Universidade de Aveiro Departamento de Eletrónica, Telecomunicações e Informática

Ano 2017

Michael Jordan Karagianis Mesquita

TÉCNICAS DE PREVISÃO EM SISTEMAS DE INFORMAÇÃO E COMUNICAÇÃO

-APLICAÇÃO ÀS REDES MÓVEIS CELULARES-

FORECASTING TECHNIQUES FOR INFORMATION AND COMMUNICATION SYSTEMS

-APPLICATION TO MOBILE CELLULAR NETWORKS-

Universidade de Aveiro Departamento de Eletrónica, Telecomunicações e Informática

Ano 2017

Michael Jordan Karagianis Mesquita

TÉCNICAS DE PREVISÃO EM SISTEMAS DE INFORMAÇÃO E COMUNICAÇÃO
-APLICAÇÃO ÀS REDES MÓVEIS CELULARES-

FORECASTING TECHNIQUES FOR INFORMATION AND COMMUNICATION SYSTEMS
-APPLICATION TO MOBILE CELLULAR NETWORKS-

Dissertação apresentada à universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sobre a orientação científica do Professor Doutor Aníbal Manuel de Oliveira Duarte, Professor Catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

…” the future is an expression of our present thoughts”.

In Unknow.

O júri

Presidente	**Professor Doutor José Rodrigues Ferreira da Rocha**
Vogal - Arguente Principal	**Professor Doutor Rui António dos Santos Cruz**
Vogal - Orientador	**Professor Doutor Aníbal Manuel de Oliveira Duarte**

Agradecimentos

Ao meu orientador Professor Doutor A. Manuel de Oliveira Duarte, pela disponibilidade e indispensável contributo para a realização deste trabalho, ao qual agradeço pela oportunidade em partilhar o seu amplo conhecimento e experiência.

Ao Doutor João Bastos pela disponibilidade no esclarecimento de dúvidas e incontáveis sugestões feitas durante o decorrer do trabalho.

À minha família, em especial a minha Mãe, o meu Pai (em memória), a Euridxé e os meus Avós pelo esforço e apoio nos momentos difíceis e por mostrar que para realizar os nossos sonhos é necessário muito trabalho.

Um agradecimento especial ao meu colega e amigo Carlos Silva pelo apoio incondicional durante o decorrer do curso.

Ao Prof. Doutor José Moreira, pelo apoio moral durante o curso.

Palavras-chave

Previsão de capacidade, Redes Móveis, Modelos de Previsão, ARIMA, LTE, UMTS.

Resumo

A rápida proliferação de tecnologias de informação e comunicação em todo o mundo aumentou a necessidade de um planeamento cuidadoso das infraestruturas. Este facto também é verdade em redes de telecomunicações móveis, data centers, servidores web, etc. Ao desenvolver e implementar soluções inteligentes para melhorar o dimensionamento e planeamento, os operadores de serviços de telecomunicações podem antecipar problemas e minimizar custos.

Esta dissertação foca-se na previsão de capacidade de redes de telecomunicações móveis de forma a maximizar recursos existentes e evitar problemas relacionados ao desempenho ou capacidade, como "*bottlenecks*" e latências. No caso das redes de telecomunicações móveis, a previsão de dados pode ser aplicada para prever o crescimento de tráfego, contribuindo para um melhor planeamento da rede. Portanto, uma previsão mal projetada pode levar os operadores de redes móveis a não estarem preparados para possíveis problemas de rede, como alcance do limite de capacidade de um RNC, isto pode se tornar muito caro em termos de OPEX.

Assim, é fundamental estudar diferentes métodos quantitativos para prever a tendência e o volume de tráfego nos RNCs. Neste trabalho são analisados alguns métodos básicos de previsão utilizados em cenários em que os dados não apresentam complexidade comportamental e alguns modelos como ARIMA e Holt Winters utilizados em dados com complexidade comportamental (presença de tendência e sazonalidade). Também se avalia a exatidão das previsões que resultam da aplicação destes modelos.

Keywords Forecasting capacity, Mobile Network, Forecasting Models, ARIMA, LTE, UMTS

Abstract

The rapid proliferation of information and communication technologies around the world has increased the need for careful planning of infrastructures. This is true in the situation like mobile telecommunication networks, data centers, web servers, etc. By developing and implementing intelligent solutions to improve the measurement and planning, telecommunication service operators can anticipate problems and minimize cost.

This dissertation focuses on forecasting mobile telecommunication capacity networks to maximize existing resources and avoid problems related to performance or capacity, such as bottlenecks and latencies. In the case of mobile telecommunication networks, the data forecast can be used to predict traffic growth, contributing to better network planning. Therefore, a poorly designed forecast may lead to mobile network operators not being prepared for possible network problems, such as reaching the limit of an RNC, which may become expensive in terms of OPEX.

Thus, it is fundamental to study different quantitative methods to forecast the trend and the volume of traffic in the RNCs. This paper analyzes some basic forecast methods used in scenarios where the data do not present behavioral complexity and some models such as ARIMA and Holt Winters used in data with behavioral complexity (presence of trend and seasonality). It also evaluates the accuracy of the forecasts determined from the application of these models.

Index

Tables Index III

Figures Index V

Acronyms IX

1 Introduction 13

1.1 Motivation 13

1.2 Objectives and methodology 14

1.3 Dissertation structure 14

2 Mobile communication technologies 15

2.1 Overview of mobile communication technologies evolution 15

2.2 Radio Access Network overview 19

2.3 Multiple Access Techniques 19

2.4 GSM 20

2.5 UMTS 23

2.6 LTE 34

2.7 Network Management 40

2.8 Key Performance Indicators 42

2.9 Mobile Data Traffic Growth 43

3 Forecasting 45

3.1 Telecommunication forecasting 45

3.2 Forecasting Categories 46

3.3 Time Series Analysis 47

3.4 Forecasting evaluation methods 51

3.5 Forecasting Models 53

3.6 Time Series Decomposition 65

3.7 Forecast validation 73

3.8 Autoregressive Models 75
3.9 Moving Average Models 76
3.10 Introduction to ARIMA Models 77
3.11 ARIMA and Exponential Smoothing analysis 88
4 Analysis and applications of Forecasting - Case study 95
4.1 Daily data analysis and forecast 95
4.2 Weekly data analysis and forecast 111
5 Forecasting Analysis Tool 125
5.1 Forecasting Analysis Tool Architecture 125
5.2 Forecasting Analysis Tool Description 125
6 Conclusion 129
6.1 Future Work 130
References 131
Appendix 1.1: Comparison of data speeds offered by various generations of mobile technology 137
Appendix 1.2: Global Growth of Smart Mobile Devices and Connections 138
Appendix 1.3: UMTS network elements and interfaces [7]. 139
Appendix 1.4: Mobile Generations Speed Evolution [16] 140
Appendix 1.5: Network Management Models 141

Tables Index

Table 1: Variants of 3G worldwide.Source: [5] 18
Table 2: 3G UMTS specification summary. Source:[5][17] 23
Table 3:Specifications for UTRAN operation for FDD & TDD. Source: [7] 27
Table 4: Comparison between HSDPA and HSUPA 33
Table 5: 3GPP Releases on UMTS networks Source: [24] 33
Table 6: LTE specifications.Source:[25] 34
Table 7: LTE specifications vs previous generations. Source: [25] 35
Table 8: LTE and LTE-Advanced capacity comparison. Source: [31] 39
Table 9: KPI categories counters 42
Table 10: Characteristic of forecasting types. Source: [48] 46
Table 11: Comparison between time-series and casual models 47
Table 12: Example of Naive method 53
Table 13: Weight towards previous observations Source: The author 59
Table 14: ARIMA (0,1,2) and ARIMA (0,1,5) information criteria 84
Table 15: Seasonal and non-seasonal model candidates AIC, AICc and MAPE 84
Table 16: Accuracy measures for Holt Winters. Source: The author 89
Table 17: Arima (1,1,0) (0,1,1)[4] Accuracy. Source: The author 91
Table 18: ARIMA (1,1,0) (1, 1, 1) [4]. Source: The author 91
Table 19: ARIMA Candidates. Source: The author 91
Table 20:Accuracy measures for ARIMA (0,1,0) (1,1,0) [4]. Source: The author 92
Table 21: Validation results. Source: The author 93
Table 22: Different window size MAPE results. Source: The Author 94
Table 23: Initial RNC data set for daily aggregated Packet Switch fill factors ADF p-values Source: The Author 96
Table 24: ARIMA Daily aggregated Packet Switch fill factors for RNC1 candidate models Source: The author 98
Table 25: RNC1 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author 99
Table 26: RNC2 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author 100
Table 27: RNC3 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author 102
Table 28: RNC4 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author 103
Table 29: RNC1 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author 105
Table 30: RNC2 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author 107
Table 31: RNC3 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author 108
Table 32: RNC4 accuracy measures for daily aggregated Circuit Switch Fill Factors 110
Table 33: RNC1 accuracy measures for weekly aggregated Packet Switch Fill Factors. Source: The author 112
Table 34:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC2. Source: The author 114
Table 35:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC3 Source: The author 115
Table 36:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC4. Source: The author 116
Table 37:Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC1. Source: The author 119

Table 38:Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC2. Source: The author 120
Table 39: Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC3 121
Table 40: Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC4 122
Table 41: Example of function block and building blocks relationship Source:[83] 147

Figures Index

Figure 1: Evolution of Mobile Cellular Networks: An overview on each generation. Source: [4] 15
Figure 2: Mobile technologies market share of subscriptions worldwide from 2016 to 2021..... 17
Figure 3: Core Network and RAN. Source: [13] 19
Figure 4: Multiple Access Techniques. Source: [14] 20
Figure 5: GSM Architecture. Source: [15] 21
Figure 6: UMTS Architecture. Source: [15] 24
Figure 7: UTRAN components and interfaces. Source: [7] 27
Figure 8: 3G network implementation on R99. Source: [5] 28
Figure 9: 3G network implementation on R4. Source: [5] 29
Figure 10: 3GPP network implementation scenario R5. Source:[19] 30
Figure 11: IMS architecture overview. Source: [20] 31
Figure 12: 3GPP network implementation scenario R6. Source: [15] 32
Figure 13: LTE architecture. Source: [15] 35
Figure 14: LTE Radio Access Network. Source: [26] 36
Figure 15:Logical node of the EPC. Source: [26] 37
Figure 16:LTE Advanced stack. Source [28] 39
Figure 17: Mobile Telecommunication cohabitation. Source [15] 40
Figure 18: Forecasts per Month of Mobile Data Traffic by 2021. Source: [38] 43
Figure 19:Time Series Example. Adapted from [51] 47
Figure 20: Trend Example. Source: The author 48
Figure 21: Seasonal Component. Source: The author 48
Figure 22: Random component. Source: The author 49
Figure 23: Time Series components. Source: The author 49
Figure 24:Data Components. Source: The author 50
Figure 25: Scatter Plot Example. Source: [41] 50
Figure 26:Application of basic forecasting methods. Adapted from: [47] 55
Figure 27: Types of MA. Source: The author 55
Figure 28: Comparison of SA and SMA. Source: The author 57
Figure 29:EMA types. Source: The author 57
Figure 30:Comparing MA weight vs ES weight. Source: [62] 59
Figure 31: SES. Source: The author 61
Figure 32:SES example using SES fitting. Source: [47] 61
Figure 33: Using Simple Exponential Smoothing in a data set with Trend. Source: The author 62
Figure 34:SES vs DES. Source: The author 63
Figure 35: Australian beer production: Additive model. Adapted from: [68] 65
Figure 36:Airline Passenger Numbers: Multiplicative model. Adapted from: [68] 66
Figure 37:Classical Decomposition Steps. Source: The author 66
Figure 38: 51-MA and 501-MA. Source: The author 68
Figure 39:Classical Decomposition using R (code snippet) 70
Figure 40: 4-MA values. Source: The author 70
Figure 41: Data trend. Source: The author 70
Figure 42:Classical Decomposition using R: Detrend 71
Figure 43:Detrend beer data. Source: The author 71
Figure 44: Seasonal Index 71
Figure 45:Seasonal Component. Source: The author 72
Figure 46: Classical Decomposition using R: Random 72
Figure 47:Random component. Source: The author 72
Figure 48: Expanding window example. Source: The author 74
Figure 49:Sliding window process. Source: The author 74
Figure 50: ARIMA as Filters. Source: The author 77
Figure 51: p-order IIR representation equivalent to AR component. Source: [74] 78
Figure 52: p-order FIR representation equivalent to MA component. Source:[74] 78

Figure 53: ARIMA (1,1,1) (1,1,1)[4] example. Source: [41]..... 78
Figure 54: Forecasting ARIMA model process. Source: The author 79
Figure 55: Time series before differentiating. Adapted from [23]..... 80
Figure 56: ACF of a time series before differencing. Source: The author..... 81
Figure 57.ADF test result before differencing. Source: The author 81
Figure 58: Time series before differentiating. Source: The author..... 82
Figure 59: ACF of a time series after differencing. Source: The author..... 82
Figure 60: ADF test result after differencing. Source: The author 83
Figure 61: Model identification using ACF and PACF plots. Source: The author..... 83
Figure 62: ARIMA (2,1,5)(3,1,3)[7] residuals diagnostic check..... 85
Figure 63: Box-Ljung test result. Source: The author 85
Figure 64: Forecast using ARIMA (2,1,5)(3,1,3)[7]. Source: The author..... 86
Figure 65:ARIMA(1,1,1)(2,0,0)[7] long horizon example. Source: The author 87
Figure 66: Observation of 24 quarterly sales period. Source: The author..... 88
Figure 67:Adjusting Holt-Winters. Source: The author..... 89
Figure 68: Holt-Winters adjustment and forecast. Source: The author..... 89
Figure 69:Differenced time series. Source: The author 90
Figure 70: ACF and PACF Plots differenced. Source: The author 90
Figure 71:Residuals ACF and PACF from ARIMA (1,1,0) (0,1,1) [4]. Source: The author..... 91
Figure 72: ARIMA (0,1,0) (1,1,1) [4]. Source: The author..... 92
Figure 73: Holt-Winter vs ARIMA fit. Source: The author 92
Figure 74: Daily aggregated Packet Switch fill factors initial data set. Source: The author..... 95
Figure 75: RNC daily aggregated packet switch fill factors ACF plots stationarity check..... 96
Figure 76: Daily aggregated Packet Switch fill factors of the differenced data. Source: The author97
Figure 77: Daily aggregated Packet Switch fill factors first difference ACF plots. 97
Figure 78:Daily aggregated Packet Switch fill factors first difference PACF plots. 98
Figure 79:RNC1 MAPE results.Source: The author..... 99
Figure 80: RNC1 daily aggregated Package Switch Fill Factors forecast. Source: The author 100
Figure 81:RNC2 MAPE results. Source: The author..... 101
Figure 82:RNC2 daily aggregated Package Switch Fill Factors Forecast. Source: The author 101
Figure 83: RNC3 MAPE results. Source: The author..... 102
Figure 84:Daily aggregated Package Switch Fill Factors Forecast for RNC3.Source: The author 103
Figure 85: RNC4 MAPE results. Source: The author..... 104
Figure 86: RNC4 daily aggregated Package Switch Fill Factors Forecast. Source: The author ... 104
Figure 87: Daily aggregated Circuit Switch fill factors initial data set. Source: The author..... 105
Figure 88: RNC1 MAPE results. Source: The author..... 106
Figure 89: RNC1Daily aggregated Circuit Switch Fill Factors Forecast. Source: The author 106
Figure 90:RNC2 MAPE results. Source: The author..... 107
Figure 91:Daily aggregated Circuit Switch Fill Factors Forecast for RNC2. Source: The author .. 108
Figure 92: RNC3 MAPE results. Source: The author..... 109
Figure 93: RNC3 daily aggregated Circuit Switch Fill Factors Forecast. Source: The author..... 109
Figure 94:MAPE results comparison for RNC4-CSFF. Source: The author..... 110
Figure 95:RNC4 daily aggregated Circuit Switch Fill Factors Forecast. Source: The author..... 111
Figure 96: RNC weekly aggregated packet switch fill factors. Source: The author..... 111
Figure 97: First difference ACF weekly aggregated Packet Switch fill factors.Source: The author112
Figure 98: RNC1 MAPE results. Source: The author..... 113
Figure 99: RNC1 weekly Aggregated Packet Switch Fill Factors Forecast. Source: The author .. 113
Figure 100: RNC2 MAPE results Source: The author..... 114
Figure 101: RNC2 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author 115
Figure 102: RNC3 MAPE results Source: The author..... 116
Figure 103: RNC3 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author 116
Figure 104: RNC4 MAPE results. Source: The author..... 117
Figure 105: RNC4 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author 117
Figure 106: RNC weekly aggregated Circuit Switch fill factor Source: The author..... 118
Figure 107: First difference ACF weekly aggregated Circuit Switch fill factors 118
Figure 108: RNC1 MAPE results. Source: The author..... 119

Figure 109: RNC1 weekly aggregated Circuit Switch Fill Factors Forecast.Source: The author .. 119
Figure 110: RNC2 MAPE results. Source: The author..... 120
Figure 111: RNC2 weekly aggregated Circuit Switch Fill Factors Forecast Source: The author .. 121
Figure 112: RNC3 MAPE results. Source: The author..... 122
Figure 113: RNC3 weekly aggregated Circuit Switch Fill Factors Forecast Source: The author .. 122
Figure 114: RNC4 MAPE results. Source: The author..... 123
Figure 115: RNC4 weekly aggregated Circuit Switch Fill Factors Forecast. Source: The author . 123
Figure 116: Forecasting Analysis Tool Architecture..... 125
Figure 117:Data Analysis component 126
Figure 118:Forecast Analysis..... 127
Figure 119: FCAPS architecture 141
Figure 120: FCAPS model functional areas and TMN logical layers. Adapted: [84] 143
Figure 121: General relationship of a TMN to a telecommunication network. Source:[85]..... 144
Figure 122: TMN Function Blocks. Source: [85] 145
Figure 123: Reference Point between Function blocks. Source: [81]..... 146
Figure 124: Relation between Functional architecture and physical architecture Source:[81]..... 147
Figure 125:TMN pyramid logical layers..... 148

Acronyms

1G	1st Mobile Generation
2G	2nd Mobile Generation
3G	3rd Mobile Generation
3GPP	3rd Generation Partnership Project
4G	4th Mobile Generation
ADF	Augmented Dickey–Fuller
AIC	Akaike Information Criterion
AMPS	Advanced Mobile Phone System
AMR	Adaptive Multi-Rate
ARQ	Automatic Repeat-reQuest
AS	Access Stratum
BSC	Base station controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CAGR	Compound Annual Growth Rate
CDMA	Code Division Multiplexing Access
CPU	Communications Processor Unit
CS	Circuit Switching
CSCF	Call Session Control Function
CS-MGW	Circuit Switched Media GateWay
CSP	Communication Service Provider
DES	Double Exponential Smoothing
EDGE	Enhanced Data Rates for Global Evolution
EIR	Equipment Identity Register
EMA	Exponential Moving Average
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
FDMA	Frequency Division Multiple Access
FIR	Finite Impulse Filter
FM	Frequency Modulation
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GMSC	Gateway MSC
GPRS	General Packet Radio Services
GRAN	Generic Radio Access Network
GSM	Global system for mobile communication
HLR	Home Location Register
HSCSD	High-speed Circuit-switched Data
HSDPA	High Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High Speed Uplink Packet Access
I-CSCF	Interrogating CSCF
IIR	Impulse Infinite Response
IMEIs	International Mobile Equipment Identities
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IS-95	Interim Standard 95
ISDN	Integrated Services Digital Network
ISP	Internet Service Provider
ISUP	ISDN User Part
KPI	Key Performance Indicator

LTE	Long Term Evolution
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ME	Mobile Equipment
MGCF	Media Gateway Control Function
MIMO	Multiple-Input and Multiple-Output
MME	Mobility Management Entity
MME	Mobility Management Entity
MPE	Mean Percentage Error
MS	Mobile Stations
MSC	Mobile Service Switching Center
MSE	Men Scaled/squared Error
MSISDN	Mobile Station International ISDN Number
NAS	Non-Access Stratum
NAT	Network Address Translation
NMT	Nordic Mobile Telephone
NSS	Network Subsystem
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expendures
OSS	Operation Suport System
PCEF	Policy Control Enforcement Function
P-CSCF	Proxy-CSCF
PDN	Packet Data Network
PDN-GW	Policy and Charging Rules Function
PLMN	Public Land Mobile Network
PS	Packet Switching
PSTN	Public switched telephone network
QAM	Quadrature amplitude modulation
QoS	Quality of Service
RAB	Radio Access Bearer
RAN	Radio Access Networks
RNC	Radio Node Controller
RNS	Radio Network Subsystem
RRC	Radio Resource Control
RRC	Radio Resource Control
SAE	System Architecture Evolution
SAE	System Architecture Evolution
SC-FDMA	Single Carrier - Frequency Division Multiple Access
SCP	Signaling Control Point/ service control point
S-CSCF	Serving-CSCF
SES	Single Exponential Smoothing
SGSN	Serving GPRS Support Node
SIM	Subscriber Identity Module
SIP	Session Initiation Protocol
SLC	Service life-cycle
SMA	Simple Moving Average
SS7	Signalling System No 7
TACS	Total Access Communications System
TDMA	Time Division Multiplexing Access
TE	Terminal Equipment
UE	User Equipment
UHF	Ultra High Frequency
UMTS	Universal Mobile Telecommunication System
USIM	UMTS Service Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VHF	Very High Frequency

VLR	Visitor Location Register
VoLTE	Voice over LTE
WCDMA	Wide-Band Code-Division Multiple Access
WMA	Weighted Moving Average

1 Introduction

The use of information and communication technology in the world continues to increase due the accelerating flow of information. In this context, the launch of increasingly appealing and attractive smartphones, and the improvement and increased speed of mobile networks, a large increase in the number of data traffic has been recorded. Thereby, mobile data is expected to grow at an overall rate of approximately 47% from 2016-2021 – almost 49 exabytes per month until 2021 and 7 times faster than 2016 [49]. It is essential that operators continue investing in the modernization of their networks and in increasing the capacity of their sites, so they can cover the demand from current users and potential users that may appear.

Mobile network operators tend to expand their infrastructure as a strategy to increase their network capacity and coverage. They maintain QoS and control congestion applying resources that meet the increasing traffic demand. To run a network efficiently, it is important to recognize the statistical behavior of the network data traffic. This is possible analyzing the real traffic.

If network providers are not able to provide services, new session requests are denied or queued, and the QoS of existing sessions is compromised. This phenomenon is known as network congestion and can be solved in two ways. First by throttling subscriber's usage during peak hours for certain areas of network, this approach does not add or re-allocate network resources; and second by network re-dimensioning which involves adding new resources or relocating resources in the network. The best way to know when to re-dimension is through forecasting.

Forecast constitutes an important and integral component in network management which helps to create scenario-based planning processes, infrastructure optimization, and planning, by applying models to forecast RNC traffic for both user (data traffic) and control plane (signaling traffic). Therefore, an accurate forecast of base-station traffic is very important for guaranteeing QoS that controls congestion and avoids overload of base stations [1].

An accurate data forecast requires the use of accurate models that can capture the statistical behavior of the traffic in analysis. In network traffic forecasting there are several models such as neural networks that can be applied. This dissertation will focus on the quantitative statistical models. Particularly the ARIMA model and smoothing models are tested. These models are based upon [2] recommendation.

Since this recommendation does not give any reference nor guidance to which model is most appropriate when forecasting telecommunications time series, in this dissertation some examinations related to performance, level of difficulty and completeness of these models (historic based models) will be made by comparing the model's accuracy.

1.1 Motivation

The motivation for this dissertation is the forecast of data in mobile networks. It is expected to leave a contribution for the implementation of quantitative methods, concretely the ARIMA model and smoothing model variations, as mentioned before. Hence, it is believed that this dissertation presents a valid and important contribution that can be implemented by a communication service providers (CSP) to prevent network problems such as capacity limit.

1.2 Objectives and methodology

The overall objective of this work is to contribute for a better understanding of data forecast models and procedures that can be applied in mobile networks to answer the needs of communication service providers. Thus, to meet the above mentioned overall objective, the following methodology were target:

- Study the precedents of mobile communication;
- Understand the functioning of the mobile networks, UMTS and LTE;
- Understand the process involved in the collection of performance indicators that give rise to time series communication data.
- Study and understand existing models and methods applied in time series forecast;
- Analyze and apply quantitative forecasting models to a real mobile network data and study its accuracy;

1.3 Dissertation structure

This dissertation is divided in 6 chapters which are explained in concise manner to reflect the work presented here. Brief description of the contents presented in each chapter and their purpose is described as follow:

- Chapter 1, "**Introduction**": Describes the motivation, objectives and methodologies that drove to the study of mobile communication forecasting;

- Chapter 2, "**Mobile telecommunication technologies**": Provides a state of art on major existing mobile telecommunication networks: their predecessors, main features, and characteristics. In addition, a detailed explanation on the UMTS and LTE networks, and their architectures are provided. Still in this chapter, a discussion around the telecommunication management network (TMN) and key performance indicator (KPI) is also provided;

- Chapter 3, "**Forecasting**": A state of art in relation to the current forecasting techniques applied in telecommunication, followed by an exemplified explanation of the models and techniques implemented in the forecast of mobile data with different behavioral complexity such as trend and seasonality;

- Chapter 4, "**Analysis and applications of Forecasting-Case study**": Describes the implementation of chapter 3 models and methods in a real case communication service provider time series data;

- Chapter 5, "**Forecasting Analysis Tool**": Describes the main functionalities of a support tool developed for analysis and data forecasting. The tool purpose is to facilitate and invigorate the modeling process.

- Chapter 6, "**Conclusion**": Gives conclusions on the degree of accuracy of the models implemented in forecasting the data as well as the results and procedures used to elaborate the forecasts; And the future work section presents future improvements related to the present dissertation such as implementation of new forecasting models and reduction of uncertainty levels in implemented models;

2 Mobile communication technologies

The purpose of this chapter is to enlighten the reader with an overview about mobile networks. Despite the objective be the forecasting this approach is highly essential to understand the work presented in this paper. Here the evolution of mobile network generations and its main features will be described. Then a much greater effort will be made to explain 3G and LTE networks, focusing in the architecture and in some network management systems (protocols, elements).

2.1 Overview of mobile communication technologies evolution

Mobile network communication has been categorized in different generations and paradigms as shown in Figure 1.The term generation is used to refer to fundamental technological implementations and transformations that occur in the type and characteristics of a specific service, non-backwards compatible with technology, and new frequency bands of a particular network system [3].

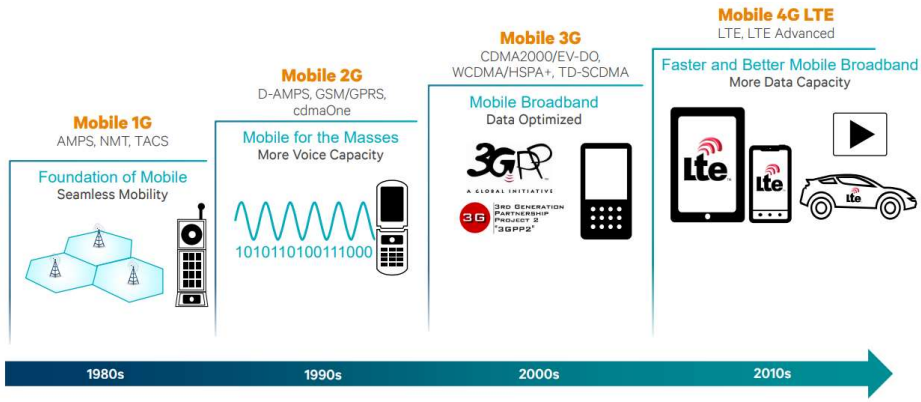


Figure 1: Evolution of Mobile Cellular Networks: An overview on each generation. Source: [4]

The evolutionary process of mobile technologies is marked by different generations where each has a particular group of characteristics and standards. Therefore, the most significant are described as follow:

2.1.1 1st Mobile Generation

This mobile telecommunications generation started in 1980, marked as a seamless mobile technology offering only mobile analogue voice services. It was considered an analogue or semi-analogue network because an analogue radio path and digital switching were used [5]. Since the 1st mobile generation, world mobile communication has suffered several transformation and experienced massive growth.

Because it was analogue, calls were extremely susceptible to interferences including: lighting, noises, and static caused by electromagnetic devices. Despite being the foundation of mobile networks, the 1st mobile generation is characterized by having several incompatible solutions and standards [6][7], along which the popular were:

- **AMPS family:** used first in North America in 1982 later in Israel, Australia and Singapore. It used the 800MHz to 900 MHz Band and the FDMA with 30 kHz wide among the mobile station (MS) and the base station (BS). AMPS have the following variants:

  - **TACS**: It is a variant of AMPS used in the United Kingdom and adopted by some middle Eastern and southern Europe countries. Both used the FM technique for radio transmission and FDMA for traffic modulation.
  - ETACS, N-AMPS (TIA/EIA/IS-91).
- **NMT:** came in two variants: NMT-450 and NMT-900, where the first operates in the area of 450 MHz and the second operates in the area of 900 MHz.
- **C-450**: used in West Germany, Portugal and South Africa.

The 1st mobile generation were designed mainly to provide voice services. Thus, some features are described as follow [6]:

- Low rate codecs which caused poor sound quality;
- Lack of security due to not supporting encryption;
- Frequency carriers on VHF and UHF bands.
- Because of the absence of standardization, it resulted in interconnection and incompatibility problems.

Among other disadvantages, 1st mobile generation solution and standards were unable to provide operations between countries despite offering hands-off and roaming capabilities [7]. Nevertheless, these standards continue to operate until being fully substituted by 2G technology.

2.1.2 2nd Mobile Generation

The 2nd mobile generation was launched in the latest 1991 based on the digital low-band data signalling. It introduced a new group of services such as: short message and lower speed data. This generation main goal was to set up a compatible and international transparency network which would allowed users to access it anywhere within a region. Unfortunately, this concept did not entirely succeeded due to regional standardization nature [5].

Generally associated to the GSM services, it uses digital modulation techniques and a combination of TDMA and CDMA [7]. This allows a single frequency to be allocated to different users. Initially GSM adopted the TDMA – the IS-95 system – which was based on CDMA technology and considered a 2G part by North America and Korea.

It is no secret that 2nd mobile generation systems by far overcome many of the deficiencies presented in 1st mobile generation. Essentially, the main difference between the 1st and 2nd mobile generation are the passage of the mobile networks from analogue transmission techniques for traffic to digital transmission techniques in the 2nd generation, and the introduction of advanced and fast phone-to-network signalling from the 2G.

This generation increased the concept of voice capacity, delivery of mobile communication to the general public, and more people had access to data in more places. It also improved sound quality, robustness, reliability, safety, better data services and efficient use of spectrum[6]. The 2nd Mobile Generation is more scalable than the 1st mobile generation because of electric components cost what makes possible to deliver more and safer signal. Also, 2nd mobile generation improved international roaming across countries especially in Europe by providing a single unified standard.

Despite the GSM and its derivates being the most widely used, different TDMA technologies were also used in the 2G, such as:

- PDC;
- iDEN;
- iS-136;
- D-AMP;

As for GSM main working band, it uses the 900 MHz and 1.8GHz bands, though different derivates exist from which the most important are:

- Digital Cellular System 1800 (DCS-1800) also known as GSM-1800;
- PCS-1900 also known as GSM-1900.

The original GSM provided a user data rate of 9.6-Kbps; 14.4-Kbps data rate was later specified [7]. Some upgrades were made to 2nd mobile generation which resulted in development of new technologies based on GSM with an improvement of several extensions and features such as mobile internet with greater speed.

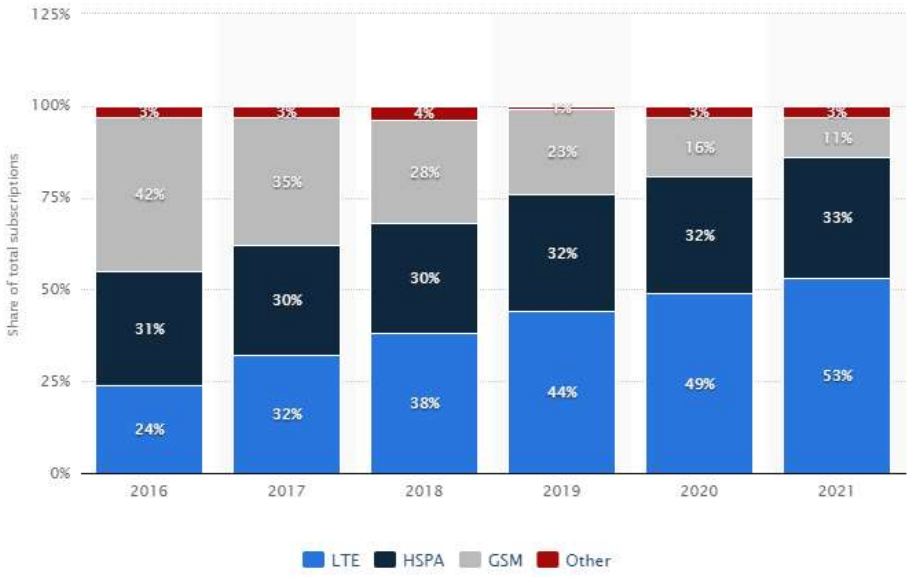


Figure 2: Mobile technologies market share of subscriptions worldwide from 2016 to 2021. Source:[8]

This new technology is called 2.5G and 2.75G, and it is considered an intermediate generation, based on GSM system which includes [7][9][10]:

- **HSCSD**: a 2.5G upgrade of GSM capable to offer speed up to 56.7kbps using reserve multiple time slots during a connection.
- **GPRS**: was the first evolution of GSM known as 2.5G. It has a packet switching in addition to CS used by HSCSD. It supported data transfer rates of 53.6kbps for Download and 26.8kbps for Upload (refer to Appendix 1.1) and a continuous connection to the network. GPRS represents a significant step towards the 3G.
- **EDGE**: known as 2.75G, it implements a higher speed packet switching method despite not requiring new technology. Its data transfer rate was about 217.6kbps for Download and 108.8kbps for Upload.

Nevertheless, the intermediate generations had a few limitations such as radio network and longer transmission delays which caused difficulty to achieve greater transmission speed.

The 2nd Mobile Generation, GSM, still dominates the worldwide mobile technologies market share in 2017 by holding 35% of worldwide subscriptions. However, this supremacy tends to incrementally be replaced by 3G systems. By the year of 2018 the market is expected to be dominated by 4G, LTE, by holding approximately 38% of the market [8].

2.1.3 3rd Mobile Generation

By the time the GSM was launched for the public, the ETSI started to develop a new system called the UMTS. It represents the successor of the 2G mobile technologies which includes GPRS and EDGE. Table 1 take a comparison of the 3G variants used worldwide and their building blocks basis.

Table 1: Variants of 3G worldwide.Source: [5]

Variant	Radio Access	Switching	2G Basis
3G(US)	WCDMA, EDGE, CDMA2000	IS-41	IS-95, GSM1900, TDMA
3G (Europe)	WCDMA, GSM, EDGE	Advanced GSM NSS and packet core	GSM900/1800
3G (Japan)	WCDMA	Advanced GSM NSS and packet core	PDC

The 3rd mobile generation is marked by UMTS systems, an evolution of the GSM and created to overcome some of its limitations. The main advance of 3G over 2G is the use of packet switching data network instead of circuit switch voice network for data transmission offering a multitude of new possibilities. Moreover, 3G was intended to provide a global wireless access telecommunication infrastructure to serve public and private networks of mobile communications users using both satellite and terrestrial systems. Also, it has the purpose of easing the convergence of existing 2G networks to 3G networks [9].

This generation allows mobile network operators to reach almost full network capacity by applying spectral efficiency meanwhile providing the final user a wider range and advanced services which includes enhanced voice and video communication and broadband connection; all together in a single mobile environment. Additionally, 3G applies data centric approach compared to 2G which was voice centric. Also, allows a combination of voice, multimedia applications and mobility [11].

The 3rd mobile generation has contributed for the popularization of the internet using mobile phones. Initially it provided information transfer at a rate of at least 200 Kbit/s later increased with releases 3.5G and 3.75G. These releases, apart from providing faster rates also provided multiple applications such as: mobile broadband to smartphones and mobile modems used for personal computers. A general comparison of the generations related to technology, voice switching, and data rates can be found in Appendix 1.1. From the table is possible to see that as the generations and technologies evolved, a significantly increase in speed also occurred.

Since UMTS is an evolution of 2G, some components of the GSM core network are reused for UMTS by simple software upgrade, as will be explained.

2.1.4 4th Mobile Generation:

The 3GPP has designated a project which consisted in developing a high-performance air interface known as LTE which marked the 4G. Introduced by 3GPP 8th release LTE represents the last step of radio technologies, an evolution of GSM/GPRS/EDGE and UMTS/HSPA designed to improve mobile network capacity and speed based on all-IP environment.

New technologies were introduced by LTE, compared to its predecessors, allowing it to operate efficiently by applying a full use of spectrum. LTE mobile network includes support to VoLTE which enhances the use of the network for both data and voice connections. To efficiently support the demand for wideband transmissions required, LTE adopted the OFDM for its multiple access technology instead of the CDMA as it solved the problem of accommodating rapidly increasing demand of data traffic [12] and to use the MIMO in the radio link at the physical layer.

One way used to distinguish 4G from 3G is the employment of all-IP network by the first in contrast to CS by the second. 4G flat-IP architecture allows all devices to communicate over IP technology.

2.2 Radio Access Network overview

Any radio telecommunication system has the purpose to provide users with the means of communication. Being one of the most important parts of a mobile telecommunication system, the radion access network (RAN) provides a connection between different devices and network components through radio connections. Therefore, some RAN will be described. Figure 3, describes the core network and RAN used from 2G up to LTE. The RAN of the illustrated technologies types are:

- **GERAN**: is the key part of GSM and joins base stations and base station controllers. When the EDGE technology is not present the RAN is called GRAN, despite this they remain identical in concept.
- **UTRAN**: is a UMTS RAN associated to the WCDMA radio access and includes three types of channel concepts: physical, transport and logical channel. This RAN is also known as UTRA and can be grouped as the RNS.
- **E-UTRAN**: introduced in 3GPP Release 8, is the RAN used for the LTE in mobile networks. It combines E-UTRA, UE and EnodeB and it provides improvements related to spectral efficiency and data rates as well as flexibility in bandwidth and frequency for the access network.

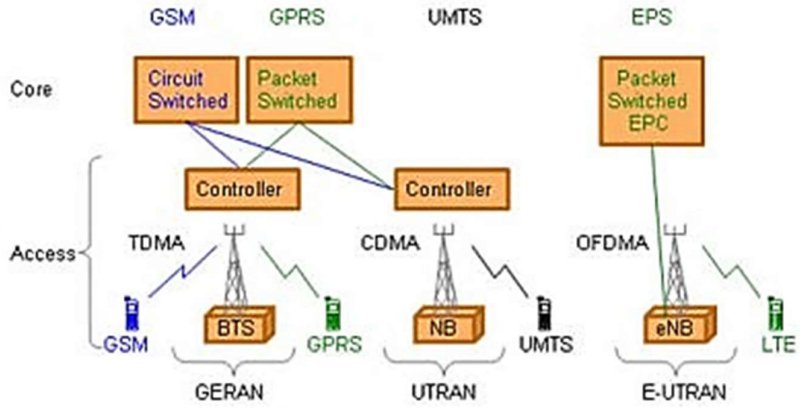


Figure 3: Core Network and RAN. Source: [13]

The RAN present in Figure 3 cannot be misled by an air interface. air interfaces are radio parts used in communication between ME and towers (BTS, ND or eNB) while RAN represents a relative large network with a higher concept level whose job is to handle routing of data or channels.

In the figure, it is possible to see GERAN and UTRAN interworking. This communication is made through the Iu interface from Release 5. This interworking allows the 3G Core Network to be used by the 2G networks (GSM/EDGE) radio interface which optimizes both radio technologies and reduces additional installation cost by the communication service providers.

2.3 Multiple Access Techniques

One of RANs common problem is to how wisely divide the spectrum among multiples users. To solve this problem several division multiple access techniques can be used allowing multiple mobile users to share the allotted spectrum efficiently. Based on the type of channel, one of the following techniques can be used [6]:

- **Frequency Channels**: splits a band frequency into small frequency channels that are assigned to different users. An example is the broadcast transmission used by FM radio

station. Each FM radio station receives an exclusive frequency. This can be illustrated in Figure 4, left graphic.

  - **OFDMA:** used to distribute information of the transmissions among parallel subsets of carriers which favours larger speeds for the downlink.
  - **SC-FDMA:** is a specification similar to OFDMA. It reduces power consumption, so the energy of connected devices also decreases. Despite the name, SC-FDMA can use carrier subsets.

- **Time-slot Within Frequency Bands:** TDMA allocates a certain amount of time and allow every user to transmit using a common frequency band. For each call, a specific timeslot is allocated. This is the technology used by 2nd Mobile generation cellular systems.
- **Distinct Codes:** CDMA assign to each user a shared code by which they can be identified. All users can transmit simultaneously in the same frequency band this allows efficient use of available power and radio resources. Similar to TDMA, the W-CDMA allows a higher transfer rate and system capacity. This is the technology of choice of 3rd Mobile generation cellular systems.

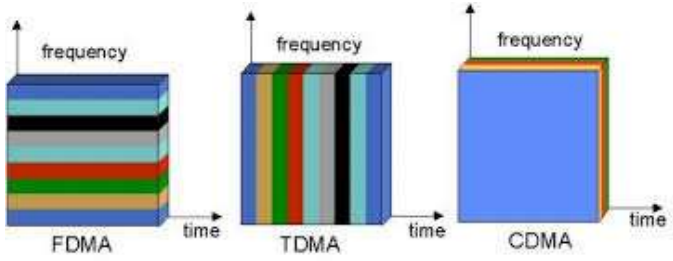


Figure 4: Multiple Access Techniques. Source: [14]

The CDMA technology compared to TDMA-Based, provides clearer voice with less background noise, decreased dropped calls providing greater reliability and enhanced security[11].

2.4 GSM

GSM has an architecture which comprehends entities with specified functions. A GSM mobile network can be extended to a fixed telephone network. Figure 5 depicts a general GSM architecture. The network system can be divided into three main elements:

- Base Station Subsystem;
- Network Subsystem;
- Mobile Station;

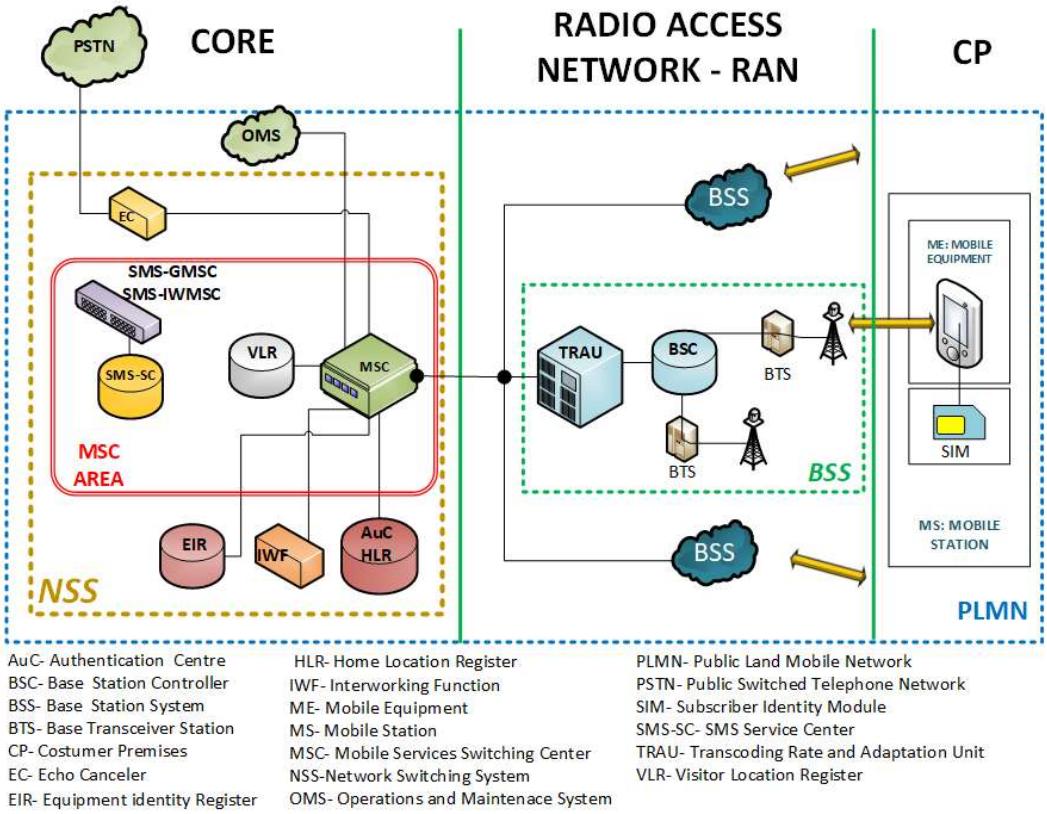


Figure 5: GSM Architecture. Source: [15]

2.4.1 Base Station Subsystem

The BSS is a radio access network that contains all the necessary equipment, hardware and software to connect the Mobile Station to the Network Subsystem (core network) and is responsible for all radio path control. The BSS consists of the following elements:

A. Base Station Controller

Considered the intelligent element of the system, it controls all the functionality of a BTS (cell) over the A-bis interface where all the BTS connects, also controls radios networks. Many BTS can connect to a single base station controller (BSC) (the exact quantity depends on the network configuration). BSC has the job of connecting the MS to the MSC. Also, it has other functions which are [7][5]:

- BTSs radio resources management which includes allocation of frequencies;
- Measurement of uplink signal time-delays with respect to the BTS clock;
- Sustain radio connections to the MS and terrestrial connections to the network subsystem (NSS);
- Intercell handovers;
- Power management.

