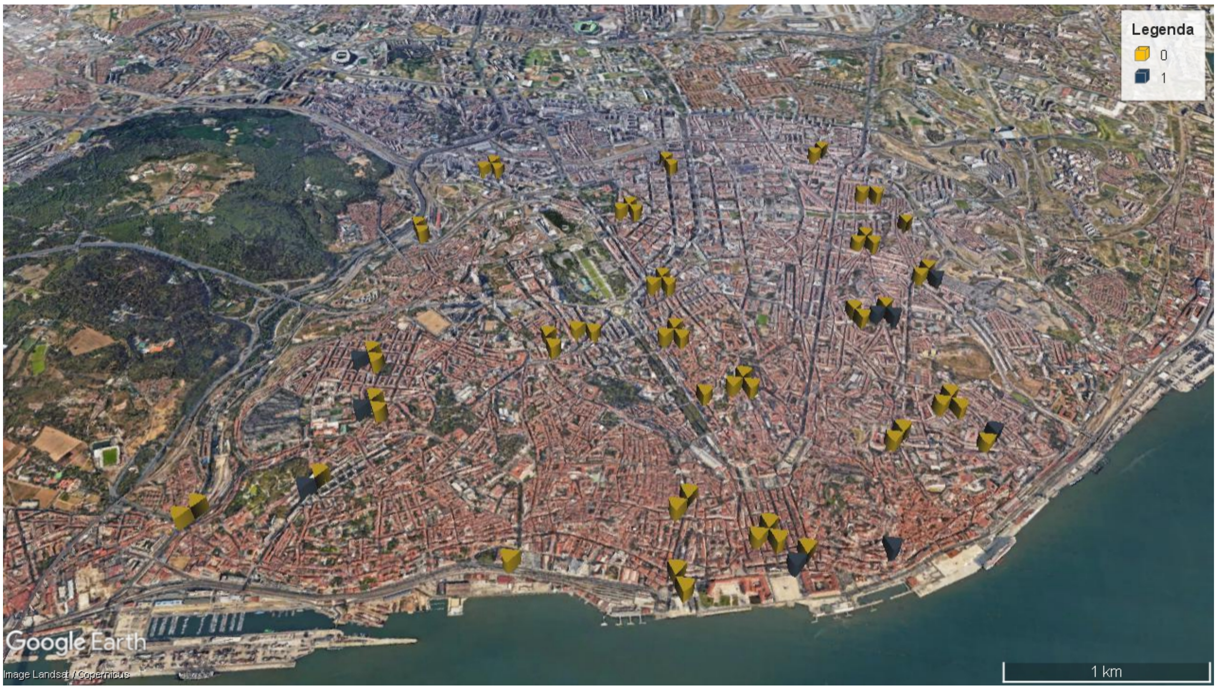


Figure 5.19: Clusters geographical distribution for Integrity KPIs (L1800).

Availability

For the Availability KPIs, the target specified in the feature extraction process, was 99.7% for the three KPIs that constitute this class. It could be argued that only the Cell_Avail_perc could be used, as this KPI presents the overall availability of the cell, thus containing the information provided by the other two KPIs. However, this KPI also provides insight about the cell sleep mode. Thus, if one would analyze the availability of the cells, based solely on the Cell_Avail_perc KPI, it could be mislead into thinking that the cell was unavailable due to a fault or a reconfiguration request when in fact it was in sleeping mode.

The optimal number of clusters obtained for each CVI is shown in Figure 5.20.

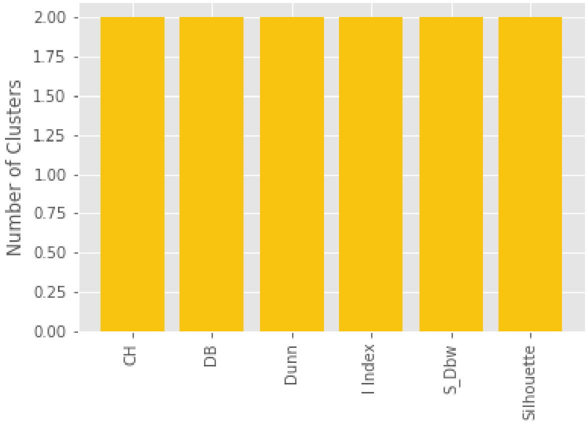


Figure 5.20: CVI results for K-means in L1800 (Availability).

In this case, all CVIs are in accordance relatively to the optimal number of clusters, which is $k = 2$. The resulting clusters can be visualized in Figure 5.21. Cluster 0 contains 44 cells while cluster 1 contains 25 cells.

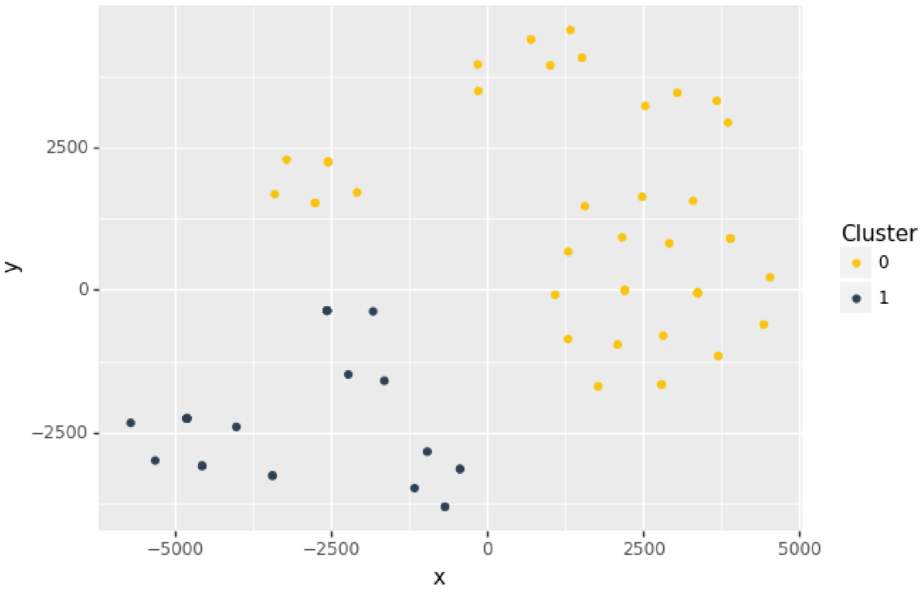
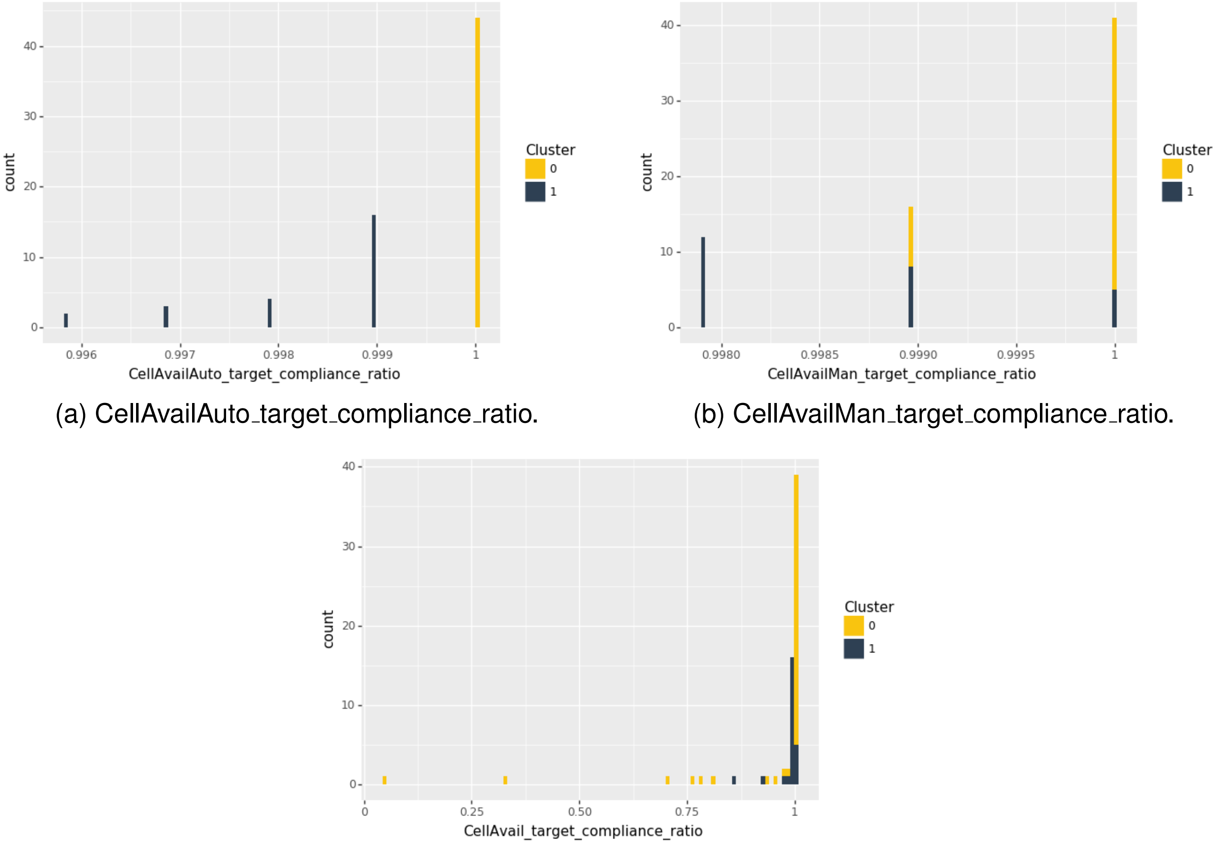