B. Base Transceiver Station

A very noticeable element of a GSM network used to provide network coverage to a designated area and allows ME to connect and communicate to a network. A BTS maintains the air interface and can have up to 3 antennas and each serve one cell. A BTS function includes[7]:

- Uplink measurements and advanced timing calculations;
- Channel encryption and decryption and speech processing;
- Broadcast and channels control;

2.4.2 Network Subsystem

The network subsystem is composed of a group of elements responsible for calls management between different network such as fixed or mobile switching center and other networks. A network subsystem in the 2nd generation contains the following main elements:

A. Mobile Switching Center

Called public land mobile network by the standards, the mobile switching center (MSC) is the central component of a NSS when referring to control which provides the necessary functionality to subscribers. The MSCs are interconnected by wires and are used to interface GSM networks to fixed networks and bridges the core network to the access network. Besides, it has the following functions [16] [9][5]:

- Responsible for the mobility management of subscribers;
- Statistics and interface towards BSS also interfaces with the external networks (such as PSTN/ISDN/packet data networks);
- BSS control functions;
- Charging (billings) and control of connections;
- Manage connections between subscribers;
- Switching of user calls.

B. Home Location Register

Contains a permanent data register of subscriber profile information. It is common to find this element implemented in the same equipment as MSC/VLR despite Figure 5 showing then separated. The next elements are find in the same equipment as HLR:

- **AuC**: stores the subscriber authentication key and encryption over the radio channel.
- **EIR**: is a database for all valid terminals on the network. It stores the IMEIs. Physically EIR is divided in several nodes.

The NSS elements are connected using a SS7 which allows to exchange the information such as calls and text messages on the network. It also helps to give correct billing to a specific user and roam users when travelling in a foreign country.

2.4.3 Mobile Station

Devices used by subscribers to connect to a GSM network are called MS. It represents the mobile part of a network and is divided into:

- SIM;
- ME.

A SIM is used to deliver mobility and subscription of services to the user. It also contains and uses the IMSI. And the ME, consists on a device used to connect to a network. The terminal consists of an equipment that allows to connect to the network through a BTS.

A MS is responsible for several tasks which includes [9]:

- Speech encoding and interleaving;
- Modulation;
- Ciphering as well as time and frequency synchronized transmitting and receiving data transfer all over the Air interface.

2.5 UMTS

The creation of the UMTS was driven by the desire of 3GPP to produce a globally applicable Technical Specifications which was the 3G Mobile Telecommunications System. Several UMTS features have been implemented based upon the GSM core network and its radio access network. This allowed some 2G to be used for UMTS.

Table 2: 3G UMTS specification summary. Source:[5][17]

3G UMTS Specification	
Parameter	**Specification**
Maximum data rate	2048kbps low range
RF channel bandwidth	5MHz
Multiple access scheme	CDMA
Duplex schemes	FDD and also TDD

The UTMS employs a spread spectrum transmission different from the GSM. Instead of TDMA transmissions UMTS uses the CDMA which has its physical layer referred as wideband CDMA. The W-CDMA is described as a wider band compared with the common CDMA which has the advantage of efficient utilization of radio spectrum giving a higher maximum data rate [11]. It is used for radio transmissions and employs a 5 MHz channel bandwidth which allows to carry over 100 simultaneous voice calls and carry data at speeds up to 2 Mbp/s, as can be seen from Table 2. These speeds were applied on the initial format. With the implementation of the HSDPA and HSUPA in later releases data transmissions were enhanced. The network system of the UTMS can be divided into the following logical components:

- Core Network;
- Radio Access network;
- Mobile Equipment.

The components are connected via an open interface .The GRAN component is expected to be capable to connect simultaneously different CNs such as GSM or a packet-data network [7]. Represented in the figure bellow, the *Iu* interface can be found between the CN and the GRAN. And the *Uu* interface is found in between the UE and the RAN.

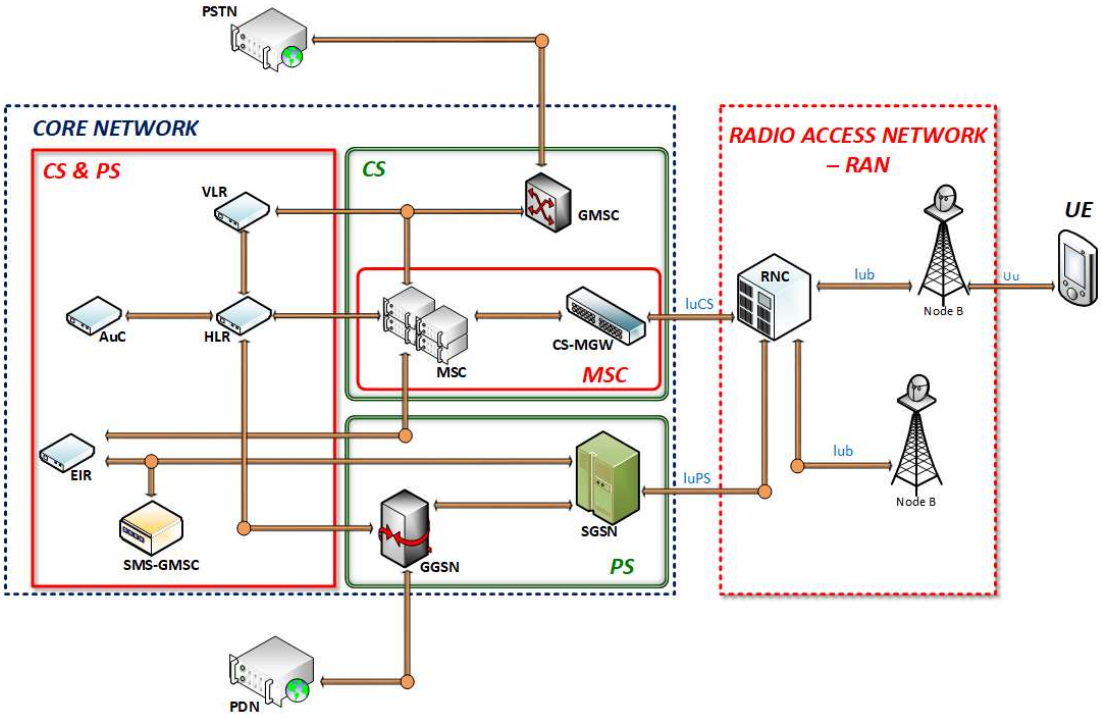


Figure 6: UMTS Architecture. Source: [15]

2.5.1 Mobile Equipment

A mobile equipment can be any device from a mobile phone to a terminal attached to a computer. In the UMTS there is a terminal equipment, which has into it a Mobile equipment and an USIM. The USIM consist on a more improved SIM card, it includes a list of preferred and prohibited PLMN as well as information about language information to be displayed when roaming.

2.5.2 Core network:

UMTS core network is interesting because it consists on a “exodus” of the GSM architecture where elements are overlaid to allow additional functions. This way UMTS core network may be divided in two different group of elements:

- **Circuit switched elements:** are used to carry data by applying CS and is based upon the GSM architecture entities. The CS contains the: MSC and GMSC.
- **Packet switched elements:** are used to carry packet data. Compared to CS this allows an improved network usage by carrying data as packets which can be routed directly to their destination. The PS contains the: SGSN and GGSN.

In addition, another group of elements mainly related to registration are shared by both groups which includes the following network entities:

- Home Location Register and Visitor Location Register;
- Equipment Identity Register;
- Authentication Center.

A. Mobile Switching Center

Similar to the GSM, in the 3G the MSC still is the main component of the circuit-switched core network. Because of the resembling it can serve the GSM-BSS and the UTRAN connections. So based on [7], a MSC can be used to serve both access networks since the GSM-MSC has been upgraded to meet 3G requirements.

UMTS GRAN supports an interface between the network and the PSTN also to different GSM NSS which includes MSC, SGSN and SCP. As for VLR, in the 3G is implemented physically with the MSC and uses a logical B-interface to communicate to each other. Using a single MSC is possible to connect more than a few set of BSs and RNC (known as BSC).

The functionalities of the MSC in a UMTS network besides interfacing a GSM networks to fixed networks is described as follows [7]:

- MS call setup coordination;
- Complex inter-MSC handovers management;
- Frequency allocation in a MSC area;
- Signal exchange between network interfaces.

B. Visitor Location Register

The visitor location register stores information about MS roaming in one or several MSC coverage area. A VLR contains almost the same information as the HLR; the only difference is the duration of the information. VLR has temporary information while HLR has permanent information storage. The information stored is about active subscribers in a VLR area. A VLR is usually implemented in the same equipment as the MSC/VLR.

The subscriber data stored in a VLR includes the following information [7]:

- The International mobile subscriber identity;
- Last known location and the initial location of the MS.
- Location area where the mobile station has been registered;
- Mobile station roaming number.

C. Home Location Register

In each HLR a permanent subscriber data profile is registered. Usually a HLR is in the same equipment as an AuC and EIR which are combined as one unit, but it is also possible to have both VLR and HLR in the same unit. The subscriber data in the HLR includes the following:

- Supplementary services parameters;
- Possible roaming restrictions;
- Authentication key;

D. Equipment Identity Register

The EIR function stores the IMEIs used in the system and decides whether a mobile equipment is allowed to a network. Each EIR has only one PLM to which it connects all the HLR. An EIR has three types of groups which are described as follows [7]:

- **White list:** contain equipment known to be in good order;
- **Black list:** has any equipment reported to be stolen;

- **Gray list:** is used to register equipment known to contain non-fatal problems that justifies barring.

E. Authentication Center

The AuC is responsible to store all the subscriber authentication key and is associated with a HLR. The data stored by the AuC is permanent and is registered upon the registry. Then it is used to generate parameter for authentication procedures. In the network, is possible to find the AuC physically located in the same unit as the HLR.

F. Serving GPRS Support Node

A Gateway GPRS Support Node represents the main element in a PS network and is used to route user data between the radio access network and the core network. It contains information related to subscriber information, VLR number, the routing area where the MS is registered among others [7]. A SGSN has the main responsibility of managing user plan and signalling plan management. Among this, it also provides the following functions:

- Interaction with other areas of the network;
- Session management;
- Mobility management;
- Billing.

It has an IuPS interface from where a connection to UTRAN is made, a Gb interface to connect to a BSS, a Gs interface to connect to MSC, a Gf interface to connect to EIR, a Gr interface to connect to HLR and a Gn interface to connect to GGSN. Please refer to Appendix 1. for a more detailed UMTS architecture.

G. Gateway GPRS Support Node

The GGSN is used to route both incoming and outgoing traffic. It corresponds to GMSC, an equivalent MSC used to route CS calls outside the mobile network. The GMSC is considered a type of MSC used specifically as a gateway to external data network such as PSTN. Furthermore, a GGSN must maintain the following information which is received from the HLR and from the SGSN[7]:

- Subscription information:
  - IMSI;
  - PDP addresses.
- Location information:
  - The SGSN address where a MS is registered.

2.5.3 UMTS Terrestrial Radio Access Network

UMTS main type of RAN is the UTRAN, also known as RNS it is equivalent to the 2G BSS. It is responsible to provide and manage all air interfaces within the network. The UTRAN system can be subdivided into the following components:

- Base Station or node B,
- Radio Network Controller.

UTRAN uses the Iu interface to connect to the core network and Uu interface (radio interface) to connect to UE, which are external connections. Internally the Iu interface connects Node Bs to the RNCs and Iur interface interconnect RNC. All the previous description is illustrated in Figure 7.

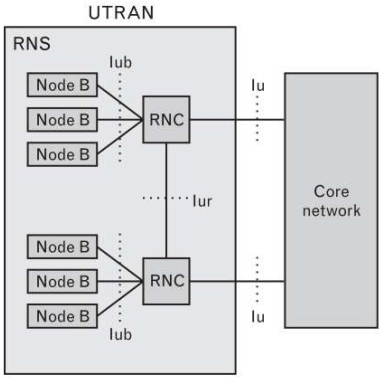


Figure 7: UTRAN components and interfaces. Source: [7]

The Iub also must manage different problems that may occur such as power control problem; therefore, some manufacturers tend to use their own proprietary solutions here [7]. Given the fact that transmissions have to be in both uplink and downlink, UTRAN uses two techniques to ensure concurrent transmissions which are:

- **FDD**: in UTMS FDD version of UTRA transmissions both uplink and downlink happen using different frequencies. Thus, the bandwidth has to be doubled in order to accommodate both transmissions. A filter also has to be added to prevent signal interfering from happen.
- **TDD**: is a time-division technology different from the FDD in UTMS TDD version of UTRA instead of a double frequency a single frequency carrier is used for uplink and downlink, but transmissions occur in different time intervals.

Table 3:Specifications for UTRAN operation for FDD & TDD. Source: [7]

Key Specification for UTRAN Operation for FDD & TDD		
Parameter	UTRA FDD	**UTRA TDD**
Multiple access method	CDMA	TDMA,CDMA
Channel spacing	5 MHz	5 MHz (and 1.6MHz for TD-SCDMA)

A. Radio Network Controller

A RNC is used to control Node Bs and are known as controlling RNC. RNC is to UMTS what a BSC is for the GSM networks. To connect between RNCs a logical Iur interface is used. Also, it is connected to the MSC and SGSN by the Iu interface, IuCS to the first and IuPS to the second respectively. A RNC can perform several functions which are described as follows [7]:

- RAN resources management which includes Iub transport;
- Allocation of traffic channels in Iu and Iub interfaces
- System information management and scheduling of system information;
- Traffic and Reporting management;

B. Node B

Node B is responsible to provide connections in between the MS and the fixed part of the network, thus managing all data over the air interface. Node B are controlled by the RNC. Generally, Node B are used to refer logical concepts, to physical refer an entity the term BS is used, an UMTS equivalent of a BTS in the GSM.

A Node B can perform the following functions[7][16]:

- Map logical resources onto hardware resources;
- Transmit RNC system information messages;
- Report uplink interference measurements and DL power Information;
- Frequency and time synchronization.

2.5.4 Release 99: A New Radio Access Network

The 3GPP has organized different UMTS specification into consecutive versions called “Releases”, which purposes is to add new functionalities and features to the system. Release 99 was the first release to be launched.

The reason for a solid GSM “presence” in the Release 99 were of much importance since the UMTS had to be compatible to existing GSM networks and both networks should interoperate. Thus, this release contains initial steps toward the UMTS which includes improvements such as redesigned radio access network, the UTRAN.

The 3GPP R99 is a GSM-based implementation, in Figure 9, a 3G GSM-evolved mobile network is depicted which represents two access networks that delivers CS and PS traffic.

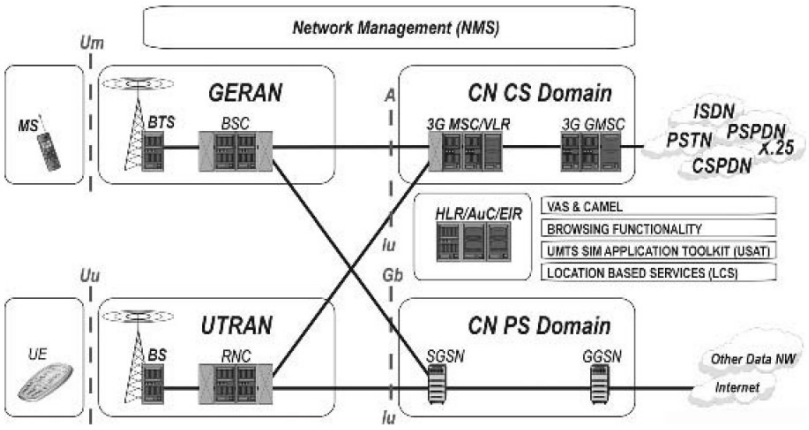


Figure 8: 3G network implementation on R99. Source: [5]

Different from the 2G, the GRAN form the UMTS (Release 99) has suffered a new set of important specifications that conforms to the Iu interface which consist on the translation of the GRAN into a new dedicated RAN called UTRAN in which 2G BSS components BTS and BSC are now called BS and RNC. The introduction of these components is a consequence of WCDMA and radio access equipment not being compatible to GSM equipment[5]. In fact, this release focus on the access network, and the changes are mainly in software improvements in order to support the new Iu interface between the MSC and the UTRAN.

Moreover, the bandwidth used by a single carrier has further increased resulting in a faster data transfer than previously possible compared to GSM. In addition, it allows the Release 99 UTRAN to send data with a higher speed. As for the frequency, blocks of 5 MHz were used in the frequency range of 1920 MHz and 1980 MHz were assigned to UMTS in Europe and Asia [16].

2.5.5 Release 4: Enhancements for the Circuit-Switched Core Network

Considered the first 3GPP upgrade system the launch of Release 4 was driven by the network operator will to associate the CN CS Domain and CN PS Domain into a new architecture which could converge all network traffic.

Release 4 improved the UMTS core CS voice and data services by introducing the Bearer Independent CS which enables the MSC subscribers physical voice circuit to be replaced by a media gateway bringing scalability to the system [5]. The introduction of GMSC-Server, MSC Server and CS-MGW comes as a substitute to the GMSC, MSC and VLR from Release 99 which gives Release 4 a higher efficiency and have flexible bearer solutions.

The CS-MGW introduced in Release 4 is a node which has the responsibility to maintain a CS connection capacity, managing all physical connection and handling the user traffic (bearer) and transcodes the user data for different transmission methods. As for the MSC server, it makes and maintains the capacity connection control and mobility management. For each MSC Servers is possible to control different CS-MGW nodes.

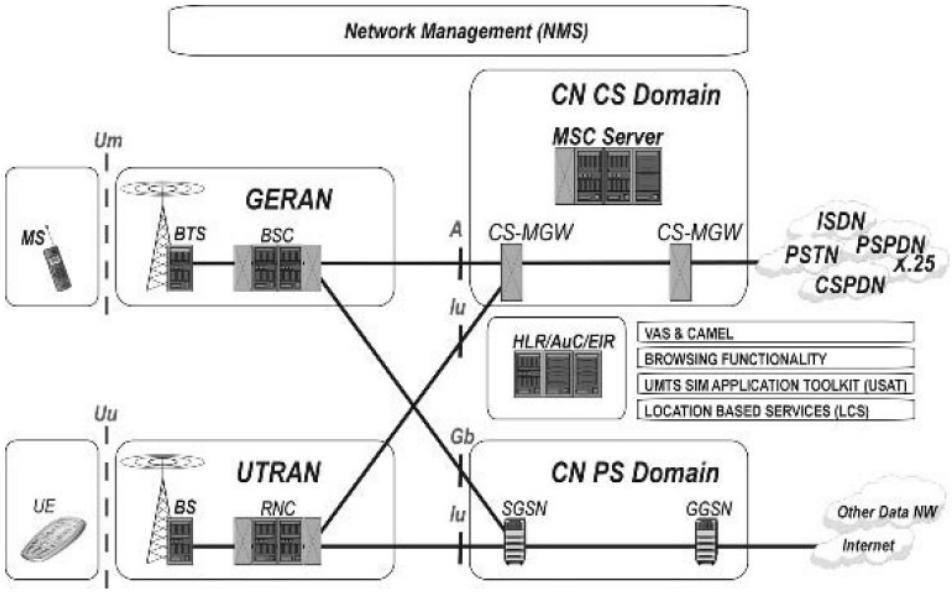


Figure 9: 3G network implementation on R4. Source: [5]

As can be seen from the figure above, in this release the UTRAN is connected to CN PS Domain and to CN CS Domain by the Iu interface. Also in Release 4 is possible to maintain a GSM GERAN through the A interface to the CN CS Domain and through the Gb to the CN PS Domain. Both, UMTS Iu interface and GSM A interface terminate into the CN CS Domain. The most important features about this release are [7] [18]:

- Use of Transcoder-Free Operation and Tandem Free Operation;
- Virtual Home Environment;
- Full support of Location Services;
- UTRA repeater and a robust Header Compression (ROHC).

2.5.6 Release 5

Release 5 brought a considerable amount of changes to the core network architecture which is marked by the introduction of the IMS as part of 3GPP, a standardized access-independent IP-based architecture which provides multimedia arrangements and mechanisms such as establishing a peer-to-peer IP communication to clients with the aim of quality of services, and provide an interworking with voice and data existing networks for fixed and mobile users.

Considered the central part of an IP core network, the IMS architecture addresses necessary functions to complete service delivery such as security, bearer control and roaming. Furthermore, its system solutions enables to use different networks including the UTMS [5]. IMS has the following functions and domains [7]:

- Separation of control and data paths;
- Implements All-IP network;
- Handles voice and data in a similar way;

The IMS core has the CSCF formed by many nodes. It replaces the MSC through which users used to communicate, and this communication are now via the SGSN and GGSN, please refer to Figure 10. The SGSN connects to GERAN base station system (BSC and BTS) through the Gb interface and to the UTRAN through the Iu interface (Gb interface). The SGSN and GGSN are interconnected through Gn interface also used to support mobility.

In Release 5, a change in the GERAN enables the BSC to be able to generate IP-based application packets, also voice transmissions are over IP and there is no need for CS. An all IP based network makes no separation between the PS and CS domains and to reduce development costs all transport technologies are uniform. Some of Release 5 features that makes it possible to provide a truly 3G services are [7]:

- High Speed Downlink Packet Access;
- Reliable end-to-end QoS for packet-switched domain and Intra Domain Connection of RAN Nodes to Multiple CN Nodes (Iu-Flex);

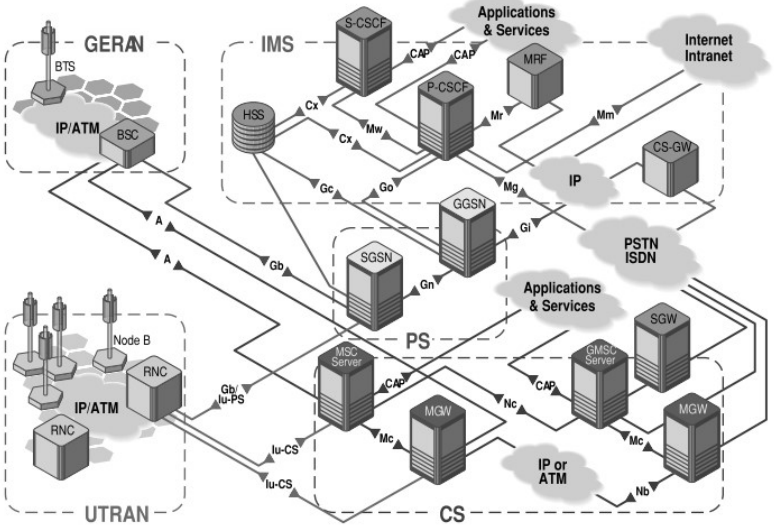


Figure 10: 3GPP network implementation scenario R5. Source:[19]

The Iur-g interfaces two BSSs or a BSC. And despite not transfer user plane the Iur-g interface between UTRAN and GERAN its used to transfer end-user-related radio functionality within the networks with the only purpose of signalling.

The SGSN and GGSN transports all IMS domains packet traffic such as voice. As for all protocol conversions within ISUP (PSTN) and the IMS call control protocols is performed by the MGCF which is an interworking management entity. This is only used when there is a user using a circuit-switched phone. The HSS is the element that connects the PS and IMS domains. Also, its used to access IMS domain function. The HLR and AuC functions are joined and played by the HSS which holed the processing files of all subscribers [7].

Essentially a CSCF is an enhanced SIP standard architecture that supports functionalities necessary for mobile networks such as call control functions in multimedia sessions, and is one of the core protocols for most voice over IP telephony services available developed initially for the fixed-line [16]. Figure 11 describes an IMS architecture.

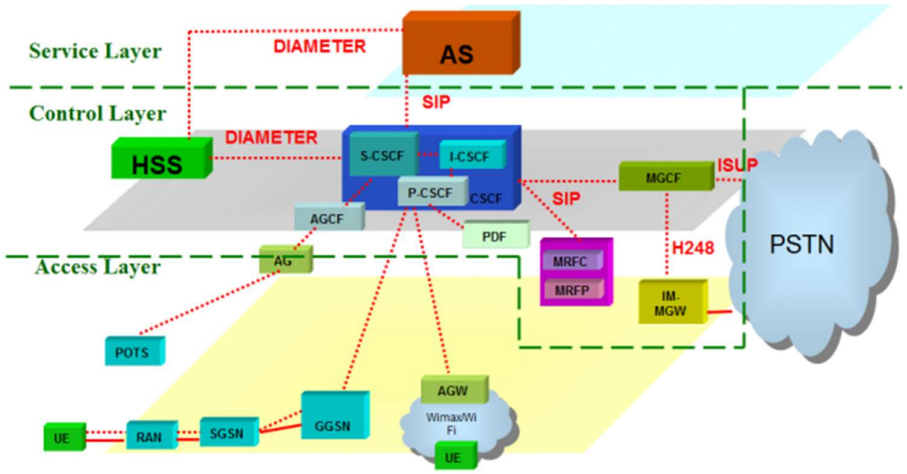


Figure 11: IMS architecture overview. Source: [20]

Being the center component or the Core Network Element in an IMS the CSCF can be divided into[7][21]:

- **Serving CSCF:** has the job of session-control services for a UE which includes decisions and establishment of routing, maintenance, and release of multimedia sessions, also generates charging information for the billing system. The S-CSCF remains the path for the subsequent SIP signals sent between a SIP dialog created by a request;
- **Proxy CSCF:** located in the same network as the GGSN, it is the first IMS entity to be contacted by the UE in a visited network session. This section is passed to the S-CSCF in the home network.
- **Interrogating CSCF:** manages all connections from an operator to a subscriber network. Also, search for the incoming SIP to request the correct S-CSCF within the network.

Apart from the IMS, the introduction of a new data transmission scheme is another important improvement of the UMTS Release 5. Also called HSDPA, this new feature improves data transmission speeds and capacity from the network to the user [16].

By employing a time-multiplexing approach to transfer data packets through a single shared channel and enhance the air interface using the 16-QAM, the HSDPA can provide a theoretical throughput of 14.4 Mbp/s with evolved terminals, which can be increase through adoption of MIMO antennas[7]. With the implementation of the HSDPA the UMTS network qualifies as 3.5G. The functional entities involved in a HSDPA includes[5]:

- Adaptive Modulation and Coding: indicates de dependency of a shared channel transport format;
- Fast Packet Scheduling;
- Hybrid ARQ: it is used as an adaptation scheme used for retransmission decisions of link layer acknowledgments.

2.5.7 Release 6

Release 5 was marked by the introduction of the HSDPA with a theoretical rate of 14.4Mbps used for download, which marked the first phase of the 3GPP High data packet specifications. Therefore, in release 6, the second phase, an uplink dedicated specification was introduced, the HSUPA. Figure 12 describes the basic Release 6 architecture.

The HSUPA comes to improve the bandwidth available per cell and per user in the uplink direction to improve the uplink speed given the fact that it has not increased since the launch of Release 99. With the introduction of the HSUPA the limited uplink speeds in networks under ideal conditions pass from the limited interval of 64–128 kbit/s to 384 kbit/s to, data rates up to 5.6 Mbit/s in theory for a single user under ideal conditions [16][22].

However, despite the introduction of the HSUPA to enhance the uplink in the RNC, the IMS and HSDPA continues to be evolved in Release 6. Indeed, in release 6 the 3GPP released the MIMO antennas, an enhancement to the HSDPA which allows an UE to reach data rates of 21.6 Mbps. Table 4 gives a comparison between the HSDPA and HSUPA.

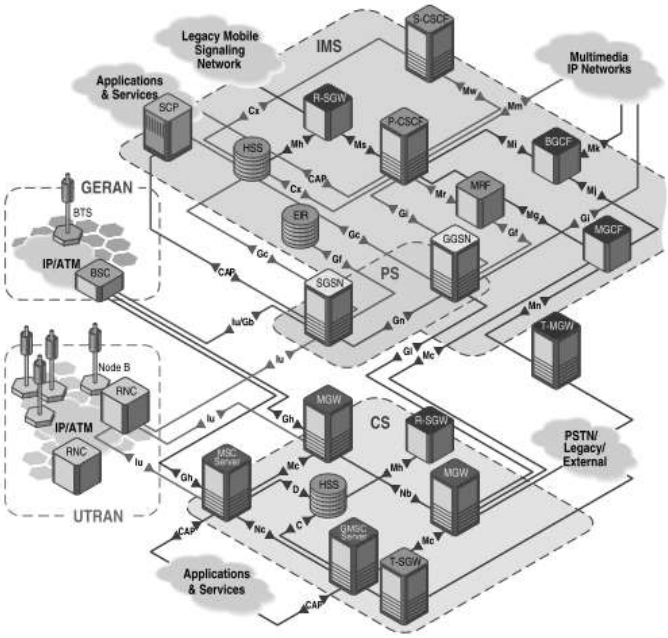


Figure 12: 3GPP network implementation scenario R6. Source: [15]

Table 4: Comparison between HSDPA and HSUPA
Source: [19]

Feature	HSDPA	HSUPA
Peak data rate	14.4 Mbps	HSUPA
Modulation scheme(s)	QPSK, 16 QAM	QPSK
TTI	2 ms	2 ms (optional)/ 10 ms
Transport channel type	Shared	Dedicated
Adaptive Modulation and Coding (AMC)	Yes	No
HARQ	HARQ with incremental redundancy; Feed Back in HS-DPCCH	HARQ with incremental redundancy; Feedback in dedicated physical channel (E-HICH)
Packet scheduling	Downlink scheduling (for capacity allocation)	Uplink scheduling (for power control)
Soft handover support(U-Plane)	No (in the downlink)	Yes

2.5.8 Release 7 and Beyond: Even Higher Data Rates

Published by the end of 2007, Release 7 is not different, it has been marked by a set of new substantial enhancements to the end-user performance, network capacity and radio technology improvements. These improvements mark a smooth evolution, compared to previous releases it includes mainly the introduction of HSPA and minor upgrades to terminal platforms and networks.

The HSPA is compatible to existing Release 5 and 6 and its capabilities put it closer to LTE. Its evolution is known as HSPA+ (also known as 3.9 G Architecture) and considered the last UMTS release, further releases are classified as LTE [23].

Besides the HSPA Release 7 has the following main features [17]:

- Policy and Charging Control: provides single policy infrastructure to handle QoS authorization and Flow-based Charging controls. Also, is applied to any IP-CAN and service network;
- HSPA+: an improvement to HSPA technology which allow to achieve higher bandwidth by applying 16-QAM an MIMO.
- IMS support of UE behind NATs;
- Extension of the HSUPA to the Time Division Duplex.

A subset of functionalities present in 3GPP releases upon the UMTS network is present in Table 5.

Table 5: 3GPP Releases on UMTS networks Source: [24]

	Release 99	Release 4	Release 5	Release 6	Release 7
Soft Handover	✓	✓	✓	✓	✓
Hard Handover	✓	✓	✓	✓	✓
HSDPA			✓	✓	✓
HSUPA				✓	✓
MBMS				✓	✓

2.6 LTE

Developed by the 3GPP the LTE is the successor to 3G UMTS, which in turn succeeds 2G GSM. It takes in account the possibility of world deployment by as many as possible entity requirements as possible. It uses the E-UTRA as its RAN which has a number of different operating bands from 700MHz to 2.7GHz and its bandwidths range from 1.4 to 20 MHz.

LTE has a very stable set of specification and enhancement which were motivated by [13]:

- The necessity to ensure the competitiveness of the 3G system for the future
- Low complexity and a continued demand for cost reduction of CAPEX and OPEX;
- A strong demand for higher data rates, low latency and QoS;

The adoption of a multi carrier approach for multiple access by the 3GPP in the LTE enables it to reach greater efficiency of radio spectral which allows improvement in time and frequency scheduling. While UMTS applies the W-CDMA, the OFDMA was adopted for downlink and the SC-FDMA (known as spread OFDMA) for uplink.

Some of the main features applied by LTE are resumed as follows [6]:

- OFDMA based;
- Frequency-Division Duplexing and Time-Division Duplexing support and interworking;
- Flexible spectrum support;
- Low Latency and High data rates;

LTE highlight specifications in Table 6 shows that it meets the requirements for high data download and upload speeds and reduced latency by providing significant improvements in the use of the available spectrum.

Table 6: LTE specifications.Source:[25]

LTE Basic specifications	
Parameter	**Details**
Peak downlink speed 64 QAM (Mbps)	100(SISO), 172(2x2 MIMO), 326(4x4 MIMO)
Peak uplink speeds (Mbps)	50 (QPSK), 57 (16 QAM), 86 (64 QAM)
Data type	All packet switched data (voice and data), No circuit switched.
Channel bandwidths (MHz)	1,4,3,5,10,15,20
Duplex schemes	FDD and TDD
Mobility	0-15 km/h (optimized), 15-120 km/h (high performance)
Latency	Idle to active less than 100ms Small packets~10ms
Spectral efficiency	Downlink: 3-4 times Rel 6 HSDPA Uplink: 2-3 x Rel 6 HSUPA
Access schemes	OFDMA (Downlink) SC-FDMA (Uplink)
Modulation types supported	QPSK, 16 QAM, 64 QAM (Uplink and downlink)

Similar to HSPA+, LTE stand for its increased rate and very reduced latency when compared to previous generations. Another important feature is the use of channel bandwidth, which can be of 1.4,3,5,15 or 20MHz. Theoretically, the bigger the bandwidth the greater the data transfer rate is. LTE technology apply MIMO to provide improved signal performance to the system by applying existing use of the radio path reflections. Despite the complexity present in the system in terms of processing and quantity of antennas required, when compared to OFDM, data throughput and

spectral efficiency created by MIMO presents a further significant improvement [26]. Some of LTE main specification is illustrated in Table 7.

Table 7: LTE specifications vs previous generations. Source: [25]

	WCDMA(UMTS)	HSPA HSDPA/HSUPA	HSPA+	LTE
Max downlink speed bps	384 k	14 M	28 M	100 M
Max uplink speed bps	128 k	5.7 M	11 M	50 M
Latency round trip time approx.	150 ms	100 ms	50 ms(max)	~10ms
3GPP releases	Rel 99/4	Rel 5/6	Rel 7	Rel 8
Approx. years of initial roll out	2003/4	2005/6 HSDPA 2007 /8 HSUPA	2008 /9	2009/1’
Access methodology	CDMA	CDMA	CDMA	OFDMA /SC-FDMA

2.6.1 Overall LTE Network Architecture

LTE architecture includes elements and interfaces which can be grouped into the following three main components:

- The User Equipment;
- The Evolved Packet Core;
- The Evolved UMTS Terrestrial Radio Access Network.

LTE UE internal architecture is similar to UMTS and GSM (Mobile Equipment). The core network is composed by many different logical parts and components that connects to one another. While the RAN contains only one node, the evolved node B (eNB). An example of LTE architecture can be seen in the Figure 13.

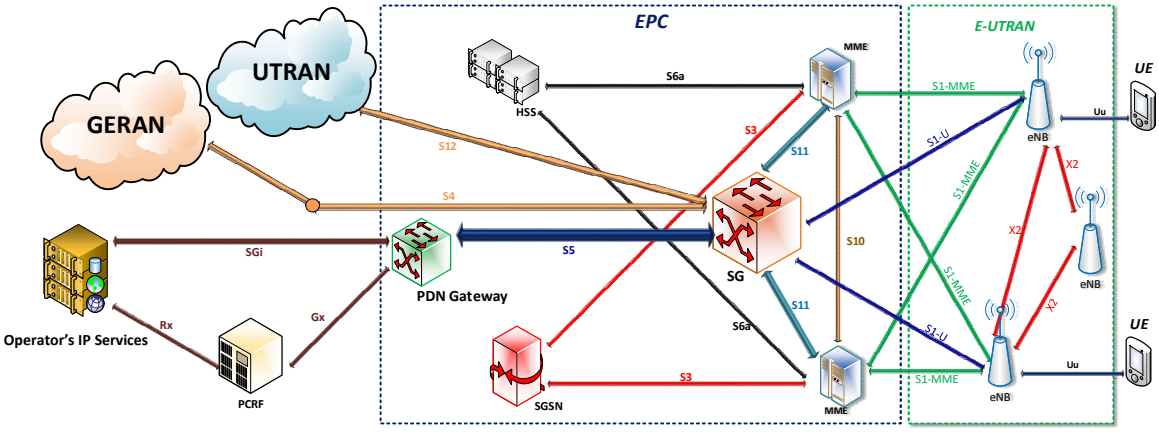


Figure 13: LTE architecture. Source: [15]

The S1 interface present in the LTE architecture can be divided into [27]:

- S1-U: used essentially for the user plane data;
- S1-MME: used to control plane information among MME and eNodeB;

For radio resource management, LTE uses the X2-interface which senses the handover command between eNodeB and also is used for data forwarding to route the data for the new eNodeB before core network tunnels are updated.

Furthermore, there is the EPC also known as the LTE core network, the evolution of the previous PS architecture used by GPRS/UMTS technology. The EPC is modelled to allow integration with other communication systems based on the IP using PS. Thus, it enables connectivity to other access technologies such as Digital Subscriber Line.

2.6.2 LTE Access Network

LTE access network consists essentially of eNB which combines the NodeB and the RNC from the 3G systems as one component. The eNB has the function of communicating with UE connected to the network providing the air interface between user plane and control plane, it also serves one or several different cells at time [12], [28].

Given the fact that eNB combines NodeB and RNC, is possible to see that RAN its composed by eNB only as said, so there is no centralized intelligent controller in the E-UTRAN. And each eNB is a base station that controls the mobiles in one or more cells. The eNBs are interconnected to each other through the X2-interface and to the core network by the S1-interface. The AS are protocols that interact across the UE and eNBs.

Similar to the eNB, there is the Home eNB considered a lower size and lower cost eNBs whose function is to serve a femtocell which often covers indoor areas such as homes, subway stations or shopping centres. The HeNB, sometimes called femtocell can be connected directly to the EPC.

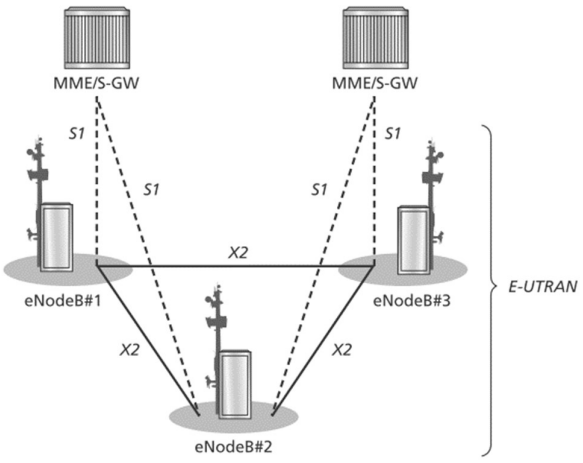


Figure 14: LTE Radio Access Network. Source: [26]

2.6.3 LTE Core Network

This section addresses the core of LTE systems, the EPC. The EPC is the current core networks architecture created by the 3GPP specifications that is both flat and all-IP-based which is based on the idea of handling the data traffic efficiently from performance and costs perspective[29] it can be said that the EPC is exclusively IP-based.

Furthermore, the EPC is compatible and can be accessed by different radios access such as UMTS. And as a consequence, with reference to [28], the multi-radio compatibility of the EPC has become very attractive to the operators since it allows to have a simple core which supports different services.

Different of the GPRS/UMTS in the LTE networks the user data, user plane, and the signalling, control plane, has been separated to make it scale independent bringing better improvements in the field of dimension and adaptation of network by network operators. This is another reason the EPC is considered flat. As depicted in Figure 15, the core of the EPC is composed by following logical nodes components:

- **Serving gateway(S-GW):** responsible in handling user data functions essentially routing and forwarding packets between UE and PDN. It also serves as a mobility point for handover[12] besides it is logically connected to other gateway, the PDN GW[29].
- **Packet Data Network Gateway (PDN-GW or P-GW):** works as a gateway to the Packet Data Network(PDN) routing packets from/to them, this is due to the fact the PDN-GW interconnects the EPC core with other external IP networks. Besides the 3GPP specify these gateways independently in practice they can be combined as single by network vendors [29].
- **Mobility Management Entity (MME):** manages the UE access and mobility establishing the path carrier for it [12]. Also, it handles security for E-UTRAN access related to end-user authentication which can be described/coupled in the following MME supported functions groups [30] :
    - **Security procedures**: further than handling end-user authentication, as explained before, can be add to MME security procedures the task of negotiating, ciphering and integrity protection of algorithms.
    - **Terminal-to-network session handling**: allows to negotiate associated parameters like the Quality of Service also the signalling procedures used to set up Packet Data context.
    - **Idle terminal location management**: enables the network join of terminals in case of incoming sessions by tracking areas of update process.

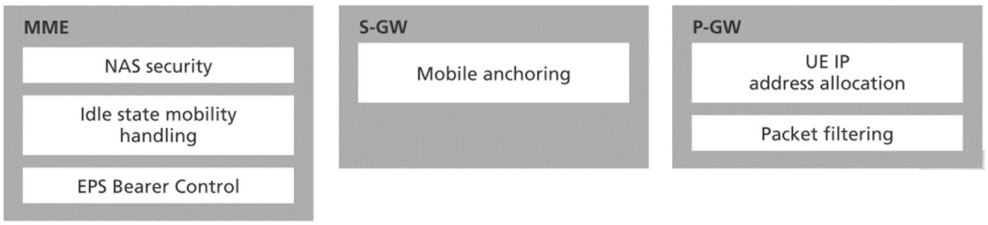


Figure 15:Logical node of the EPC. Source: [26]

The EPC also includes PCRF and HSS as logical nodes which are described as follows:

- **Policy and Charging Rules Function (PCRF):** its responsible for the management of policies, decision making and charging rules. It also provides QoS authorization and sends QoS information to user's sessions. The PCRF ensures that a certain data flow will be treated in the PCEF (provides policy enforcement and charging functionalities) in accordance with the user's subscription [26].
- **Home Subscriber Server (HSS)**: contains information of the PDNs where the users can connect an information about the identity of a MME upon which a user is registered. It contains the HLR and the AuC present in 2G/GSM and 3G/UMTS networks.

2.6.4 Releases 8 and 9 LTE

LTE first version was documented in Release 8 of the 3GPP specification which the previous sections were based. The following releases only enhanced the technology. In this release, the maximum data rate was of 150 Mbps. Particularly, Release 9 introduces broadcast mode support based on Single Frequency Network type transmissions.

Nevertheless, some implementations of LTE consist on nonradio aspects such as system architecture evolution which includes the Evolved Packet Core network which when combined comprise the Evolved Packet System (EPS)[28]. Release 8 initial left behind enhancements were included to LTE in release 9. Some of these improvements include the following listed elements [27]:

- **PWS (Public Warning System)**: which should include different types of alerts to the user such as alerts related to natural disasters or other critical situations;
- **Femto Cell**: consisted on the implementations of small cells inside offices or homes which were connected to networks providers using a landline broadband connection.
- **Self-Organizing Networks (SON)**: allows a significant cost reduction by applying means of self-installation to reduce manual work.

2.6.5 Releases 10 and 11 LTE

In Release 10 and 11, the first LTE-Advanced release was introduced which brought clear increase in data capabilities for LTE networks when compared to 3GPP release 8 specifications. It changed the previous data rate of 300 Mbps for downlink and 75 Mbps for Uplink from release 8 and 9 to a peak data rate of 1Gbps for downlink and 500 Mbps for uplink (refer to Table 8). These data rate could be improved under good signal conditions by combining up to five a 20 MHz carrier which forms 100 MHz bandwidths.

The set of improvements introduced with Releases 10 and 11 LTE Advanced are as follows [27]:

- **Carrier aggregation improvements**: allows a UE to transmit and receive using multiple carriers in downlink and uplink direction.
- **Evolved MIMO operation**: which support up to eight transmiters and eight receiver antennas in the downlink direction.
- **Heterogeneous network operation**: Improved support for co-operation from multiple transmitters or receivers. It allows the combination of large macro cells with small cells networks.
- **Enhanced Uplink multiple access**: by implementing clustered SC-FDMA in uplink.
- **SON Improvements**: enhances SON features introduced in release 8 and 9.

Table 8: LTE and LTE-Advanced capacity comparison. Source: [31]

Parameter		LTE	LTE-Advanced
Scalable bandwidths		1.4-20 MHz	20-100 MHz
Peak data reate downlink	DL	300 Mbps	1Gbps
	UL	75 Mbps	500 Mbps
Transmission bandwidth	DL	20 MHz	100 MHz
	UL	20 MHz	40 MHz
Peak Spectrum Efficiency [bps/Hz]	DL	15	30
	UL	3.75	15

2.6.6 LTE-Advanced Protocol Stack

The protocol stack of the LTE is divided into two components: the user plane and the control plane which its depicted in Figure 16. Both user plane and control plane have two Layers of protocols which are: Access Stratum and Non-Access Stratum.

The AS of the eNB, is responsible to provide E-UTRAN the necessary protocols for the user control and user plane. The protocols stack for both the control plane and the user plane in the U-ETRAN accomplish the same goals besides the control plane doesn't allow header compression function. The user and control protocols are as follows:

- Packet Data Convergence Protocol (PDCP);
- Radio Link Control (RLC);
- Medium Access Control (MAC);
- Physical Layer (PHY) protocols;

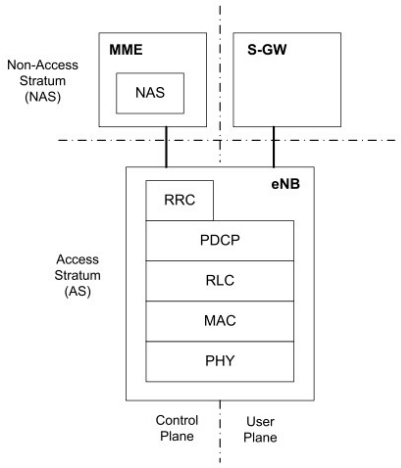


Figure 16:LTE Advanced stack. Source [28]

All data management taken place during handover that belongs to a user who is moving in a E-UTRAN is performed by the eNodeB if any centralized controller node is present. As for data protection in such a situation, the PDCP become in control of the handover and is used for control plane signal as well. For further detail on roaming architecture of an LTE network please refer to [26] section 3.3.

Also, the control plane includes Radio Resource Control protocols also known as layer 3 by the AS protocol layer. The RRC layer is the main control function inside the AS and is used to establish radio

bearers and configure lower layer between eNodeB and UE. Also it is used to cover all messages such as reports of measurements, connection configuration and RRC reconfiguration [26].

The Non-Access Stratum (NAS) is a set of transparent access network protocols that are part of the EPS. It is the higher layer of the control plane and often used to grant non-radio signal among the User Equipment (UE) and the Mobility Management Entity (MME) for an LTE/E-UTRAN access[32]. It also supports management of sessions procedures to be maintained while an IP connectivity between UE and PDN-GW is being established. The NAS may have the following functions[28]:

- Enables mobility connection management between UE and the core network as the user moves;
- Identity management (Authentication, Registration);
- Location registration management.

The current state of mobile telecommunications networks is characterized by cohabitation between 2, 3 and 4 generation specifically by GSM, UMTS and LTE technologies. Figure 17 represents the current state of this cohabitation. It can be seen a sharing of equipment such as the RAN by the technologies.

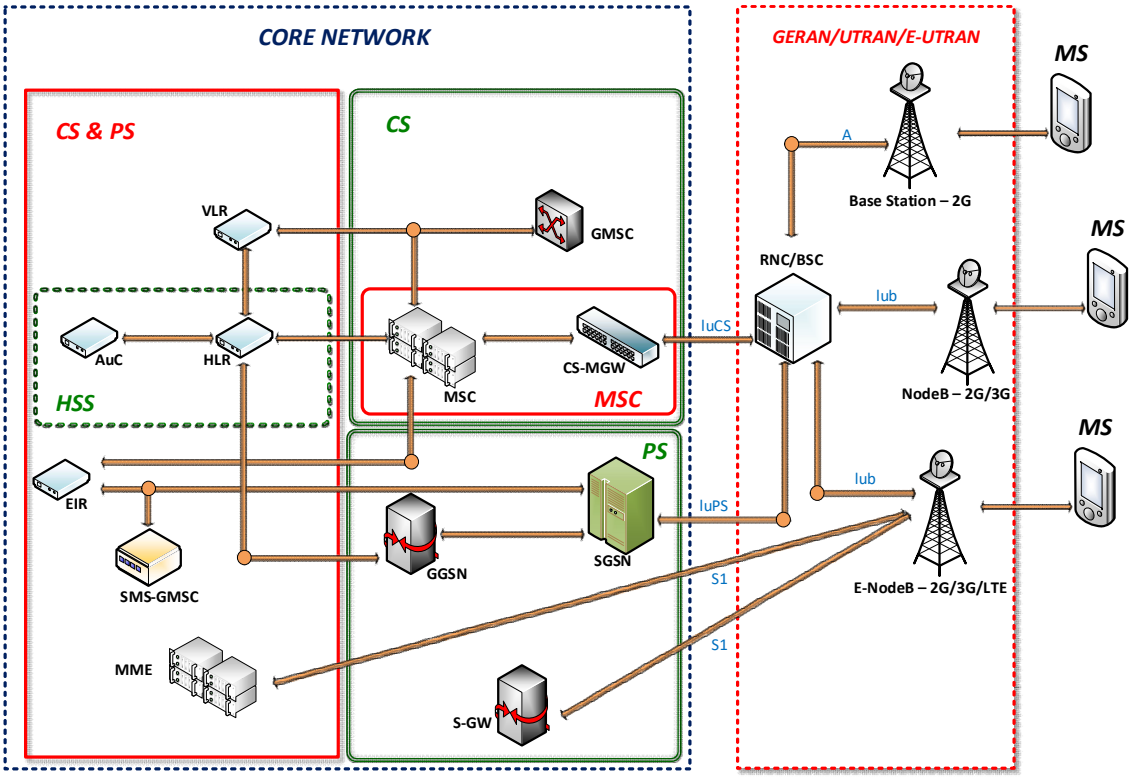


Figure 17: Mobile Telecommunication cohabitation. Source [15]

2.7 Network Management

In Telecommunication context, Network Management consists on a set of capabilities of a given system and elements to allow an exchange and processing of information which can be used to assist communication service providers in conducting their business efficiently. It has the element

management systems and an actual network management systems and includes the service management and customer care [33]. It encompasses three main areas of concerning:

- Administration of network systems: which deals with network control;
- Provision: simplify service requirements related to network configuration;
- Maintenance and Operation: keeps the network and its services working by reducing the impact on user experience;

In network management area is possible to find several models which helps to incorporate actions and operational activities to the network. Some of the existent network management models includes:

- Telecommunications Management Network;
- FCAPS;
- Information Technology Infrastructure Library;
- Cisco Lifecycle Services;
- Enhanced Telecom Operations Map (eTOM).

A more exhaustive description on network management models can be found in Appendix 1.5.

2.7.1 Network Management in UMTS

UMTS network management architecture has its base on the TMN and contains definitions only for general management frameworks and concepts as discussed on the previous section.

Some of UTMS main management functions are describe as follows [7]:

- Software management;
- QoS management;
- Performance management;
- User equipment management;

A. Performance management

Performance management function is used to assess the current configuration of the network for data colleting. This function allows to collect the following type of data:

- User data and control signalling traffic levels in the network and the usage of resources by network nodes;
- Network configuration and verification;
- Resource-access measurements and availability;
- QoS measurements experienced by the user;

B. QoS Management

Includes both QoS policy provisioning and QoS monitoring, where the first is used to configure and maintain network elements with the QoS policies created based on network performance. And the second, is used to collect all statistics data related to performance which are then used to generate reports to be used when making changes to the network.

C. Software management

This function is divided into software-management process, used to manage new software releases or correction; And software-fault-management process used to handle software faults identified when network monitoring.

D. User equipment management

Also known as “user tracing”, it allows a network operator to trace a particular subscriber activity and report to network management system.

2.7.2 Operation Suport System

A lot of different and simultaneous network operation are being employed by various communication service providers. The OSS can be used when managing complex service and network planning. It includes a set of different applications required by communication service providers to run their “back-office” functions[34]. By applying the OSS applications all those functions can be automated reducing the total time taken to complete a certain repetitive function (such as routing costumer). Some OSS main functions include:

- Analysis of networks and services: used to help in decision-making and optimal network design.
- Identification of focus points: by using network record, it helps the identification of products sold in a certain region;
- Simplification and automation of operational tasks: enable the construction of products form network resources building blocks.
- Gather data: helps collecting data about the network state which can be used when design and implementing strategic plans.

2.8 Key Performance Indicators

KPI is one of the most important performance measurement indicators used for scanning the whole network or specific parts at specific time with the purpose of finding problems or analyzing the network [35] such as: performance, trend analysis, errors and bottlenecks detection.

Since KPI are nothing more than measurement of parameter used for monitoring, there a many parameter that can be used. However, there is not a criterion to choose the right parameter which means, it is up to the communication service providers to choose the KPI that suits its needs. The communication service provider chooses the parameter of its interest to target the indicator that best describes its network behavior. The network behavior is tracked by different counters distributed in different nodes of the network to pick parameter measures of interest which then is managed by an OSS.

There is a wide variety of KPI, however they are divided into the following [36][37]:

Table 9: KPI categories counters

KPI	Counter
Accessibility	• Radio access success ratio; • IU paging access ratio; • AMR RAB setup success;
Mobility	• Soft handover success ratio; • Inter/intra frequency hard handover success ration;
Availability	• Worts cell ratio; • Congested cell ratio; • Average CPU load;
Traffic	• CS equivalent traffic; • PS UL throughput; • PS DL throughput;

In the case of this dissertation, the air interfaces either of UTRAN (in the case of UMTS) or E-UTRAN (in the case of LTE) would be the area of interest to take measures.

Specifically, the RNC, a critical part of the mobile network will be the place where counters will take performance measures using traffic KPI counters. Traffic KPI counters are essentially used to evaluate all CS service equivalent traffic from a specific RNC [36] These measures are provided by the OSS which gather measures by defined time intervals. In fact, in this dissertation forecast models will be applied to traffic KPI data set counters made available by a communication service provider.

An interesting aspect of the KPI is the ability to estimate them, which helps communication service provider simplify their network planning making it possible to find areas affected by a problem and invest more time in it, instead of losing time with non-identified problematic area. A good estimated KPI help communication service provider provide a good subscriber experience, an important factor, since users are only interested in not having problems that will interrupt their voice or data call.

2.9 Mobile Data Traffic Growth

World mobile data traffic has been registering a significant huge increase in the last few years, which has been driven by a large and constant introduction of new technological devices such as smart mobile devices (refer to Appendix 1.2), also a higher number of mobile broadband subscribers and a supply of higher data rates at affordable prices by mobile network operator due to competition among then.

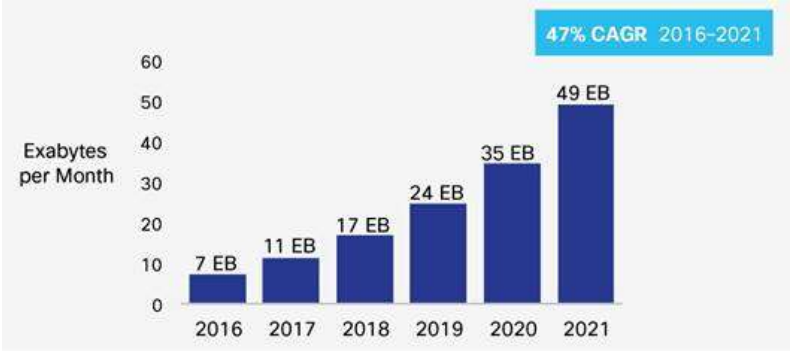


Figure 18: Forecasts per Month of Mobile Data Traffic by 2021. Source: [38]

The mobile data is expected to grow an overall rate of approximately 47% from 2016-2021, almost 49 exabytes per month until 2021 which represents an increase of 7 times over 2016 [38]. This growth is illustrated in Figure 18, so it is possible to see that smart mobile devices traffic will dominate even more as time passes.

Nowadays mobile devices trend is assumed to be growing much faster than any other device due to several factors such as always-online applications. As the growth is expected to continue, besides the need for more network installations, spectra and radio optimization and deployment of small cell, it's also necessary to understand a network behaviour to better understand when it will be reaching its limit to allow it to expand without causing any increase in the network operator revenues.

This fast growth has led communication service provider to a challenging task of providing sufficient data capacity for mobile users and avoiding problems in the network such as congestion. Thus, this kind of problems can be prevented by applying estimation of RNC traffic load in both the user (data traffic) and control plane (signalling traffic) which can be made by applying forecasting techniques.

3 Forecasting

Forecast may serve many different needs such as: economics (stock exchange forecast), outcome of political elections, telecommunication networks performance, etc. This section addresses issues related to mobile network behaviour (capacity, throughput, etc.). An extensive literature already exists in the field [39][40]. In broad terms, the existing literature can be divided in two major approaches:

- Historic data (time series) based forecasting;
- Diffusion of innovation forecasting;

In this dissertation the historic data based forecasting approach will be considered. Essentially, time series based forecasting consists of predicting, as accurately as possible, the future of an event given all past data available [41]. Or, in other words, forecast consists on estimating as accurately as possible values that are not yet known based on the known values.

Some literature [42][43][44] use the term **prediction** when referring to events independently of their time of occurrence and **forecast** as referring to an event associated with its time of occurrence.

An outsider from the forecasting world may doubt the credibility and efficacy of forecasting certain aspects from the future. However, the art of forecasting can be an important support for efficient and effective planning[42][40]. The use of forecasting is crucial in any organization daily planning activity. It is a fundamental business function since almost all concerns are based on forecasting. Inadequate forecast may have bad consequences on business decisions making. It can let an organization unprepared to meet nearly future demands [45].

Forecasting has evolved and nowadays there is a vast amount of tools that can be applied in different situations (e.g. finance, weather, equipment life cycle, etc). However, there are still domains where current forecasting technologies have not yet been fully exploited. Out of them is planning in information and communication system, namely in mobile network. According to [41] the predictability of a given event may depend on several factors which includes:

- How well the factors that contribute to it are understood;
- How much data are available;
- Whether the forecasts can affect the thing being forecasted.

3.1 Telecommunication forecasting

Every product or service has a life-cycle which can be divided into the following segments: introduction, growth, saturation, and decline. Thus, it is important to understand and forecast each segment of Service life-cycle (SLC) as a matter to improve business planning purposes

The utilization of forecasting techniques may lead the mobile operators to be prepared to solve network problems such as reduced QoS, bad traffic service almost reaching capacity limit, queuing delays, packet loss and blocking of new connection. Telecommunication forecasting are usually based on indicators such as:

- User growth,
- Traffic growth
- Market share,
- Volume - Pricing,
- Average revenue per user (ARPU) forecasting and forecasting of revenue in total.

In this dissertation, the use of forecast will be applied to help communication service providers in the management of the network degradation of Quality of Service (QoS). Given the fact communication service providers are suffering revenue losses because of the existing difficulty in knowing when a NE has their capacity reached. The application of forecasting models makes the communication service provider able to know the estimated time when a capacity upgrade is needed.

3.1.1 Forecasting Methods in Telecommunication

Nowadays, it is possible to find a wide variety of existing methods that are used for forecasting. Those methods can be divided into the following categories:

- Qualitative methods;
- Quantitative methods

In telecommunications forecasting, there are several software tools and methods available, thus according to [1], the following are the most often used:

- New telecommunications service penetration forecasting by using growth models which includes logistic model and Bass growth model;
- Statistical forecasting models based on seasonal variations elimination;
- Cross-section models for forecasting based on the relations between different services;
- Consensus method by seeking expert opinions to make a forecast or the Delphi method;
- Simulation or scenario methods;

In this dissertation, will be tested, analysed and applied several quantitative statistical forecasting models.

3.2 Forecasting Categories

As said in the previous section the forecasting methods can be divided in two basic categories which are described as follows:

- **Qualitative forecasting:** based primarily on the forecasters judgmental opinions which can lead to subjective and nonmathematical human opinion thus its often-called judgmental method or subjective forecast.
    - Some methods that can be cited are [1][46]: Judgmental method, Delphi method and Scenario method, which is somehow related to the expert experience;
- **Quantitative forecasting:** Are methods based on analytical and statistical models of an observed phenomenon. It uses consistent mathematical modelling to generate objective forecasts taking in consideration a lot of information at one time. Also known as objective forecast.
    - Quantitative important subcategories are [47][1]: Time series methods and causal methods.

Generally, quantitative forecasting problems uses either time series data (collected at regular intervals over time) or cross-sectional data (collected at a single point in time)[41]. The forecast methods defined above can be categorized by their weakness and strengths as described in the table below:

Table 10: Characteristic of forecasting types. Source: [48]

	Qualitative Methods	**Quantitative Methods**
Strengths	May incorporate Latest changes both in the environment and inside information	Consistent and objective which allow to consider and data at one time
Weaknesses	Reduces forecast accuracy	Depends on quantifiable data

Since this dissertation will be working on quantitative methods, the qualitative part of the forecasting will not be addressed, for further information please refer to [48].

3.2.1 Quantitative Methods

The quantitative methods have the advantage of being consistent and objective over the qualitative because it is based on mathematics. These methods can be grouped in two categories [1][49]:

- **Time-series models:** consists of a set of observations taken at regular intervals over a specific period. It takes into consideration only the history of the variable to forecast.
- **Causal models:** identifies the causal relationships between the forecast variable and a set of predictors (causal factors). Makes use of regression models and various techniques for the evaluation of their applicability, as well as the reliability of forecasting results.

The group described above are part of the quantitative method which makes them both mathematical but apart this similarity they differ in the form of generating forecasts. Time series models easily generate forecasts over causal models which require a more complex process such as model building.This basic difference can be observed in the table below [46]:

Table 11: Comparison between time-series and casual models

Time-series models	Causal models
Can generate a forecast based on patterns on data; All information necessary to generate a forecast is contained in the data;	The variable to forecast is related to other variables; Can be very complex and not very accurate;

3.3 Time Series Analysis

Time Series Analysis consists of a set of several quantitative data observations taken over different periods of time at regularly spaced intervals [44][50][51]. Time series analysis are among one of the most important problems analysts have to face. Time series analysis has many business and social applications which can be used in a wide range situation such as modelling, to produce forecasts of a company sales or an agricultural production, product demand, create a stock fitting model, analyse market trends, etc.

The following picture describes a time series plot example where can be seen a constant set of observations over time. The plot content is from the U.S. annual production of blue and gorgonzola cheeses time series which shows a linear trend with a positive slope trend over the years.

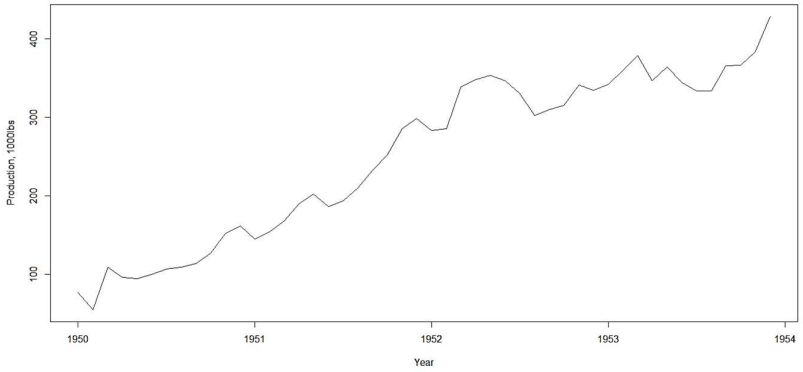


Figure 19:Time Series Example. Adapted from [51]

3.3.1 Time series components

A pattern is a set of components that may compose a data. Time series are divided into the following four different components [48]:

- **Trend**: is the overall direction of a time series i.e. upwards, downwards etc. (may be associated with the cycle component resulting in the trend-cycle component which contains both trend and cycle [41]). A time-series is said to have a trend when data shows a long-term increasing or decreasing behaviour over time as can be illustrated by the Figure 20.

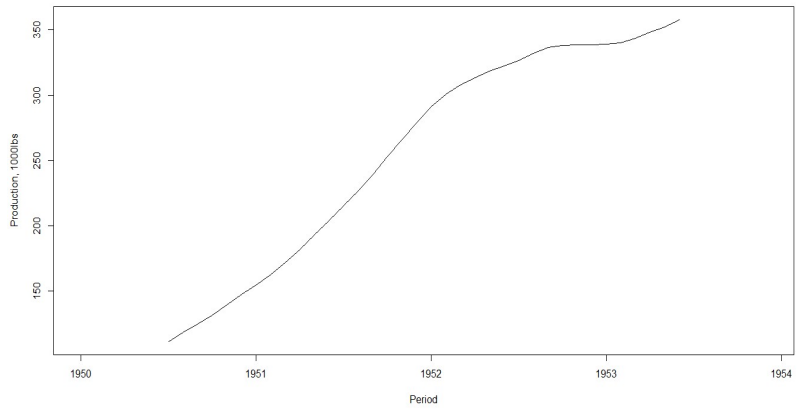


Figure 20: Trend Example. Source: The author

- **Seasonality**: exists when a times series often repeats a certain behaviour e.g. monthly, quarterly or quarterly at a constant level base. This type of series is sometimes called "periodic" despite not repeating over each time period [47]. Figure 21 depicted an example of the seasonal component.

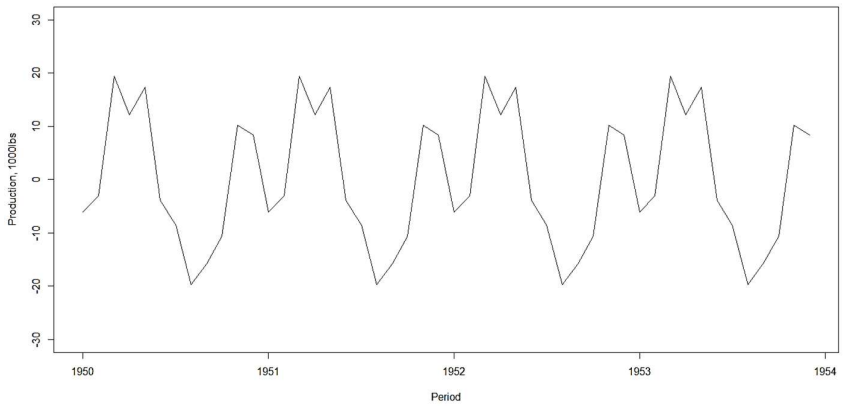


Figure 21: Seasonal Component. Source: The author

- **Irregular (remainder or random)**: it's the noise left after the extraction of all components, not essentially white noise [52]. It may sometimes be referred as random variation which consists of unexplained variation that may occur and cannot be predicted. An example is given in the Figure 22, there a random sequence is plotted from where is obvious to see no pattern present.

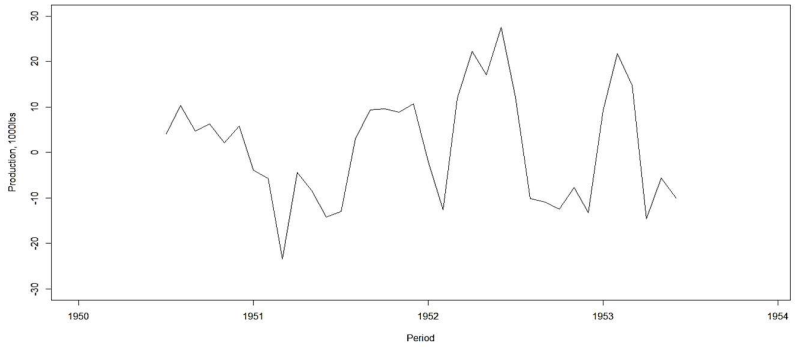


Figure 22: Random component. Source: The author

- **Cycle**: long-term business cycles. It manifests by exhibiting unfixed rises and falls over certain period. Generally, where there is a cycle component, is also possible to find seasonality, thus some literature may treat both components together.

The concepts of seasonality and cycle may sometimes generate a misunderstanding. When referring to seasonality it usually has fixed and known-length and magnitude in other hand the cyclic pattern varies which makes it very difficult to forecast comparing to other.

Figure 23 helps visually perceive the relationship between the components that can be found in a time series. In relation to the observed values the trend represents the actual behaviour of the series which is increasing. The noise is a fraction of values in the series with a very small amplitude and finally the seasonality which illustrates the seasonal activities of the data.

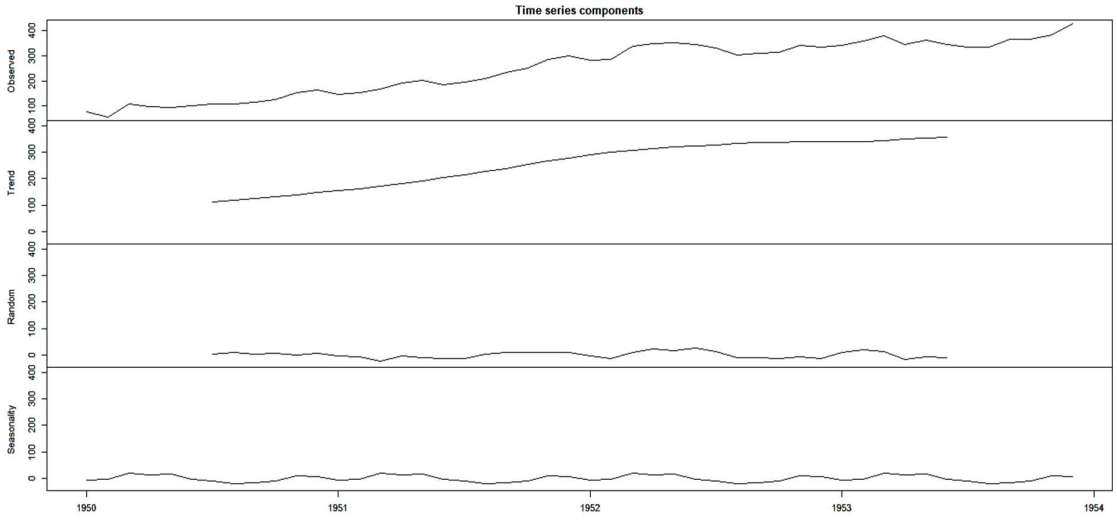


Figure 23: Time Series components. Source: The author

Another component to be found in a time series besides the patterns presented before is the level. The level constitutes the average of all values in the series [51]. In addition, the schematic in Figure 24 shows the pattern components.

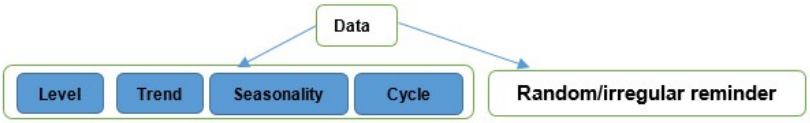


Figure 24:Data Components. Source: The author

Some authors may not separate the trend and cycle component for long-term behaviour time series given the difficult on separating them [53][54].One of the main forms of identifying time series patterns consists of plotting the data for examination using graphics, topic addressed next. All previous components when add up produces the final series. If a time series is decomposed the components are obtained and when re-composed, the result is the initial time series.

3.3.2 Graphics

Often when working with time series analysis the first thought is about the complex statistical and mathematical models leaving the graphic representation part aside. Graphing the data set is one of the most important steps in time series analysis. Indeed, it allows and helps identify peculiar patterns such as outliers present in the data. It can be set as an objective of graphic representation, the display of data as accurately, clearly and consistent as possible.

Nevertheless, there are two common types of graph often applied to represent time series data which are:

- **Time plots**: plot values against time (can only display time on the x-axis) which are them linked to each other by a straight line helping in visualizing the data and its analysis. Also known as time series graph (line plots). An example of this type of graphic is depicted in Figure 19.
- **Scatter plots**: plot two different variables value against each other. Scatterplots are very important when analysing variables relationships. Though it helps find different types of correlations among variables. An example of a scatter plot is represented in Figure 25.

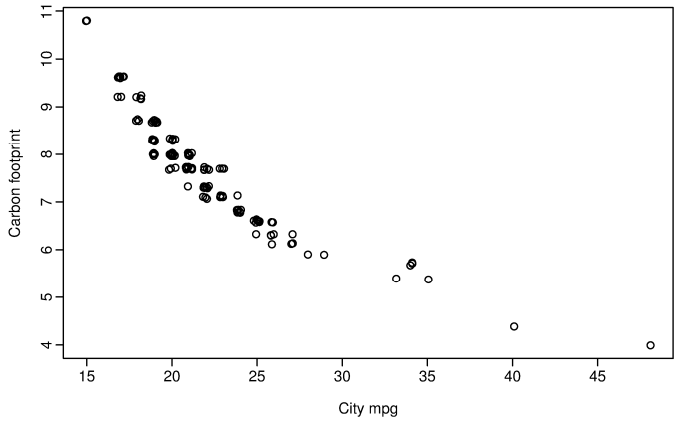


Figure 25: Scatter Plot Example. Source: [41]

The type of graphic varies accordingly to the type of data to be plotted, as a matter of fact, as the method of forecast is determined by the type of data so the graph to use is [41].

However, considering that scatter plot is used when there is more than one variable, it will not be used in this dissertation since the data being used contains one variable. In addition, to support the identification of time-series patterns the use of plots may help the forecaster draw an attention to the occurrences present in the data such as atypical events.

3.4 Forecasting evaluation methods

A model can be well fitted to a historical data set in analysis, but this will not guarantee that when its forecasts values are compared with future values, these will be good. This section presents some basic methods used to give some additional information about a forecast model quality by using measures to evaluate its accuracy and goodness of fit.

The use of forecasting accuracy methods helps answering the question “how good is a forecast?”. This dissertation will be analysing forecast accuracy in two separate aspects first by measuring the forecast accuracy and second by comparing the methods in study accuracy result’s. For an actual data observation x_t in a t period and its forecast F_t, the error defined as:

$$e_t = x_t - F_t \qquad \text{Equation 1}$$

For each t time periods observations there will be t error terms. The most commonly forecast measures of accuracy are [55]:

- Mean error and Mean Absolute error-MAE;
- Mean scaled error and Mean absolute scaled error-MASE;
- Mean percentage errors and Mean absolute percentage errors;

A. Mean Absolute error

The mean absolute error (MAE) consists on measuring the difference between two continuous variables, the forecast and the data observation. MAE measures accuracy based on forecast bias, for large deviation from zero will mean the presence of bias, and the forecast is not good. Also, MAE tells the size of error to be expected from the forecast on average. This method is based on the mean error which can be found embedded in most error based methods.

The MAE can be represented by the following equation:

$$\frac{1}{n}\sum_{k=1}^{n}(|x_t - \widehat{x_t}|) \qquad \text{Equation 2}$$

B. Mean scaled error MSE

The mean scaled error (MSE) is a non-negative estimator used to measure the average squares errors. MSE values closer to zero indicates how good a certain model fits. It can be represented by the following equation:

$$\frac{1}{n}\sum_{k=1}^{n}(x_t - \widehat{x_t})^2 \qquad \text{Equation 3}$$

The difference between the MAE and MSE is that MAE is more sensitive to small deviations near zero and less sensitive to larger deviations.

C. Mean absolute percentage errors

The Mean Absolute Percentage Error is one of the most used forecast accuracy measures applied in different literatures [47][51][45] and also present in the [2] recommendation.

Statistically, MAPE is defined as the average of percentage errors and is represented by the following equation:

$$\frac{1}{n}\sum_{k=1}^{n}\left(\left|\frac{x_t-\widehat{x_t}}{x_t}\right| \mathrm{X100}\right) \text{ or } mean\left(\left|\frac{x_t-\widehat{x_t}}{x_t}\right|\right)\mathrm{X100} \qquad \text{Equation 4}$$

When evaluating using MAPE, the model which minimizes is considered ideal for forecasting the data. A null MAPE means a perfect model fit. However, a special attention has to be taken for time series that contain zeros values in the observation. Percentage errors cannot be computed when the value is zero. To find the MPE remove the module from the equation above.

The MAPE comes to answer the problem of MAE and other methods of depending on the scaling of the variable an inconvenient when comparing accuracy [55]. Although there are other variations of the MAPE method such as symmetric MAPE, the use of MAPE is preferred [56].

D. Akaike Information Criterion

Introduced by Akaike in 1969, the AIC a widely used measure of statistical model, is a method or criteria used to select a model from a set of models by quantifying the goodness of fit and simplicity. In case all models are considered poor, AIC selects the best among the poor's without worrying about.

Despite helping find the best model among a set of models it does not give any indication about the absolute quality of the selected model. This method is often applied to select ARIMA models by comparing the result of their AIC. For example, if one compare AIC1 for MODEL1 and AIC2 for MODEL2 the one with the lowest AIC value is preferred.

The AIC is represented by the following equation:

$$\mathrm{AIC} = \ln(\widehat{\sigma}_{\varepsilon}^{2}) + \frac{2\mathrm{k}}{\mathrm{T}} \qquad \text{Equation 5}$$

Where:

k → Sum of ARMA parameters $(p+q)$
T → Observations in analyse
$\widehat{\sigma}_{\varepsilon}^{2}$ → Residuals Variance

A variation of the AIC used for smaller sample data set is the AICc describe as follows [41].

E. Corrected Akaike's Information Criterion

AICc is defined as an AIC corrected for finite size samples which allows a greater penalty for extra parameters.

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2(\mathrm{k}+2)(\mathrm{k}+3)}{\mathrm{T}-\mathrm{k}} \qquad \text{Equation 6}$$

As with the AIC, the AICc should be minimized and the model with the lowest value is preferred.

3.5 Forecasting Models

Forecasting estimation is a difficult exercise and involves a multiplicity of forces. Also, it requires one to be familiarized with different types of methods and technics to better deal with different types of situations. Along the years, several mathematical methods of forecasting have been developed to solve different types of situations such as: demand, economic, technological forecast and so on. From those, there is no easier method nor simplest formula that allows to forecast the future. Instead one of each method has its own behaviour thus it is necessary to study and test several methods to select the most suitable for each situation needs.

From the existing forecasting methods the most mathematical existent models and also recommended by [2] are:

- Basic methods (naïve, seasonal naïve and simple average)
- Smoothing models;
- ARIMA models;
- Decomposition;

Besides being a recommendation from [2], another reason for the above chosen methods is due the fact of it using historical data and to performs better in relation to more statistically sophisticated and complex methods (for further information please refer [57]).

Another model that can also be used is the linear regression (linear trend model) and curve fitting. However, in this work it will not be implemented because it captures only events with intercepts-slopes and discards other components such as seasonality and cyclicity. For further details please refer to [6].

3.5.1 Basic Forecasting methods

This section looks on basic forecasting methods, their characteristics and applications on time series data.

A. Naïve Method

This method forecasts data based on the assumption that forecast period value is the same as the most recent value. For example, given the Table 12 about the internet users in Europe & Asia [58], to forecast the user quantity for the year 2016 when applying the naïve method just take the previous observed value which is 651396608, value for the year 2015. In other word for a certain period h, the $t + h$ forecast value will be the previous h value.

Table 12: Example of Naive method

Internet user in Europe & Central Asia						
Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Quantity	525517184	568226496	597322112	628357376	651396608	-
Forecast	-	525517184	568226496	597322112	628357376	651396608

Therefore, mathematically would be represented as follows:

$$F_{t+h} = A_t \qquad \text{Equation 7}$$

Where:

F_{t+h} the forecast for $t + h$,
A_t the actual value for t period.
h the forecast horizon.

B. Seasonal Naïve Method

Sometimes a naïve method can be modified to consider seasonality. In this case the forecast is set to be equal to the last observed value from the same season [41]. For example, given a seasonal data, the forecast value of the month of February will be the same as the previous February value, and the forecast for March will be the same as the previous March and so on. The mathematical expression for the seasonal naïve method is represented next.

$$\mathrm{y}_{\mathrm{T+h-km}} \qquad \text{Equation 8}$$

Where:

m seasonal period, $k = \lfloor (h-1)/m \rfloor + 1$

The naïve methods are simple and well working with a bit variation from a period to another. It can often be used in evaluation of more sophisticated forecasting models [48]. This dissertation recommendation is to apply this method when the data is represented by a slight set of information.

Another variation of the naïve method is the naïve method with drift, as the name suggest it consists on allowing the forecasting values to increase and decrease over time, which means adding or subtracting an average change to the forecast, please for further information refer to [46][47].

C. Simple Mean

Using simple mean or average method, the forecast of *h* periods on time are made by the average of all data set.

$$\hat{F}_{T+h} = \bar{F} = \frac{\Sigma A_T}{T} = \frac{A_t + A_{t-1} + A_{t-2} + \cdots + A_T}{T} \qquad \text{Equation 9}$$

Where:

T : total periods to be average (all periods available).
h : quantity of periods ahead.
$\hat{F}_{T+h}$ estimation of F_{T+h} based on previous data values.

In the simple mean method two pieces of information are carried which are the mean and the total number of observation it was based. A variation of the simple mean is the simple moving average which will be addressed next.

In Figure 26 an application of the previous explained method is applied to the data about the Australian quarterly beer production from 1992-2006 period. From the figure the drawn conclusion is that sometimes one of the basic methods may not serve to forecast, but can be used to benchmark, to compare more complex methods, such as combined forecasting methods.

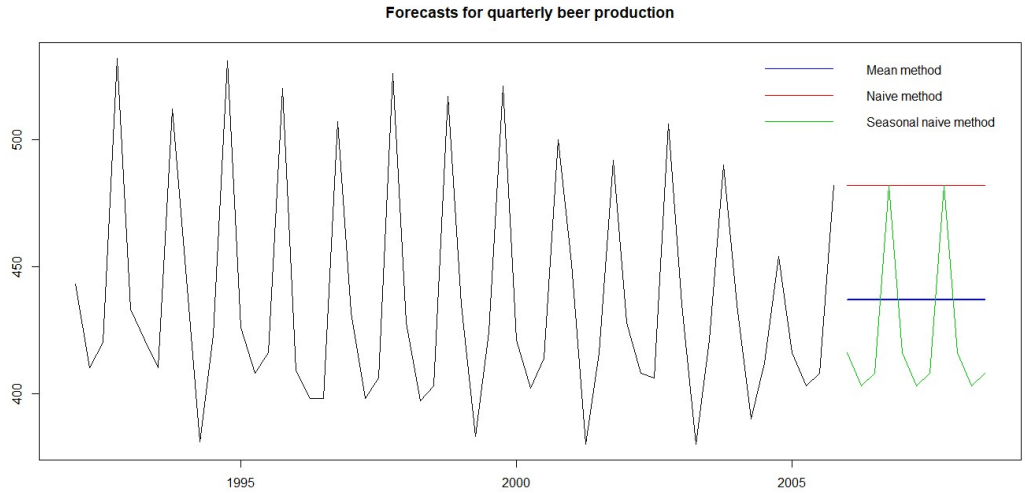


Figure 26:Application of basic forecasting methods. Adapted from: [47]

3.5.2 Smoothing models

The smoothing models are another group of models a forecaster can take before recurring to more complex methods. The moving average concept is the best known smoothing model and comes from the fact that as new data become available old data are dropped causing the average to move across time (simply putting, is “an average that moves across time”). It works on the principle that the time series being used is locally stationary and have a slow varying mean [59].

In general, it can be used as a tool to: filter noise, smooth time series, lagging indicator and trend direction identification. The difference between the various types of moving averages comes from the weight assigned to the values. The Figure 27 depicts existing types of Moving Averages or smoothing models, which are:

- Simple Moving Average (equally weighted);
- Weighted Moving Average;
- Exponential Moving Average.

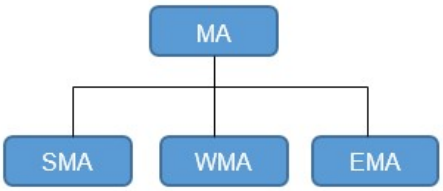


Figure 27: Types of MA. Source: The author

In those models is possible to notice the term “moving average” is applied for all. The SMA and EMA are most suitable when analysing time series. A description about the three methods will be given. Thus, more emphasis will be put on the SMA and EMA, since it is the most commonly used MA and are fundamental to the understanding of this dissertation.

3.5.3 Weighted Moving Average

Weighted Moving Average is generally used when there is no trend component present in the time series. It gives older data usually less importance thus more recent observation is given more weight [47]. WMA can be determined by the following equation:

$$\hat{F}_{t+h} = \frac{W_1 A_t + W_2 A_{t-1} + W_3 A_{t-2} + \cdots + W_n A_{t-n+h}}{n} \quad \text{Equation 10}$$

Where:

W_n : Number of weights to assign. $W_1 + W_2 + \cdots + W_n = 1$
$\hat{F}_{T+h}$: estimation of F_{T+h} based on previous data values.
h : quantity of periods ahead.
n : quantity of periods to be average.

This method requires a lot of historical data. It is not a good choice for trend forecasting given that increasing n makes forecast less sensitive to forecast changes [59] [60].

3.5.4 Simple Moving Average

This method is called moving average because it uses a constant number of observations to forecast. For example, as new data becomes available and old data are discarded it keeps moving through time [48]. Simple putting, is an average that moves with time.

$$\hat{F}_{t+h} = \frac{A_t + A_{t-1} + A_{t-2} + \cdots + A_{t-n+h}}{n} \quad \text{Equation 11}$$

Where:

n : quantity of periods to be average.
h : quantity of periods ahead.
$\hat{F}_{T+h}$: estimation of F_{T+h} based on previous data values.

It is a particular case of simple mean. The SMA also uses the mean from the data to forecast, the only difference from the simple mean is that SMA includes only the n most recent data.

Given the similarity to the simple average, they use the same mathematical representation, the difference is that the MA divisor takes only a part of the data to calculate the average. One characteristic of the SMA is that the more periods a it has, the greater the data to forecast will be lagged [61]. Another characteristic of the SMA is that for long n values, the model become equivalent to simple average.

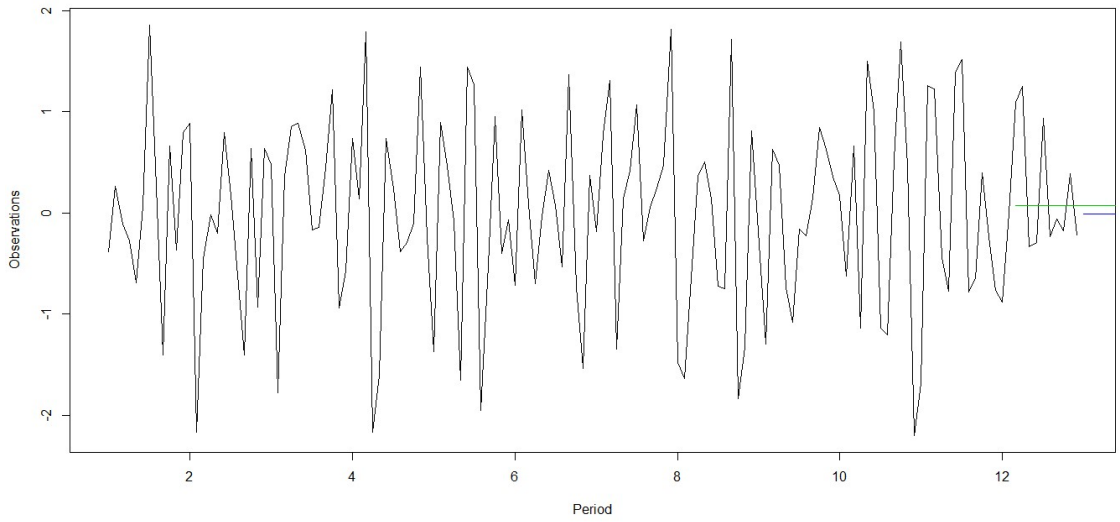


Figure 28: Comparison of SA and SMA. Source: The author

Using the above time series, created using R, the plot represented in Figure 28 was created. In the plot, a comparison between the simple average (in short blue line) and SMA (in long green line) methods are depicted. The SMA applied has an order of 50 which means it has a sliding window of 50. Should be noted that for long-term forecast the SMA is a long straight line just as simple average. Nevertheless, their forecast is different, the SMA uses equal weighted average of n recent values and simple average uses all set value. A deeper approach about the SMA will be given.

3.5.5 Exponential smoothing

The Simple Moving Average described in previous section, has a behaviour of treating the last n observations equally and ignoring all predecessor observations thus weighting then equally. It also disregards previous calculations when computing current values.

One form to get around this is by utilizing exponential smoothing models (exponentially weighted moving average, not to be confused with weighted moving average). It is a widely-used forecasting time series technique which assigns exponentially decreasing weights as time increases (as the observation get older) and takes all the previous observations into account [60]. There are three types of exponential smoothing models which are represented in Figure 29.

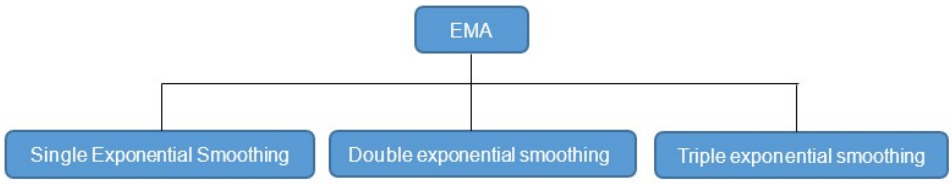


Figure 29:EMA types. Source: The author

A. Single exponential smoothing

The single exponential smoothing method also known as Brown's Simple Exponential Smoothing is ideal when forecasting data with no apparent trend or seasonality[41]. Using the SES, the next forecast can be expressed directly from the previous forecast and previous observations, as depicted by the next expression.

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \quad \text{Equation 12}$$

Where:

$\hat{y}_{t+1}$: represents the forecast value of y in time $t+1$. Read as: "***the next forecast***".
$\hat{y}_t$: represents the forecast value of y in time t. Read as: "***the previous forecast***".
$(1-\alpha)$: represents the decrease in weight within time.
α : denotes the exponential smoothing (controls) constant and assumes $0 < \alpha < 1$ values.

The previous Equation 12 is considered a special form of $ARCH(q)$ model obtained by declining the weight α exponentially through time. The $ARCH(q)$ is represented next:

$$\sigma_t^2 = \alpha_0 + \alpha_0\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2 = \sum_{i=1}^{q} \alpha_i\epsilon_{t-i}^2 \quad (Arch) \quad \text{Equation 13}$$

Expanding the Equation 12 by first substituting for y_{t-1}, the following equation is obtained:

$$\hat{y}_t = \alpha y_{t-1} + (1-\alpha)[\alpha y_{t-2} + (1-\alpha)\hat{y}_{t-2}] \quad \text{Equation 14}$$
$$\hat{y}_t = \alpha y_{t-1} + \alpha(1-\alpha)y_{t-2} + (1-\alpha)^2\hat{y}_{t-2}$$

If the substitution continues, the forecast value on the right side of the equation will be represented as follows:

$$\hat{y}_t = \alpha y_{t-1} + \alpha(1-\alpha)y_{t-2} + (1-\alpha)^2 y_{t-3} + \cdots + \alpha(1-\alpha)^j y_{t-j} + \cdots + \alpha(1-\alpha)^{t-2} y_1 \quad \text{Equation 15}$$

The previous equation is the reason why is called exponential smoothing and considered a recursive process [60].

Note:

If $\alpha = 1$, the SES model is equivalent to a random walk model without drift. And if $\alpha = 0$, the SES model is equivalent to a simple mean model.

The SES forecast can be represented using an alternative representation called component form, which comprises the use of forecast and smoothing equation to represent the components present in a time series ,for more details refer to [41].

$$\hat{y}_{t+1} = l_t \quad \text{Equation 16}$$

Where:

l_t : is read current t level.

Representing the Equation 12 in the component form, the following equation is obtained:

$$l_t = \alpha y_t + (1-\alpha)l_{t-1}\ (2) \quad \text{Equation 17}$$

1. Age of the data in the simple exponential smoothing forecast

The average age (amount of lag) of data being used can be determined in a SES relative to the period which the forecast is computed using the following equation:

$$\frac{1}{\alpha} \quad \text{Equation 18}$$

It allows to find how many lags exist relative to the current period. For example, $\alpha = 0.2$ the lag is 5 periods.