Figure 5.21: Clustering visualization for K-means in L1800 (Availability).

Through the visualization of the histograms for each feature, in Figure 5.22, the situation explained above is verified.



(a) CellAvailAuto_target_compliance_ratio.

(b) CellAvailMan_target_compliance_ratio.

(c) CellAvail_target_compliance_ratio.

Figure 5.22: Histograms of Availability features for K-means in L1800.

It can be observed that even though both clusters would present a score close to 1 for the features

CellAvailAuto_target_compliance_ratio and CellAvailMan_target_compliance_ratio, the same would not be verified for cluster 0 in regard to CellAvail_target_compliance_ratio. In other words, it could be perceived, through the analysis of the feature CellAvail_target_compliance_ratio, that some cells exhibit availability problems. However, through the analysis of the other two features, it can be inferred that the cells that presented a lower value for the CellAvail_target_compliance_ratio, were in fact in a sleeping mode state, which is not related with a performance issue but rather with a mechanism to improve energy efficiency.

Thus, the overall score for the Availability should be computed taking into account only the features CellAvailAuto_target_compliance_ratio and CellAvailMan_target_compliance_ratio. In this case it is straightforward to see that both clusters would have a score near 1, even for a value of $time_{threshold}$ as high as 0.995.

5.2.3 L2600

The clustering results for the L2600 frequency band, using K-means, are presented in this section. The PM dataset contains 121 cells operating on this frequency band.

Accessibility

Following the procedure detailed in the previous sections, the optimal number of clusters for each CVIs were attained. Figure 5.23 illustrates the optimal number of clusters for each CVI.

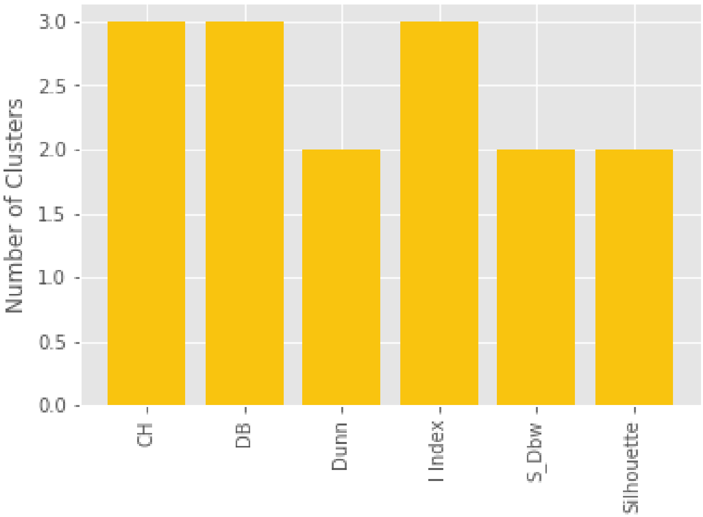


Figure 5.23: CVI results for K-means in L2600 (Accessibility).

It can be verified that there is a tie between $k = 2$ and $k = 3$. Adopting the same line of thought presented in Section 5.2.2, for the clustering of Accessibility features, it was selected $k = 3$. The clusters can be visualized in Figure 5.24.

Clusters 0 and 2 are the most representative of the dataset, containing 80 and 35 cells respectively. There are only 6 cells belonging to cluster 1. The scores and respective classification for each cluster are presented in Table 5.10.

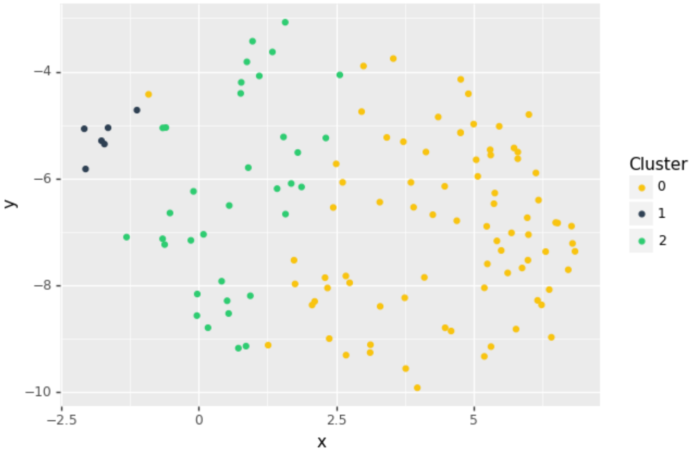


Figure 5.24: Clustering visualization for K-means in L2600 (Accessibility).

Cluster	Number of Cells	Score	Classification
0	80	0.89	Above average
1	6	0.57	Average
2	35	0.79	Above average

Table 5.10: Cluster classification for K-means in L2600 (Accessibility).