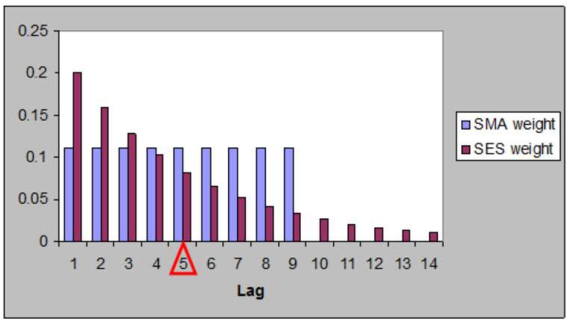


Figure 30:Comparing MA weight vs ES weight. Source: [62]

Figure 30 shows a graphic representation where a comparison between exponential smoothing lag and moving average lag is presented. Given the fact the SES places more weight on most recent observation it is somehow superior and responds faster to changes occurring in recent data in relation to SMA forecast. That is, as depicted in the above figure, the weight assigned to a data decreases exponentially as observations goes further in the past. The oldest observation gets the smallest weights.

2. Understanding the values of α

The understanding of the α parameter is essential when working with SES, thus in this section will be explained how it works. As said early the α denotes the exponential smoothing constant, thus some authors [47][60] refer as memory decay rate. The parameter α constitutes a function of value representing the speed at which previous observations are "smoothed" (damped) [60]. In other words, the parameter α controls the rate at which the weight decreases.

The best way to obtain the α parameter is to estimate it from the observed data. The following table show an example of weights for different values of α.

Table 13: Weight towards previous observations Source: The author

$\boldsymbol{\alpha}$	$\boldsymbol{(1-\alpha)}$	$\boldsymbol{(1-\alpha)^2}$	$\boldsymbol{(1-\alpha)^3}$
0.9	0.1	0.01	0.001
0.5	0.5	0.25	0.125
0.2	0.8	0.64	0.512

From Table 13 can be noticed that:

- As time goes back $[(1-\alpha), (1-\alpha)^2, \dots, (1-\alpha)^n]$ the weights attached to the observations decrease exponentially.
- By choosing an α close to 1 will cause a faster smoothing (previous values are damped faster) thus making more responsive forecasts to more recent levels. For higher values of α the faster the method "*forgets*".
- By choosing an α close to 0 will cause a slower smoothing resulting in a damping curve[1].

[1] A smooth curve almost a straight line.

The value for α can be chosen by minimizing the value of MSE [63]. Or using nonlinear optimizer, a searching method that minimizes the sum of squares of residuals, such as Marquardt [64]. However, many statistical software programs (such as the Solver from EXCEL or SES() function from R) can help find the value of α that minimizes the MSE. Besides, is also possible to determine the value of α by fitting the observed data to an ARIMA(0,1,1) model [65] which is explained in the next section.

3. Initialization step

The initial value for the exponential smoothing has a very important role in the calculation of subsequent values. One method to find out the initial value consists on initializing the model as follows:

$$\hat{y}_1 = l_0 = y_1 \qquad \text{Equation 19}$$

Given the Equation 12 the value $\hat{y}$ for the period $t+1$ is found by the previous t value. Since there are no values for $t = 0$ the following is assumed:

$$\hat{y}_1 = y_1 \qquad \text{Equation 20}$$

An alternative approach to estimate the value of l_0 suggested by [41] consists on using optimization.

4. Forecasting with Single Exponential Smoothing

For forecasting data SES Equation 12 will be used as base. Writing the previous equation, to obtain the next forecast by adjusting the previous forecast found into the direction of the error, the following equation is obtained:

$$\hat{y}_{t+1} = \hat{y}_t + \alpha\epsilon_t \qquad \text{Equation 21}$$

Where:

ϵ_t : represents the previous period t forecast error.

New forecast is represented as a function of previous forecast plus a certain adjustment error to the previous t observation. When simple exponential smoothing is represented by adjusting the error it is called **error correction form**. Equation 21 can also be represented by the component form as represented next:

$$l_t = l_{t-1} + \alpha e_t \qquad \text{Equation 22}$$

The Figure 31 represents an application of the SES method to a time series created using R and depicted in black line. From the figure is possible to see that the data doesn't show a trend behaviour.

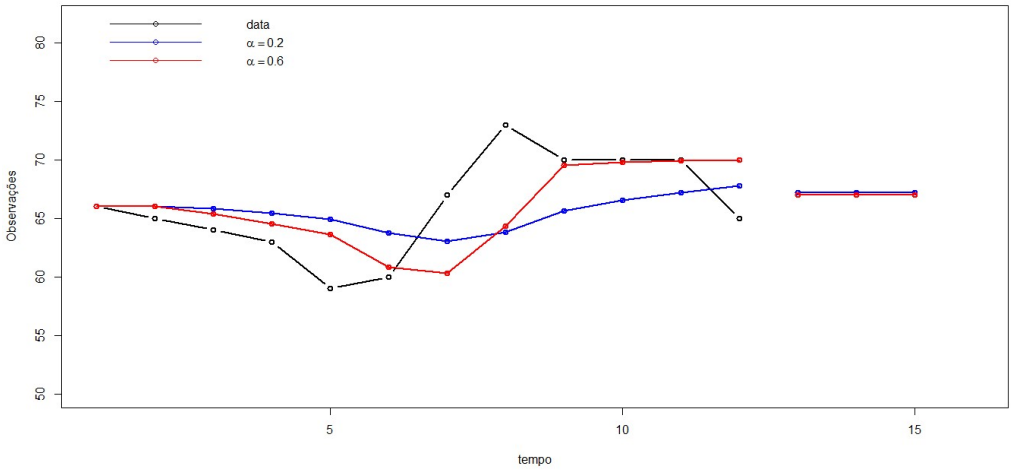


Figure 31: SES. Source: The author

Still in the figure, is clearly visible that for larger values of $\boldsymbol{\alpha}$, the greater the adjustment takes place in the direction of previous data point, which means its better fitted to the time series. Since this model uses only the last observed value and the forecast for the current instant, the forecasts for larger horizons are constant along time [53], thus the forecast for h next periods can be written as:

$$\hat{y}_{t+h} = \hat{y}_t \qquad \text{Equation 23}$$

Therefore, theoretically this method can only be used when the data has no trend nor seasonal behaviour.

5. Simple Exponential Smoothing in R

Using the fpp library, in R is possible to apply the SES. By default, if the alfa parameter is not provided the ses() function provides one, which is determined using the MSE. In the Figure 32 a simple exponential smoothing applied to oil production in Saudi Arabia, represented by the green lines is depicted. The original data from the period 1996-2007, is represented by the black line. In the case of the data in analysis, the $\boldsymbol{\alpha}$ estimated were 0.89.

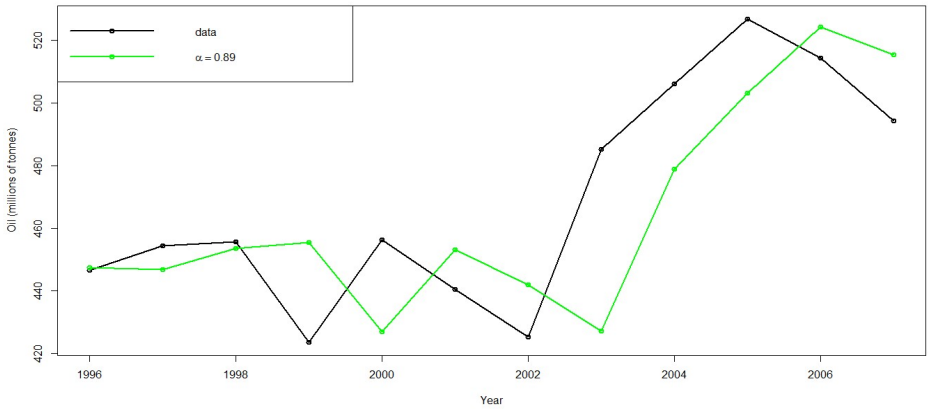


Figure 32:SES example using SES fitting. Source: [47]

From the smoothing methods explained, it was verified that SES is more suitable to be applied when the data has no trend or seasonality pattern. And different from the MA, the SES provides a greater weight to recent time series observations, thus, for recent values the SES is more sensitive and has less lag when compared to SMA.

B. Exponential Smoothing with trend

The SES model previously discussed assumes a time series with no trend, which can be enough when making one-step-ahead forecasts. If the time series has a fluctuating growing behaviour, it may be necessary to adjust the simple exponential smoothing to consider those cyclical or trend patterns. The adjustment is necessary because the SES is slow when responding to trends presence [62]. In those cases, is necessary to add another smoothing parameter to account the trend component.

From Figure 33 can be seen the line in black has a trend. By applying the simple exponential smoothing (line in blue) one can notice that the exponential line doesn't follow the original data line where there is trend resulting in an inadequate of simple exponential smoothing fitting.

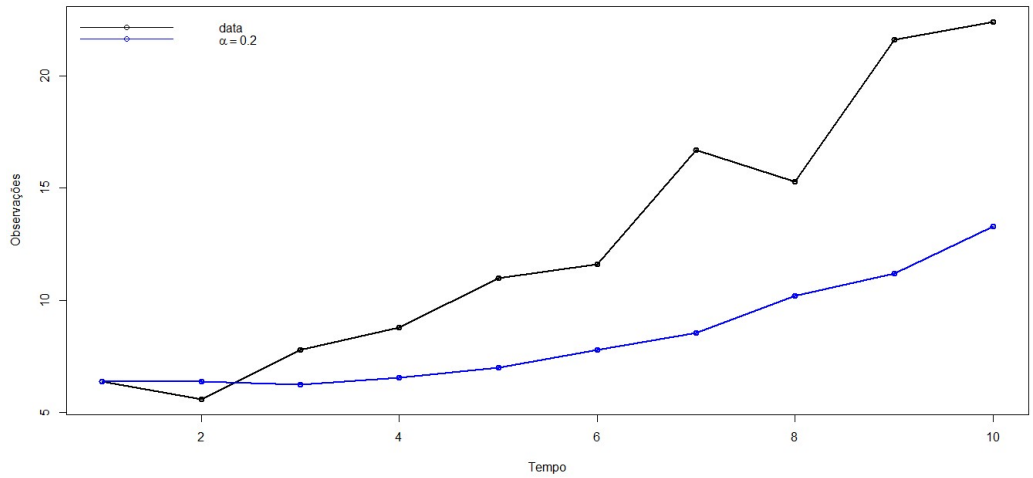


Figure 33: Using Simple Exponential Smoothing in a data set with Trend. Source: The author

If the data has a trend but no seasonality the concept of Double Exponential Smoothing is used. Simply explaining, DES consists on nothing more than exponential smoothing applied to level and trend. DES creates forecast from a combination of exponential estimates of trend and the level [65]. It is accomplished by applying two different weights ($\alpha \; and \; \beta \; smoothing \; parameters$) to update two observations at time. There are two types of DES models which are [59][61]:

- Brown's Double Exponential Smoothing (also known as Brown's linear exponential smoothing);
- Holt-Winters double exponential smoothing.

Will be addressed the Holt-Winters model.

C. Holt's double Exponential Smoothing

The issue about the Brown's model is due to the fact that it uses a single smoothing parameter which places a certain limitation on the data pattern the model is able to fit resulting on a dependent level and trend variation [59]. This issue is resolved by the Holt's double Exponential Smoothing method by including two smoothing constants, one for trend and other for the level.

The Holt's double exponential smoothing method consist of one forecasting equation and two smoothing equation, one for the level component and another for the trend component. Those equations are represented next:

Level equation: $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

Trend equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

Equation 24

Since the Holt's double method takes into account two smoothing equation its forecasting is given by the following equation:

Forecast for h period-ahead: $\hat{y}_{t+h|t} = l_t + hb_t$ Equation 25

Where:

l_t : represents an estimation of the level of the series at period t.
b_t : represents the estimation of the trend or slope of the series at period t.
α : represents the smoothing parameter for the level component and assumes $0 \leq \alpha \leq 1$ values.
β : represents the smoothing parameter for the level component and assumes $0 \leq \beta \leq 1$: values.

Important aspects to understand:

- Different from SES here because of the presence of trend component the one step ahead forecast for period t is given not only by the previous l_{t-1} but by the addition of l_{t-1} to a previous weighted average of the estimated trend b_{t-1}.
- The trend equation b_t is composed by a weighted average from the estimation trend at period t on level $l_t - l_{t-1}$ plus the previous estimated trend b_{t-1}.
- If $\alpha = \beta$ the holt models become equivalent to Brown´s DES.

1. Double Exponential Smoothing in R

When adjusting the model represented in Figure 33 from using the alpha parameter of 0.2 to an alpha of 0.9773, the parameter that most fitted the data using the SES model, it was allowed to reduce significantly the MAPE from previous 29.98424 to 15.2267 thus obtaining the adjusted plot represented in Figure 34.

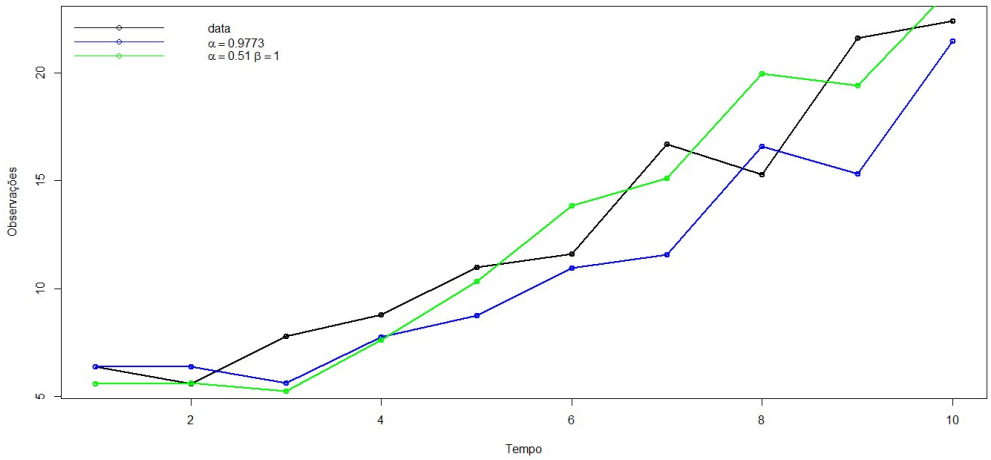


Figure 34:SES vs DES. Source: The author

As it can be seen in Figure 34 the DES lines in green fits / follows the trend better, different from the SES represented in blue lines which despite of fitting the data, was not ideal for the data in analyses. Although the model is adjusted, it does not account the trend component, so it is necessary to use the DES in order to adjust the data and get better forecasting.

Similar to the SES, for the DES it can be fitted and forecasted by using the fpp library. Therefore, the generated double smoothing model from the fpp library resulted in an α estimated of 0.4821 and the β estimated of 1, considered very reasonable. But, Given the fact that it could be improved, the improved model has an α of 0.51 and β of 1 which has a slight better MAPE of 14.14209 when compared to a MAPE of 14.16547 of the first parameters found.

D. Triple exponential smoothing

The DES is useful to apply when in the presence of data that show trend behaviour but in case the data shows also a seasonal behaviour applying DES will not have good results. When in the presence of data containing seasonality in case the DES doesn't work is necessary to introduce a third equation which will take care of the seasonal component. These set of equations forms the triple exponential smoothing also known as Holt-Winters method.

This method has two variants from what concerns the seasonal component [41]. The first is the additive and the second is multiplicative which differ on its seasonal variation. The first has an almost constant variation through time and the second has a changing variation. This dissertation will address the multiplicative variation for further information on the additive approach please refer to [66]. Both variation uses the same constants α, β, γ to represent the level, trend and seasonality smoothing parameter. The multiplicative approach equation is represented next.

1. Multiplicative model

The Holt-Winter multiplicative approach uses three smoothing equations one for level, one for trend and one for seasonality, as represented by the following equations:

Level equation:	$l_t = \alpha \frac{y_t}{S_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1})$	Equation 26
Trend equation:	$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$	
Seasonal equation:	$S_t = \gamma \frac{y_t}{l_t} + (1-\gamma)S_{t-m}$	

Since the Holt-Winters method takes into account three smoothing parameter its forecasting is given by the following equation:

Forecast for h period-ahead: $\hat{y}_{t+h|t} = (l_t + hb_t)S_{t-h-m}$ Equation 27

Where:

l_t : level component;
b_t : trend component;
S_t : seasonal component
m : seasonal period
$0 < \alpha < 1, \quad 0 < \beta < 1, \quad 0 < \gamma < 1 - \alpha$

Similar to previous methods, when applying the Holt-Winters one should choose the model whose constant result in the least MSE. The methods based on the exponential model are not restricted to those presented in this section of the dissertation. It is still possible to find a variety with more than 10 combinations of tendency and seasonality methods based on exponential smoothing. For further details about other exponential smoothing methods, please refer to section 7.6 of [41].

3.6 Time Series Decomposition

Times series can present a series of patterns, such as trend, seasonality that can be useful to categorize some behaviours that may occur. Decomposing a time series involves separating it into several distinct components. Given the fact that each one of the components found represents the underlying pattern categories, is much easier to forecast since different regular patterns embedded in the observed time series are identified [67].

The general mathematical representation of the decomposition approach is represented by the following expression:

$$y_t = f(S_y, T_t, I_t) \quad \text{Equation 28}$$

Where:

Y_t : the time series value at period t;
S_t : the seasonal component at period t;
T_t : the trend-cycle component at period t;
I_t : the irregular component at period t;

3.6.1 Additive and Multiplicative Models

There are two basic models of time series [41][44]:

- **The additive model:** can be used when the data seasonal variation is considered constant over time, its amplitude remains also constant over time. The following equation represents the model:

$$y_t = T_y + S_t + I_t \quad \text{Equation 29}$$

A graphic representation of an additive model is represented in Figure 35 by the Australian beer production dataset. In the plot it is possible to observe a trend and seasonal variation that remains almost constant over the time.

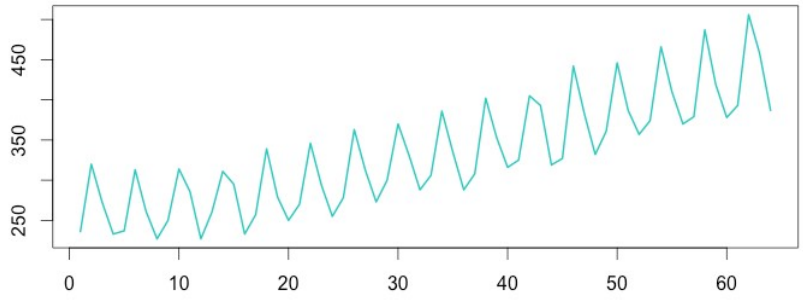


Figure 35: Australian beer production: Additive model. Adapted from: [68]

- **The multiplicative:** often used when the data presents seasonal variation which increases over time and its amplitude component grows over time.

$$y_t = T_y * S_t * I_t \quad \text{Equation 30}$$

A graphic representation of a multiplicative model is represented in Figure 36. It is possible to observe the presence of a trend and seasonal variation, an amplitude that increases over time.

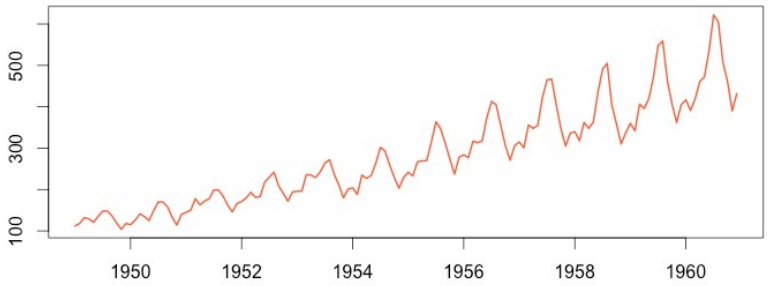


Figure 36:Airline Passenger Numbers: Multiplicative model. Adapted from: [68]

An approach used to override the multiplicative model consists on the transformation of the data until the variation in the series appears to stabilize over time, after that an additive model is used [41].

3.6.2 Procedures used in time series decomposition

Decomposition procedures consists on a set of steps used to help in the identification and categorization of some pattern and behaviours that may be present in a time series [69].An explanation on how to decompose an additive time series model using the decomposition method will be given. As for decomposition of multiplicative model please refer to [41] section 6.3. To decompose an additive time series the steps represented in Figure 37 must be considered:

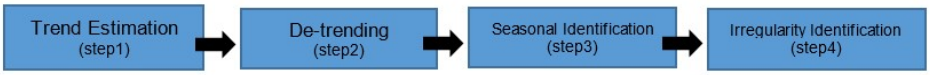


Figure 37:Classical Decomposition Steps. Source: The author

A. Estimate the trend

Trend estimation involves a set of methods and procedures used to identify the presence of trend signal in a time series and help justify statements about tendencies present in the data. A trend exists when there is a long-term increase or decrease in the data and it does not have to be linear[47]. A time series with trend is said to be non-stationary. There are many procedures such as median filtering, Bandpass filtering, curve-fitting, Logistic or Gompertz growth curve ,and Hodrick Prescott Filter used to decipher trend embedded into a time series. The most used are:

- **Moving Averages:** using MA is possible to estimate a time series trend thus using MA the trend will be described without taken into consideration an equation [69].
- **Regression equation:** another approach consists on using an RE equation to model the trend.

Of the presented methods, the MA will be implemented.

1. Moving average smoothing applied to time series decomposition

It is considered one of the classical methods of time series decomposition, dated in the 1920s. It continues to be the basis of time series methods and is widely used in many different trend analyses. The principle of using the moving average is, as observations are near in time it is likely to be close in value.

An analogy used to explain the concept of Moving Average is the ironing of clothes, where the idea consists of removing all wrinkled parts from the time series. It measures the average of a subset of numbers. The result after applying the MA is a stable trend through adjacent values for a period. It also can be used to filter random "**white noise**" from the data by smoothing the series to reduce its variation [47] thus emphasizing hidden components.

A MA can be represented by the number of data points included in each average which is called ***order***, the higher the order, the smoother the curve. The number of data points included affects directly the smoothness of the resulting data. To eliminate the randomness of a constant trend data it is recommended the use of a higher order and if there is a need to capture recent changes over the time series its recommended to use a lower order [53].

2. Simple moving average

Using the SMA method from the previous section will help decomposing a time series specifically identifying the trend component. Re-writing the SMA equation presented in the previous equation, the following equation also known as arithmetic average of any odd order is obtained:

$$\widehat{T_t} = \frac{1}{m}\sum_{j=-k}^{k} y_t + j \qquad \text{Equation 31}$$

Where:

$$m = 2k + 1$$

The above equation represents the simple moving average smoothing, it estimates the trend-cycle at time $\boldsymbol{t}$ that is obtained by averaging values of the time series within $\boldsymbol{k}$ periods of t [41]. It is important to note that ***m*** has to be an odd number in order to have a symetric **half-width**. For each value from the data series an equal weight is applied. For even order the moving average of moving average procedure should be used. For further information consult [41] on section 6.2.

Note:

> One disadvantage of the SMA is that it can not estimate the trend which are close to the beginning and end of the series. This happens because the MA model has to be fed before starting to work.

The cardinality of a time period which are not included into the beginning and end of the time series can be found by the following formula:

$$k = (m-1)/2 \qquad \text{Equation 32}$$

The $\boldsymbol{k}$ factor calculated can also represent the **half-width** of a MA, the number of periods in each side of MA being calculated. The following example takes moving averages of different time periods. The first figure an order of 51-MA and for the second an order of 501-MA is shown. In Figure 38 the MA are shown in red line and the actual time series in black line.

To calcute the MA presented in the following figures, time series about Internet traffic data, in bits, from an ISP where used. The MA calculated were of order 51 and 501, this difference in the order is to emphasize the effect of applying a moving average with a higher order. For example, if order 5 or 10 were used, visually, would be more difficult to identify the trend behaviour.

By applying the 51-MA, some slight variation of trend can be notice. However, there are still a certain difficulty in finding out the exact trend behaviour. For the 501-MA the behaviour is already different,

and it can be notice the trend is smoother when compared to the original series shown in black. Notwithstanding, it captures the main movement of the time series without all the minor fluctuations.

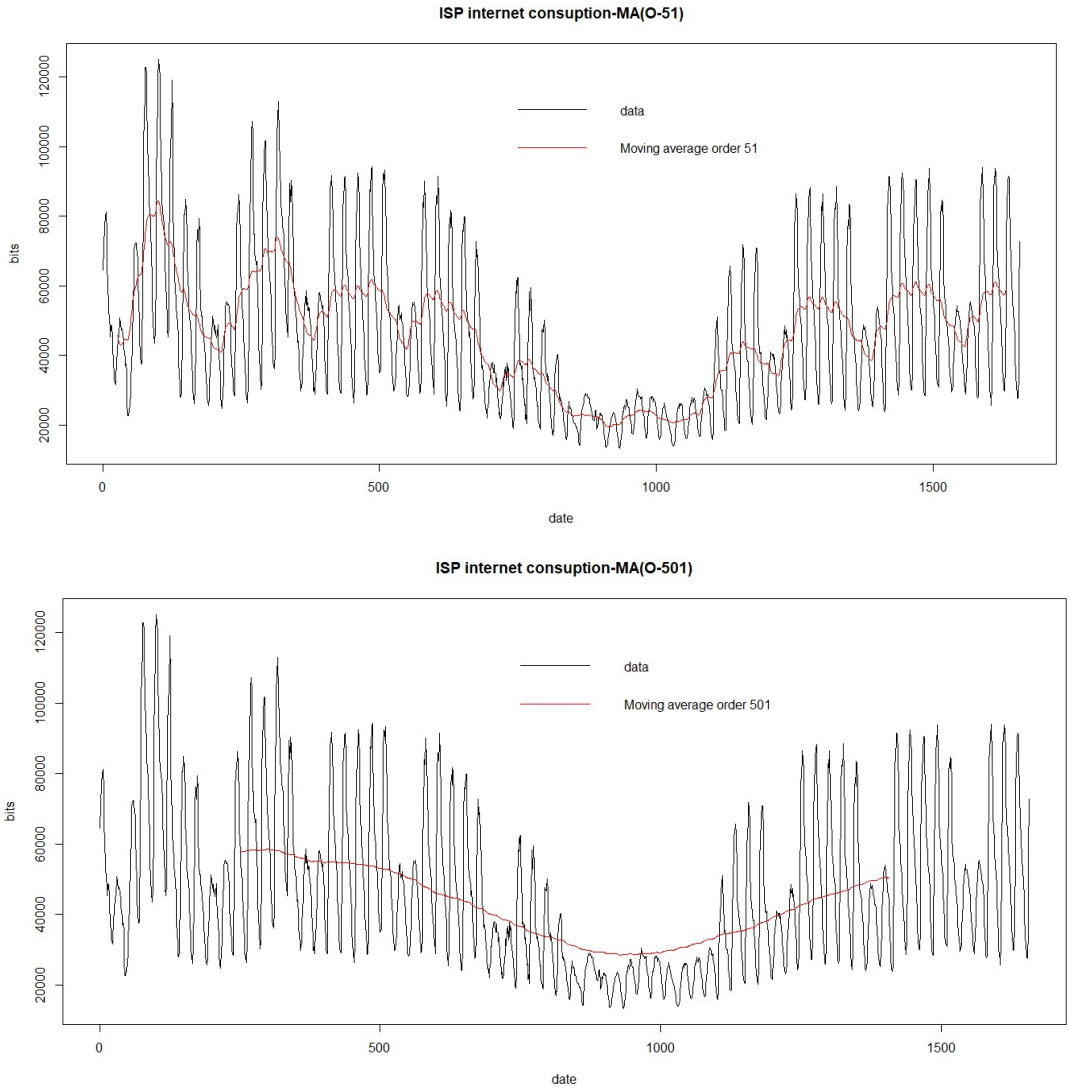


Figure 38: 51-MA and 501-MA. Source: The author

An observation to be considered is the fact that MA doesn't estimate periods T_t where t is close to the end of the series [41], that's why the MA lines in the previous figure doesn't extend to the edges of the graph on both side. Also, can be notice that higher order of the MA results in a smoother curve.

B. De-trending a time series

Detrending consists on removing the trend pattern component from a time series. If the time series is an additive model the de-trending is made by subtracting the trend estimate component from the series leaving the seasonal and irregular components.

$$Y_t - T_y = S_t + I_t \qquad \text{Equation 33}$$

C. Seasonality component identification

Seasonality component identification captures level shifts that repeats systematically within the same period (e.g. month or quarter) over time [70]. To remove seasonality the de-trend component from the previous step will be used.

Assuming a constant seasonal component from year to year, there is only a need to calculate one value for each month. This is made by taking all the values for a given month and take the average. For example, to get a seasonal effect for January, is necessary to average the de-trended values for all Januarys in the series [69]. The value that comes out of this average is called the *seasonal indices* [47].

After removing the seasonal component, the time series is said to be seasonal adjusted or seasonal stationary. A seasonal adjusted value removes the seasonal effect from a time series which allow trends to be seen more clearly. A time series that is not possible to identify its calendar or seasonal effects is considered "*de facto"* seasonal adjusted. Time series to which seasonal adjustment method have been applied shouldn´t have signs of residual seasonality.

Another form to identify seasonality is to use periodogram plot or look into ACF plot of the data and search for significant lags that have the same peak cycle.

D. Irregular component identification

The irregular or remainder component is found by subtracting the trend and seasonality from the original series. To obtain the irregular component is indispensable to first identify the trend and seasonal components.

$$Y_t - (T_y + S_t) = I_t \qquad \text{Equation 34}$$

One of the main problems of the classical decomposition is that for longer time series, it becomes unable to capture all seasonal changes that may occur over time. Other problems associated to the classical decomposition can be found in [41]. However, the use of decomposition is an important tool to be applied by forecaster to understand the data, also serves as an initial step taken before applying other forecast methods.

3.6.3 Decomposing time series in R - classical decomposition approach

In this section, a practical example of decomposing a time series by applying the classical decomposition method is given. For a better understanding, an additive model is used. The data set to be decomposed will be the quarterly Australian beer production from 1956 to 1972 [41]. These data set has a constant seasonal variation over time, ideal to demonstrate an additive decomposition, a characteristic to be found in telecommunication data sets.

Using the classic decomposition method explained in the previous section, the first step consists of detecting the trend, to do that the Moving Average will be used. Other methods can also be used in this step. Before continuing, as MA works similar to a moving window, is important to know the exactly size of the seasonality present in the data, so it can be used as the MA order. The seasonality period can be detected using different methods.

In the case of this example, since the data consists on a quarterly record per year, the seasonality is also four. Once identified the seasonality, the next step is to apply the seasonality period as a moving average, in this case a 4-MA, as depicted in the R code represented in Figure 39.

```
library(fpp)#package
library(forecast)#package
timeseries_beer<-  tail(head(ausbeer, 64),64) #extract first 64 observation,

findfrequency(timeseries_beer)

trend_beer <- ma(timeseries_beer, order = 4, centre = T)

View(trend_beer)
```

Figure 39:Classical Decomposition using R (code snippet)

Calculated the 4-MA, the "trend" of the data for the period in analysis is also determined. By observing the 4-MA values represented in Figure 40 is possible to note some missing values at the beginning and end which is due the fact that when using MA it has to be fed in order to produce average data points.

	trend_beer
1	NA
2	NA
3	255.250
4	254.375
5	257.375
6	260.000
7	262.750
8	264.625
9	265.375
10	264.625
11	262.375

Figure 40: 4-MA values. Source: The author

Figure 41 represents the 4-MA result which allows to identify the trend. The trend obtained is positive and is represented by a blue line. The initial data set values are represented in black line.

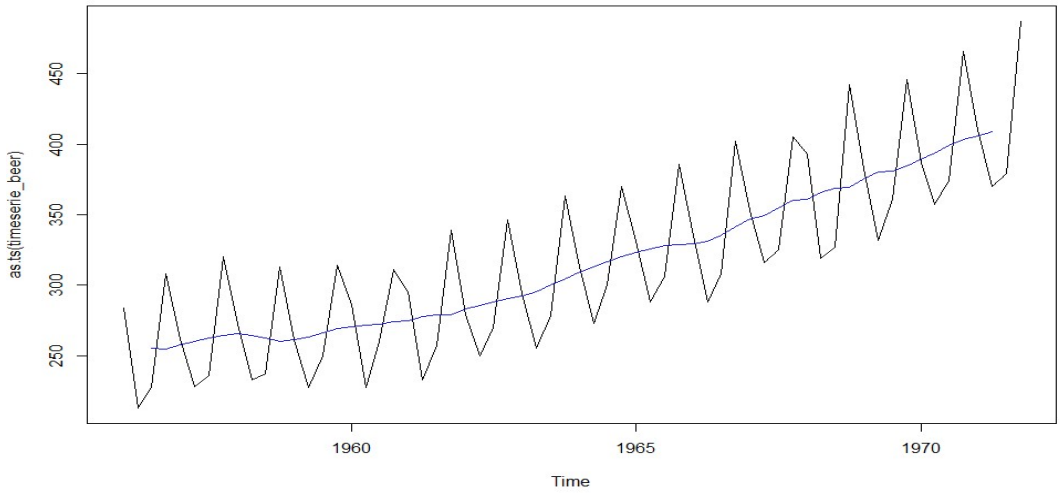


Figure 41: Data trend. Source: The author

The missing values at the end of the MA are replaced by an average of the region, where the order of the average is the quantity of missing values at each end.

For example, the 1st missing value represented in Figure 41 is replaced by the average of the 3rd and 4th values ((255.5200+254.375)/2), and the 2nd missing value is replaced by the average of 4th and 5th. The missing value at the end of the data was also replaced by the average of the region. Thus, to remove the trend from the data, the values of the 4-MA are subtracted from the initial data as represented by Figure 42.

```
detrend_beer = timeseries_beer - trend_beer
```

Figure 42:Classical Decomposition using R: Detrend

The result is represented in Figure 43.

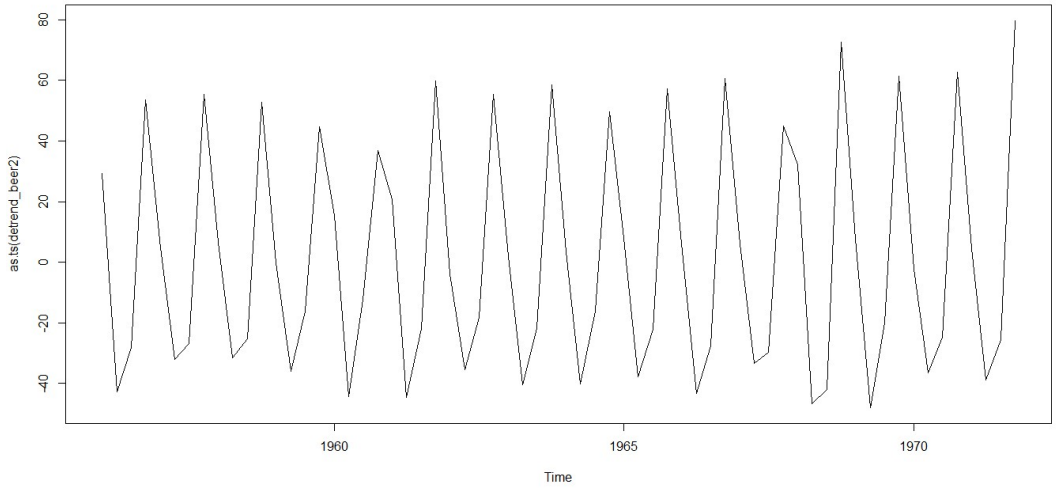


Figure 43:Detrend beer data. Source: The author

From the plot, is possible to observe that after removing the trend it clear exposes the seasonality present in the data. The seasonality or seasonal index can be found by averaging the values for each quarter. For example, the seasonal index for the first quarter is the average of all detrended first quarter values present in the data. After calculating all the seasonal index, the seasonal component is obtained by binding all the seasonal index for each year present in the data.

One way to find the seasonal index for quarter data consists on finding the transposed of a matrix of 4 rows then averaging the columns. Thereby the average values are replicated 16 times which is the number of 4 quarter sets as represented by Figure 44

```
m_beer = t(matrix(data = detrend_beer, nrow = 4))

seasonal_beer = colMeans(m_beer, na.rm = T)
```

Figure 44: Seasonal Index

In the Figure 45 the obtained seasonal data is plotted.

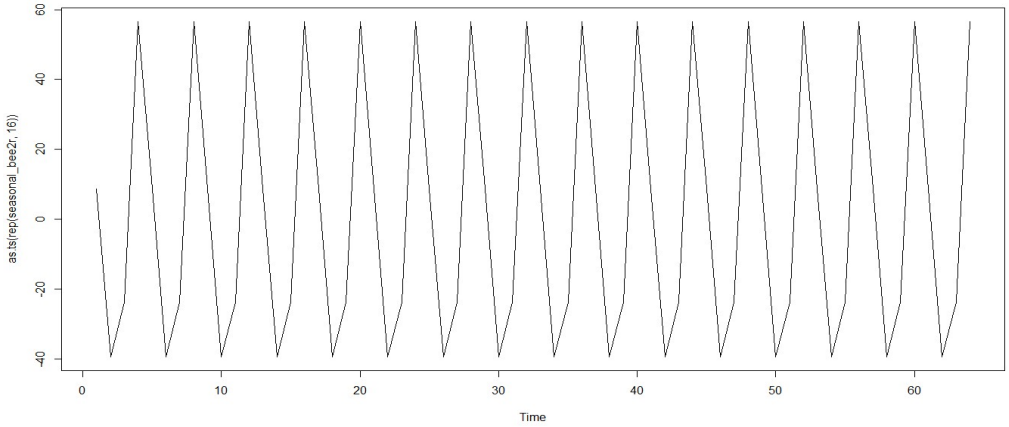


Figure 45:Seasonal Component. Source: The author

With the seasonal component found, the random component can be obtained. To find the random component subtract the trend component and seasonal component from the original data as represented in Figure 46.

```
random_beer = timeserie_beer - trend_beer - seasonal_beer
```

Figure 46: Classical Decomposition using R: Random

The result of this operation is represented in theFigure 47.

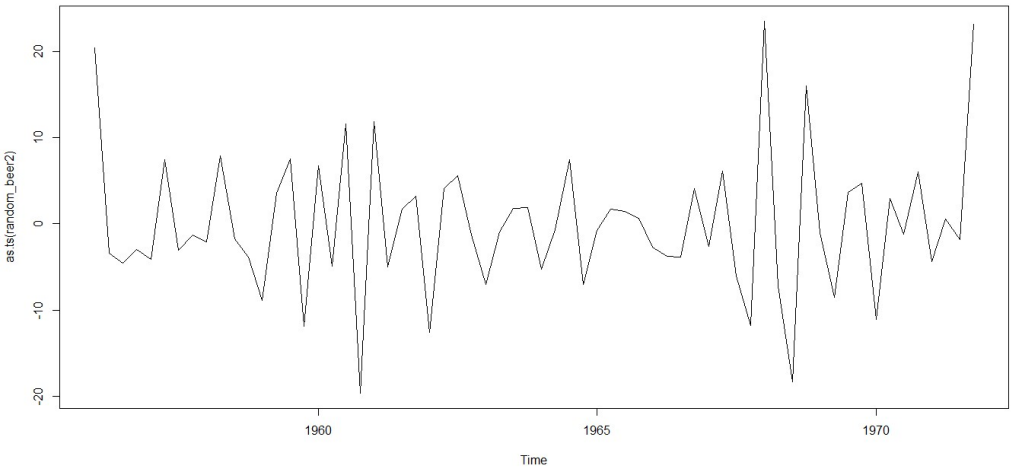


Figure 47:Random component. Source: The author

Although the examples contain all components that compose a time series, not always a time series can present all the components. Several attempts to use decomposition to forecast data using directly decomposition, where the components are projected and recombined into the future to determine the forecast has been used by several years [47]. Due to its limitations, in practice it rarely works well. However, in this dissertation the decomposition of time series was used for understanding behaviours and identifying possible forecasting models to use.

3.6.4 Other models for decomposition

In addition to classical decomposition, there are other methods that can be used to split a time series into its components pattern. An example of such methods is the Seasonal and Trend decomposition using Loess (STL) decomposition, X-12-ARIMA decomposition, etc.

STL consists on filtering procedures used for decomposing a seasonal time series into three components: trend, seasonal and remainder using Loess, a method for fitting a smooth curve, presented by Robert Cleveland, William Cleveland, Jean McRae and Irma Terpenning in the Journal of Official Statistics in 1990 [71].

The X-12-ARIMA decomposition was developed by the US Bureau of Census. Based on classical decomposition, the X-12-ARIMA overcomes some of its predecessor drawbacks. One of the relevant features present in this method is the ability to handle various situations such as day variation and holidays effects [41].

3.7 Forecast validation

Generally, the interest is in forecasting the future. In order to do that, is necessary to validate that future. Forecast validation is divided into three main groups [72]:

- In sample forecasting;
- Out of sample forecasting;
- Pseudo out-of-sample forecasting.

Out-of-sample method consists on fitting a model that, based on data up to, and including today, determines a forecast of tomorrow's value (Y_{T+1}) and waits until new input values, records the forecast error (simple error or MAPE), re-estimates the model (using the new input value), then makes a new forecast (Y_{T+1}, Y_{T+2}), etc. However, this process is very time consuming.

In this dissertation, a pseudo out-of-sample variation will be tested, a variation of out-of-sample. This consists on using an historical data as a starting point. For further information about other forecast evaluation groups please refer to [59] [72].

Out-of-sample and pseudo out-of-sample validation method are used to test models and compare their forecasting performance. When applying out-of-sample method the data set is divided into the following groups:

- **Validation:** also known as test period or test set, is used to test the forecast determined using the estimation or fitting period to see how accurate the candidate model is.
- **Estimation:** also known as adjustment period or training set, is used to help chose among candidate's models, the most fitted or adjusted. Consists on suppressing some of the initial data set for model identification and estimation to then forecast.

In other word, out-of-sample means, training the model on the training set and test its accuracy on a separate data set, the test set. To measure out-of-sample error is necessary to look out to the test set error, not the train set error.

Out-of-sample forecasts are computed by applying one of the following methods:

- **Expanding window:** is a recursive forecasting method. Initially a data set starting from $t = 1,..,N$ is taken to estimate models then an $h - step$ forecast is produced starting from N period. After the first $h - step$ forecast has been determined the data set is increased by one, which means a new input value from the source, then a new model is estimated, in other words re-estimate old model plus new incremented input value. An illustration of an expanding window is represented in Figure 48.

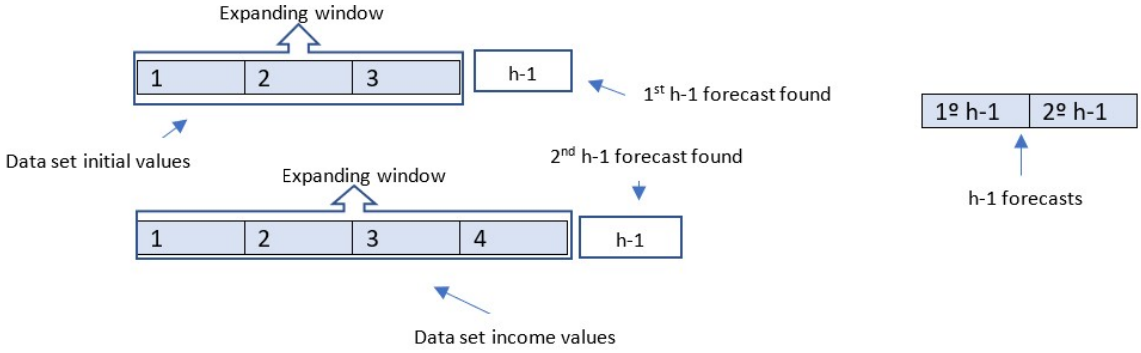


Figure 48: Expanding window example. Source: The author

- **Moving window:** also known as a sliding window it is based on rolling the forecast window. For this method, first is specified a window width T, which will be used in the initial data set to estimate the models and forecast $h - steps$ ahead starting from the initial T. Then the defined window T is moved one-time period, a new model estimation is made using the old window data minus one value that is discarded, plus a new input data value. After modelling the new window another $h - step$ forecast is made. An example of a moving window is represented in Figure 49.

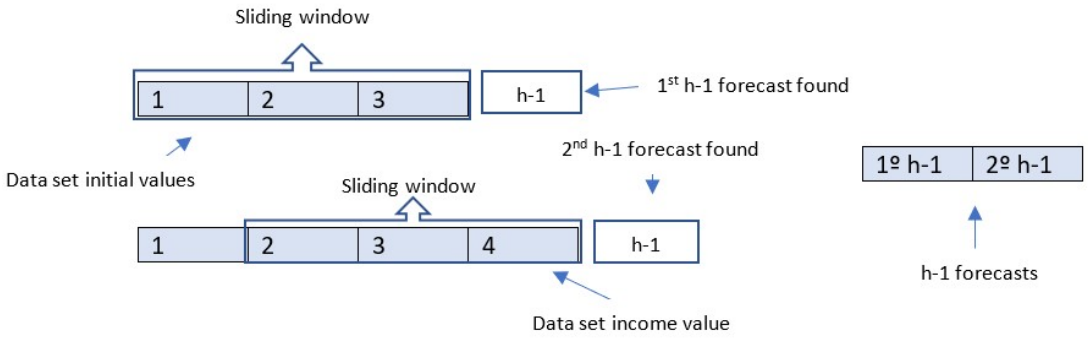


Figure 49:Sliding window process. Source: The author

Generally, the use of a small windows size means the forecast will be determined based on a smaller data set and a longer window size, a longer data set. The use of longer windows implies the risk of including old data which are no longer representative of the time series current behaviour.

3.8 Autoregressive Models

A model is considered autoregressive when a value from the time series is regressed on its previous values. Forecasting speaking, given a traffic demand represented by Xt, at time t if it can be expressed as a linear combination of its previous equidistant observations then an autoregressive process is present. The term autoregression arises as an indication that the regression is between a variable against itself [41]. An autoregressive model can have different orders, to obtain an $p - order$ model the following equation is used:

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-1} + \ldots + \Phi_p X_{t-p} + a_t \quad \text{Equation 35}$$

Where:

c	: constant
a_t	: white noise at time t;
$\Phi_k \; k = 1, \ldots p$	: autoregressive parameters.
p	: autoregression order

Note:

In autoregressive models the order represents the number of nearly previous values that are used to forecast the value at present time.

The variance of a_t (or ε_t in the case of MA) error term will change the scale of the time series not its patterns. This is also true for MA models.

3.8.1 First order auto-regression

The following equation shows an example of an auto regressive model of $1 - order$.

$$x_t = c + \phi_1 x_{t-1} + \varepsilon_t \quad \text{Equation 36}$$

The preceding time series period happen to be the predictors, for example, to predict the value on x_t period is fundamental to have the value on x_{t-1} period. The ϕ_1 represents a coefficient used to minimize the error of the model. When in the presence of an $AR(1)$ model some conditions may exists, and is necessary to be aware. These conditions are described as follows [41]:

- $\phi_1 = 0$, the result is a white noise (ARIMA (0,0,0) equivalent).
- $\phi_1 = 1 \; and \; c = 0$, the result is a random walk (ARIMA (0,1,0) without constant).
- $\phi_1 = 1 \; and \; c \neq 0$, the result is a random walk with drift (ARIMA (0,1,0) whit constant).

Therefore, the model defined by Equation 36 is a first order autoregression model and is represented as $AR\ (1)$. This demonstrates that it is only possible to forecast the present t period value by knowing the previous $t - 1$ period value.

3.8.2 Second order auto-regression

In addition, the Equation 37represents a second-order autoregression model, where different from the $AR(1)$, in the $AR(2)$ its t period values are obtained by the $t - 1$ and $t - 2$ periods.

$$x_t = \alpha_0 + \alpha_1 x_{1_{t-1}} + \alpha_2 x_{2_{t-2}} + \varepsilon_t \quad \text{Equation 37}$$

Therefore, the value of the series at any time t being a linear function with values $t - 1, t - 2, \ldots, t - k$, is a multiple linear regression of k^{th} autoregression order that can be written as $AR(k)$[73].

3.8.3 Random walk model

A random walk is a special case of ARIMA. It consists on a process where the current value of a variable is composed of past value plus an error term defined as a white noise.

$$x_t = x_{t-1} + \varepsilon_t \quad \text{Equation 38}$$

The previous formula can be deducted from the first order auto regression by applying a $\emptyset_1 = 1$. The principle of the random walk consists on the best forecast of x for next period. It takes the current observation plus a random value or movement represented by ε_t thus it does not allow to forecast the changes from $x_t - x_{t-1}$ what makes the change in x random. Random walks can be characterized typically by having[47]:

- long periods of apparent trends up or down;
- sudden and unpredictable changes in direction.

3.9 Moving Average Models

The moving average consists on a combination of the model previous errors ε_t.

$$X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \quad \text{Equation 39}$$

Or

$$X_t = c + \sum_{k=1}^{q} \theta_q \varepsilon_{t-k} + \varepsilon_t \quad \text{Equation 40}$$

Where:

c	: constant
ε_t	: white noise error terms at time t;
$\theta_k \; k = 1, \ldots q$	: moving average parameters.
q	: moving average order

The X_t value can be assumed as a weighted moving average of previous forecast errors. But a caution has to be taken to not confuse the MA models with the MA smoothing. MA is essentially used for forecast values and MA smoothing is applied when making trend-cycle estimation of past values [41]. Since ε_t values are not observed values, the MA is not a really regression.

In addition, MA (q) refers to errors lags, its combinations and the MA is a data smoothing technique. However, it is possible to obtain the SMA, one of the MA techniques through the ARIMA MA (q) as follows:

$$X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \quad \text{Equation 41}$$

Considering the MA(q) equation above, if $c = 0$ and $\varepsilon_t = 0$. For $\theta_1 \ldots \theta_n = 1/n$. The equation above will result in a SMA:

$$X_t = \frac{1}{n} * \varepsilon_{t-1} + \frac{1}{n} * \varepsilon_{t-1} + \ldots + \frac{1}{n} * \varepsilon_{t-q} = \sum_{q=1}^{n} \frac{1}{n} * \varepsilon_{t-q} \quad \text{Equation 42}$$

3.10 Introduction to ARIMA Models

ARIMA models sometimes called Box Jenkins model, are one of the most commonly used time series models with normal application to both assumed stationary and non-stationary series. The ARIMA are part of a flexible forecasting group of models that uses historical data to make forecast. A time series is stationary when its probability distribution remains constant over time [50]. These models describe the autocorrelations of data, different from the exponential which aimed to describe trend and seasonality. There are two types of ARIMA models [73]:

- Non-seasonal ARIMA Models;
- seasonal ARIMA Models.

ARIMA models decomposes data into an Autoregressive process responsible for the memory of past events, an Integrated process which involves the process of transforming the data to become stationary; and the Moving Average of forecast errors. If a model involves only the autoregressive part, it is referred as an AR model. Otherwise, if a model involves only the moving average part it is referred as an MA model. An ARIMA without any difference is known as ARMA.

3.10.1 Non-seasonal ARIMA Models

The combination of Autoregressive, Integrated and Moving Average components differencing, creates a non-seasonal ARIMA model which can be represented by the following equation:

$$X'_t = c + \Phi_1 X'_{t-1} + \Phi_2 X'_{t-1} + \ldots + \Phi_p X'_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad \text{Equation 43}$$

In Arima models is possible to express the current value X_t as a function of past values or past errors. When forecasting a value after the end of the series, a value that does not yet exists, is necessary to have on the right side of the Equation 43 observed values of the series. All conditions applied to AR and MA models also applies to ARIMA. When combining complex ARIMA models the backshift notation is applied to help ease the work [41].

$$\left(1 - \Phi_1 B - \cdots - \Phi_p B^p\right)(1 - B)^d x_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\, \varepsilon_t \qquad \text{Equation 44}$$

Time series analysis has a close relation with digital processing signal, though the previous components that form the three parameters of the ARIMA models, the AR(p), I(d) and MA(q) can be seen as a set of different filters that separate signals from noise and extrapolate to obtain forecasts. The process is represented in Figure 50.

Figure 50: ARIMA as Filters. Source: The author

For the AR component, it can be represented by a IIR. This filter has its output going for infinite periods after the input has finished. In the AR there is some residue that remains looping around. The Figure 51 represents a variation IIR filter.

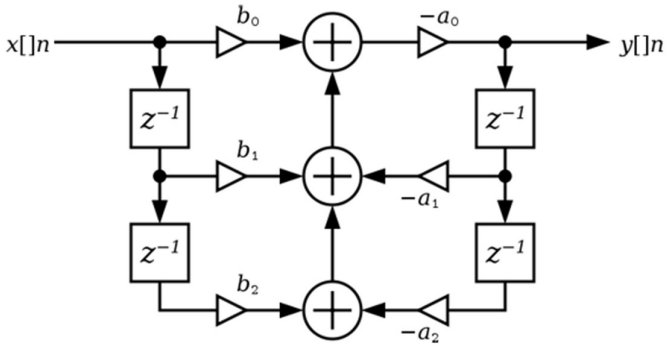


Figure 51: p-order IIR representation equivalent to AR component. Source: [74]

For the MA component, it can be represented by a FIR. This filter has its output going to zero after the input has finished. FIR is more stable when compared to IIR. The Figure 52 represents a FIR filter. For further details on FIR and IIR filters please refer to [60].

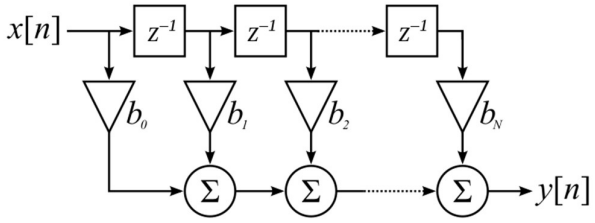


Figure 52: p-order FIR representation equivalent to MA component. Source:[74]

3.10.2 Seasonal ARIMA Models

Seasonal ARIMA models are used to specify time series with seasonal behaviour using a seasonal structure. Different from the non-seasonal models, the seasonal ARIMA models is specified by two sets of parameter order:

- $(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$ used to describe the non-seasonal component of the data set;
- $(\boldsymbol{P}, \boldsymbol{D}, \boldsymbol{Q})$ used to describe the seasonal component of m periods from the data.

Although, seasonal ARIMA process of determining (P, D, Q) parameter is analogous to when determining (p, d, q) seasonal component parameter, seasonal ARIMA in general requires a complex model structure specification. The Figure 53 represents an example of seasonal ARIMA (1,1,1)(1,1,1)[4] model equation using the backshift notation :

$$(1-\Phi_1 B)(1-\Phi_1 B^4)(1-\Phi_1 B)(1-B^4)y_t = (1+\theta_1 B)(1+\theta_1 B^4)\varepsilon_t$$

$(\text{Non}-\text{Seasonal AR}(1))$ $(\text{Seasonal AR}(1))$ $(\text{Non}-\text{Seasonal difference})$ $(\text{Seasonal difference})$ $(\text{Non}-\text{Seasonal MA}(1))$ $(\text{Seasonal MA}(1))$

Figure 53: ARIMA (1,1,1) (1,1,1)[4] example. Source: [41]

3.10.3 Modelling ARIMA procedure

There are several procedures that can be used to model Arima models. In this work are applied those considered important for the accomplishment of the defined objectives. The process of modelling a time series using ARIMA method can be grouped in three main steps [75]:

- Identification of the model;
- Estimation of model parameters;
- Forecasting future values

A fourth step which can be included consists on preparing the time series data for model building. It includes the identification of stationarity and transformation to stationary. Figure 54 depicts the general process of fitting and forecasting a time series data set using ARIMA models.

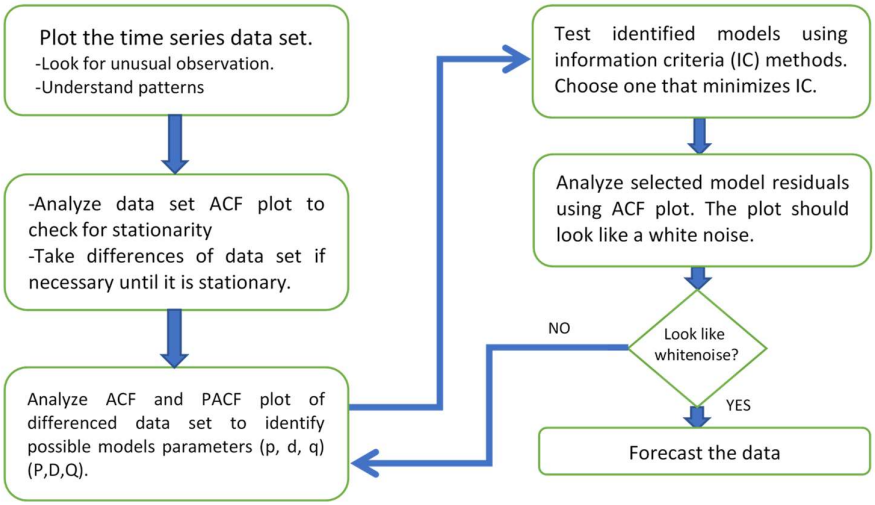


Figure 54: Forecasting ARIMA model process. Source: The author

The schematic represented in Figure 54 summarizes the steps used in this work to process an ARIMA model. However, other methods can be used in each modelling step. The previous model process can be done almost automatically using the auto.arima () function of R forecast package. The function uses the Hyndman and Khandakar algorithm. However, the auto.arima() has some limitations [41]. In the following sections some comparison will be made between the manual modelling process and the modelling using the auto.arima() function.

A. Stationarity and differencing

A stationary time series is one whose properties do not depend on the time at which it is observed, it has a mean, variance, co-variance and autocorrelation constant overtime. One example of stationarity are the white noise time series. It doesn't matter when is observed it will look much the same at any period [41].

Time series with trends, or with seasonality are not considered stationary because the trend and seasonality both influence the value of the time series at different times. Existing statistical forecasting methods works on the precondition of “stationarized” time series. Making a time series stationary is a very important step to apply when fitting a model such as ARIMA for example [76].

In case the time series is not stationary it has to be transformed by applying some mathematical transformations techniques, such as differencing, regression, transformation of data using square roots and Box-Cox, to stabilize the variance.

The identification of stationarity may be confusing at some point. However, there are several methods to help in the identification, which includes [44][46]: Augmented Dickey–Fuller, Kwiatkowski–Phillips–Schmidt–Shin (KPSS), ACF graphs to check significant lags.

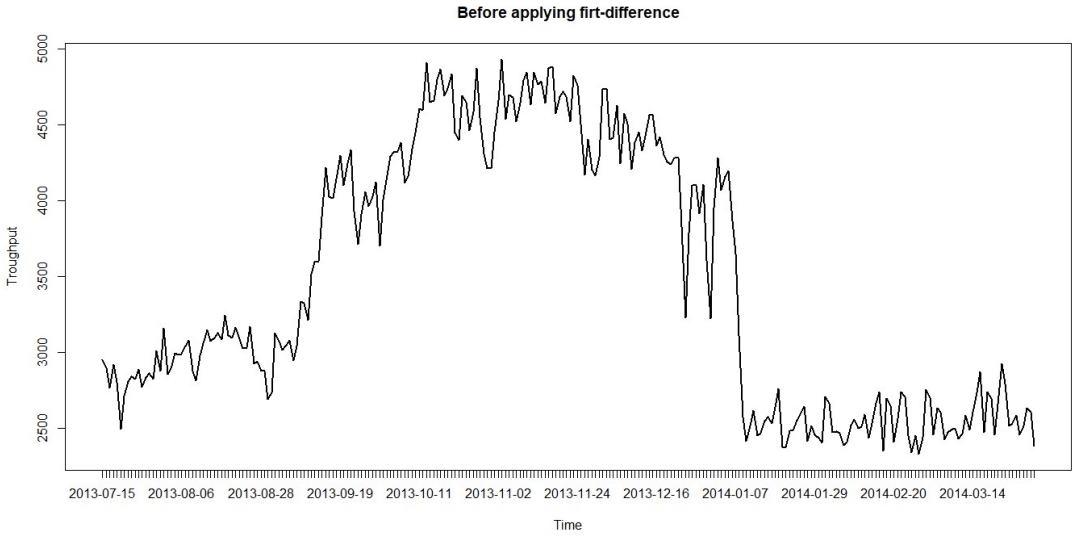


Figure 55: Time series before differentiating. Adapted from [23]

In this dissertation, the approach involves the visual identification of stationarity, recurring to time series visualization, ACF plots and the ADF test. To illustrate the process of identification and transformation of a time series stationarity, the data referring to a communication service provider weekly PS throughput will be applied. Considering the plot represented in Figure 55, of first it can be ruled out the hypothesis stationarity because:

- In general, the series presents cycles of seasonality in some periods. Specifically, the seasonality is of 7 periods.
- In the interval of 2013-07-15 up to 2013-10-11 in the time series, can be observed a tendency accompanied by cycles of seasonality which contradicts the principle of stationarity.

In addition to using plots for visual identification of stationarity as seen in the ARIMA model process schematic, another approach consists on using ACF plots which display the existent correlation between a series and its lags. Using ACF plot for stationarity identification consists in analysing the rate at which the lag decreases. For example, if the lag drops very fast to zero it means, the series is stationary and if drops slowly to zero, the series is non-stationary [47][46][75].

Using the ACF plot in the time series in analyses, before differencing, can be verified that, the lag decreases slowly which is an indication of non-stationarity, as can be seen in Figure 56.

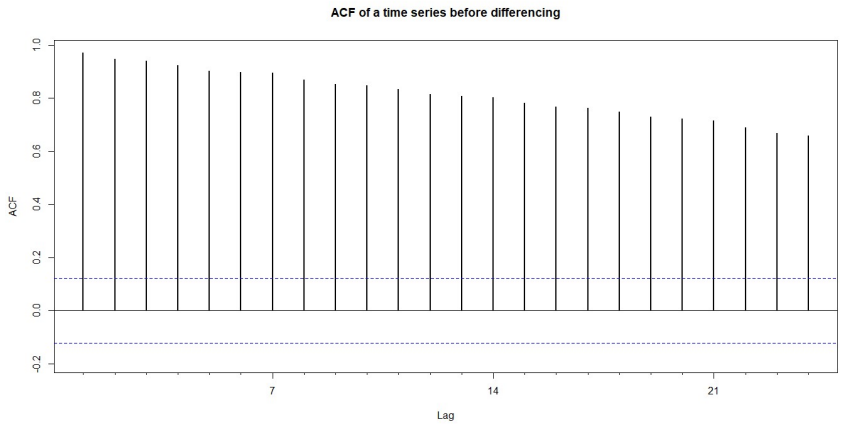


Figure 56: ACF of a time series before differencing. Source: The author

Also, can be applied the Augmented Dickey–Fuller (ADF), a unit root statistical hypothesis test which helps checking for stationarity and determines if differencing is necessary. Since this test is based on hypothesis, for null-hypothesis and for large p-values is stated that, the series is non-stationary and for small p-values the series is considered stationary. If p-value is greater than 0.05, difference application is necessary.

To obtain the ADF test result showed in the next figure the adf.test() function from the fpp library in R was used. For further details about the ADF test please refer to [50]. By calculating ADF, the test results returned a p-value of 0.8087, a value considered high. This means the null hypothesis is being accepted which means the series is not stationary and is necessary to apply a difference.

```
        Augmented Dickey-Fuller Test
data:  datats
Dickey-Fuller = -1.4471, Lag order = 6, p-value = 0.8087
Alternative hypothesis: stationary
```

Figure 57.ADF test result before differencing. Source: The author

Therefore, the set of steps involved in the process of stationarity identification offers important evidences about the most appropriate model to be used when forecasting. Often the technique used to accomplish a stationary time series is differencing.

B. Differencing

Since an ARIMA model is used to describe stationary time series, it is necessary to determine differences, in case it is not. Differencing consists on computing differences between consecutive observations. By applying difference a time series will be stabilizing its mean, eliminating trend and seasonality [41] [59]. Since the original data does not have constant properties over time, when applying difference is expected that the change from one period to another might. To differentiate a time series, the following equation is applied:

$$x'_t = x_t - x_{t-1} \qquad \text{Equation 45}$$

The above equation can also be represented by the backward shift operator:

$$x_t = (1 - B)^d x_t \qquad \text{Equation 46}$$

Where:

d : is the number of differences to apply in order to have stationarity

Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, eliminating trend and seasonality thus making the time series stationary. After differencing, the differenced data will have less points compared to the original data. If it is necessary to difference again, it will be called second order difference.

Some authors [62][73] recommend different orders of differencing, is this job recommendation that a differencing should be made as far it doesn't cause the introduction of unnecessary dependency levels.

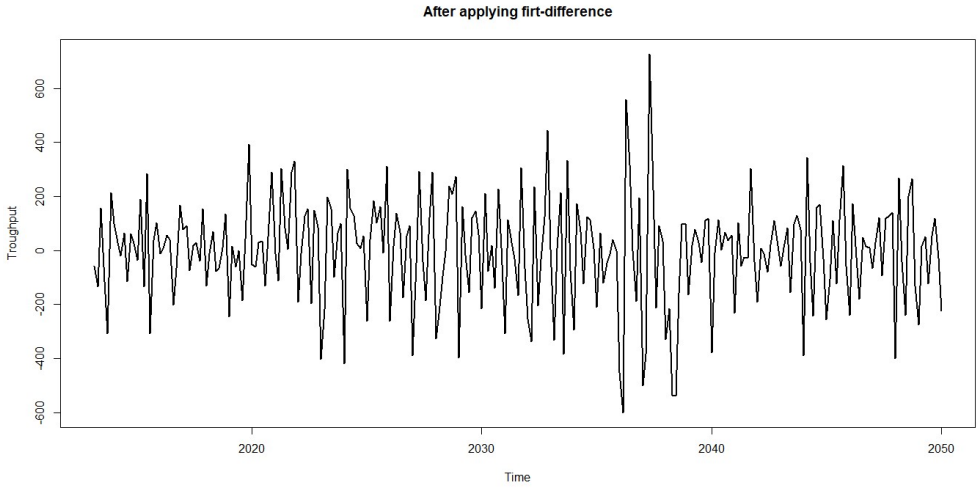


Figure 58: Time series before differentiating. Source: The author

Applying first difference to the series under analysis allowed to obtain the plot represented by Figure 58. In the plot, is possible to verify no presence of tendency nor seasonality elements which makes it stationary, that can be proven by the ACF plot represented in Figure 59. In the ACF plot, is shown a little significant lag that quickly drops to zero.

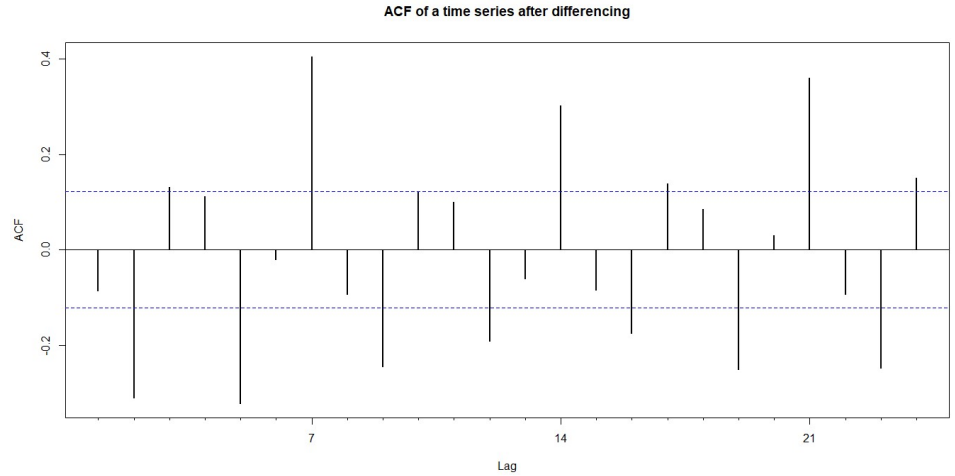


Figure 59: ACF of a time series after differencing. Source: The author

To confirm if the series needs any more differencing, the ADF test can be applied once more. Since the returned p-value represented in Figure 60 is 0.01, this suggests the series is stationary and there is no need for further differencing.

```
                Augmented Dickey-Fuller Test
data:  datats
Dickey-Fuller = -5.8489, Lag order = 6, p-value = 0.01
Alternative hypothesis: stationary
```

Figure 60: ADF test result after differencing. Source: The author

Other existing type of difference is the seasonal difference. It consists on applying s periods of seasonal multiple differences between a value and its lag. For further details on seasonal difference please refer to [41].

C. Identification of the model

To model a time series using ARIMA models is crucial to make the identification of p, d, q parameter to form the model. The identification of d is the first step to be taken, indeed by doing the difference the d value is already known. Although not addressed in any previous section, before modelling or choosing a model there is a need to analyse the influence of external variables. Using time series plot of the observed series might help in the identification of some behaviour such as upward and downward linear trend which suggest the necessity to apply difference.

As for the p and q parameter, both are identified by applying the ACF and PACF plots. These functions can help identify the number of lagged AR and MA terms on the precondition that the time series is already stationary.

The ACF plot in Figure 61 represents the existing autocorrelations by measuring the relationship between two values of the same variable, e.g. X_t and X_{t-1}, at different time t [51]. Using ACF functions is possible to express the correlation between time series values. And the PACF express unexplained correlation of two variables and a specified set of variables [59]. In other words, it shows the correlation between, a variable and its current lag that previous lags did not explain.

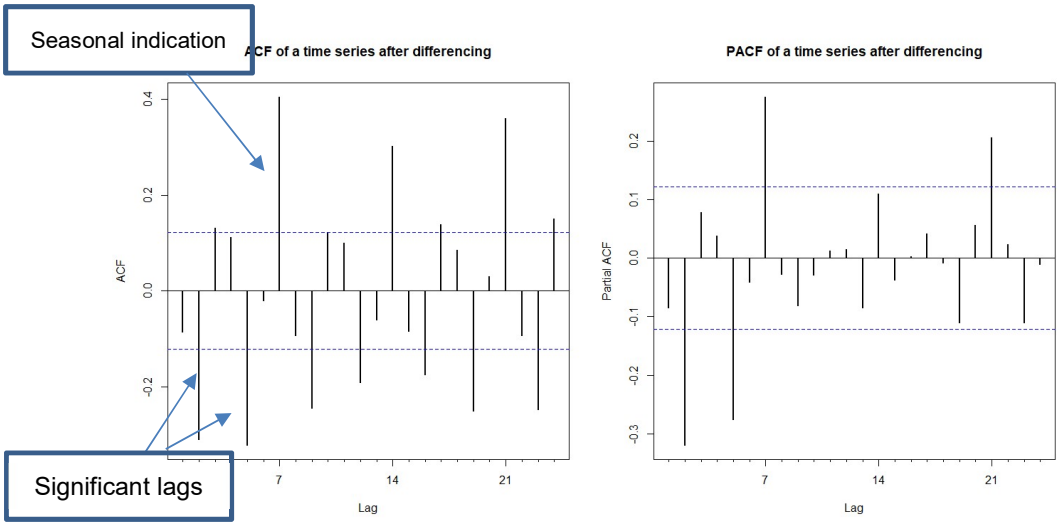


Figure 61: Model identification using ACF and PACF plots. Source: The author

Through the ACF and PACF plots of the differentiated data represented in Figure 61, is possible to determine the MA (q) and the AR(p) components as stated above. For the non-seasonal component, the goal is to look at significant lags [59]. This means, looking at the order where the lag cuts off. If both p and q are positive, then the PACF and ACF plots will not help finding the most suitable values of p and q [41]. An example of a positive ACF plot is represented in Figure 56.

The ACF plot represented in Figure 61, suggests an MA of $q = 2$ order because, the significant lag starts on order 2 and it cuts off. Still in the MA, using the same ACF plot another possible MA candidate is $q = 5$. With these, the initial candidate models are ARIMA (0,1,2) and ARIMA (0,1,5). Since only one difference of the data was taken, the d parameter is 1.

For these candidates, information criteria methods (AIC and AICc) will be used to evaluate the best candidate model. Though for the models the results in Table 14 were obtained.

Table 14: ARIMA (0,1,2) and ARIMA (0,1,5) information criteria

Model	AIC	AICc
ARIMA (0,1,2)	3445.07	3445.16
ARIMA (0,1,5)	3443.34	3443.68

From the AIC and AICc result of the initial candidate models is possible to see that, ARIMA (0,1,5) is the best model between the two. It has the minimum information criteria results. Thus, for the AR component, the PACF plot indicates $p = 2$ and $p = 5$ as possible candidates. The models are ARIMA (2,1,0) and ARIMA (5,1,0). However, it is still possible to improve the modelling and obtain other possible variation.

Because the data set presents seasonality, it is necessary to model and include the seasonal component. For the seasonal component, the AR and MA model components will be found at seasonal lags of the ACF and PACF lags [59][41]. This means, restrict attention to the seasonal lags when identifying the seasonal component. The rest of the procedure is similar to the non-seasonal model.

The significant lags in the orders 7, 14, 21 for both ACF and PACF plots are checked, because the data set shows a weekly seasonal behaviour. Thus, the significant spike at lag 7 in the ACF plot represents a good indication of a possible MA (1) model for the seasonal component. In the case of this data set, the PACF also has a significant spike at lag 7, also a good indication of AR (1) model. Thus, a possible candidate model can be, for example, an ARIMA (1,1,1)(1,1,1)[7].

Different candidate models were tested with AR and MA terms for both non-seasonal and seasonal components. The candidate models are found in the table below.

Table 15: Seasonal and non-seasonal model candidates AIC, AICc and MAPE

Model	AIC	AICc	MAPE
(2,1,5)(3,1,3)[7]	3295.99	3297.76	3.2330
(1,1,1)(2,1,1)[7]	3302.90	3303.24	3.5093
(5,1,2)(3,1,3)[7]	3304.49	3306.27	3.3218
(2,1,1)(3,1,3)[7]	3304.99	3305.90	3.4345
(2,1,1)(3,1,0)[7]	3306.30	3306.70	3.4834
(1,1,1)(3,1,0)[7]	3310.32	3310.66	3.5030
(5,1,5)(3,1,3)[7]	3311.49	3313.81	3.3124

Comparing possible model candidates represented in Table 15, the determined ARIMA (2,1,5)(3,1,3)[7] presents an AIC of 3295.99, the lowest among all. Consequently, this low value is a good indication that the model should be chosen.

An important point that must be noticed when comparing Information Criteria methods is that, all models must have the same order of differencing [41]. The objective of this dissertation is to provide the most accurate possible forecast, for that, the eligibility criteria will be not only based on information criteria method but an intersection between the minimum information criteria and minimum error. However, the preference will always be the candidates with the lowest information criteria value.

After choosing the model, an ACF and PACF plot of the residual is also analyzed to perform a residual diagnostic check, in order to infer the residuals and check for unmodeled autocorrelation that may remained [59][77]. As said, the residual diagnostic check can be made using ACF and PACF plots represented by Figure 62. When either of the plots present any autocorrelation in the residuals, other model must be considered. In case the plots behave as a white noise or all lags are within significant limits, is an indication that the chosen model is good.

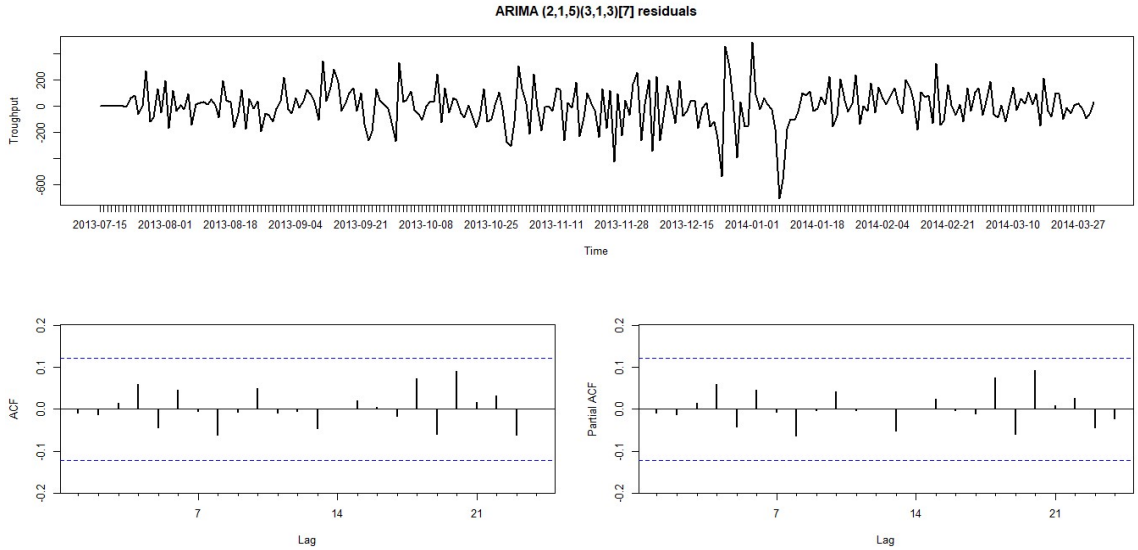


Figure 62: ARIMA (2,1,5)(3,1,3)[7] residuals diagnostic check

For a more formal check, the portmanteau test can be performed, specifically the Box-Pierce (Ljung) to test possible residual correlation. A good model must have zero correlation or nearly zero, between its residuals or can be incurred the risk of forecasting them. When performing this test, one has to look for p-values, an indication of non-zero autocorrelation possibilities within the first m lags [78]. For further detail on portmanteau test please refer to [79].

Through the application of portmanteau test to the residues of the candidate model, it can be confirmed indeed that, its residues present a behavior similar to a white noise. The result of the p-value represented in Figure 63 is 0.2164, well above 0.05, suggesting a “non-significance” of the residues.

```
        Box-Ljung test
data:  residuals(teste_arima_manual)
X-squared = 9.5369, df = 7, p-value = 0.2164
```

Figure 63: Box-Ljung test result. Source: The author

Performing the modeling using the auto.arima function, the candidate model obtained is an ARIMA(1,1,1)(2,0,0)[7], which has an AIC of 3410.57, an AICc of 3410.81 and a MAPE of 3.585373. When compared with the results of the ARIMA (2,1,5)(3,1,3)[7] it shows a great difference. This

indicates that, due auto.arima function limitation, the provided model is not as good as the ARIMA (2,1,5)(3,1,3)[7].

D. Forecasting future values

This is the last step of the ARIMA modelling. Thus, for forecasting future values the following steps can be used [41]:

- Isolate ARIMA equation in order to have the X'_t on the left side and rest of the variables on the right side of the equation, as represented in Equation 43;
- Rewrite the equation by replacing all t by $t + h$ where h is forecast horizon;
- Replace the determined input future observations by their forecasts, future errors by zero, and past errors by the corresponding residuals.

For the selected ARIMA model, the forecast result plot can be seen in the Figure 64. The figure represents a $h = 15$ periods forecast of an ARIMA (2,1,5) (3,1,3)[7]. The forecast periods appear to be slighting decreasing. In the plot, the dark grey areas represent 80% prediction intervals and the light grey represents 95% prediction interval. Those intervals are used to represent the uncertainty of the forecast and increase its length as forecast horizon is increased.

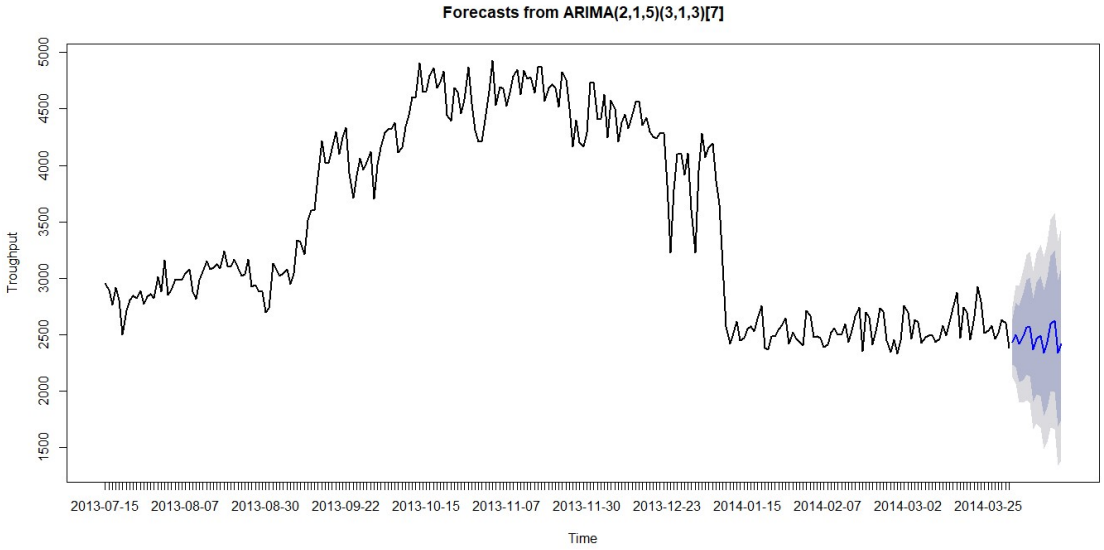


Figure 64: Forecast using ARIMA (2,1,5)(3,1,3)[7]. Source: The author

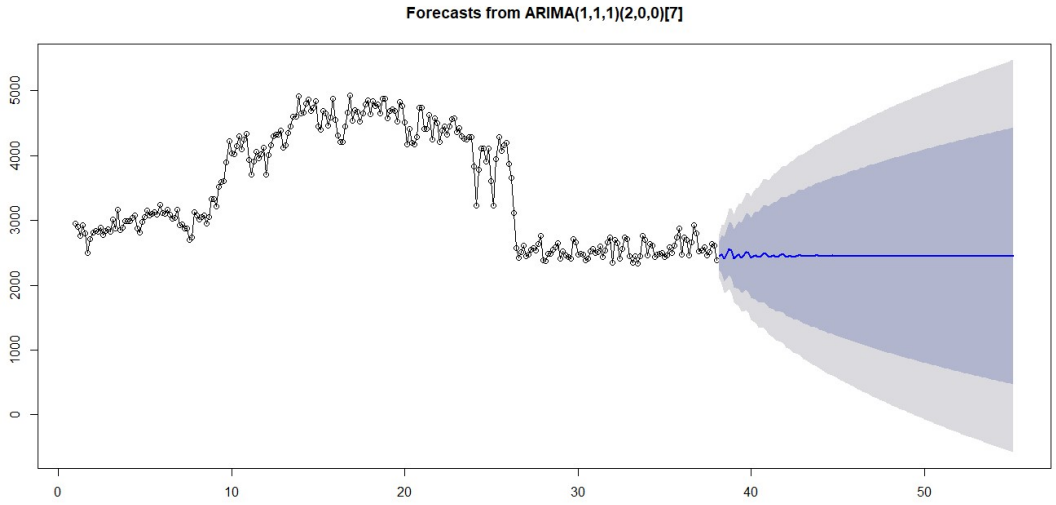


Figure 65:ARIMA(1,1,1)(2,0,0)[7] long horizon example. Source: The author

Note:

As shown in Figure 65, for stationary models the forecast of future values with a high forecast time horizon will eventually become the mean of the last observation. That is, as the forecast horizon increases at some point the forecast will pass to a naïve/random walk method. To illustrate this situation the ARIMA(1,1,1)(2,0,0)[7] is used.

3.10.4 Exponential Smoothing equivalence in ARIMA

SES is a forecasting method that seems not to require a model for the data at first. However, SES is equivalent to an ARIMA (0,1,1) model with one nonseasonal difference, a MA(1) and no constant. Nevertheless, the application of exponential smoothing should be implemented with a certain care, because it might not be well modelled by an ARIMA (0,1,1) in some situations [65]. Considering an ARIMA (0,1,1) with $\mu = 0$. Taking the first difference: $x_t - x_{t-1}$, the following model is obtained:

$$x_t - x_{t-1} = \omega_t + \theta_1 \omega_{t-1} \qquad \text{Equation 47}$$

For:

$\omega_t = x_t - \hat{x}_t$, we have:

$$\hat{x}_t = (1 + \theta_1)x_{t-1} - \theta_1 \hat{x}_{t-1} \qquad \text{Equation 48}$$

Considering an $\alpha = (1 + \theta_1)$ the following SES equation is obtained.

$$\hat{x}_t = \alpha x_{t-1} + (1 - \alpha)\hat{x}_{t-1} \qquad \text{Equation 49}$$

3.11 ARIMA and Exponential Smoothing analysis

ARIMA and Holt-Winters are one of the most used tools for time series analysis which has in common the attempt to forecast values from a variable based on current values. This section approach is to analyse both methods. Using the [62] data about 24 quarterly sales period from 1990 up to 1995, depicted in Figure 66, a comparative analysis will be made between one of exponential smoothing methods, the Holt Winters and the ARIMA, to verify the accuracy of the models.

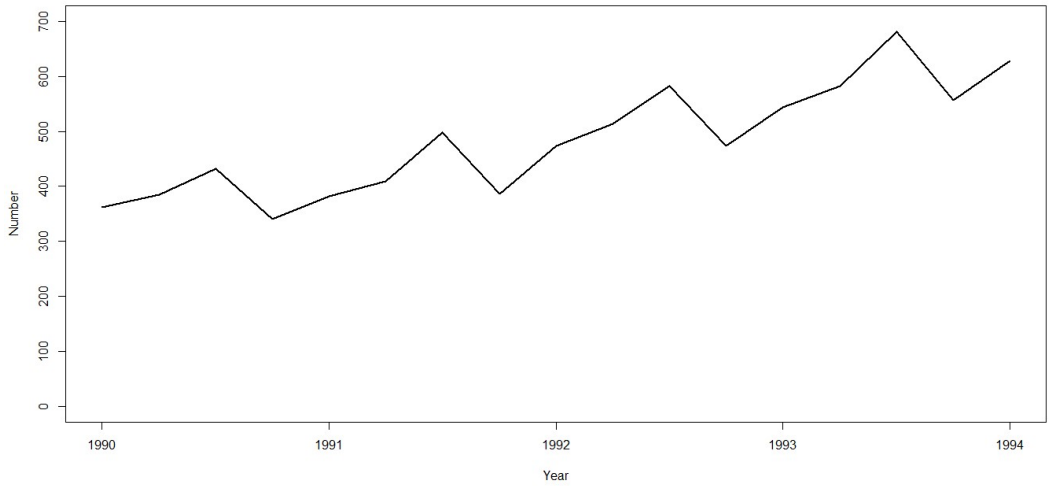


Figure 66: Observation of 24 quarterly sales period. Source: The author

Looking at the data above, can be observe the presence of some cycle and multiplicative tendency involved. With this it can be concluded initially that, the series follows a multiplicative pattern in the trend and seasonal component.

For this comparative analysis the series has been divided into two sets using a function created in R. For the first set called the “train set”, about 70% of the data is used to construct both the exponential model and the ARIMA model, in the second set called the “test set”, the rest of the data, about 30% of the data is used to help test the accuracy of the models.

Through the preliminary analysis made to the data, was identified the presence of a multiplicative seasonality, thus the holt winter method with multiplicative seasonality can be used. The fact of using Holt Winters method doesn’t require difference calculations despite the time series not being stationary, an advantage over ARIMA models. Adjusting the multiplicative Holt-Winters model to the time series in analysis, the adjustment represented in blue dotted line in Figure 67 were obtained.

In Figure 67 it is also possible to see the additive Holt Winters model depicted in red dotted line. Although it seems better adjusted than the multiplicative, it does not present good accuracy results, compared to the multiplicative.

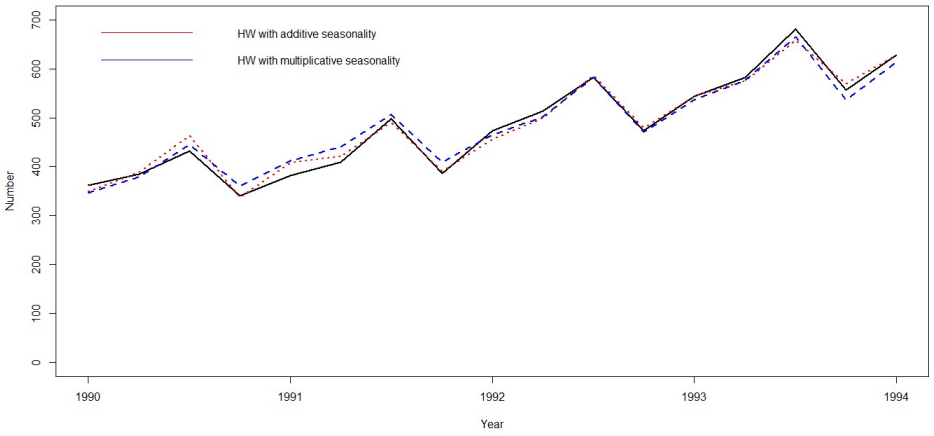


Figure 67:Adjusting Holt-Winters. Source: The author

After adjusting the model using Holt Winters model, the forecast in Figure 68 were determined. The forecast is for a $h = 7$, which represents years ahead. The value of forecasted periods was 7 because it is the size of the test set data that will be used to find the MAPE. However, this value can be changed to meet the interests of a communication service provider.

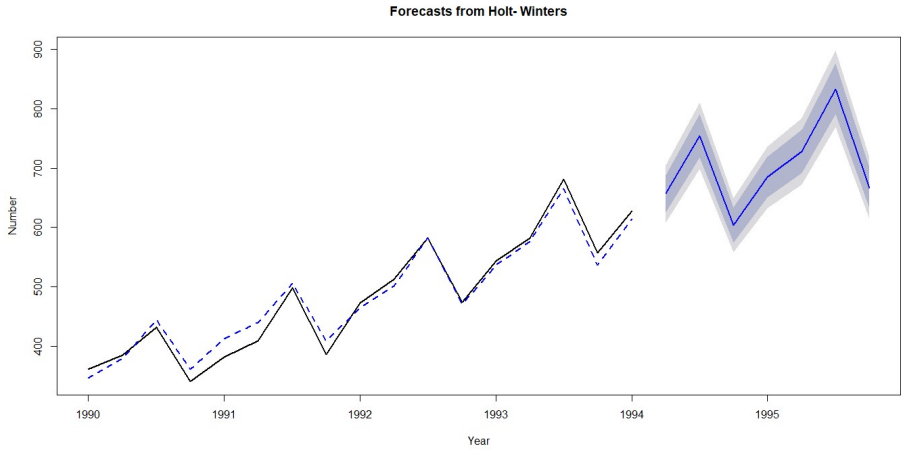


Figure 68: Holt-Winters adjustment and forecast. Source: The author

The forecast result was put to an accuracy test with the test set and resulted in Table 16 error accuracy measures. From the measures presented in the table the MAPE will be the term of comparison to the ARIMA model.

Table 16: Accuracy measures for Holt Winters. Source: The author

ME	RMSE	MAE	MPE	MAPE
1.74343	31.13781	23.92504	-0.06036231	3.480681

To model the ARIMA, as seen in the previous section, first is necessary to check the stationarity of the series. The series under analysis presents features of a non-stationary series as described at the beginning of this section through the help of Figure 66. The stationarity criteria were identified using visualization of plot but as explained in previous section the use of ADF test and ACF plots

could also be carried. Since the time series is non-stationary it was transformed to stationary. By applying first difference transformation the Figure 69 resulted.

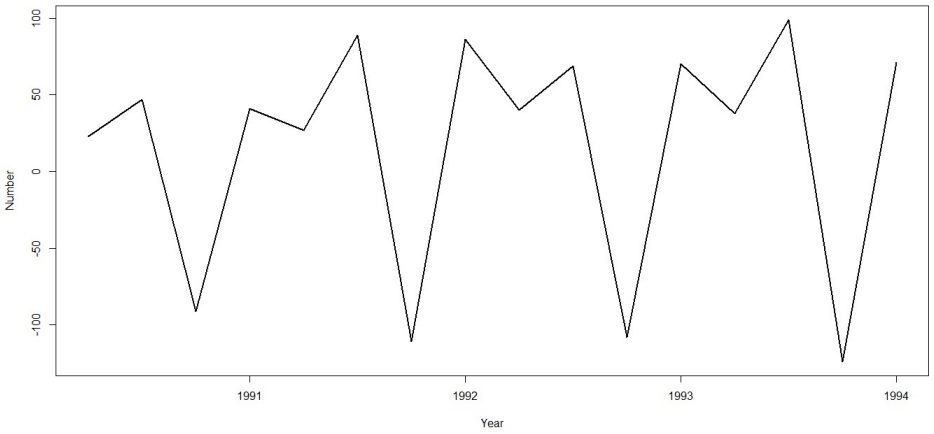


Figure 69:Differenced time series. Source: The author

With the initial data series now stationary, using the ACF and PACF plots of the training first difference data set from Figure 70, the ARIMA model will now be found. Looking for both ACF and PACF plot is possible to find an almost exponential declining behaviour in the ACF suggesting a MA (0) and a significant spike at lag 1 on the PACF plot suggesting an AR (1) for the non-seasonal component of the model.