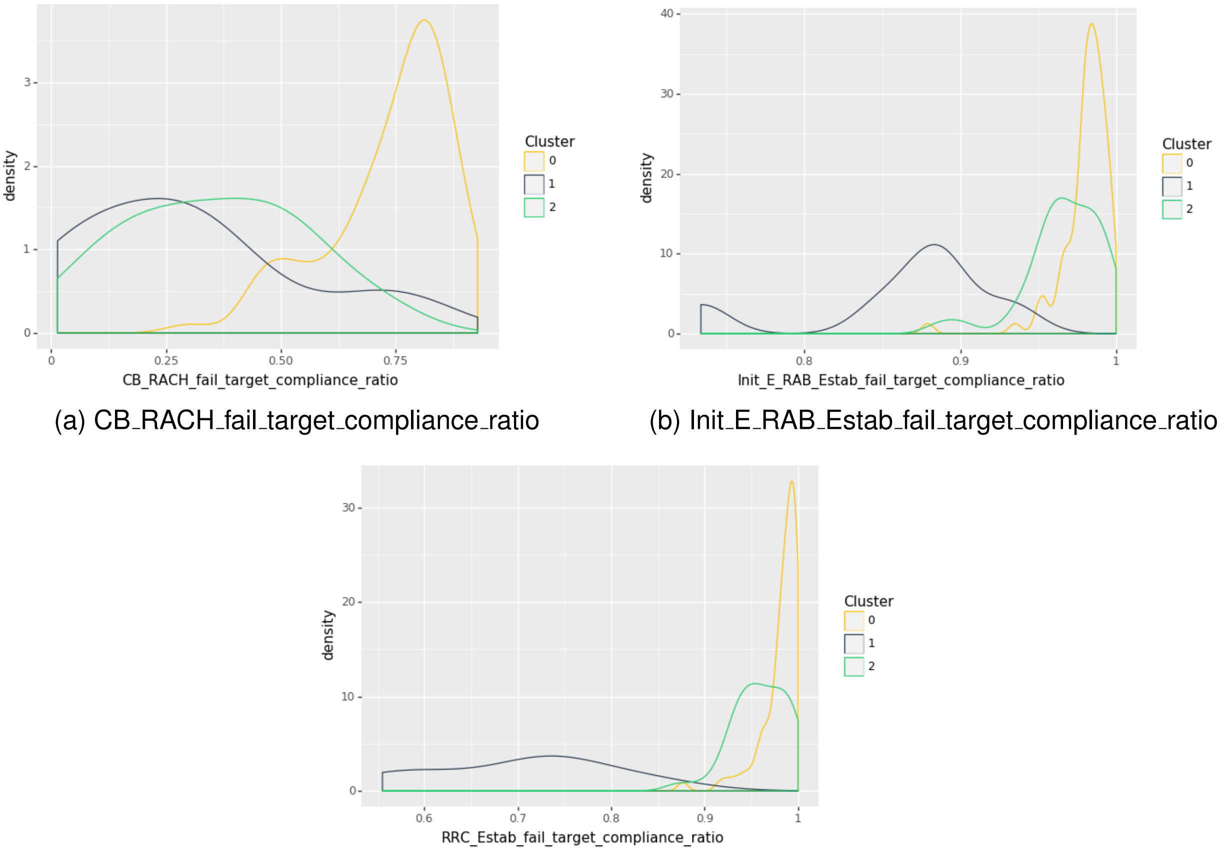
Cluster 0 and cluster 2 were both classified with above average performance, attaining a 0.89 and 0.79 score, respectively. Cluster 0 only scored 0.57, thus having an average performance. Figure 5.25 shows the histograms regarding the features that better explain the obtained partitioning.

It can be observed that the main difference regarding clusters 0 and 2 lies on the CB_RACH_target_compliance_ratio feature. Regarding cluster 1, it can be verified that its behaviour differs from the other two clusters for both the Init_E_RAB_target_compliance_ratio and RRC_Estab_fail_target_compliance_ratio. Moreover, the cluster 1 also presents a distinct behaviour for the CB_RACH_target_compliance_ratio feature when comparing to the one of cluster 2.

The scores of each cluster can be mainly explained through the ECDFs for the RRC_Estab_fail_target_compliance_ratio and CB_RACH_target_compliance_ratio. These are shown in Figure 5.26.

It can be verified that for the CB_RACH_target_compliance_ratio feature, both cluster 0 and cluster 1 present a score of 0, since their probability of having a cell that is compliant with the target, for at least 80% of the time (*i.e.* $time_{threshold} = 0.8$), is zero. Regarding the RRC_Estab_fail_target_compliance_ratio, the cluster 1 performs worse than clusters 0 and 2, which have a score of 1.

The location of the cells of each cluster is presented in Figure 5.27.

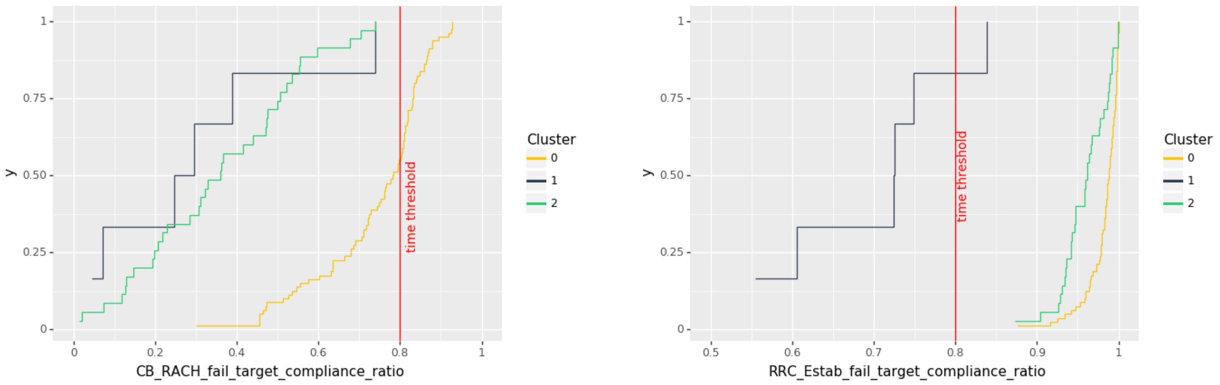


(a) CB_RACH_fail_target_compliance_ratio

(b) Init_E_RAB_Estab_fail_target_compliance_ratio

(c) RRC_Estab_fail_target_compliance_ratio

Figure 5.25: Histograms of Accessibility features for K-means in L2600.



(a) CB_RACH_fail_target_compliance_ratio

(b) RRC_Estab_fail_target_compliance_ratio

Figure 5.26: ECDFs of Accessibility features for K-means in L2600.