As for the seasonal component of the model, the lags of seasonal order from the plots initially indicates a significant spike at lag 4 in the ACF plot suggesting a seasonal MA (1) component and in the PACF plot because there is not a significant spike at lag 4 a seasonal AR (0) component is suggested. Figure 70 represents the previous description.

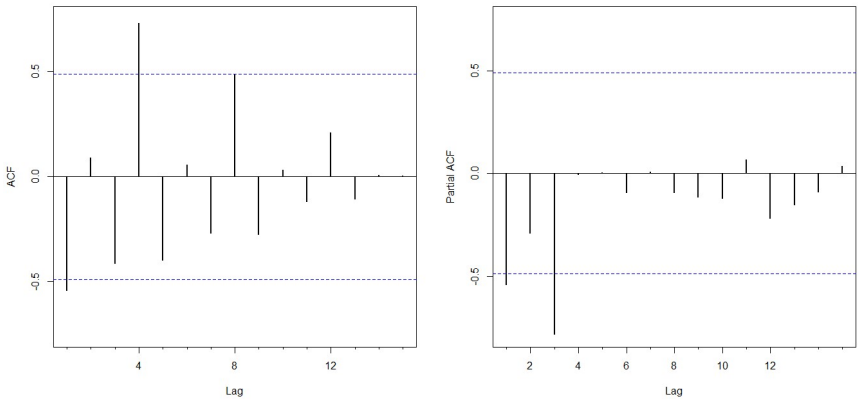


Figure 70: ACF and PACF Plots differenced. Source: The author

Though, with these components $non-seasonal\ (p\ =\ 1, d\ =\ 1, q\ =\ 0)$ and $seasonal\ (P\ =\ 0\ , D\ =\ 1, Q\ =\ 1)$, the candidate model to be formed is an ARIMA (1,1,0)(0,1,1)[4].This model resulted in the accuracy test represented in Table 17, executed between the forecast and the test set, presented in Table 17 and an AICc result of 115.21.

Table 17: Arima (1,1,0) (0,1,1)[4] Accuracy. Source: The author

ME	RMSE	MAE	MPE	MAPE
-12.506475	44.05624	39.764398	-2.332874	5.909926

Therefore, taking the previous model, the ACF and PACF plots of its residuals are represented in Figure 71, it shows that the model residues behave almost like a white noise. There are some significant spikes remaining, which indicates a necessity for further adjustments. An adjustment from the AR (0) seasonal component to an AR (1) seasonal component were made, thus, forming the candidate ARIMA model (1,1,0) (1,1,1) [4].

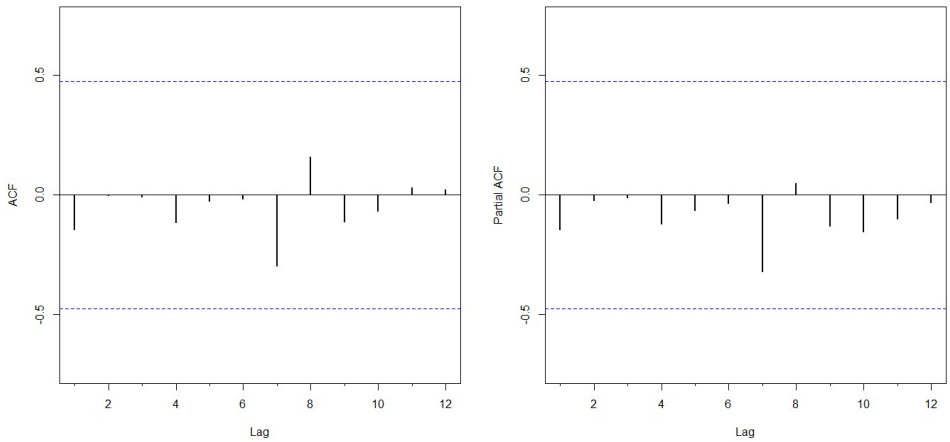


Figure 71:Residuals ACF and PACF from ARIMA (1,1,0) (0,1,1) [4]. Source: The author

For the ARIMA (1,1,0) (1, 1, 1) [4] the AICc determined was 117.82 and the accuracy measures for it is represented in Table 18.

Table 18: ARIMA (1,1,0) (1, 1, 1) [4]. Source: The author

ME	RMSE	MAE	MPE	MAPE
-13.406338	40.96645	36.650648	-2.423531	5.480605

Comparing the ARIMA (1,1,0) (0,1,1) [4] and ARIMA (1,1,0) (1,1,1) [4] AICCs, a partial conclusion is that the ARIMA (1,1,0) (0,1,1) [4] model is the best fitted model because it has the lowest AICc between the two. In addition to the above identified model's other candidate models described in Table 19 were also identified.

Table 19: ARIMA Candidates. Source: The author

Model	AIC	AICc	MAPE
$(\mathbf{0},\mathbf{1},\mathbf{0})(\mathbf{1},\mathbf{1},\mathbf{1})[\mathbf{4}]$	110.28	113.28	5.33
$(\mathbf{0},\mathbf{1},\mathbf{3})(\mathbf{1},\mathbf{1},\mathbf{0})[\mathbf{4}]$	113.98	123.98	5.65
$(\mathbf{3},\mathbf{1},\mathbf{0})(\mathbf{0},\mathbf{1},\mathbf{1})[\mathbf{4}]$	115.22	125.22	6.09
$(\mathbf{0},\mathbf{1},\mathbf{0})(\mathbf{1},\mathbf{1},\mathbf{2})[\mathbf{4}]$	112.28	118	5.92

To these candidate models, as previously proceeded, the models with lower information criteria among the ARIMA candidates will be seek. From all, the model with the lowest AICc is the ARIMA (0,1,0) (1,1,1) [4], with an AICc of 113.28. Also, was used the R forecast library auto.arima() function to model. The candidate model was an ARIMA (0,1,0) (1,1,0) [4] that has an AICc of 108.67 and a MAPE of 5.79. The results of other accuracy measures are represented in Table 20.

Table 20:Accuracy measures for ARIMA (0,1,0) (1,1,0) [4]. Source: The author

ME	RMSE	MAE	MPE	MAPE
-13.835556	43.26743	38.86750	-2.4940539	5.798748

Since the objective is to have as most as possible accurate forecast when comparing both MAPEs, the ARIMA (0,1,0) (1,1,1) [4] model has the lowest, suggesting a more accurate and best forecast performance. Figure 72 represents the forecast of an ARIMA (0,1,0) (1,1,1) [4].

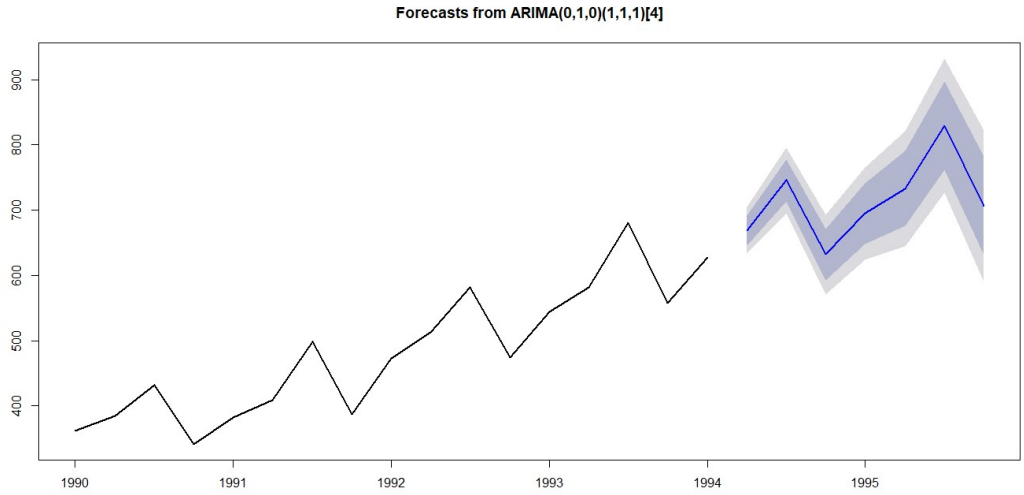


Figure 72: ARIMA (0,1,0) (1,1,1) [4]. Source: The author

For the result of accuracy test between Holt-Winter and ARIMA (0,1,0) (1,1,1)[4], the Holt Winter has a relative better MAPE. Despite this, when fitting the model, the Holt-Winter fitted the data as good as the ARIMA, as can be seen in Figure 73. However, they have small difference in the MAPE and with the help of the above figure is difficult to conclude which of then is good.

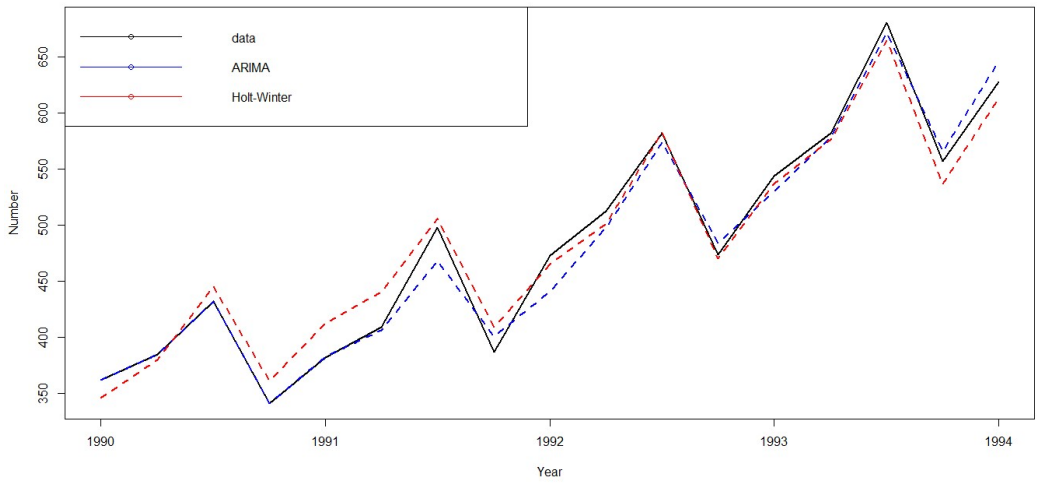


Figure 73: Holt-Winter vs ARIMA fit. Source: The author

So far, the forecasts in this section were made using the in-sample forecast. This type of forecast consists on estimating the model using all data, including T observation, and use what has been forecasted to make new forecasts. However, the determination of the MAPE was found using the test set and a set of forecasted values.

Now, forecasting will be made using the pseudo out-of-sample method. For this, the expanding window was tested first then the sliding window. For the expanding window, also the initial data where separated into two groups, the training set and the test set. But, different from previous forecast, the in-sample, in the out-of-sample for each h-step forecast the initial data set is incremented with a new input value, new observation. In the end, the forecasted values are examined with the test set to determine the MAPE and other errors.

For the expanding window method implementation, were applied the same ARIMA (0,1,0) (1,1,1) [4] previously fitted. The model resulted in a MAPE of 4.9458. The error obtained is relative low when comparing with the in-sample forecast MAPE of 5.334547. This suggests a better forecast performance when comparing both methods.

For uniformity, a size of 17 was used for the slide window because it is the size of the initial data set which result in a 1-step forecast MAPE of 5.159724, a value better than the in-sample MAPE but not as good as the expanding window MAPE. Table 21 presents the 1-step forecast values using the previous methods and also the MAPE of the forecast value with the test set values.

Table 21: Validation results. Source: The author

Expanding window forecasted 1-step values (A)							**MAPE**
669.1810	784.8291	659.2327	657.6087	694.8664	694.8664	676.4460	4.9458

Test data set values (B)							
707	773	592	627	725	854	661	

In-sample forecast (C)							**MAPE**
669.1805	745.4847	631.8202	694.7233	733.1118	829.3395	706.6020	5.334547

Sliding window- 17 forecasted 1-step values (D)							**MAPE**
669.1810	782.7030	658.8522	657.5401	696.1965	798.6759	680.1626	5.159724

Still in the sliding window, other sizes of windows were tested for the fitted ARIMA (0,1,0) (1,1,1)[4] model. In order to maintain the maximum uniformity possible, 7 periods are forecasted using 1-step horizon. First the size 4, with this window size it is not possible to forecast because the model ARIMA (0,1,0) (1,1,1)[4] cannot be used since there is not sufficient seasonal cycle to include the difference implied by the model seasonal component.

It is believed that this behavior is due the fact that, the model ARIMA (0,1,0) (1,1,1)[4] includes the seasonal component D(1), an integration, which makes it necessary to include at least two seasonal observation cycle in the initial data set.

For the slide window size tests, the test set size remained the same. A size 8, size 12 and size 16 windows were those tested. For the 8-sized window two cycles were included, the data is about quarter of sales, which resulted in a MAPE of 3.41. As for the other windows sizes, the 12-size window registered the lowest MAPE value of 2.96 and the 16-size window registered a MAPE of 5.15 as represented in Table 22. It can be observed that the windows size that include more cycles, that better describe the behavior of the series, have smaller error, until a certain window size. However, it is believed that this relationship may vary between different data set.

Table 22: Different window size MAPE results. Source: The Author

Window Size	MAPE
8	3.41
12	2.96
16	5.15

About the size of the windows, it will depend on several factors such as the size of the data set or its behavior, seasonality, trend, etc. The error resulting from applying a particular window size is also a factor to take into account. The more stable a window size is, the smaller the error is generated, the more ideal is the window size.

4 Analysis and applications of Forecasting - Case study

When modelling a time series there is a lot of approaches to consider. This section applies the forecasting techniques studied so far to carry out an analysis in a telecommunication context case study. For this, a detailed description and initial interpretation of an actual communication service provider data will be made, followed by the application of forecasting models and analysis of the errors generated by the applied models.

4.1 Daily data analysis and forecast

The data set used for the analysis and application of the forecast models were provided by a communication service provider. This type of data is obtained from the OSS performance management software. It contains information regarding a set of RNCs belonging to a 3G network. The data set describes a KPI traffic retrieved by 4 RNCs, where counters related to Circuit Switch and Packet Switch fill factors are measured. These measures are commonly used by communication service provider to provide a based proportion related to RNC licenced capacity being used for voice and data traffic at a specific moment.

For what regards daily data, the data set contains 312 observations of Circuit Switch and Packet Switch fill factor of 4 RNCs, which corresponds to about 10 months of observations.

Particularly, each group of data will be analyzed and described as follows. In Figure 74, the daily PS fil factor observations of the 4 RNC is represented.

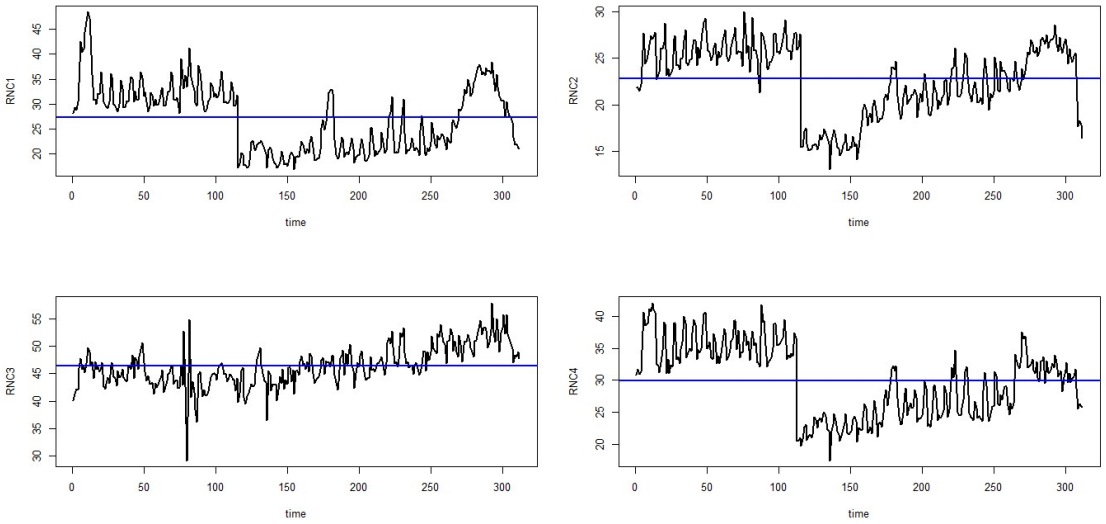


Figure 74: Daily aggregated Packet Switch fill factors initial data set. Source: The author

Regarding the trend RNC 2 and RNC 4 have slight positive trend between period 0 to 120. For other periods an increased positive trend. In the period 120, for RNC1, RNC2 and RNC4 there is an abrupt reduction in the observation values. This reduction can have several causes which includes: failure in the RNC and interruption of services or even abandonment of services by customers. In this case, an upgrade of the service capacity license for the packet switch fill factor happened.

The data represents a moderate amount of traffic for seasonal variation component with high weekends data traffic values. This suggest higher data usage during weekends.

Through the ACF plots of the Packet Switch fill factor represented in Figure 75, an exponential decrease in lag and not an abrupt drop to zero is verified. As seen in previous sections, this behavior in ACF plots is an indication of non-stationarity. Another indication is that the observations values do not obey to a certain extent the average throughout the time series, meaning the mean is not constant as shown by the blue line in Figure 74.

In Figure 74, the oscillations in the plots of ACF is a strong indication of seasonality. Specifically, these oscillations are characterized by repetitive peaks of 7 lags, a weekly seasonality indication. More details will be presented throughout this section.

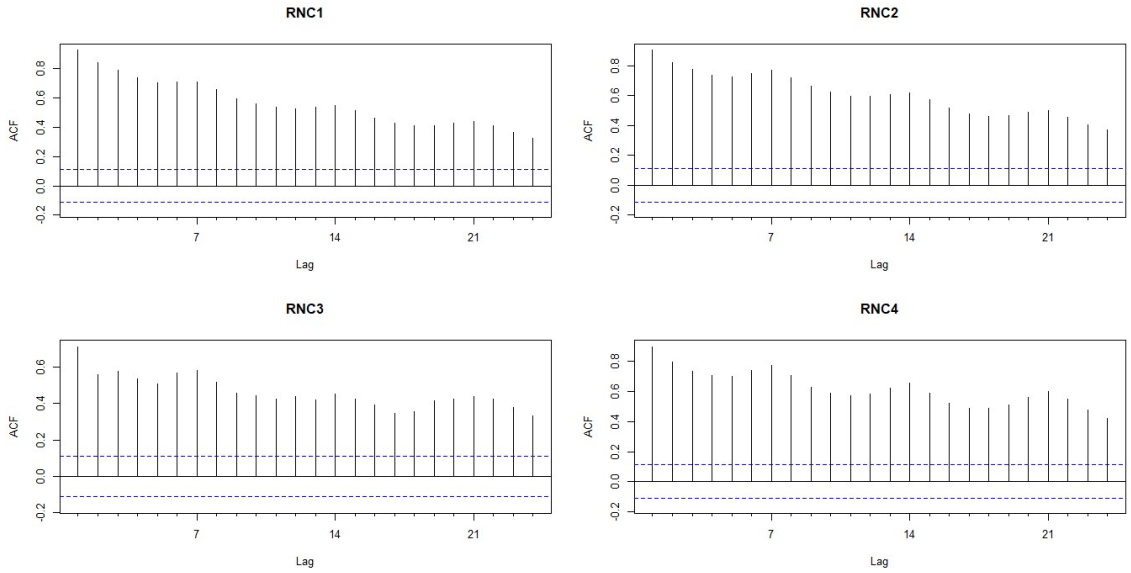


Figure 75: RNC daily aggregated packet switch fill factors ACF plots stationarity check. Source: The author

To ensure the presence of stationarity in the data is correctly determined, the Augmented-Dickey-Fuller-test was resorted. The result of the applied test is shown in Table 23.

Table 23: Initial RNC data set for daily aggregated Packet Switch fill factors ADF p-values Source: The Author

Packet switch fill factors	
RNC	**ADF p-value**
1	0.4848
2	0.7662
3	0.0728
4	0.7196

The p-value results are all greater than 0.05, consequently the null hypothesis is accepted and is confirmed that the time series are not stationary. Since the ARIMA method will be used and one of its implementation precondition is a stationary time series, a first difference will be applied to the time series. In Figure 76, the result of the application of the first difference to the data is represented. In it a strong indication of stationarity is verified. Also verified by the constant average value represented by the blue line.

The stationarity of the first difference data were confirmed by the application of Augmented-Dickey-Fuller-test, which p-values results were less than 0.05. The null hypothesis is rejected, the hypothesis of stationarity is accepted.

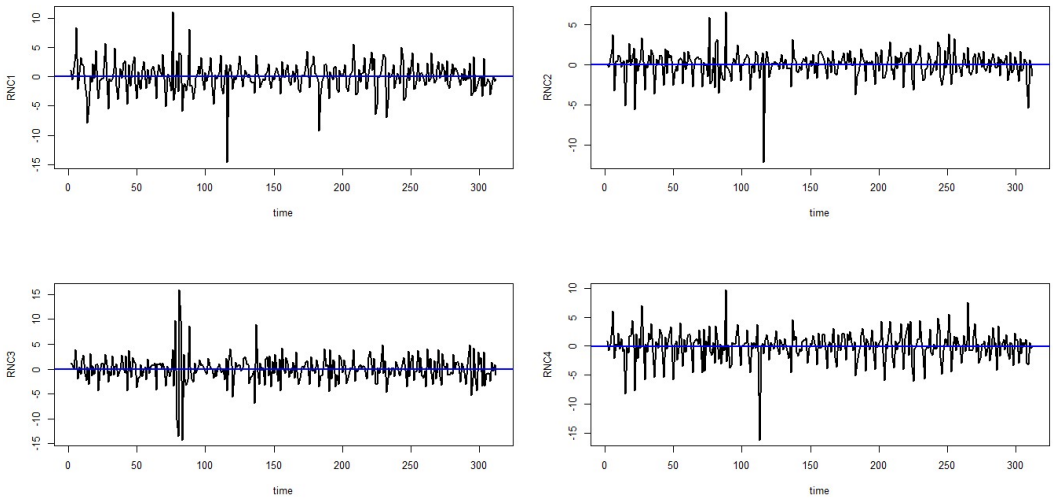


Figure 76: Daily aggregated Packet Switch fill factors of the differenced data. Source: The author

After transforming the time series data to stationary, the parameters of the ARIMA model are identified. Looking at the plots of daily Packet Switch fill factor in Figure 74, there is less to be said about the component of seasonality. To better identify the seasonal component, the ACF were used. In the initial ACF plot, represented in Figure 75 it was possible to observe cycles repetition of 7 lags and with the first difference ACF plot, represented in Figure 76 it is confirmed. The presence of seasonality indicates that the ARIMA model is seasonal.

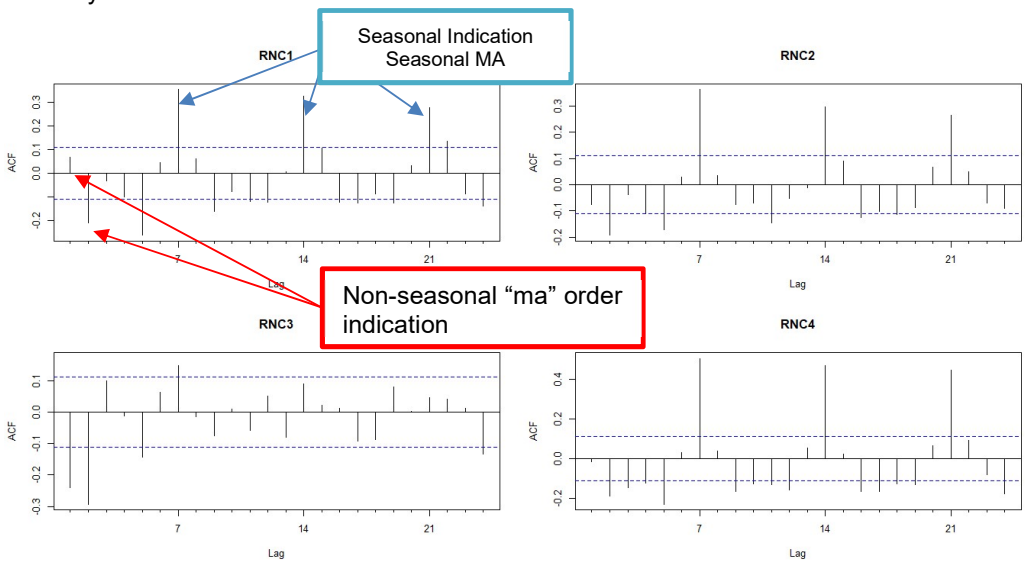


Figure 77: Daily aggregated Packet Switch fill factors first difference ACF plots. Source: The author

Based on the ACF plot of the first difference above is possible to observe that for RNC1, RNC 2 and RNC 4 the significant lags 7, 14, 21 have peaks, an indication of weekly seasonality because this peak repeats every 7 days. For RNC 3, the conclusion was also the presence of a weekly seasonality. Despite only having lag 7 as significant, were also noticed peaks in lag 14 and 21, thus the assumption of the RNC 3 having weakly seasonality. As will be seen, the Packet Switch traffic seasonal variations do not appear to be as strong as the Circuit Switch traffic.

Based in the significant lag present in the ACF plots represented in Figure 77, the possible candidates for the seasonal part of the ARIMA model parameter Q that represents the Moving Average component are: Q=1, Q=2 and Q=3 for all 4 RNCs. The Moving Average seasonal parameters are indicated by blue arrows. For the RNC data set under analysis, the candidate parameters are the same. This does not mean the chosen model will be the same for all RNCs.

In the same ACF plots, the non-seasonal part of the model is also identified. For the RNC1 the possible candidates of the non-seasonal parameter q of the Moving Average component are q=1 and q=2. The Moving Average non-seasonal parameters are indicated by red arrows.

Based on PACF plot depicted in Figure 78, the seasonal part of the ARIMA model possible candidates parameter P of the Autoregressive component are: P = 1 e P = 2, for RNC 1, RNC 2 and RNC 4. As for RNC 3 there is no candidate parameter. In these case P=0. The Autoregressive seasonal parameters are indicated by a purple line.

For the RNC1 the possible candidates of the non-seasonal parameter p of the Autoregressive component are p=1, p=2 e p=5. The Autoregressive non-seasonal parameters are indicated by a green line. As the data were differentiated once the parameter d is 1.

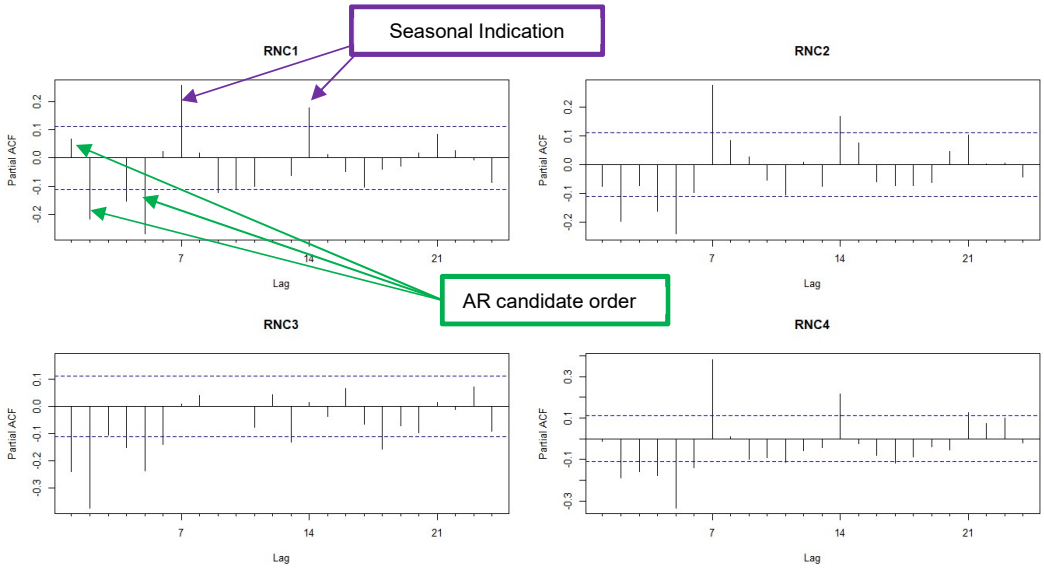


Figure 78:Daily aggregated Packet Switch fill factors first difference PACF plots. Source: The author

Must be noticed that the possible candidates described are not definitive. As will be seen, it can be changed to accommodate the candidate with the lowest information criteria. It is indicated in Table 24 ARIMA model candidates for RNC 1, chosen based on the lowest value of information criteria.

Table 24: ARIMA Daily aggregated Packet Switch fill factors for RNC1 candidate models Source: The author

Model	AIC	AICc	MAPE(%)
(5,1,0)(1,1,1)[7]	1342.46	1342.95	5.64
(2,1,1)(1,1,1)[7]	1349.2	1349.48	5.56
(2,1,2)(1,1,1)[7]	1346.05	1346.43	5.52
(2,1,2)(1,1,2)[7]	1347.49	1347.98	5.57
(2,1,1)(1,1,2)[7]	1346.27	1346.65	5.57

Of the candidate models, ARIMA (2,1,1) (1,1,2) [7] is considered the ideal. Although there are others models with better information criteria such as ARIMA (5,1,0)(1,1,1)[7]. Nevertheless, ARIMA (2,1,1) (1,1,2) [7] presents better results in the residue analysis.

To determine the forecasts, it was decided to use the expanding window with an initial window for the daily data of 192 days successively increased by one observation. For forecasting, horizons of 1, 5 and 10 days ahead were applied. These forecasts are considered short term forecast, a few days ahead, no more than 2 weeks. To test the accuracy of the determined forecasts the MAPE, MAE and RMSE were applied. The test set consisted of the last 120 days, almost 40 % of the data.

The accuracy test results for RNC1 ARIMA model can be seen in Table 25.

Table 25: RNC1 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author

RNC1			
Model	**RSME**	**MAE**	**MAPE(%)**
h = 1			
Random Walk	2.18	1.67	6.66
Holt Winters	2.03	1.57	5.95
Seasonal ARIMA	1.89	1.48	5.57
h = 5			
Random Walk	4.12	3.23	12.82
Holt Winters	4.35	3.54	14.07
Seasonal ARIMA	3.57	4.37	14.14
h=10			
Random Walk	5.11	4.27	16.41
Holt Winters	6.13	4.69	17.13
Seasonal ARIMA	5.18	4.35	16.59

In the table, is also represented error results of other models. Comparing the MAPE results from the application of these models, is possible to verify that the random walk presents poor performance for forecasts horizons of 1 day ahead. However, for horizon of 5 and 10 days, the the random walk results are better.

The Holt Winters model presents an error behavior similar to the chosen seasonal ARIMA, thier errors increase greatly as the forecast horizon increases from horizon 1 to horizon 5, compared to Random walk. From horizon 5 to 10 all three methods have an almost equal error increase rate. This situation is represented in Figure 79.

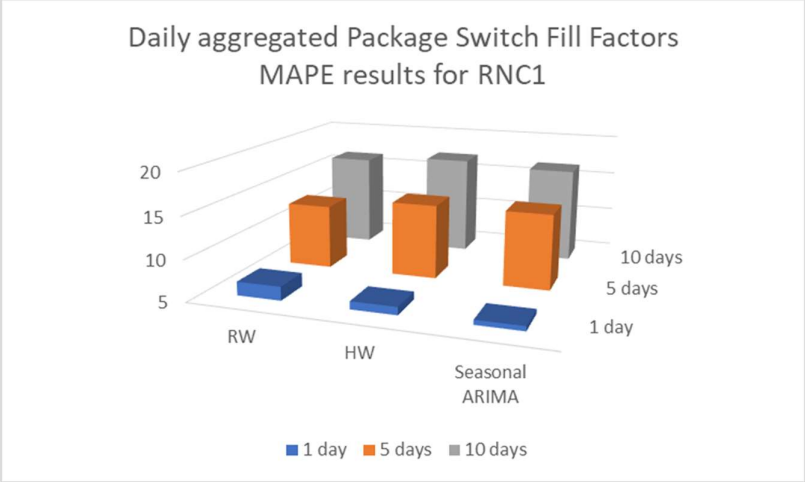


Figure 79:RNC1 MAPE results.Source: The author

The three models forecast results for horizon 1 is represented in Figure 80. The Holt-Winter has some delays after the occurrence of unexpected drops or rises of values in the data (e.g. in period 240 and between period 280-290). Also, this situation is verified for the Arima model. In general, all the three methods have a good adjustment to the data, despite the errors difference.

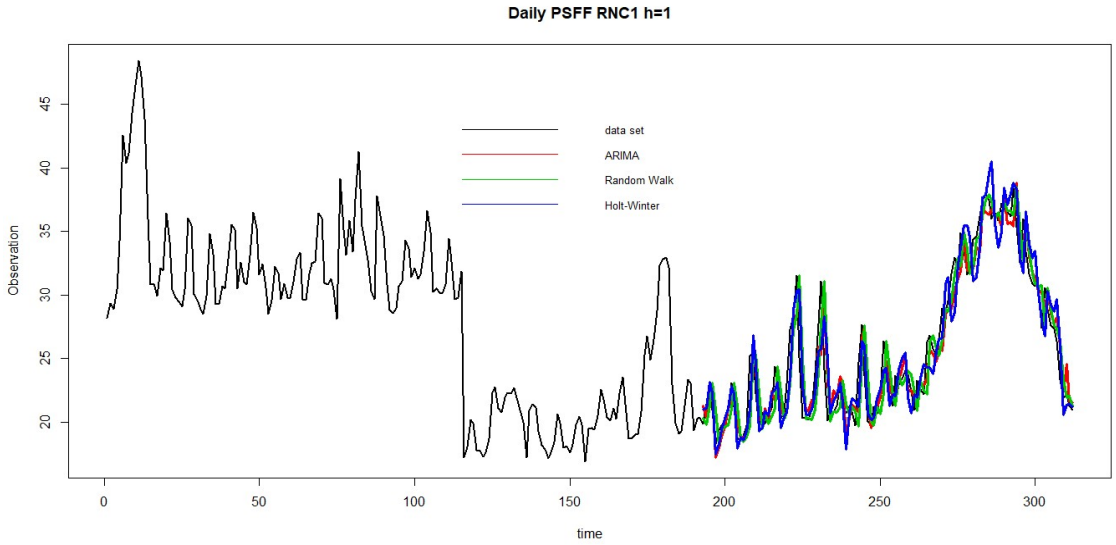


Figure 80: RNC1 daily aggregated Package Switch Fill Factors forecast. Source: The author

The same modeling process to choose the ARIMA candidate model applied to RNC 1 is applied to the rest of RNCs. Since the process is similar will be determined the accuracy measures and plot of the determined forecast.

For the RNC 2 the best candidate model was ARIMA (1,1,1) (0,1,2) [7]. For the given model, the errors generated by its forecast and other methods forecast were determined. The resulting error values are represented in Table 26.

Table 26: RNC2 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author

RNC2			
Model	**RSME**	**MAE**	**MAPE(%)**
h = 1			
Random Walk	1.41	1.11	4.95
Holt Winters	1.23	0.95	4.22
Seasonal ARIMA	1.12	0.88	3.90
h = 5			
Random Walk	2.59	1.93	8.72
Holt Winters	2.59	2.04	9.24
Seasonal ARIMA	2.69	2.12	9.59
h=10			
Random Walk	2.83	2.19	9.93
Holt Winters	2.75	2.10	9.44
Seasonal ARIMA	2.81	2.20	9.91

For horizon of 1 day, the seasonal ARIMA model has better results compared to other models. Similar to RNC1, for this RNC all three methods have their MAPE substantially increased from horizon of 1 day to horizon of 5 days, and slightly increased from horizon of 5day to horizon of 10 days.

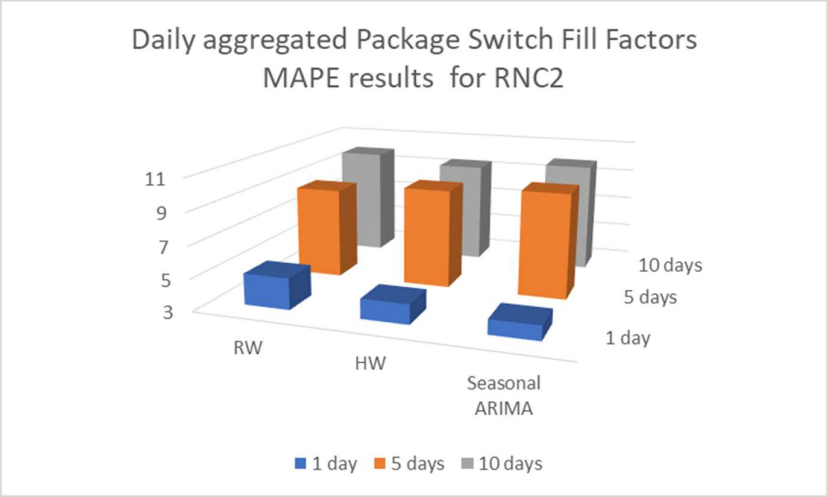


Figure 81:RNC2 MAPE results. Source: The author

In Figure 82, despite differences in the adjustment to the real observed value, the MAPE performance of the three methods when determining forecast values do not present a considerable difference with the real value. Though the highest MAPE is 4.95, for horizon of 1-day forecast. This value is acceptable

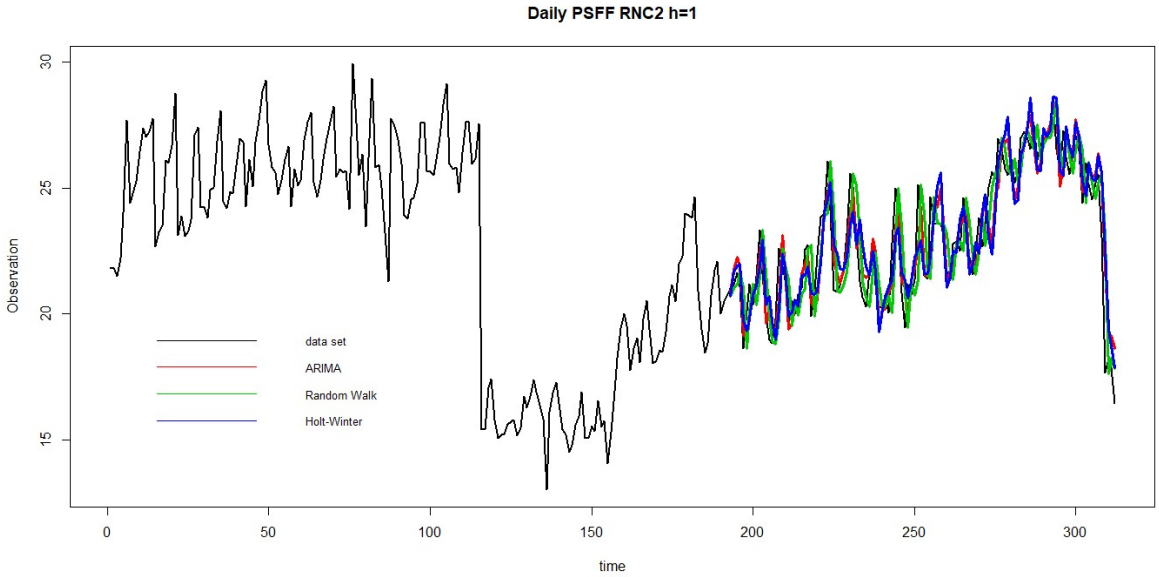


Figure 82:RNC2 daily aggregated Package Switch Fill Factors Forecast. Source: The author

The best ARIMA candidate for RNC 3 data was the ARIMA (1,1,2)(0,1,1) [7], with an AIC of 1396.66 and AICc of 1396.86. This model errors results with those of Random Walk and Holt Winters is represented in Table 27.

Table 27: RNC3 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author

RNC3			
Model	**RSME**	**MAE**	**MAPE(%)**
h = 1			
Random Walk	2.17	1.76	3.55
Holt Winters	2.16	1.78	3.61
Seasonal ARIMA	2.13	1.73	3.49
h = 5			
Random Walk	3.27	2.60	5.29
Holt Winters	3.08	2.61	5.33
Seasonal ARIMA	3.22	2.67	5.55
h=10			
Random Walk	3.33	2.79	5.66
Holt Winters	3.29	2.78	5.61
Seasonal ARIMA	3.11	2.58	5.33

Different from the previous RNCs, in RNC 3 there is little difference between the model's errors. For horizon of 1 day and horizon of 10 days forecast horizon Arima has better MAPE results and for horizon of 5 days, the Random Walk. Still, the errors do not increase with foreground forward horizons, as seen in RNC1 and RNC2 cases.

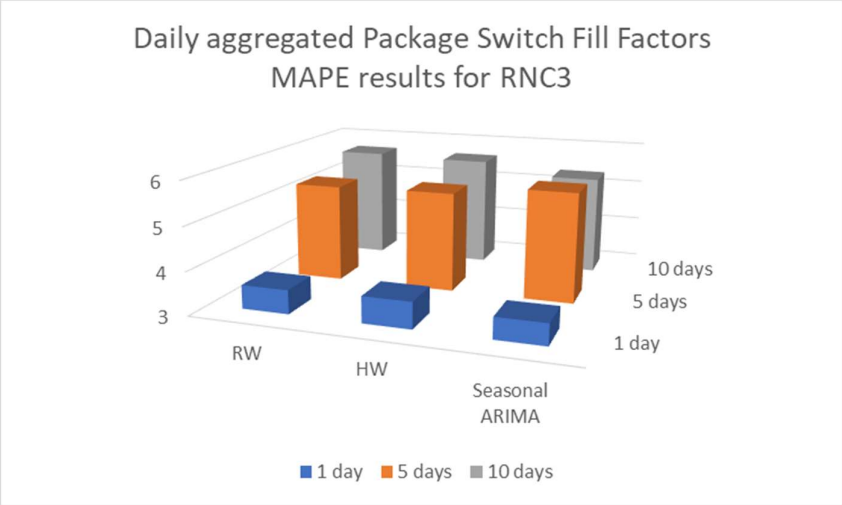


Figure 83: RNC3 MAPE results. Source: The author

Figure 83 represents the previous statement. In the figure, is possible to see that from horizon of 1 day to horizon of 5 days, the almost double. And the MAPE values differences between models is almost the same. As depicted in Figure 84, for RNC 3, between period 220 to 250, both Holt-Winters and ARIMA models have a lot of error. It is believed that this error is due to the slowness of both methods to recover from an abrupt change in data.

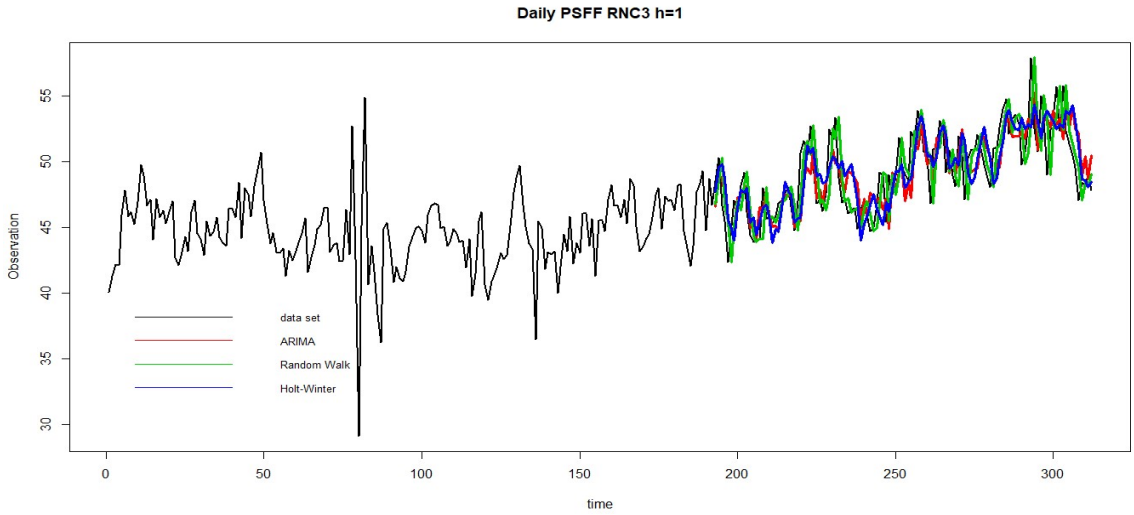


Figure 84:Daily aggregated Package Switch Fill Factors Forecast for RNC3. Source: The author

For RNC 4 the best candidate was ARIMA (1,1,1)(2,1,3) [7] with an AIC of 1251.4 and an AICc of 1251.89. A comparison of ARIMA model errors with those of Random Walk and Holt Winters is presented in Table 28.

Table 28: RNC4 accuracy measures for daily aggregated Package Switch Fill Factors Source: The author

RNC4			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	2.35	1.74	6.12
Holt Winters	1.91	1.42	4.89
Seasonal ARIMA	1.82	1.33	4.56
h = 5			
Random Walk	3.99	3.26	11.41
Holt Winters	4.28	3.60	12.61
Seasonal ARIMA	4.14	3.51	12.27
h=10			
Random Walk	4.44	3.46	12.09
Holt Winters	4.31	3.41	11.84
Seasonal ARIMA	4.31	3.43	11.87

Similar to previous RNC the ARIMA model has better results for horizon of 1 day. For horizon of 5 days and horizon of 10 days it has the second lowest MAPE. Similar to RNC1 and RNC2, the three methods have their MAPE substantially increased from h=1 to h=5. The MAPE behavior for the three models is shown in Figure 85.