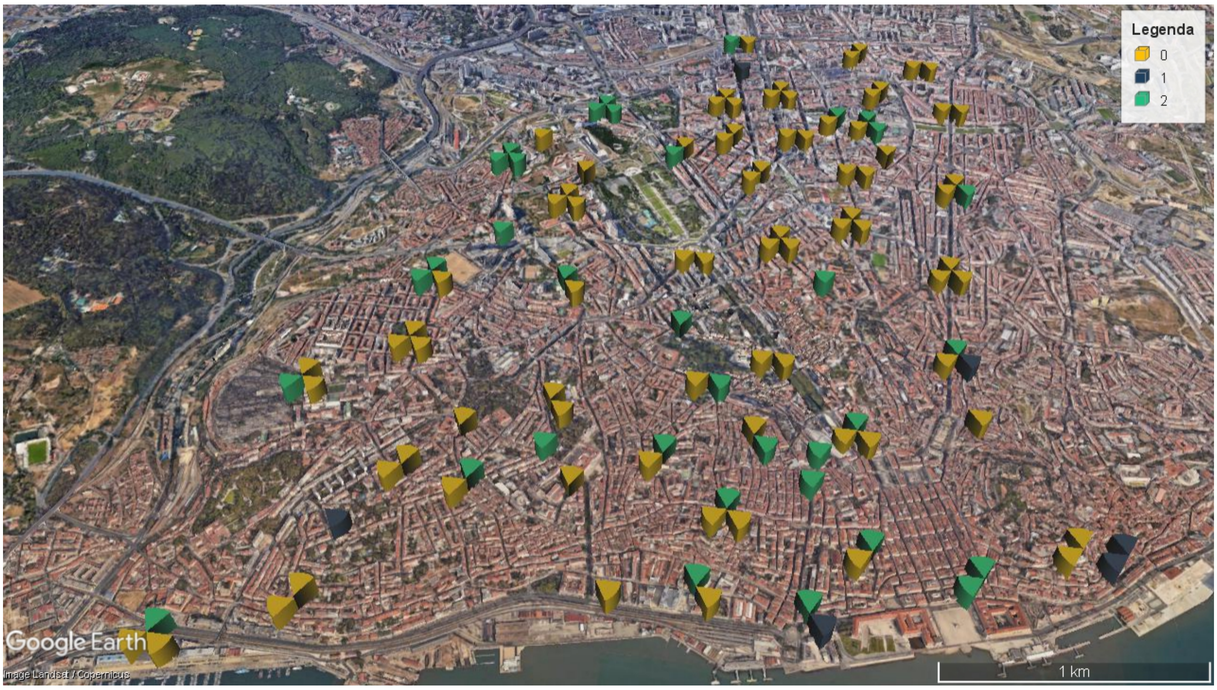


Figure 5.27: Clusters geographical distribution for Accessibility KPIs (L2600).

Integrity

Regarding the Integrity KPIs, the optimal number of clusters with respect to the CVI is presented in Figure 5.28.

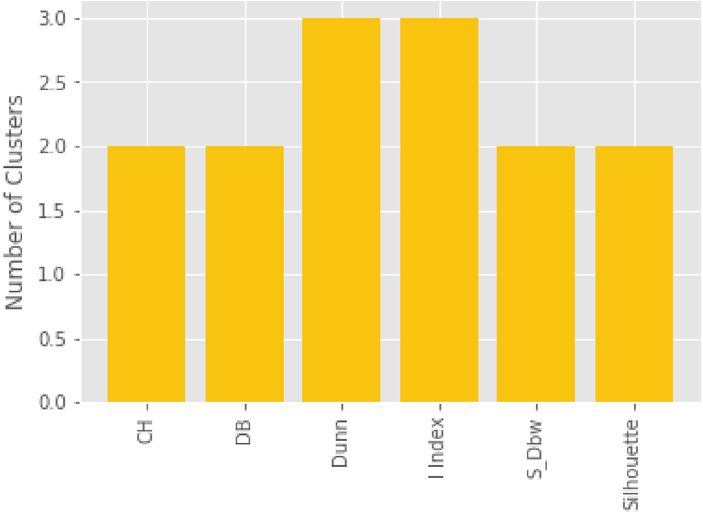


Figure 5.28: CVI results for K-means in L2600 (Integrity).

It can be verified that the most voted optimal number of clusters was $k = 2$. The obtained partitioning for $k = 2$ can be visualized in Figure 5.29.

Cluster 0 contains 101 cells, which is approximately 83% of the cells in the data set. The remaining 20 cells belong to cluster 1. The scores and classification of each cluster can be observed in Table 5.11.

It can be verified that cluster 0 has a nearly ideal performance. Conversely, cluster 0 is classified as average. The differences in the performance classification of the two clusters can be explained through

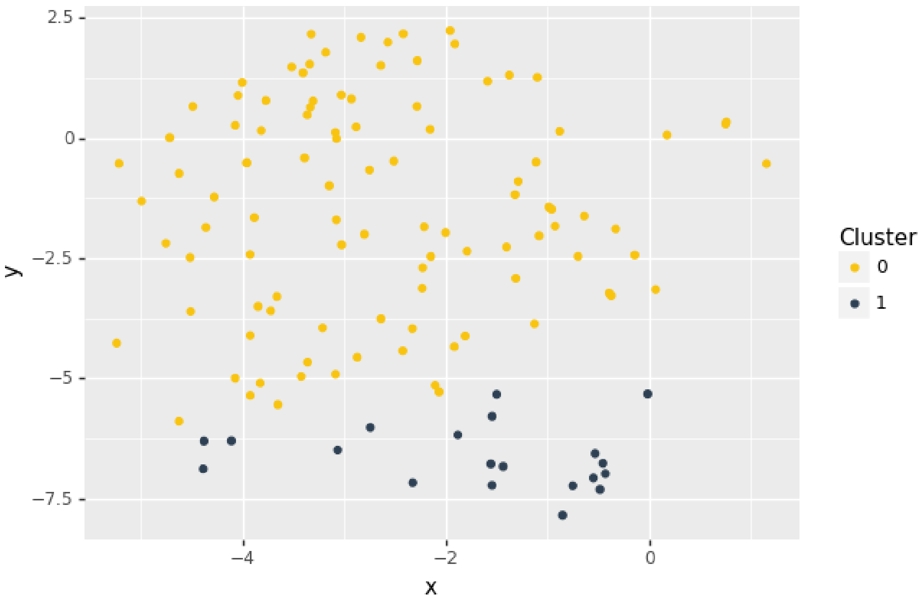
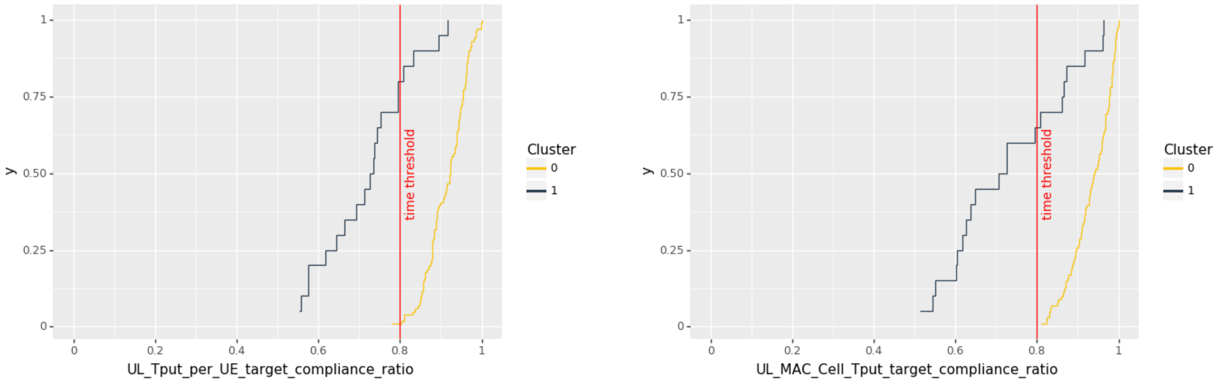


Figure 5.29: Clustering visualization for K-means in L2600 (Integrity).

Cluster	Number of Cells	Score	Classification
0	101	0.99	Above average
1	20	0.56	Average

Table 5.11: Cluster classification for K-means in L2600 (Integrity).

the ECDFs of the features related to the uplink transmission. These ECDFs are shown in Figure 5.30.