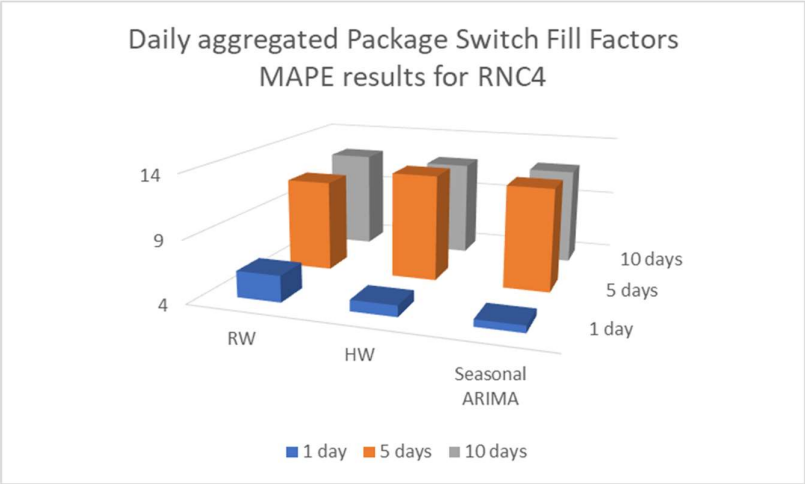


Figure 85: RNC4 MAPE results. Source: The author

Both ARIMA and Holt-Winter models forecast do not follow abrupt occurred changes in the data. However, both models are accurate. In the forecasts represented by Figure 86, ARIMA and Holt-Winters produces considered amount of error between period 230 and 240. Despite this, the quality of the forecast is maintained through time.

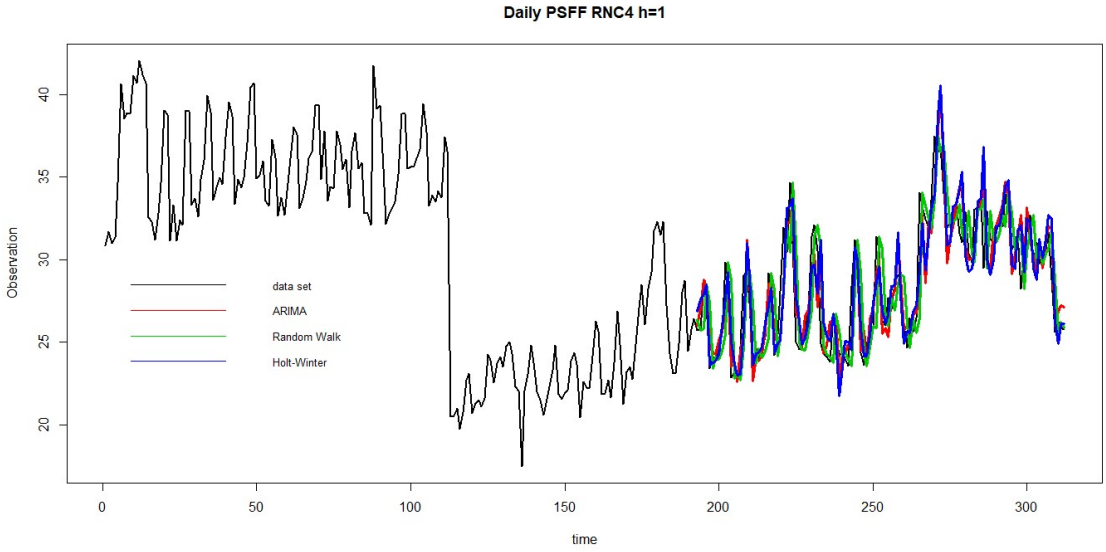


Figure 86: RNC4 daily aggregated Package Switch Fill Factors Forecast. Source: The author

The daily Circuit Switch fill factor data of the four RNC is represented in Figure 87. The voice traffic compared to the data traffic previously analysed has a well-defined seasonality. However, this seasonality is characterized by high values during the week working days and substantially lower values during the weekends, this suggests a high weekly voice traffic utilization.

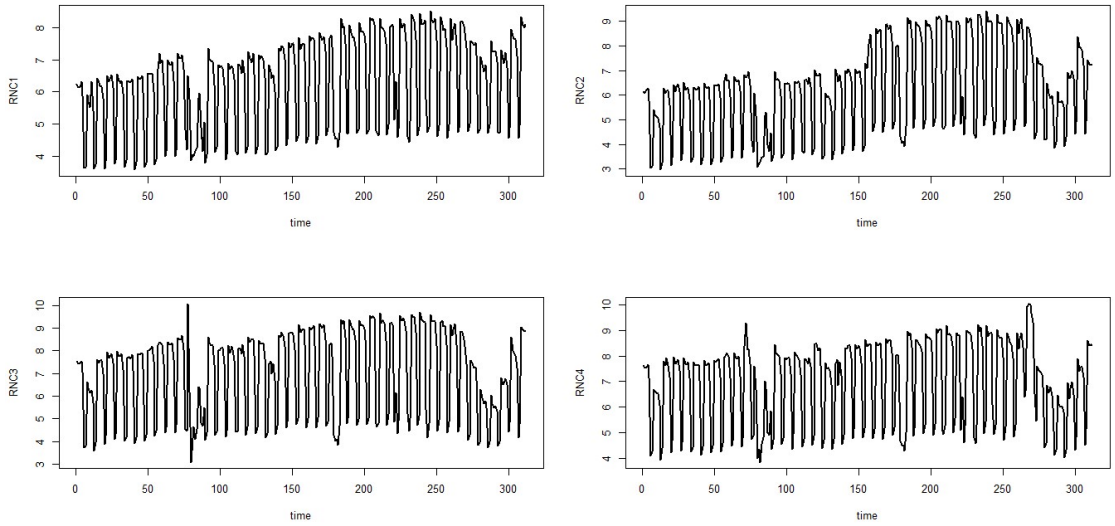


Figure 87: Daily aggregated Circuit Switch fill factors initial data set. Source: The author

Using ARIMA modeling procedures applied in previous sections, for RNC1 daily Circuit Switch fill factor data, the best candidate was ARIMA (2,1,2) (1,0,0) [7] with an AIC of 676.4 and an AICc of 676.67. ARIMA forecast errors and other models forecast errors is represented in Table 29.

Table 29: RNC1 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author

RNC1			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	1.61	1.02	16.50
Holt Winters	0.65	0.43	6.96
Seasonal ARIMA	0.79	0.48	7.59
h = 5			
Random Walk	2.29	1.86	31.25
Holt Winters	2.42	2.07	34.14
Seasonal ARIMA	2.28	1.89	31.18
h=10			
Random Walk	2.47	2.12	34.69
Holt Winters	2.22	1.85	30.02
Seasonal ARIMA	2.24	1.83	29.81

From the table can be verified that Random Walk presents higher values of error in relation to other models for horizon of 1 day and horizon of 10 days. When compared to other models ARIMA has minor errors for horizon of 5 days and horizon of 10 days. Holt-Winters has the worst error result for horizon of 5 days. From horizon of 1 day to horizon of 5 days the Mape is considerable increased. This is represented by Figure 88 MAPE results.

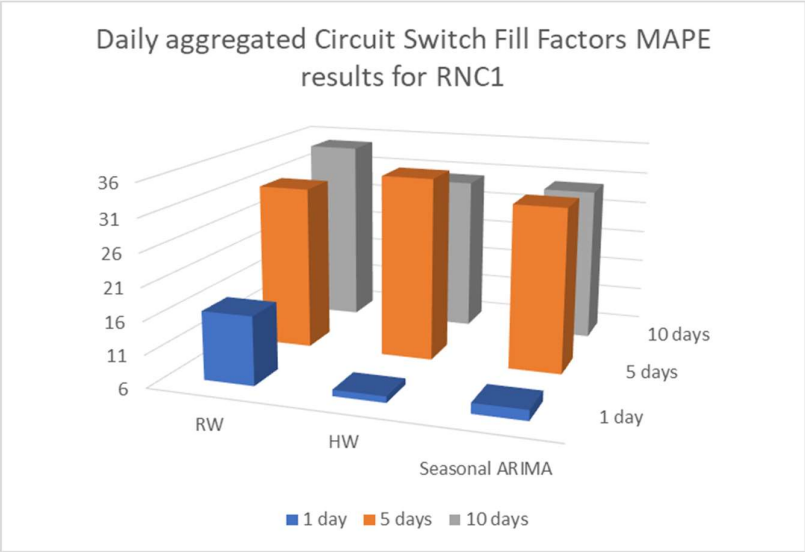


Figure 88: RNC1 MAPE results. Source: The author

Figure 89 shows the models forecast. From the figure, is possible to see that the Random Walk accurately fits forecast values to the real values despite the high MAPE. The Holt-Winters and Arima has significant abrupt peaks in periods 210, 220 and 260. Despite the differences in the error result all three methods has a good performance, except the Random Walk in horizon of 1 day. It means the percentage of error when applying MAPE for horizon of 1 day will be greater than other methods.

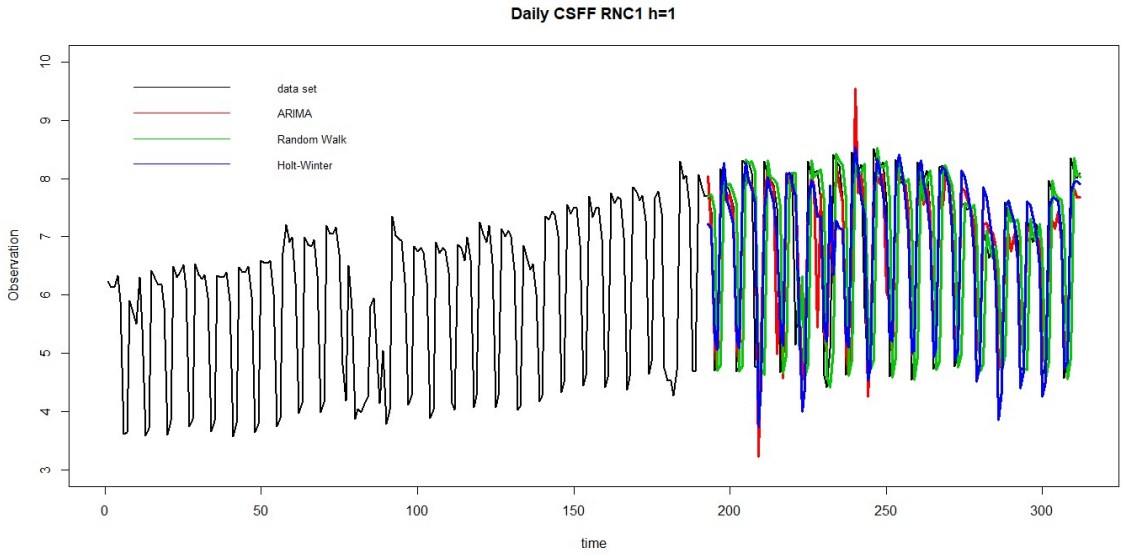


Figure 89: RNC1Daily aggregated Circuit Switch Fill Factors Forecast. Source: The author

The best candidate model for RNC 2 data was the ARIMA (2,1,2)(2,0,0)[7] with an AIC of 777.77 and an AICc of 778.03. A comparison of ARIMA model errors with those of Random Walk and Holt-Winters is represented in Table 30.

Table 30: RNC2 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author

RNC2			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	1.92	1.19	19.62
Holt Winters	0.96	0.67	11.33
Seasonal ARIMA	1.06	0.68	10.67
h = 5			
Random Walk	2.79	2.20	37.2
Holt Winters	2.96	2.52	41.94
Seasonal ARIMA	2.95	2.45	40.42
h=10			
Random Walk	3.05	2.61	42.82
Holt Winters	2.73	2.29	37.34
Seasonal ARIMA	2.82	2.37	38.68

For this RNC, all three methods have consiradable high MAPE results for horizon of 5 days and horizon of 10 days. Similar to previous analysis, for horizon of 1 day and horizon of 10 days the Random walk method has the highest MAPE values. The ARIMA has the lowest MAPE for horizon of 1 day and the second lower MAPE for horizon of 10 days. Figure 90 represents the above models MAPE for horizon 1, 5 and 10. This helps understand how the errors deviates from one another.

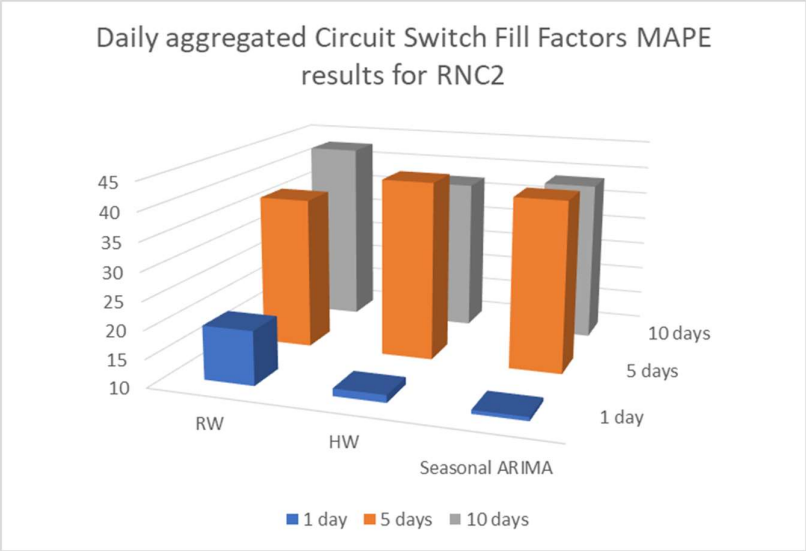


Figure 90:RNC2 MAPE results. Source: The author

Both ARIMA and Holt-Winters has some peak along the forecast, which mean a bad adjustment to new observed values. Despite those peaks, both models have better MAPE values compared to Random Walk.

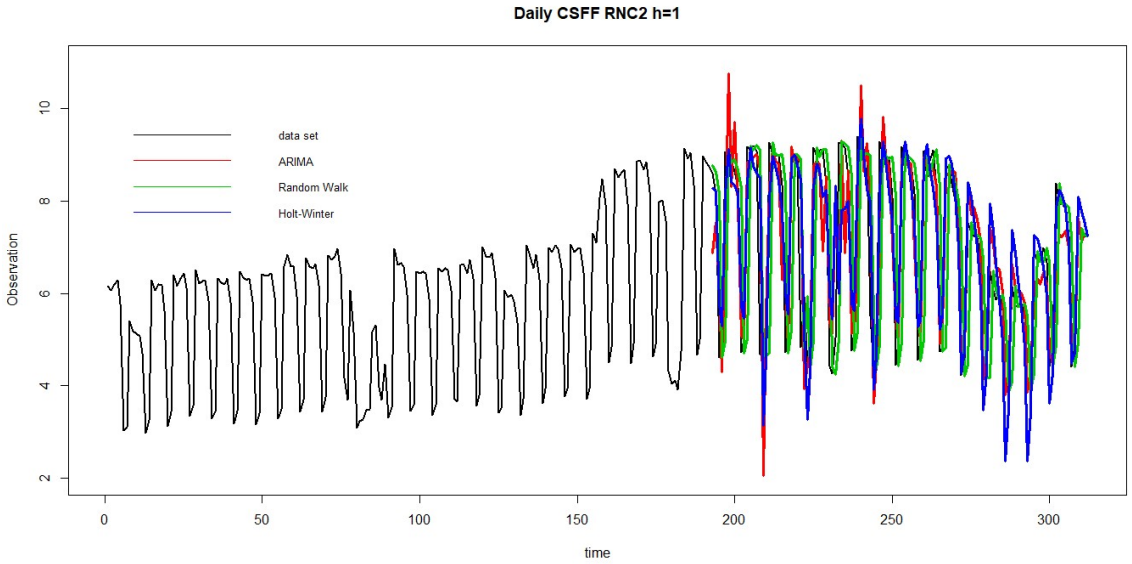


Figure 91:Daily aggregated Circuit Switch Fill Factors Forecast for RNC2. Source: The author

For the RNC 3 data set, the best candidate was ARIMA (2,1,2)(2,0,0)[7] with an AIC of 859.15 and an AICc of 859.52. A comparison of errors for all three models is presented in Table 31.

Table 31: RNC3 accuracy measures for daily aggregated Circuit Switch Fill Factors Source: The author

RNC3			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	2.06	1.27	20.78
Holt Winters	1.02	0.70	11.83
Seasonal ARIMA	1.14	0.77	11.93
h = 5			
Random Walk	3.00	2.38	40.10
Holt Winters	3.20	2.73	45.26
Seasonal ARIMA	3.16	2.65	43.45
h=10			
Random Walk	3.29	2.81	45.96
Holt Winters	2.98	2.47	40.19
Seasonal ARIMA	3.03	2.57	41.71

In this data set, similar to RNC2 for all three methods the MAPE has high values for horizon of 5 days and horizon of 10 days. This situations is also true for horizon of 1 day MAPE for Random Walk method. The MAPE behavior for the three models is shown in Figure 92. Recuring to this figure, is possible to see an enormous increase of MAPE values from horizon of 1 day to horizon of 5 days.

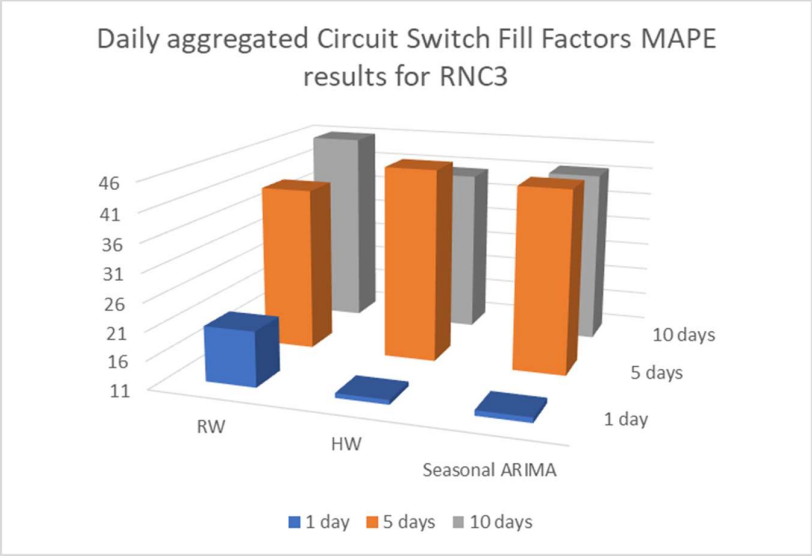


Figure 92: RNC3 MAPE results. Source: The author

Similar to previous RNC, for this RNC there is also some peaks along the forecast. Despite those peaks their MAPE is better than the Random walk. Must be notice that the figure bellow represent horizon of 1-day forecast. For horizon of 5 days ahead the forecast tends to have higher MAPE values.

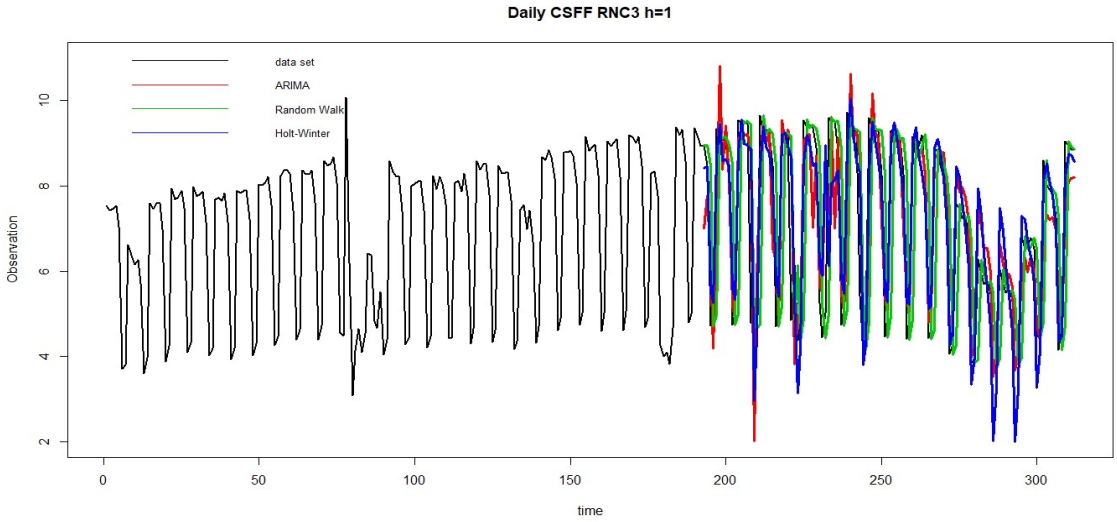


Figure 93: RNC3 daily aggregated Circuit Switch Fill Factors Forecast. Source: The author

For the RNC 4 data the best candidate was ARIMA (2,1,2)(2,0,0)[7], an AIC of 755.01 and an AICc of 755.39 were determined. A comparison of ARIMA model errors with those of Random Walk and Holt Winters is represented inTable 32.

Table 32: RNC4 accuracy measures for daily aggregated Circuit Switch Fill Factors

RNC4			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	1.73	1.10	17.40
Holt Winters	0.83	0.56	8.98
Seasonal ARIMA	0.96	0.66	9.93
h = 5			
Random Walk	2.52	2.03	32.29
Holt Winters	2.72	2.32	36.18
Seasonal ARIMA	2.66	2.23	34.81
h=10			
Random Walk	2.80	2.38	37.21
Holt Winters	2.55	2.10	32.59
Seasonal ARIMA	2.57	2.15	33.29

This RNC has a similar behaviour to RNC1. The MAPE behavior for the three models is shown in Figure 94. And the forecast for horizon of 1 day is represented in Figure 95.

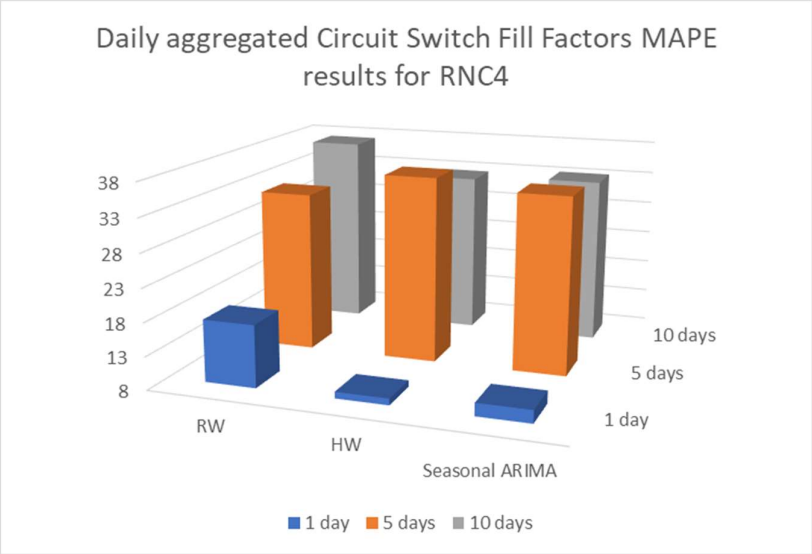


Figure 94:MAPE results comparison for RNC4-CSFF. Source: The author

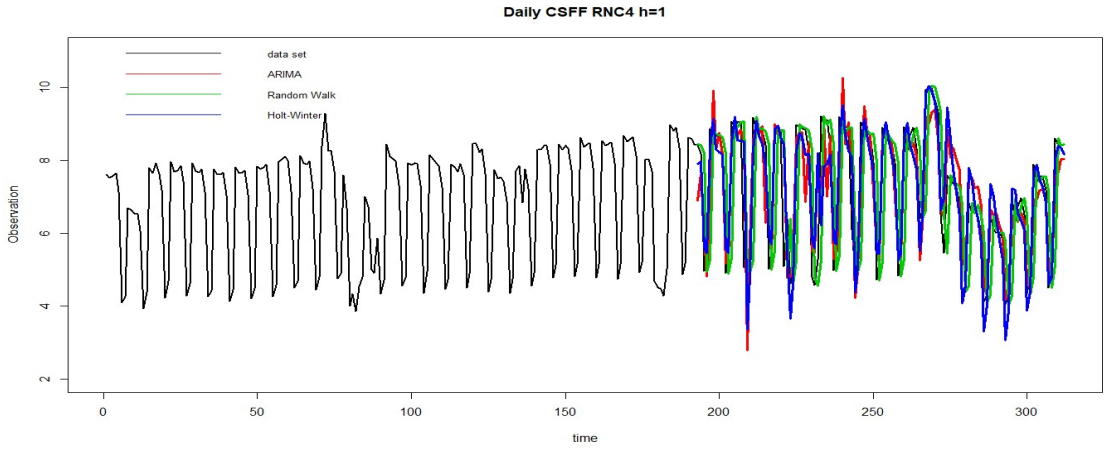


Figure 95:RNC4 daily aggregated Circuit Switch Fill Factors Forecast. Source: The author

4.2 Weekly data analysis and forecast

For the weekly data, were selected data set with a total of 146 observations of Circuit Switch and Packet Switch fill factor of 4 RNCs corresponding to almost three years. Of this total, exactly 86 observations are used for modeling and the remaining 60 observations are used to determine how well the models fit the data, by determining accuracy error. The 86 observation is also the size of the expanding window applied for the weekly data, almost 60% of total.

Different from daily data, the weekly data will be used in order to determine long term forecasts, up to several weeks ahead. Thus, forecast of 1, 2 and 4 weeks ahead successively increased by one observation will be determined. In Figure 96, the weekly RNC PS fil factor observations in analysis is represented.

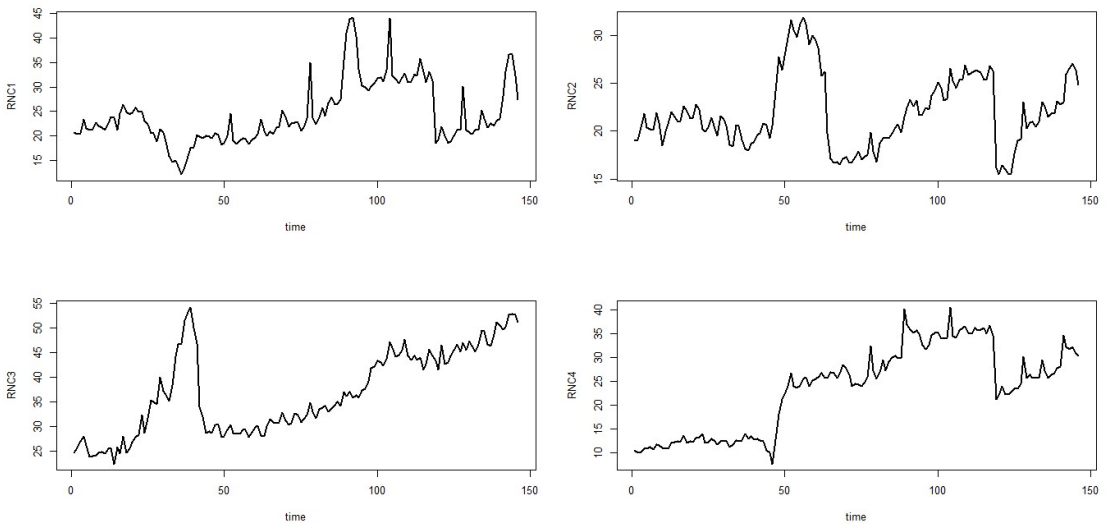


Figure 96: RNC weekly aggregated packet switch fill factors. Source: The author

Regarding the trend, for RNC1 there is an increasing tendency in observations between periods 50 and 150. RNC 2 trend is very noticed from period 50 to 150. As for RNC3 and RNC4, the trend is identified along the time series. Due to afore mentioned upgrade on the capacity license, a steep decrease can be verified in RNC1, RNC2 and RNC4 around period 130. Also, large gradual variations are observed.

In general, for the 4 RNC time series it is verified that the weekly seasonality is not present. This lack of seasonality is confirmed by the ACF plots shown in Figure 97. In the figure the lags do not have a repeating pattern. Since the data under analysis do not have the seasonal component, the model to be used is a non-seasonal ARIMA. Therefore, during the modeling process the seasonal part were not included. The ARIMA model with the lowest information criterion value is chosen, similar to previous analyzes.

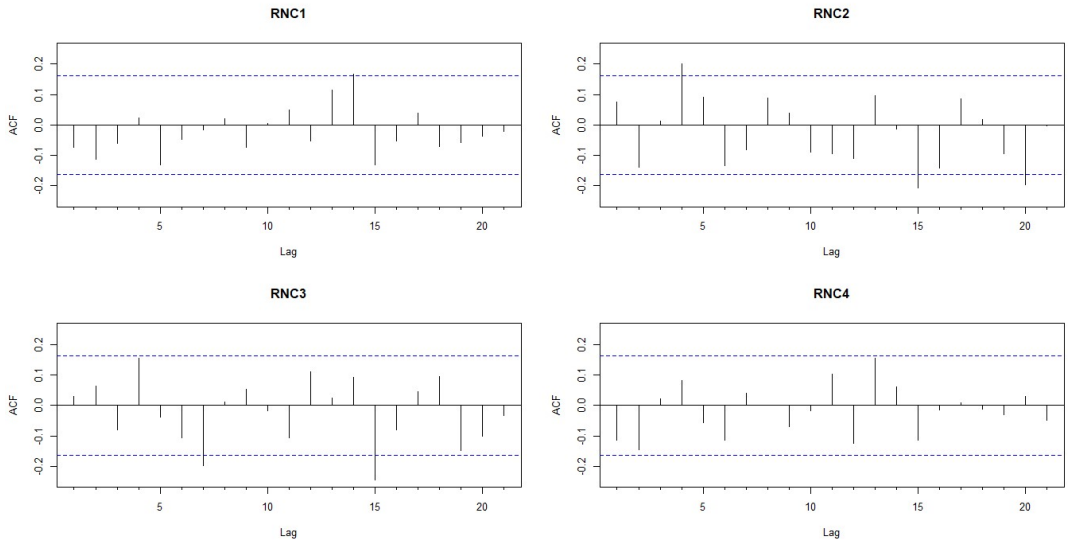


Figure 97: First difference ACF weekly aggregated Packet Switch fill factors. Source: The author

The best ARIMA candidate model for RNC1 was the ARIMA(1,1,1) model, with an AIC of 733.11 and an AICc de 733.28. An accuracy table between the chosen model, Random walk and Holt-Winter models is represented in Table 33.

Table 33: RNC1 accuracy measures for weekly aggregated Packet Switch Fill Factors. Source: The author

RNC1			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	3.951459	2.586875	9.055661
Holt Winters	4.146371	2.76155	9.526881
ARIMA	4.098037	2.717918	9.417961
h = 2			
Random Walk	5.656086	3.915226	13.76977
Holt Winters	5.852848	4.115914	14.25493
ARIMA	5.75441	3.901424	13.54079
h=4			
Random Walk	7.598717	5.335732	18.38721
Holt Winters	7.831512	5.626464	19.1571
ARIMA	7.445966	5.339964	18.68303

The MAPE error horizon of 1 week suggest the random walk is the best model to be applied. However, the other models error values are also very low which means they can also be used to forecast horizon of 1 week ahead. From horizon of 1 week to horizon of 2 weeks there is a slight increase in MAPE, less than the increase registered in the daily PS and CS fill factor analysis. For horizon of 2 weeks and horizon of 4 weeks the difference between the models MAPE is of decimals. The MAPE behaviour of RNC1 models in represented in Figure 98.

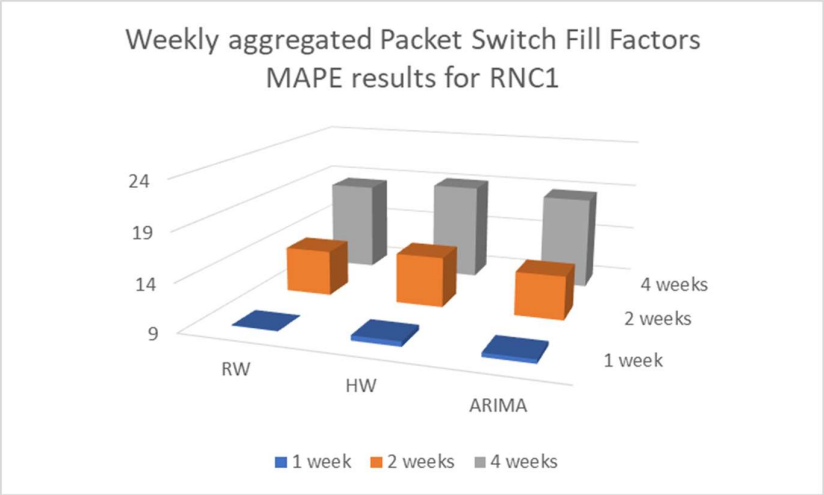


Figure 98: RNC1 MAPE results. Source: The author

Despite the MAPE, the quality of the forecast performance in the RNC1 can be confirmed by the plot represented in Figure 99. In the plot is possible to see an almost perfect adjustment of forecast value determined by one of each implemented models with the real values.

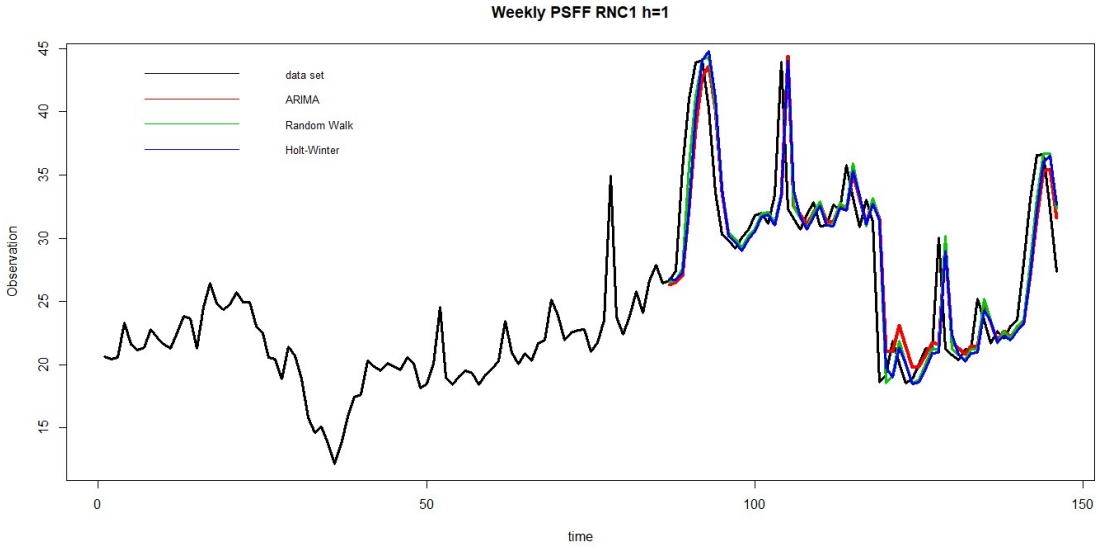


Figure 99: RNC1 weekly Aggregated Packet Switch Fill Factors Forecast. Source: The author

For RNC2, the best candidate model was the ARIMA(1,0,1), with an AIC of 546.06 and an AICc of 546.35. For horizon of 1 week, similar to RNC there is almost difference between the MAPE results.

Table 34:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC2. Source: The author

RNC2			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	1.767145	1.025211	4.835144
Holt Winters	1.759095	1.0115	4.777243
ARIMA	1.736717	1.049176	4.932306
h = 2			
Random Walk	2.524408	1.552946	7.517301
Holt Winters	2.500784	1.517627	7.365299
ARIMA	2.431432	1.592833	7.648459
h=4			
Random Walk	3.391078	2.274842	11.23192
Holt Winters	3.32503	2.163333	10.76843
ARIMA	3.08404	2.184394	10.52442

For 2 steps ahead forecast ARIMA's MAPE is slight higher compared to other two models. For other horizons the difference is practically insignificant. The behaviour of the methods is represented in Figure 100.

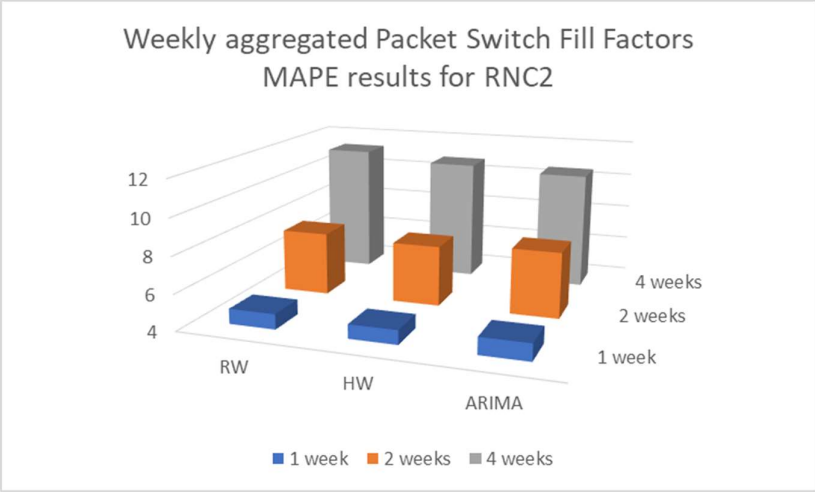


Figure 100: RNC2 MAPE results Source: The author

Similar to previous RNC the forecast values is an almost perfect adjustment for one of each implemented models. In period 120 for Arima and Holt-Winter there is a big difference from the real values and the forecast values, due the fact of those method being slow to recover from abrupt changes, as said.

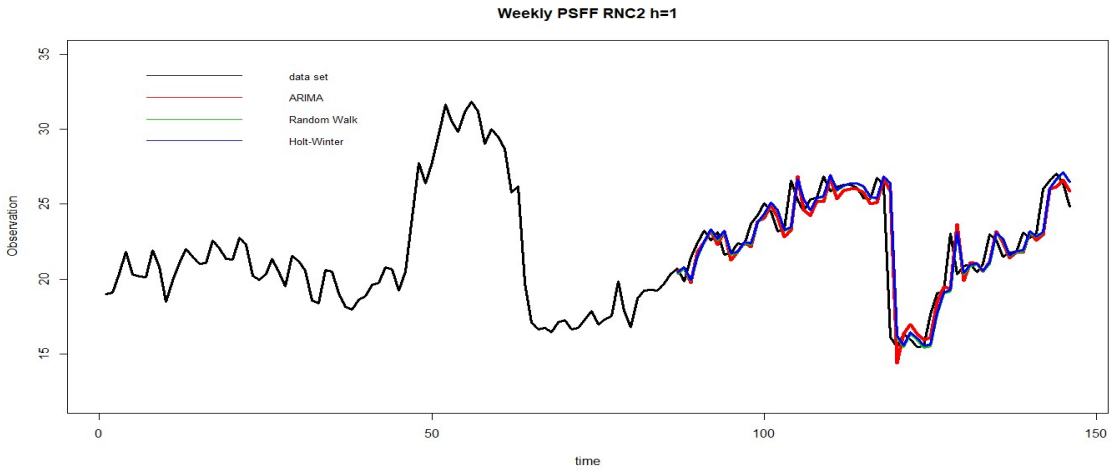


Figure 101: RNC2 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author

Applying the ARIMA model to RNC3, the best candidate was ARIMA(1,1,2), with an AIC of 637.04 and an AICc of 637.33. For this RNC, the models MAPE does not increase to much with forecast horizon.

Table 35:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC3 Source: The author

RNC3			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	1.749566	1.425154	3.215531
Holt Winters	1.768263	1.439285	3.246488
ARIMA	1.845183	1.503298	3.372777
h = 2			
Random Walk	2.216924	1.803079	4.025212
Holt Winters	2.259481	1.84161	4.114356
ARIMA	2.341517	1.909081	4.252762
h=4			
Random Walk	2.132381	1.724173	3.897375
Holt Winters	2.216924	1.803079	4.025212
ARIMA	2.413122	1.982221	4.440416

The forecast horizon MAPE for RNC3 show minor difference from one model to another. The behaviour of the methods is represented in Figure 102.

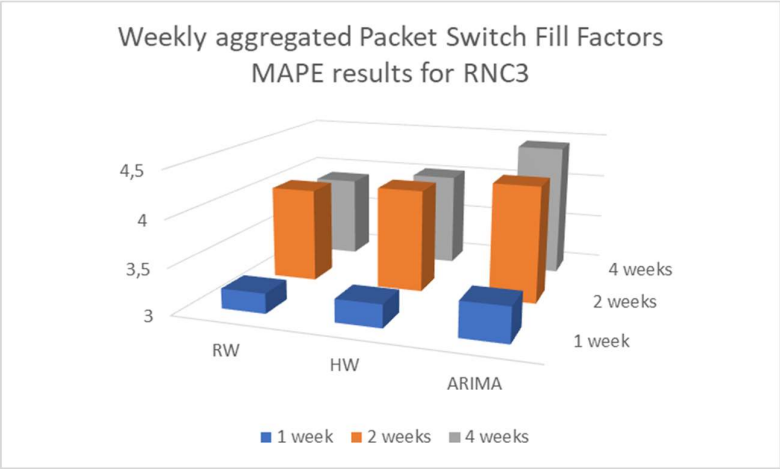


Figure 102: RNC3 MAPE results Source: The author

The horizon of 1 week has an almost perfect adjustment to the original values, as in Figure 103.

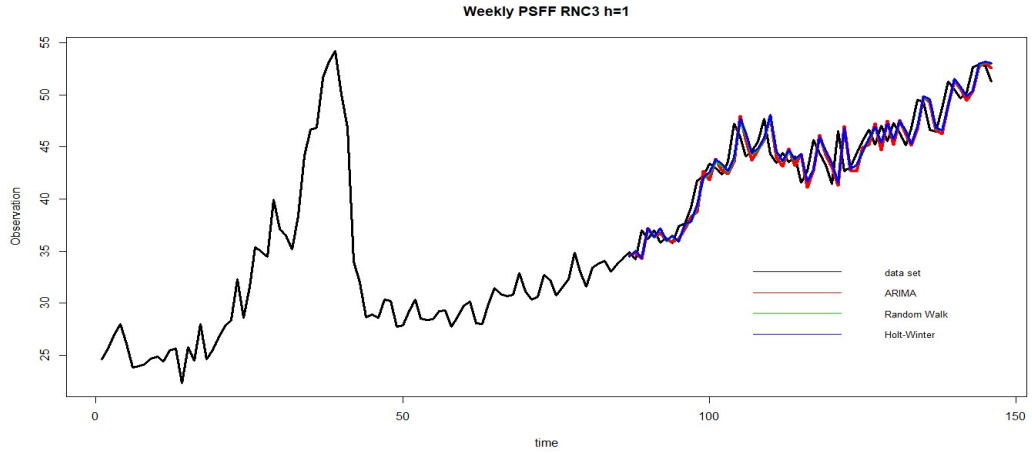


Figure 103: RNC3 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author

For RNC4, the best candidate model were the ARIMA(2,1,0), with an AIC of 652.46 and an AICc of 652.63.

Table 36:Accuracy measures for weekly aggregated Packet Switch Fill Factors for RNC4. Source: The author

RNC4			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	3.046666	1.751966	5.806886
Holt Winters	3.066027	1.772766	5.861059
ARIMA	2.988895	1.6492	5.442864
h = 2			
Random Walk	3.978309	2.475652	8.443125
Holt Winters	3.993547	2.524555	8.597688
ARIMA	3.82246	2.309399	7.881677
h=4			
Random Walk	4.815287	3.247821	11.49087
Holt Winters	4.924649	3.400349	12.01468
ARIMA	4.583799	3.042654	10.8024

The MAPE behaviour for the three models is represented in Figure 104. For horizon of 1 week the MAPE of the three models is similar and for horizon of 2 and 4 weeks, Arima models presents better results.

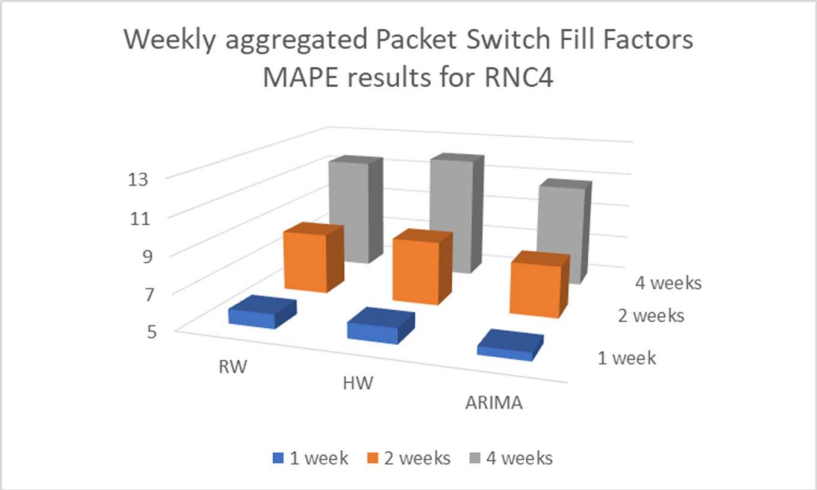


Figure 104: RNC4 MAPE results. Source: The author

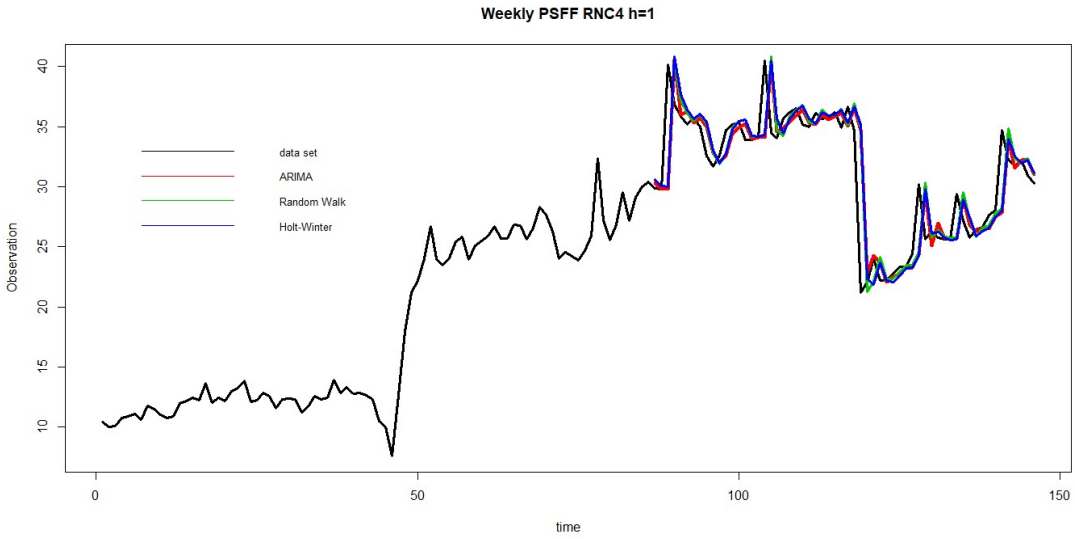


Figure 105: RNC4 weekly aggregated Packet Switch Fill Factors Forecast. Source: The author

Daily Circuit Switch fill factor weekly data are shown in Figure 106. Similar to Packet Switch fill factor data, the daily data plots indicate no presence of weekly seasonal component which can be confirmed by the ACF plot represented in Figure 107. In the plot, there are large and gradual variations in the traffic pattern. This behaviour was verified in the Packet Switch analysis, and is due to the redistribution of traffic among the RNCs.