(a) UL_Tput_per_UE_target_compliance_ratio

(b) UL_MAC_Cell_Tput_target_compliance_ratio

Figure 5.30: ECDFs of Integrity features for K-means in L2600.

It can be observed that the probability of a cell belonging to cluster 0 being compliant with the $time_{threshold}$, for each one of the features presented, is approximately 1, while for cluster 1, the score for each feature is lower than 0.5.

The deployment of the cells on the urban area can be visualized in Figure 5.31.

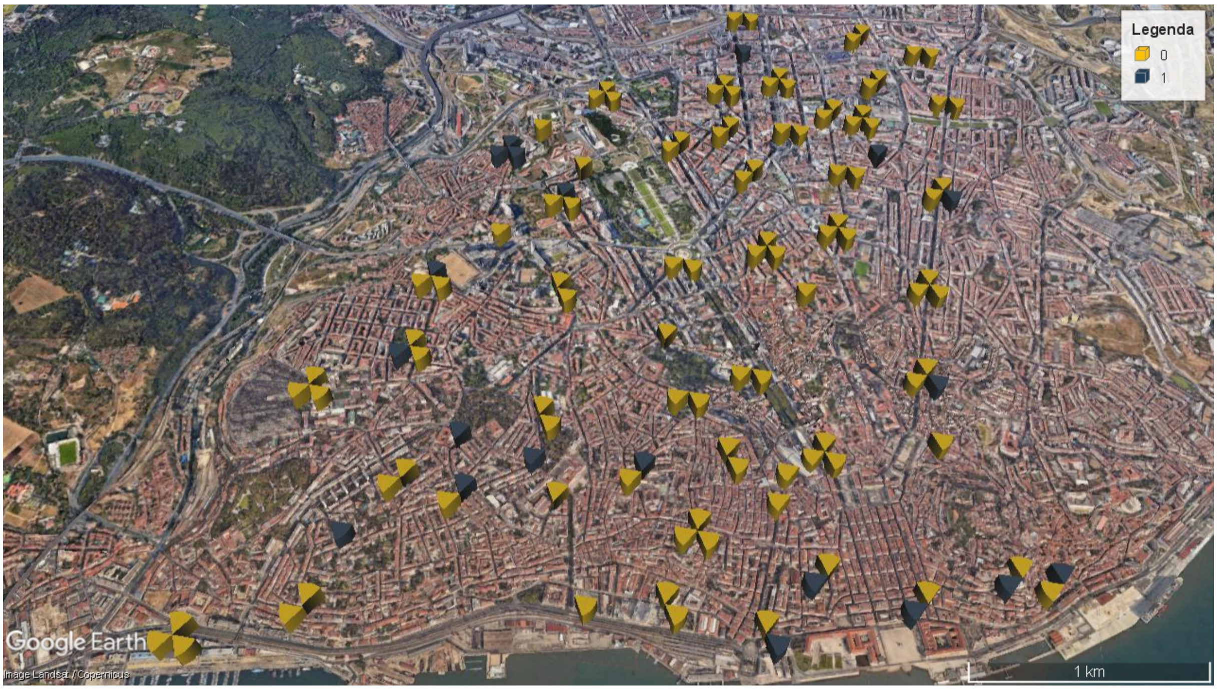


Figure 5.31: Clusters geographical distribution for Integrity KPIs (L2600).

Given the fact that cluster 0 contains most of the cells operating in the 2600 MHz frequency band, it comes as no surprise that there are areas predominantly occupied by cells of cluster 0. Regarding the cells of cluster 1, they are sparsely located over the demarcated geographic region.

5.3 Clustering Algorithms Comparison

This section presents an overview over the results obtained with the three clustering algorithms tested: K-means, EM with GMM, and Spectral Clustering. Only the Accessibility and Integrity classes were tested, as the Availability class, due to the reduced number of features and nature of the same, is quite straightforward to analyze.

Apart from the number of clusters, the EM with GMM algorithm, from the Scikit-learn library, also includes the cov_type which is a parameter that can be tuned to specify the type of covariance parameters to be used [29]: spherical, diagonal, tied or full.

Spectral Clustering requires the number of clusters to be specified beforehand, similarly to both K-means and GMM. Additionally, another input parameter, $gamma$, was configured. This parameter is the kernel coefficient for the RBF [29]. The set of values considered for $gamma$ was $[0.01, 0.1, 1]$.

The same approach, using the election mechanism, was used for both EM with GMM and Spectral Clustering, with the difference that the optimal configuration now includes two input parameters instead of just one, as in K-means.

5.3.1 L800

For the L800 frequency band, the results attained with each clustering algorithm are summarized in Table 5.12.

KPI Class	Algorithm	Clusters	Number of cells	Score	Classification
Accessibility	K-means	2	79	0.32	Below average
			140	0.57	Average
	EM with GMM	2	15	0.2	Unsatisfactory
			204	0.5	Average
	Spectral	2	89	0.33	Below average
			130	0.58	Average
Integrity	K-means	2	89	0.11	Unsatisfactory
			130	0.67	Average
	EM with GMM	2	68	0.09	Unsatisfactory
			151	0.6	Average
	Spectral	2	90	0.11	Unsatisfactory
			129	0.68	Average

Table 5.12: Results comparison for L800.

It can be verified that there are no significant differences between the results from different clustering algorithms.

The most distinct one corresponds to the Accessibility class, where the clusters resulting from the EM with GMM were much more unbalanced with respect to the number of cell that they contain. Furthermore, it was able to identify a more specific cluster, containing less cells, that is characterized by an unsatisfactory performance. However, there were quite a few cells that were included in the cluster with average performance that would have been in a below average performing cluster if K-means or Spectral Clustering were used.

5.3.2 L1800

Regarding the L1800 frequency band, the results obtained are condensed in Table 5.13.

It can be observed that the Spectral Clustering algorithm was not able to partition the data into clusters with specific performances, since all clusters are classified as above average.

For the Integrity class, both K-means and EM with GMM managed to identify one cluster that presents average performance, thus being able to partition the data into two clusters with different performances.

Regarding the Accessibility class, both K-means and and EM with GMM partitioned the cells into three clusters. Even though all the clusters attained for EM with GMM have the same classification, it can be inferred that this result is similar to the one of K-means, as the cluster that presents a different classification only differs in one cell.

KPI Class	Algorithm	Clusters	Number of cells	Score	Classification
Accessibility	K-means	3	4	0.7	Average
			26	0.8	Above average
			39	0.85	Above average
	EM with GMM	3	5	0.76	Above average
			31	0.86	Above average
			33	0.83	Above average
	Spectral	2	18	0.81	Above average
			51	0.85	Above average
Integrity	K-means	2	9	0.72	Average
			60	0.99	Above average
	EM with GMM	2	6	0.58	Average
			63	0.99	Above average
	Spectral	2	20	0.85	Above Average
			49	1	Above Average

Table 5.13: Results comparison for L1800.

5.3.3 L2600

The outcome of each clustering algorithm regarding the Accessibility and Integrity classes of KPIs are shown in Table 5.14.

KPI Class	Algorithm	Clusters	Number of cells	Score	Classification
Accessibility	K-means	3	6	0.57	Average
			35	0.79	Above average
			80	0.89	Above average
	EM with GMM	2	14	0.71	Average
			107	0.86	Above average
	Spectral	3	8	0.63	Average
			32	0.8	Above average
			81	0.88	Above average
Integrity	K-means	2	20	0.56	Average
			101	0.99	Above average
	EM with GMM	2	15	0.53	Average
			106	0.97	Above average
	Spectral	2	21	0.57	Average
			100	0.99	Above average

Table 5.14: Results comparison for L2600.

Regarding the Integrity class, the results are, once again, very similar for the considered algorithms. However, for the Accessibility class, it can be observed that the optimal number of clusters for both Spectral Clustering and K-means is three, which suggests that they were able to identify two clusters with different behaviours, even though they have the same classification.

Through the comparison of the results obtained for each clustering algorithm, shown in Tables

5.12, 5.13 and 5.14, it can be verified that, using the proposed mechanism to find the values of the input parameters which give the optimal partitioning, there were no significant differences between the clustering algorithms. As such, the K-means algorithm was considered the best out of the three tested algorithms, since it has less input parameters to tune, thus making it easier to use.

In addition, the proposed election mechanism with multiple CVIs to acquire the optimal configuration parameters, and therefore the optimal number of clusters, predominantly selects $k = 2$ for the optimal partitioning. This selection allows to capture the overall performance of the network, by identifying a cluster mostly composed of the best performing cells and a cluster mostly composed of the worst performing cells. However, it fails to find clusters of cells with more specific behaviours.

Lastly, it can be interesting to compare the results for the different frequency bands. The L800 frequency band shows an overall worse performance, for the Integrity and Accessibility classes, with the obtained clusters being classified as unsatisfactory, below average or average. On the other hand, the L1800 and L2600 frequency bands show a better performance, with all clusters having either average or above average performance.

5.4 CM Independence Evaluation

This section presents the process used to correlate the performance of the clusters with the respective configuration parameters and, consequently, finding the optimal cell configuration.

The CM independence evaluation is performed after the clustering and PM analysis stages, as seen in Figure 4.1. Thus, the cells are labelled with the cluster to which they belong and those clusters have already been classified with respect to their performance for the KPI class being evaluated.

As briefly explained in Section 4.5, firstly, it is applied an independence test, for each configuration feature, to test if there are distinct configurations regarding that feature for cells belonging to different clusters. In this regard, the Fisher's exact test was used.

Then, in case a correlation between a CM feature and the clusters is found, *i.e.* the values for that feature are dependent of the cluster, an engineer should evaluate that result to conclude if that configuration parameter as, in fact, any impact on the performance of class of KPIs being evaluated. If yes, then it should be straightforward to understand which values of the configuration feature are associated with the clusters that exhibit better performances.

5.4.1 Fisher's Exact Test

Fisher's exact test of independence is used when one has two nominal variables and wants to test, with a level of certainty defined through the significance level, whether the proportions of one variable change depending on the other variable. The null hypothesis then corresponds to the relative proportions of one variable being independent of the value of the other variable. This test can be used in the problem of connecting the configuration parameters with the performance of each cluster because both the labels that identify the cluster and the configuration features can be considered nominal variables, since the optimal number of clusters k is forced to be small (between 2 and 8) and the observations of each one of the CM features can be classified into a small number of categories, as shown in Figure 5.32.

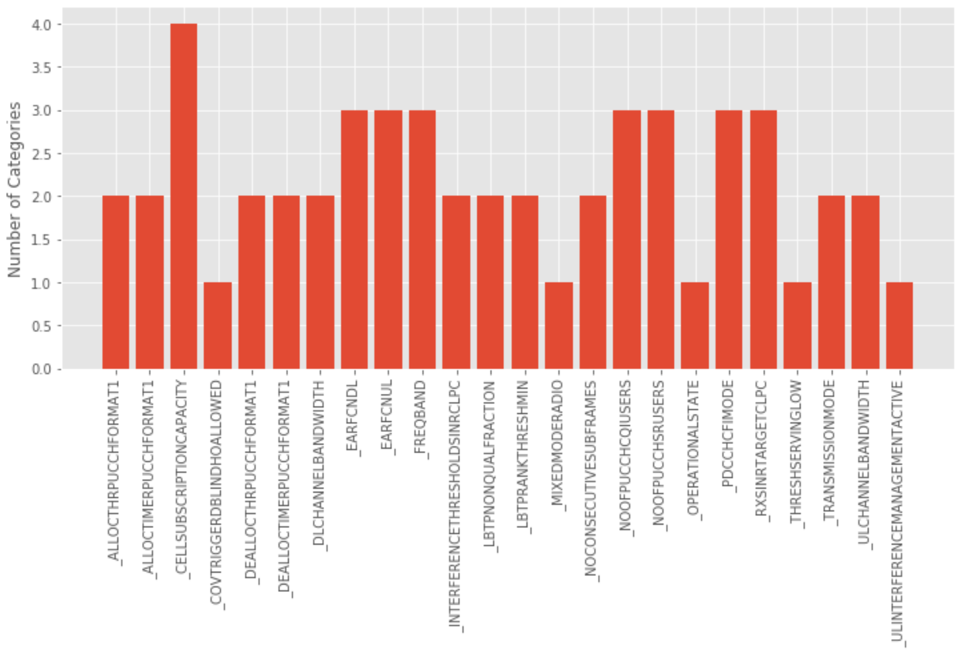


Figure 5.32: Number of categories per configuration feature.

Furthermore, the null hypothesis in this context is that, for each configuration feature, its relative proportions are independent of the labels (*i.e.* clusters). When this null hypothesis is rejected it means that different clusters have different proportions regarding the values of the configuration parameter being evaluated which, in turn, might indicate that there is a direct correlation between that configuration parameter and the performance of each cluster. In such case, that should be further investigated by an expert to identify if there is, in fact, a direct influence of the parameter value on the behaviour of the cells regarding the KPI class being evaluated and also what is the configuration that results in a better performance.

As a consequence of the clustering being performed individually for each frequency band, some CM features present only one category for the frequency band under analysis. This is the case of the following features: EARFCNDL, EARFCNUL, FREQBAND, DLCHANNELBANDWIDTH, ULCHANNELBANDWITH, ALLOCTHRPUCCHFORMAT1, ALLOCTIMERPUCCHFORMAT1, DEALLOCTHRPUCCHFORMAT1, DEALLOCTIMERPUCCHFORMAT1. Moreover, it can be seen in Figure 5.32 that the features MIXEDMODERADIO, ULINTERFERENCEMANAGEMENTACTIVE, COVTRIGGERDBLINDHOALLOWED, THRESHSERVINGLOW and OPERATIONALSTATE already have only one category, independently of the frequency band used. Therefore, this features are, evidently, independent of the clusters and are not considered in this evaluation.

5.4.2 Use Case: Cell Subscription Capacity

The use case presented in this section is related to the clustering results obtained using K-means, for both the 800 MHz and 2600 MHz frequency bands, when evaluating the performance of the respective cells regarding the Accessibility KPI class. The significance level used for the Fisher's exact test was 0.05. If the *p*-value obtained for a CM feature when testing the null hypothesis is lower than the significance level, the null hypothesis is rejected with a confidence interval of 95%, otherwise is accepted. If the null hypothesis is rejected than it can be inferred that there is a correlation between the performance level of a cluster and the configuration of the cells.

Let us consider the 800 MHz frequency band first. For this frequency band and accessibility features, two clusters were obtained. As presented in Table 5.2, cluster 1 has an average performance with a score of 0.57, while cluster 0 has below average performance with a score of 0.32.

Fisher's exact test was then applied for each CM feature with respect to the obtained clusters. The results are presented in Table 5.15.

CM Feature	*p*-value	Independent?
CELLSUBSCRIPTIONCAPACITY	0.000004	No
LBTPNONQUALFRACTION	0.058	Yes
LBTPRANKTHRESHMIN	0.058	Yes
RXSINRTARGETCLPC	0.14	Yes
INTERFERENCETHRESHOLDSINRCLPC	0.31	Yes
NOCONSECUTIVESUBFRAMES	0.41	Yes
NOOFPUCCHCQIUSERS	0.54	Yes
NOOFPUCCHSRUSERS	0.54	Yes
PDCCHCFIMODE	0.75	Yes

Table 5.15: Fisher's exact test results for L800 (Accessibility).

From Table 5.15 it can be verified that the only feature that is dependent on the cluster is the CELLSUBSCRIPTIONCAPACITY. Figure 5.33 shows the percentage of cells, for each one of the clusters, with respect to the configuration they have for the CELLSUBSCRIPTIONCAPACITY feature.

It can be seen that cluster 0, that has a lower score, contains a higher percentage of cells with the CELLSUBSCRIPTIONCAPACITY feature set to 75000 when compared to cluster 1.

For the 2600 MHz frequency band and accessibility features, a similar situation occurs. For this frequency band the optimal number of clusters was three. From those three clusters, two presented above average performance with the remaining having an average performance, as seen in Table 5.10.

The *p*-values, with respect to each CM feature, obtained using Fisher's exact test are presented in Table 5.16.

Figure 5.34 shows the percentage of cells with respect to the values they present for the CELLSUB-

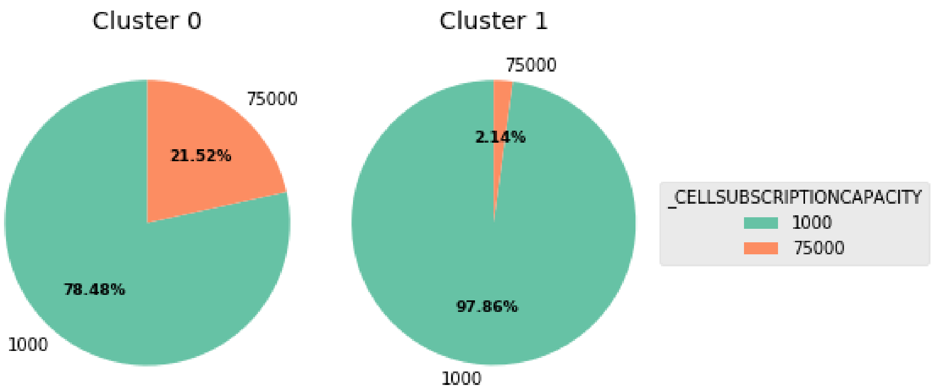


Figure 5.33: Proportions for CELLSUBSCRIPTIONCAPACITY per cluster in L800 (Accessibility).

CM Feature	***p*-value**	**Independent?**
CELLSUBSCRIPTIONCAPACITY	0.017	No
PDCCHCFIMODE	0.34	Yes
TRANSMISSIONMODE	0.34	Yes
RXSINRTARGETCLPC	0.53	Yes
LBTPNONQUALFRACTION	0.66	Yes
LBTPRANKTHRESHMIN	0.66	Yes
NOCONSECUTIVESUBFRAMES	0.67	Yes
INTERFERENCETHRESHOLDSINRCLPC	1	Yes

Table 5.16: Fisher's exact test results for L2600 (Accessibility).

SCRIPTIONCAPACITY, per cluster.

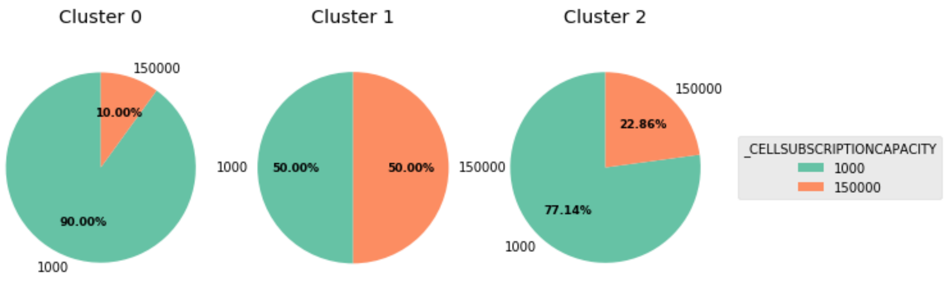


Figure 5.34: Proportions for CELLSUBSCRIPTIONCAPACITY per cluster in L2600 (Accessibility).

Taking into account the scores presented in Table 5.10, it can be verified that the higher the score of the cluster, the lower is the percentage of cells with the CELLSUBSCRIPTIONCAPACITY feature set to 150000.

Taking these two cases into accounts, it seems that setting a higher value for CELLSUBSCRIPTION-

CAPACITY results in a lower performance regarding the accessibility.

The CELLSUBSCRIPTIONCAPACITY is a feature that impacts the load balancing in a cell. Let $cellSubscriptionCapacity$ be the variable associated with the CELLSUBSCRIPTIONCAPACITY feature. The $cellSubscriptionCapacity$ is used to compute the $SubscriptionRatio$ as follows:

$$SubscriptionRatio = \frac{\sum qciSubscriptionQuanta}{cellSubscriptionCapacity} \tag{5.1}$$

where $qciSubscriptionQuanta$ is a weight given to an established E-RAB based on its CQI. Thus, the $SubscriptionRatio$ can be viewed as the load in the cell. A higher value of $cellSubscriptionCapacity$ will cause the value of $SubscriptionRatio$ to drop, meaning that the cell will try to accommodate more users which may cause accessibility issues.

The CELLSUBSCRIPTIONCAPACITY feature is only taken into account when another CM feature, that controls the load balancing process, is active. Since there was no information available regarding the feature that activates the load balancing mechanism, it was assumed in this work that that feature was, in fact, activated.

Chapter 6

Conclusions

This chapter is divided into two sections. In Section 6.1, a summary of the work carried throughout this thesis is presented as well as some conclusions drawn. Section 6.2 aims to give insight about the possible steps that can be taken to further improve and evaluate the methodology presented.

6.1 Summary

The main goal for this thesis was to create a system that assesses the performance of a LTE network, by analyzing the KPIs collected from that network, and through that evaluation is able to find groups of cells that exhibit similar performances. Ideally, the system should be capable of finding the groups of cells that exhibit an undesired level of performance, in regard to the requirements of the network operator, if they exist. This system is based on unsupervised learning techniques.

Additionally, this thesis also aimed to find the configuration parameters that were associated with the groups of cells that presented a desired level of performance. In that regard, Fisher's exact test was used.

Chapter 2 provides a technical overview of LTE networks. This allowed to better understand both the PM and CM data that was available.

Chapter 3 focuses on both the PM and CM data available. A brief explanation of each KPI considered in this work, as well as of each CM feature available, is given. Furthermore, the preprocessing steps considered to remove any artifacts and null values that the data may contain are presented.

Chapter 4 introduces the ML concepts and techniques to develop the desired system to evaluate the network performance and presents the proposed system.

The proposed method to evaluate the performance of the network is based on a feature-based approach, where the KPIs for each cell go through a feature engineering stage before applying the clustering algorithm. The optimal input parameters for the clustering algorithm are selected through an election mechanism using multiple CVIs. A scoring system for the clusters attained when applying this method is also proposed. Lastly, it is presented the line of thought used to correlate the configuration parameters with the performance of the attained clusters.

Chapter 5 presents the results obtained using the method proposed in Chapter 4. To assess the performance of the network, the method is applied individually to each frequency band and KPI class. Three different clustering algorithms were tested: K-means, EM using GMM and Spectral Clustering.

In this chapter, the considered targets, used in the feature engineering step for each KPI, were also presented. Since the dataset to which the clustering algorithms are applied depends on these targets, it is straightforward to understand that the definition of these targets is a key aspect of the methodology proposed. Thus, the targets for each KPI should be specified by the mobile network operator according to the desired level of performance for the network.

Moreover, it was also presented a possible qualitative classification for the clusters based on their score. Yet again, the score depends on a target, $time_{threshold}$, that should be set in agreement with the requirements of the network operator regarding the level of performance of the network. It was verified, for the 1800 MHz frequency band and Integrity KPIs, that a slight change in the value of $time_{threshold}$ results in a very distinct classification for one of the clusters obtained using K-means.

Regarding the clustering results using K-means, for the Accessibility and Integrity classes, it was observed that the optimal number of clusters is given by the election mechanism with multiple CVIs is predominantly two, with one cluster mainly containing the best overall performing cells, for the KPI class and frequency band being evaluated, while the other is predominantly composed by the poorest performing cells. When a tie occurs in the election mechanism, the rationale is to use the set of parameters that correspond to a higher number of clusters, as this allows to have more granularity over the performance of the cells due to more groups with distinct behaviours being identified.

Through the visualization of the attained clustering, using t-SNE, it is possible to infer about the separability of the data. For both the Integrity and Accessibility classes of KPIs in L800 it was observed that it did not exist a clear separation between the two attained clusters.

Spectral Clustering and clustering using GMM were also tested regarding the Accessibility and Integrity in the three frequency bands. It was verified that there were no significant differences in the results obtained with both EM with GMM and Spectral Clustering, when compared to the ones obtained with K-means. Therefore, given the simplicity in tuning the input parameters for the K-means algorithm, this was considered as the best out of the three.

Regarding the performance per frequency band, it was observed that the L800 exhibits a worse performance, for the Integrity and Accessibility classes, with the obtained clusters being classified as unsatisfactory, below average or average. The clusters obtained for both the L1800 and L2600 frequency bands are classified with either average or above average performance, thus it can be inferred that the cells operating in these frequency bands exhibit better overall performance when compared to the ones operating in L800.

The obtained results showed that the system is able to find different groups of cells regarding their performance and most importantly, is able to detect clusters of cells that show a performance level that is below the desired. However, it mostly captures the overall performance regarding the features being evaluated, having trouble to find clusters with more specific behaviours.

Lastly, this chapter presented the results for linking the performance of the clusters with the configuration of the cells that constitute them. Only a use case, related with the feature CELLSUBSCRIPTIONCA-

PACITY, was detected. For the L800 and L2600 frequency bands it was verified that the clusters with better performance, regarding the Accessibility features, are constituted by a higher percentage of cells with a lower configuration value for CELLSUBSCRIPTIONCAPACITY, while for the clusters with lower performance the opposite happens. The limited amount of CM features available and the fact that most of those CM features present a constant value for this network constituted an obstacle towards evaluating the value of the proposed method to find the optimal configuration parameters.

The main challenge faced in this work was to evaluate the goodness of the proposed model, as in unsupervised learning there is no ground truth against which the results can be compared.

6.2 Future Work

In this thesis, a methodology to evaluate the performance of an LTE network and correlate the performance with the configuration parameters was proposed.

In terms of future work for the methodology proposed, a different combination of clustering algorithms and validation metrics could be tested with the end goal of obtaining clusters with more specific behaviours or performances. Regarding the analysis between the configuration features and the performance of the clusters, it would be interesting to test the method proposed with a dataset that not only contained more cells but also more CM features. Moreover, the proposed method should be tested with an active approach, where the CM features can be modified, instead of the passive analysis performed in this work.

Bibliography

[1] A. Gómez-Andrades, P. Muñoz, I. Serrano, and R. Barco. Automatic root cause analysis for lte networks based on unsupervised techniques. *IEEE Transactions on Vehicular Technology*, 65(4): 2369–2386, 2016.

[2] H. Holma and A. Toskala. *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*. Wiley Publishing, 1st edition, 2009. ISBN 978-0-470-99401-6.

[3] I. T. S. Sesia and M. Baker. *LTE – The UMTS Long Term Evolution*. Wiley Publishing, 2nd edition, 2011. ISBN 978-0-470-66025-6.

[4] Alcatel-Lucent. The LTE Network Architecture — A Comprehensive Tutorial. Online, 2009. Available at: `http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf`.

[5] H. Holma and A. Toskala. *WCDMA for UMTS: HSPA Evolution and LTE*. John Wiley & Sons, 2007. ISBN 978-0-470-68646-1.

[6] C. Cox. *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. John Wiley & Sons, 2012. ISBN 978-1-119-97038-5.

[7] H. Holma and A. Toskala. *LTE for UMTS: Evolution to LTE-Advanced*. John Wiley & Sons, 2011. ISBN 978-0-470-66000-3.

[8] N. Shankar and S. Nayak. Performance management in network management system. *International Journal of Science and Research (IJSR)*, 4(5):2505–2507, 2015.

[9] Cisco. Performance Management Best Practices for Broadband Service Providers. Online, 2008. Available at: `https://www.cisco.com/en/US/technologies/collateral/tk869/tk769/white_paper_c11-478096.pdf`.

[10] Huawei. eNodeB V100R005C00 KPI Reference, 2012.

[11] 3GPP. Technical Specification Group Services and System Aspects; Telecommunication management; Configuration Management (CM); Concept and high-level requirements. Technical Specification (TS) 32.600, 3rd Generation Partnership Project (3GPP), 06 2010. URL `https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2440`. Version 10.0.0.

[12] Ericsson. Key Performance Indicators - User Guide, 2016.

[13] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006. ISBN 0387310738.

[14] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[15] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[16] P.-Y. Zhou and K. C. Chan. A model-based multivariate time series clustering algorithm. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 805–817. Springer, 2014.

[17] M. J. Zaki, W. Meira Jr, and W. Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[18] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16 (3):645–678, 2005.

[19] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 144–153. ACM, 2006.

[20] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[21] Shawe-Taylor, John and Cristianini, Nello. *Kernel Methods for Pattern Analysis*, pages 40–42. Cambridge University Press, Cambridge, UK, 2004.

[22] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[23] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 911–916. IEEE, 2010.

[24] M. Halkidi and M. Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 187–194. IEEE, 2001.

[25] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[26] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

[27] F. J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[28] J. H. McDonald. *Handbook of Biological Statistics*, volume 2. Sparky House Publishing, 2009.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[30] Google Inc. Google Earth Pro. Online, 2018. available: `https://www.google.com/earth/download/gep/agree.html`. Last accessed on 14/10/2018.

[31] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.