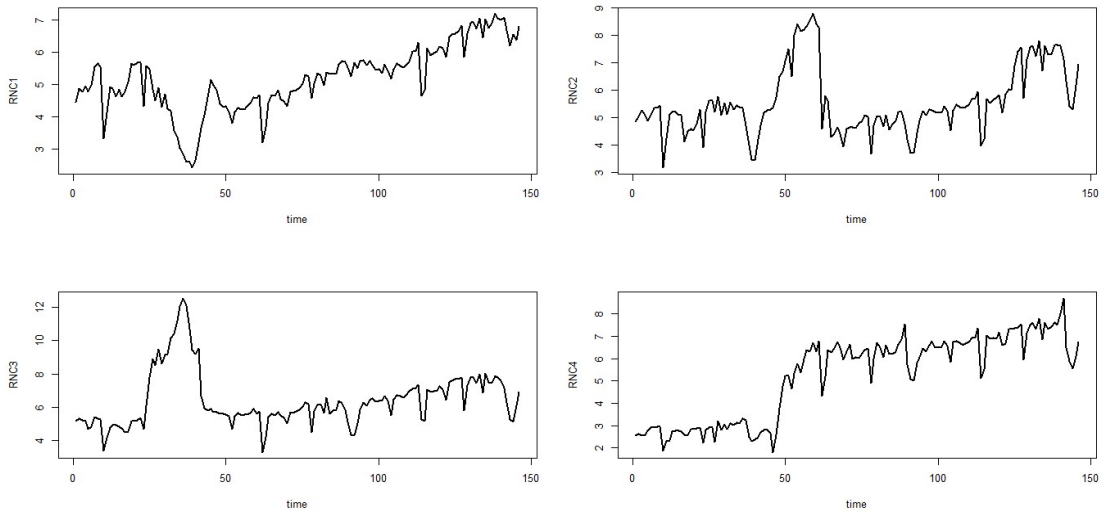


Figure 106: RNC weekly aggregated Circuit Switch fill factor Source: The author

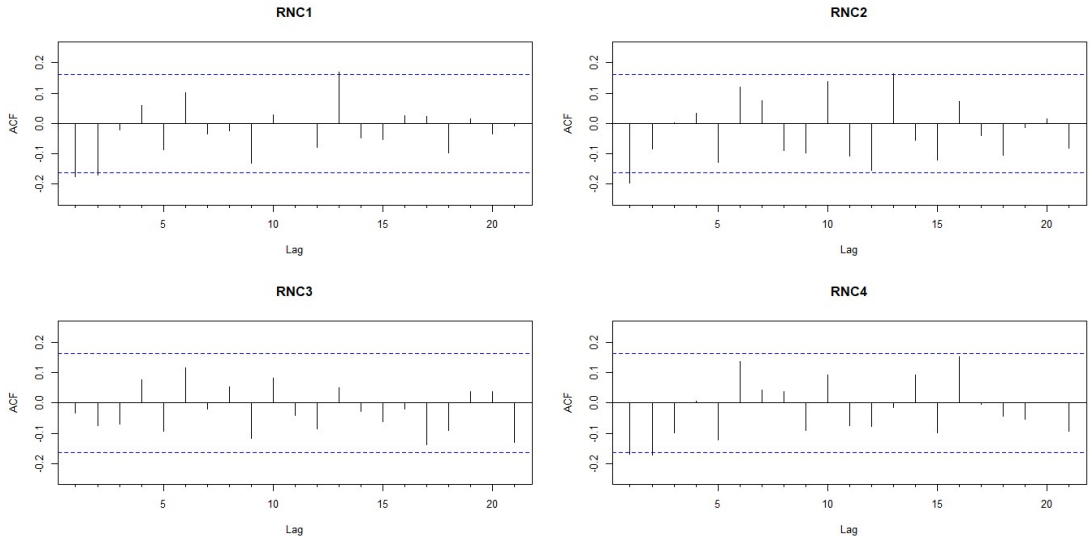


Figure 107: First difference ACF weekly aggregated Circuit Switch fill factors Source: The author

For RNC1, the best ARIMA candidate model was ARIMA(1,1,1), with an AIC of 171.37 and an AICc of 171.54. From the table, for horizon of 1 weeks the is not difference between the models. From horizon of 2 week to horizons of 2 and 4 weeks there is an increase of 1 % which means that the error does not increase abruptly with time.

Table 37:Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC1. Source: The author

RNC1			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	0.4003563	0.2681974	4.523618
Holt Winters	0.3986929	0.2572627	4.349268
ARIMA	0.3825955	0.2675251	4.468655
h = 2			
Random Walk	0.4796465	0.3221739	5.506883
Holt Winters	0.4896373	0.3250218	5.575995
ARIMA	0.4527375	0.3372511	5.700323
h=4			
Random Walk	0.4919648	0.3616223	6.08378
Holt Winters	0.5343246	0.3940799	6.657663
ARIMA	0.4935547	0.4139602	6.845679

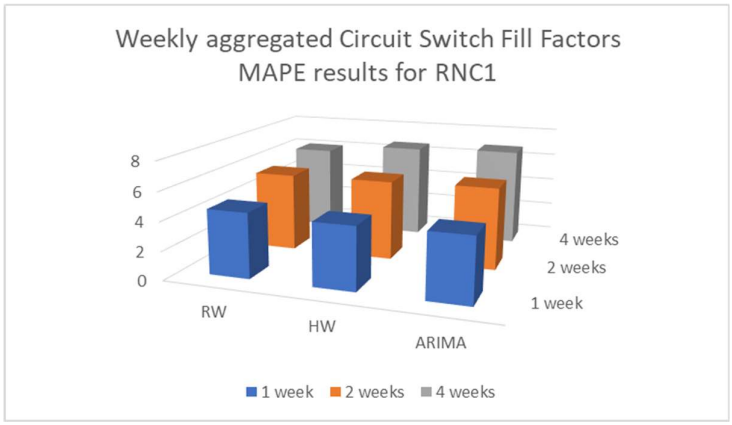


Figure 108: RNC1 MAPE results. Source: The author

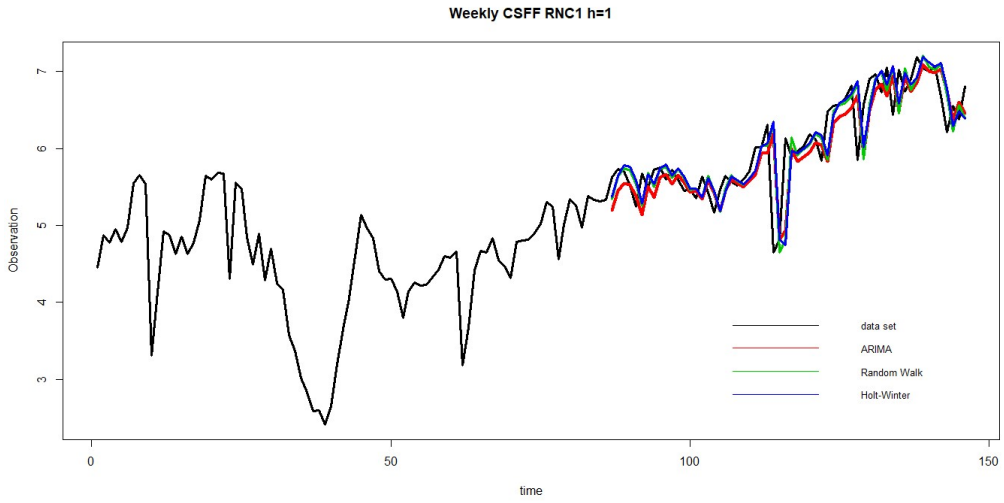


Figure 109: RNC1 weekly aggregated Circuit Switch Fill Factors Forecast.Source: The author

For the RNC2, the best candidate model was ARIMA(0,1,1), with an AIC of 289.31 and an AICc of 289.4. Initially the ARIMA MAPE suggest a slight poor MAPE for 1 step ahead forecast horizons but as the horizons are increased the MAPE becomes better outperforming the other two models.

Table 38:Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC2. Source: The author

RNC2			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	0.6162607	0.4206756	7.494132
Holt Winters	0.6158817	0.4353886	7.817694
ARIMA	0.6097473	0.4399272	7.887177
h = 2			
Random Walk	0.823285	0.6082964	11.03999
Holt Winters	0.8071678	0.6003567	11.00648
ARIMA	0.7865351	0.591651	10.81461
h=4			
Random Walk	0.934397	0.6722383	12.33352
Holt Winters	0.9253773	0.6682177	12.30553
ARIMA	0.8832321	0.6569835	11.99718

The forecast values fit the real value very accurately and the MAPE results for h=1 are equall for the thre methods. For h=2 and h=4, ARIMA has the lowest MAPE value, but the difference to other is low.

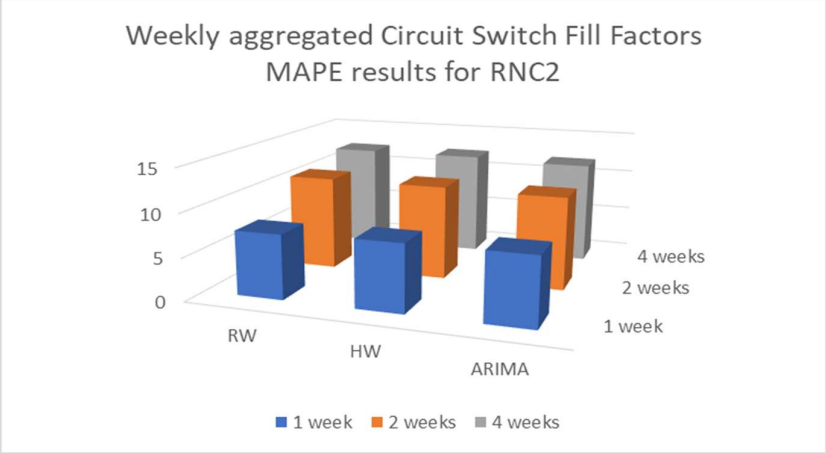


Figure 110: RNC2 MAPE results. Source: The author

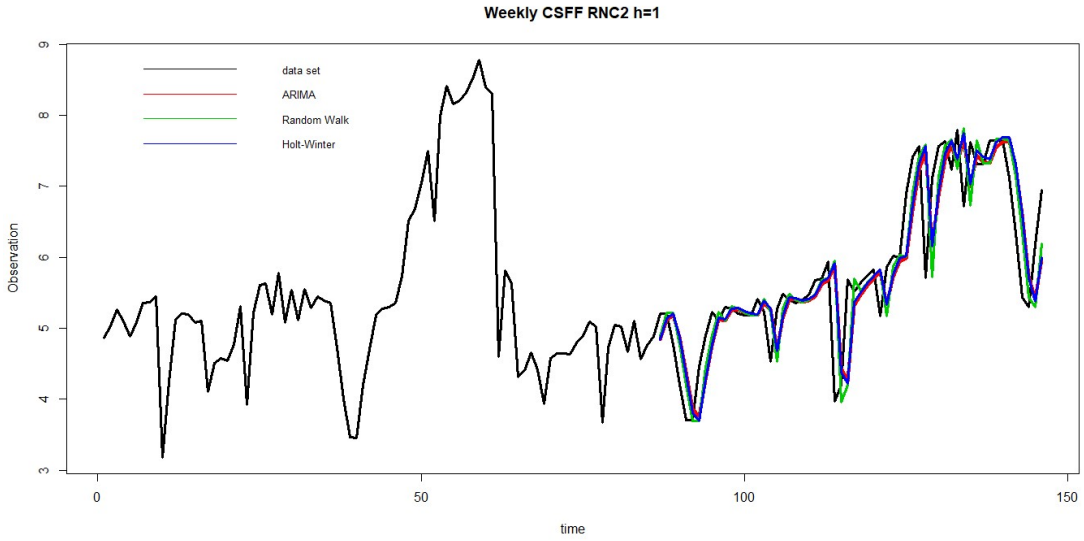


Figure 111: RNC2 weekly aggregated Circuit Switch Fill Factors Forecast Source: The author

For RNC3, the best candidate model was ARIMA(1,1,1), with an AIC of 315.37 and an AICc of 315.54.

Table 39: Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC3

RNC3			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	0.6909901	0.4735285	7.511735
Holt Winters	0.6957601	0.4766038	7.59448
ARIMA	0.6852371	0.4727998	7.488869
h = 2			
Random Walk	0.9293115	0.6776589	11.0976
Holt Winters	0.9422683	0.6738465	11.1247
ARIMA	0.9213209	0.6761906	11.04249
h=4			
Random Walk	1.074329	0.7902578	13.17784
Holt Winters	1.109351	0.8058158	13.58037
ARIMA	1.049185	0.7695041	12.82177

This RNC MAPE result has the same behaviour as RNC1 and RNC2 for h=1. ARIMA model has better results than the other models.

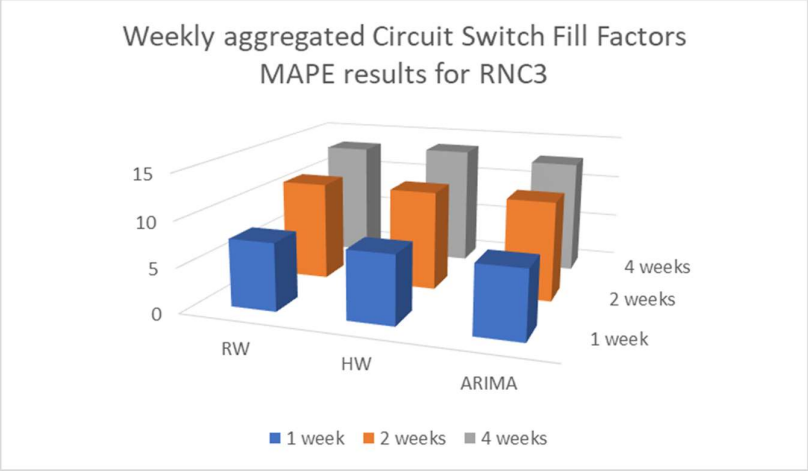


Figure 112: RNC3 MAPE results. Source: The author

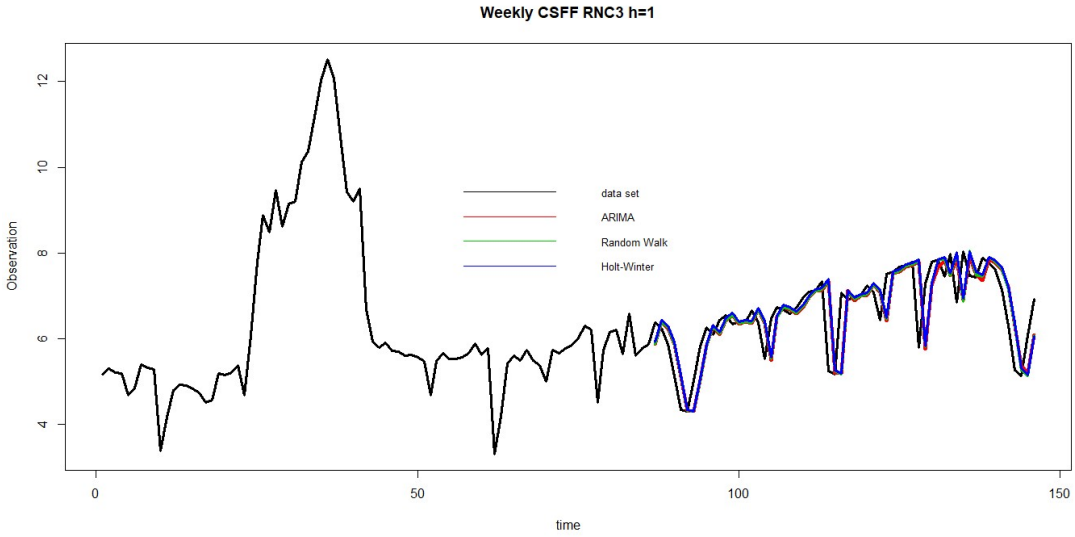


Figure 113: RNC3 weekly aggregated Circuit Switch Fill Factors Forecast Source: The author

For RNC4, the best candidate model was ARIMA(1,1,1), with an of AIC de 255.26 and an AICc of 255.43.

Table 40: Accuracy measures for weekly aggregated Circuit Switch Fill Factors for RNC4

RNC4			
Model	**RSME**	**MAE**	**MAPE**
h = 1			
Random Walk	0.6882756	0.4490945	7.014258
Holt Winters	0.687518	0.4607332	7.323478
ARIMA	0.6433285	0.4469067	6.978758
h = 2			
Random Walk	0.9014466	0.6228519	9.910453
Holt Winters	0.8693971	0.5910228	9.574023
ARIMA	0.7910602	0.5791512	9.181664
h=4			
Random Walk	0.9871433	0.6992996	11.24028
Holt Winters	0.9528188	0.7005873	11.34307
ARIMA	0.8099895	0.6154182	9.748822

For RNC4, ARIMA has the best MAPE results for all forecast horizons. The adjustment of the forecast values to the real values is almost perfect without significative peaks.

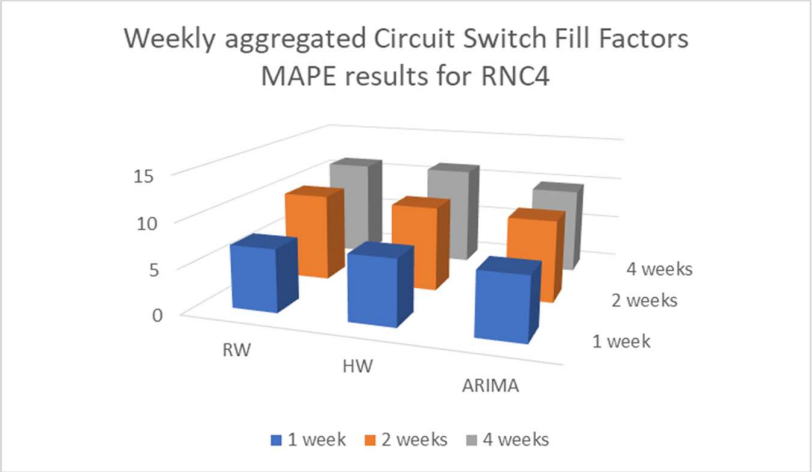


Figure 114: RNC4 MAPE results. Source: The author

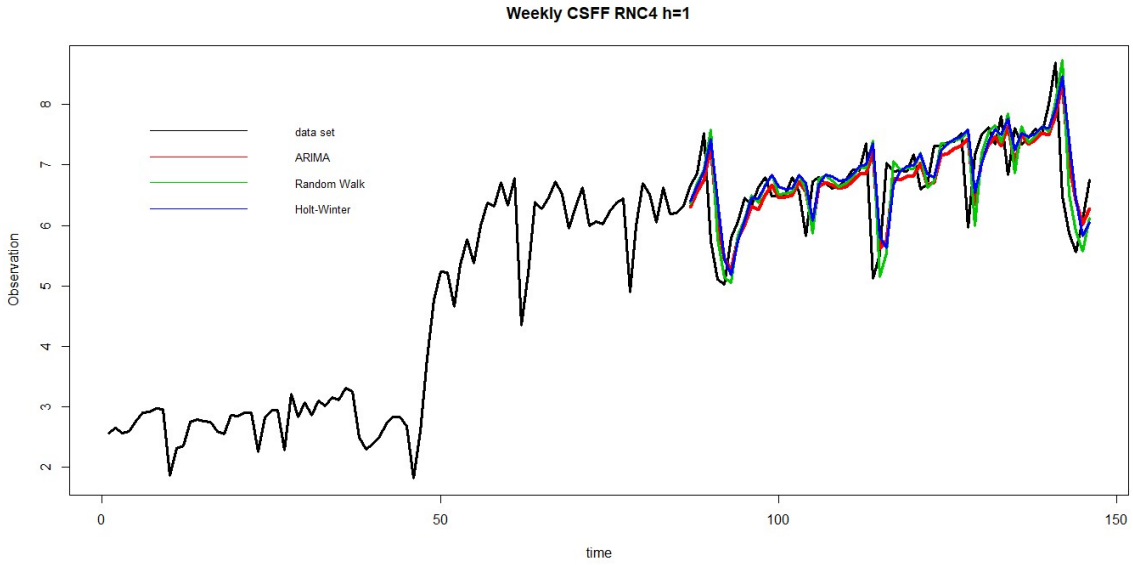


Figure 115: RNC4 weekly aggregated Circuit Switch Fill Factors Forecast. Source: The author

5 Forecasting Analysis Tool

This section describes the main functionalities of a support tool in the analysis and data forecasting. Although it is not part of the objectives set at the beginning of the development of this work, the idea of developing a forecast analysis tool arises as a consequence of the work carried out and the difficulty found in the modeling part of the ARIMA. Though not yet completed, the purpose of the tool is to facilitate and invigorate the modeling process.

5.1 Forecasting Analysis Tool Architecture

A view of the tool architecture is represented in the Figure 116. The tool architecture is divided into three modules:

- **User Interface:** this component is where a user can do the analysis and choose the models to determine forecasts. The interactive user interface is provided by the shiny framework;
- **Forecast Engine:** contain R forecasting libraries and created functions used to determine user forecasts;
- **Data store:** is the component where the data to be analyzed is located;

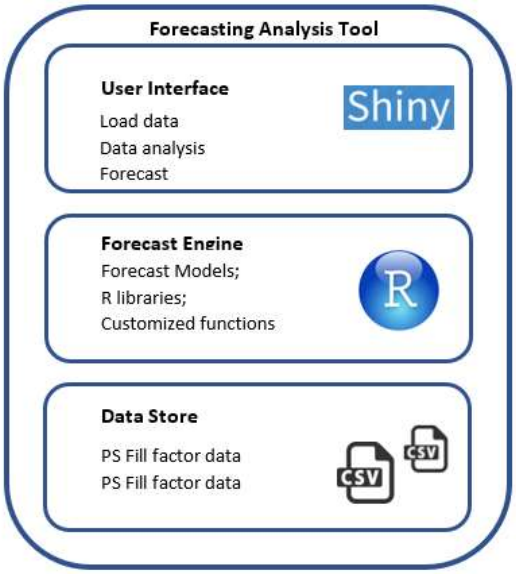


Figure 116: Forecasting Analysis Tool Architecture. Source: The author

5.2 Forecasting Analysis Tool Description

The shiny framework was used to develop the tool. It is an interactive web applications framework that allows to build apps straight from R. The application presents 3 main menus: data analysis, forecast analysis and tutorial. However, it should be noted that some of the features of the application are not complete.

The data analysis menu represents the component of the application reserved to familiarize the analyst with the data. This component allows to load data to the tool and visualize it in a table or in scatter and line charts. An example of this component functionality is represented by Figure 117.

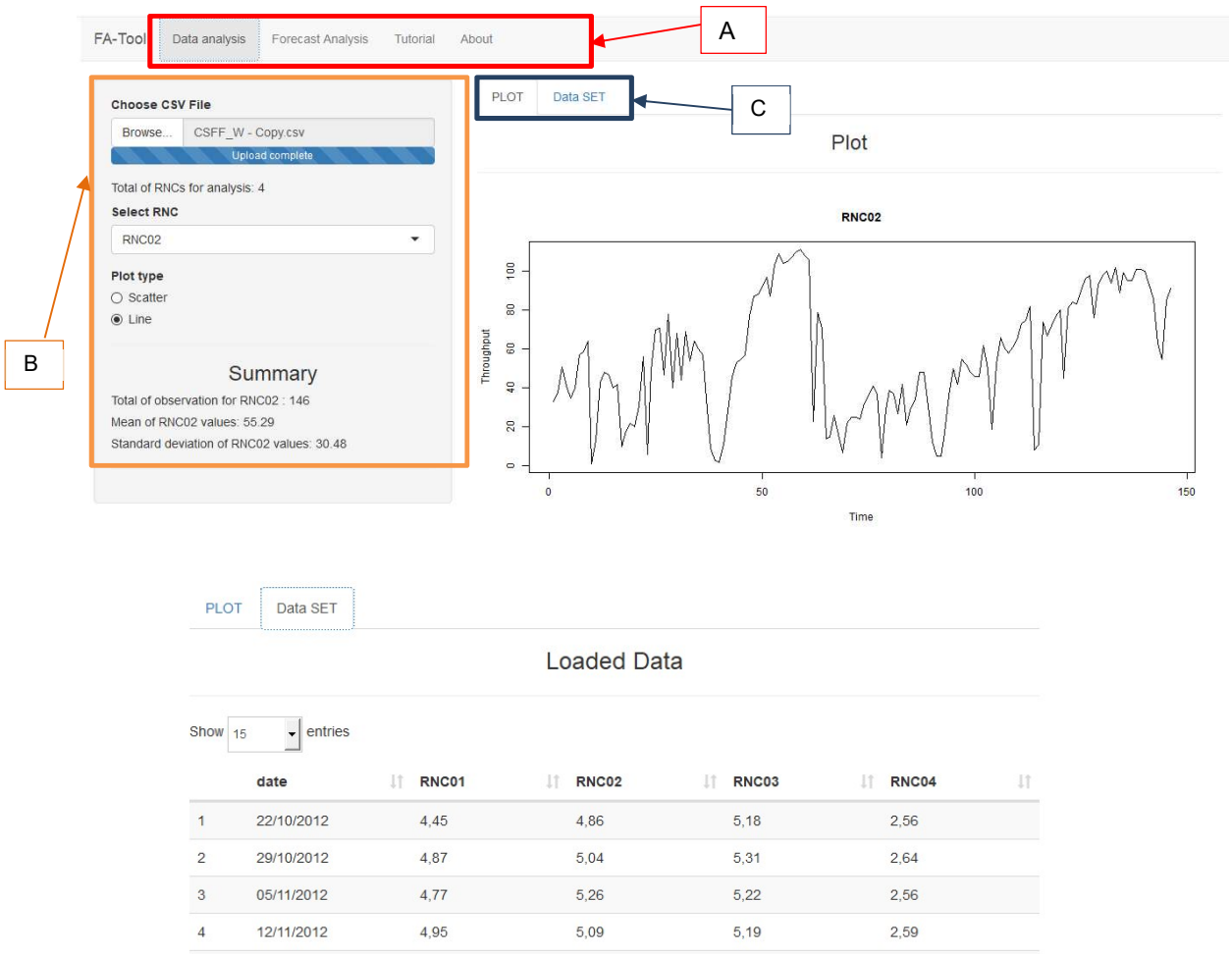


Figure 117:Data Analysis component. Source: The author

Figure 117 top, gives an overview of the main functions of data analysis components. Item “A” represents all the menu available in the tool. The Tutorial menu provides a little explanation about ARIMA modelling and how to use the tool. And the about give information about the developer and other additional information. Both menus are not implemented.

Item “B” allow to load data to be analyzed by the tool and a total of RNC data present in the loaded file. In item “B” the user can select which RNC to analyze and the type of plot to display the data. Also, a summary of information about the data. In item “C”, is possible to choose between plotting the data or display it in a table, Figure 117 bottom.

In the forecast analysis component allows to analyze the forecasts of the data. It is represented by Figure 118. In item D, is represented the menu responsible for the forecast analysis. It allows to choose the forecast model, in the case of the figure, the ARIMA model were chosen. Also, it allows to make the stationarity check to the data and take differences to analyze both ACF and PACF plots.

In case the data show seasonal behavior, the user can choose and provide the parameter into the input fields. For this example, the top figure shows a case of a non-seasonal option. In the case of a seasonal, as said, the user has also to provide the parameters. Figure 118 bottom shows a case of a seasonal ARIMA model. For both models is possible to adjust the forecast horizon.

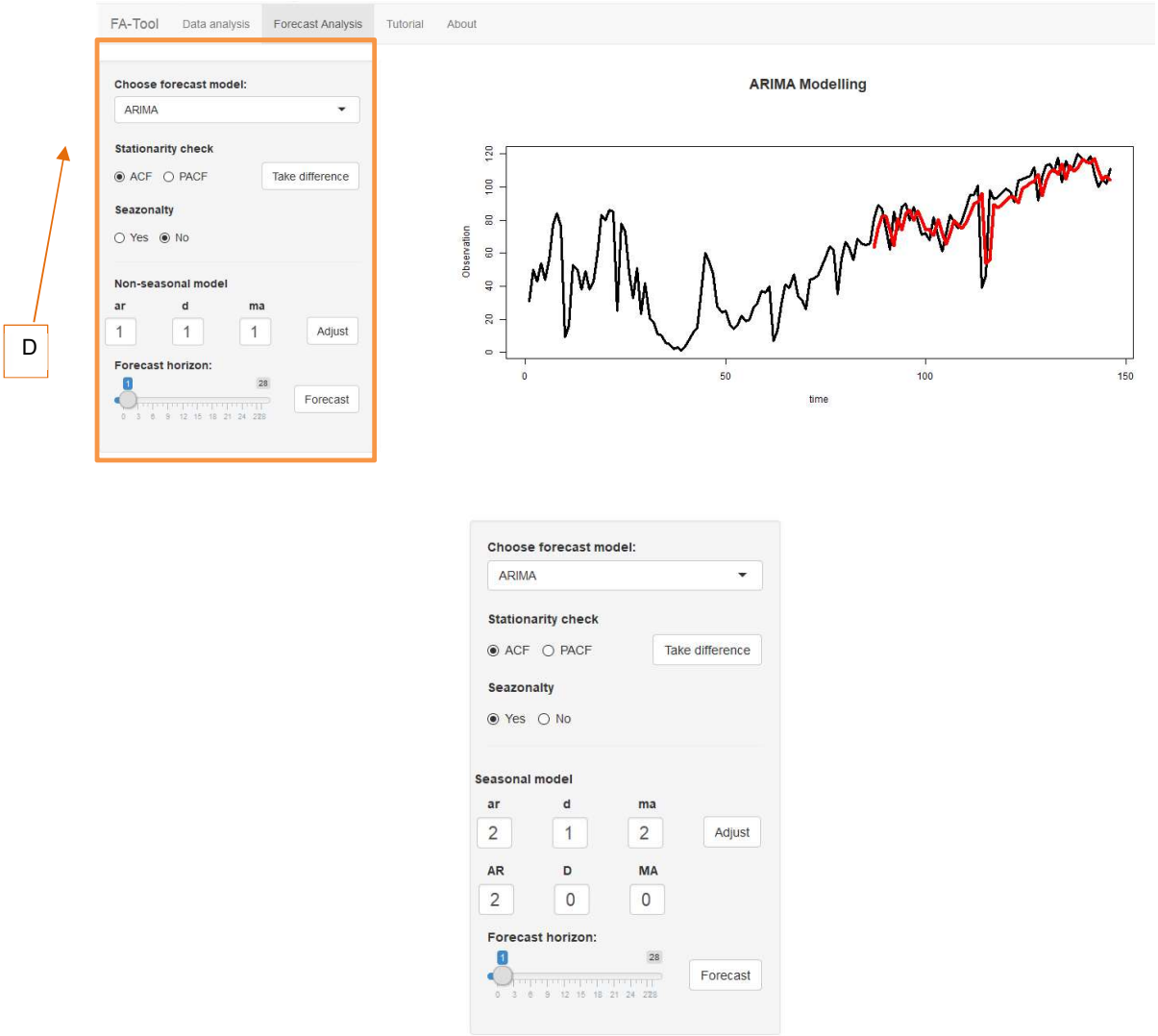


Figure 118:Forecast Analysis. Source: The author

6 Conclusion

The information and communication technology area has been developing a lot since its creation, also in this context the mobile telecommunications. With the creation of the Internet this growth has been even more accentuated. This may create some difficulties on the part of mobile operators such as: poor monitoring of traffic demand growth or maximizing resources use. The use of forecasts can contribute to the prevention of these problems. Therefore, it was the main objective of this work to study forecast models, their procedures and applications in mobile cellular networks data.

To accomplish the afore mentioned objective, as seen in chapter 1, a study on the precedents involved in the area of mobile cellular networks was first carried out in order to create a framework base. Aspects such as the generations of existing mobile cellular networks, their technologies and architecture have been addressed. This aspect allowed to perceive the location in the network of the components where CS and PS traffic readings are made. Subsequently, the process involved in the collection of performance indicators and the main categories of performance indicators was also extremely relevant.

Being the study on data forecast, were addressed concepts involved in data forecasting process in chapter 3. Fundamentally, quantitative models that use historical data for forecast, such as ARIMA, and exponential models, among others, were studied. Subsequently, small applications and demonstrations using these, and other models were tested, which contributed to create some sensitivity in the identification of probable behaviors resulting from the application of each model.

The case study in chapter 4 presents a practical application of the models studied to a communication service provider data. This data represents a real case CS and PS daily and weekly traffic. To this data, three models were applied, ARIMA, Random Walk and Holt-Winters, and forecasts were made for different time horizons. To the forecast, an accuracy test was performed, where MAPE, RMSE and MAE were verified. Due to the advantages of MAPE, which includes the fact of be scale independent, which makes it possible to compare performance between different data sets, it was used as the main performance criterion among the applied methods.

The MAPE results showed that the differences in quality between the horizons forecasts of weekly PS and CS fill factor data of the RNC in analysis when applying the three methods are not statistically significant. Specifically, due to practically nonexistence of differences between the MAPE of the different models, can be concluded that Random Walk, Holt-winters and ARIMA can be applied to data without seasonality and with a non-accentuated trend.

For the daily PS and CS fill factor data there are some variations in the accuracy of the implemented methods. These variations in the values of the MAPE in turn also change with the horizon to be used. Despite this, the MAPE values do not present significant differences. Compared to the daily PS and CS fill factor forecast, the weekly PS and CS fill factor forecast MAPE is much better.

In chapter 5, presents a tool developed to assist in the forecasting of data. The idea of developing a forecast analysis tool arises as a consequence of the work carried out and the difficulty found in the modeling the ARIMA. Though not yet completed, the purpose of the tool is to facilitate and invigorate the modeling process.

Although telecommunication data are short-lived, since most communication service providers do not record historical data, some limitations may exist in the forecast of annual seasonality. However, in general, the results of applying the models to the data with weekly seasonality show that is possible to apply the models to annual seasonal data.

More than the case study results, the fundamental contribution of this work was to show that it is possible to use statistical models to forecast data and analyse those model's accuracy to provide useful information for communication service provider decision-making. Also, it proves the possibility of using based historical information to forecast future time series behaviour.

However, it should be taken into account that although the work focuses were on mobile cellular networks, and has been applied specifically to the data of a 3G mobile network, the procedures and models applied here can certainly be applied to other different scenarios, such as forecast of latency in information systems.

6.1 Future Work

The proposal of this work consisted mainly in the use of forecasting models to prevent capacity problems in mobile networks. Being the area of forecasting in general and mobile data forecasting, very vast. It would be impossible to consider this work as finalized in terms of scientific research. Thereby, some important ideas that can be further developed and implemented in the future, are summarized here.

The time series forecast for telecommunication determined here was limited to a certain set of methods and models. However, the forecast in general presents several models and methods. Thereby, it would be interesting to investigate the hypothesis of implementing other methods and models for forecasting data in mobile networks, as well as the use of different types of KPI other than those used, to carry out the forecast of other parts of a network.

Given the limitations of automation in the time series forecasting process, such as the auto.arima () function, it is also future interest of this work, to improve the automation of these processes and to include the monitoring component in them.

Another aspect that should be further studied is the integration of the collection process, processing and transformation of obtained data and application of forecasting methods in real time or near real time. In a general, this work could be used as a base for other studies besides mobile network capacity such as: information systems dimensioning, forecasts on data server latency, etc.

References

[1] M. Sokele, “Analytical Method For Forecasting Of Telecommunications Service Life-Cycle Quantitative Factors Analitički Postupak Predviđanja Kvantitativnih Čimbenika Životnog Vijeka Telekomunikacijske Usluge,” 2009.

[2] International Telecommunication Union, “Models for forecasting international traffic,” *Blue Book*, vol. E.507. ITU, 1993.

[3] P. Sharma, “Evolution of Mobile Wireless Communication Networks-1G to 5G as well as Future Prospective of Next Generation Communication Network,” *Int. J. Comput. Sci. Mob. Comput. - not index*, vol. 2, no. August, pp. 47–53, 2013.

[4] QUALCOMM, “The Evolution of Mobile Technologies: 1G 2G 3G 4G LTE,” 2014. [Online]. Available: https://www.qualcomm.com/media/documents/files/the-evolution-of-mobile-technologies-1g-to-2g-to-3g-to-4g-lte.pdf. [Accessed: 25-Aug-2017].

[5] H. Kaaranen, “UMTS Networks : Architecture, Mobility and Services,” *John Wiley Sons*, p. 401, 2005.

[6] M. D. O. Duarte, D. Teixeira, B. D. O. Cruz, and J. P. V. Dias, “Cellular Telecommunications network: planning, dimensioning and economics,” Aveiro, 2015.

[7] J. Korhonen, *Introduction to 3G mobile communications*, Second edi. London: Artech House, 2003.

[8] The Statistics Portal, “Mobile technologies market share of subscriptions worldwide from 2016 to 2021,” *Statista*, 2017. [Online]. Available: https://www.statista.com/statistics/206655/forecast-of-the-distribution-of-global-mobile-broadband-subscriptions-by-technology-in-2016/. [Accessed: 25-Aug-2017].

[9] P. Nicopolitidis, M. S. Obaidat, G. I. Papadimitriou, and A. S. Pomportsis, *Wireless Networks*. Chichester, UK: John Wiley & Sons, Ltd, 2002.

[10] D. A. Gratton, *Developing practical wireless applications*. Elsevier Digital Press, 2007.

[11] V. Pereira and T. Sousa, “Evolution of Mobile Communications: from 1G to 4G,” *Dep. Informatics Eng. Univ. Coimbra*, p. 7, 2004.

[12] T.-T. Tran, Y. Shin, O.-S. Shin, and etall, “Overview of enabling technologies for 3GPP LTE-advanced,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2012, no. 1, p. 54, 2012.

[13] M. Nohrborg, “LTE.” [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/98-lte. [Accessed: 04-Jan-2017].

[14] Unknown, “Multiple Access Techniques,” 2017. [Online]. Available: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTL4c1GCvCWApD3bL6JoSVuPMCJ5tklerVa0ryu9pvLVXHWyyYcLw.

[15] Y. Ceita, *Shared Solutions for Telecommunication Network*. Universidade de Aveiro, 2017.

[16] M. Sauter and Wiley InterScience (Online service), *Communication systems for the mobile information society*. John Wiley, 2006.

[17] C. Kappler, *UMTS networks and beyond*. John Wiley & Sons, 2009.

[18] 3GPP, “Release 4,” 2017. [Online]. Available: http://www.3gpp.org/specifications/releases/76-release-4. [Accessed: 05-Sep-2017].

[19] R. Kreher and T. Rüdebusch, “UMTS Basics,” in *UMTS Signaling*, I. Tektronix, Ed. West Sussex: John Wiley & Sons, Inc., 2007, p. 146.

[20] J. HERNANDEZ, “Mobile Unified Communications Network Architecture – an Operator

Perspective," 2015. [Online]. Available: http://www.trefor.net/2015/09/14/mobile-unified-communications-network-architecture/. [Accessed: 16-Sep-2017].

[21] G. Camarillo and M. A. García-Martín, *The 3G IP Multimedia Subsystem (IMS) The 3G IP Multimedia Subsystem (IMS) Merging the Internet and the Cellular Worlds*, Second Edi. West Sussex, 2006.

[22] D. Fox and Wiley InterScience (Online service), *Testing UMTS : assuring conformance and quality of UMTS user equipment*. John Wiley, 2008.

[23] M. D. F. Montenegro, "Capacity forecasting for radio access networks," Universidade de Aveiro, Aveiro, 2015.

[24] C. Johnson, *Radio access networks for UMTS : principles and practice*. John Wiley & Sons, 2008.

[25] I. Poole, "LTE Long Term Evolution Tutorial & Basics." [Online]. Available: http://www.radio-electronics.com/info/cellulartelecomms/lte-long-term-evolution/3g-lte-basics.php. [Accessed: 09-Oct-2017].

[26] Alcatel-Lucent, "The LTE Network Architecture A comprehensive tutorial," unknow, 2013.

[27] J. Reunanen, J. Salo, and R. Luostari, "LTE Key Performance Indicator Optimization," in *LTE Small Cell Optimization*, Chichester, UK: John Wiley & Sons Ltd, 2015, pp. 195–248.

[28] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes, "The evolution to 4G cellular systems: LTE-Advanced," *Phys. Commun.*, vol. 3, pp. 217–244, 2010.

[29] Frédéric Firmin and 3GPP MCC, "The Evolved Packet Core." [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/100-the-evolved-packet-core.

[30] P. Lescuyer and T. Lucidarme, *Evolved packet system (EPS) : the LTE and SAE evolution of 3G UMTS*. J. Wiley & Sons, 2008.

[31] M. F. L. Abdullah and A. Z. Yonis, "Performance of LTE Release 8 and Release 10 in wireless communications," in *Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, 2012, pp. 236–241.

[32] F. Firmin and 3GPP, "NAS." [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/96-nas. [Accessed: 07-Jan-2017].

[33] E. Metsälä and J. Salmelin, *LTE backhaul : planning and optimization*. John Wiley & Sons, Ltd, 2016.

[34] J. Pullen, "The Guide to Modern OSS." [Online]. Available: http://www.donriver.com/document/TheGuidetoModernOSS.pdf. [Accessed: 29-Oct-2017].

[35] A. Capone, "Smartphones Analysis – Estimating Rnc Unit Load," MILANO, 2011.

[36] Huawei technologies, "Huawei Ran KPI for performance Management," unknow, 2006.

[37] I. Ray Khastur, LTE Optimization Consultant at P.I.Works, "KPI in LTE Radio Network (Huawei Based)," 2015. [Online]. Available: http://www.slideshare.net/RayKhastur/kpi-in-lte-radio-network-huawei-based. [Accessed: 09-Jan-2017].

[38] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper - Cisco," 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. [Accessed: 29-Aug-2017].

[39] X. Hu and J. Wu, "Traffic Forecasting Based on Chaos Analysis in GSM Communication Network," in *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, 2007, pp. 829–833.

[40] A. M. Álvarez-Socarrás, A. Berrones, G. J. Moreno, J. A. Rodríguez-Sarasty, and M. Cabrera-

Ríos, “Practice Summary: Enhancing Forecasting and Capacity Planning Capabilities in a Telecommunications Company,” *Interfaces (Providence).*, vol. 43, no. 4, pp. 385–387, Aug. 2013.

[41] R. J. Hyndman and G. Athanasopoulos, “Forecasting : principles and practice,” 2013. [Online]. Available: https://www.otexts.org/fpp/8/1. [Accessed: 11-Oct-2016].

[42] C. N. Konstantinopoulou, K. A. Koutsopoulos, G. L. Lyberopoulos, and M. E. Theologou, “Core network planning, optimization and forecasting in GSM/GPRS networks,” in *IEEE Benelux Chapter on Vehicular Technology and Communications. Symposium on Communications and Vehicular Technology. SCVT-2000. Proceedings (Cat. No.00EX465)*, pp. 55–61.

[43] A. Akbar Mulani, S. Muraraka, and K. Sujatha, “TELECOMMUNICATION DATA FORECASTING BASED ON ARIMA MODEL,” *Internation Journal of Etectrical and Electronics Engineers*, vol. 8, no. 2, p. 8, 2016.

[44] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, vol. 102. Springer New York, 2006.

[45] L. K. VANSTON and R. L. HODGES, “Technology forecasting for telecommunications,” *Telektronikk*, vol. 4, p. 12, 2004.

[46] J. S. Armstrong, *Principles of forecasting : a handbook for researchers and practitioners*. Kluwer Academic, 2001.

[47] S. Makridakis, R. J. Hyndman, and S. C. Wheelwright, *Forecasting : methods and applications*. Wiley, 1998.

[48] R. D. Reid and N. R. Sanders, *Operations management : an integrated approach*. John Wiley, 2010.

[49] E. H. Hassani, “Causal Method and Time Series Forecasting model based on Artificial Neural Network,” *Int. J. Comput. Appl.*, vol. 75, no. 7, pp. 975–8887, 2013.

[50] J. and E. W. S. M. P.F. Miller, “Introduction to Time Series Regression and Forecasting,” *University of Pennsylvania*. [Online]. Available: http://www.ssc.upenn.edu/~fdiebold/Teaching104/Ch14_slides.pdf. [Accessed: 09-Mar-2017].

[51] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. Wiley Series in Probability and Statistics, 2008.

[52] G. van de Ven, “Removal of Trend & Seasonality,” 2010. [Online]. Available: https://www.stat.berkeley.edu/~gido/Removal of Trend and Seasonality.pdf. [Accessed: 03-Mar-2017].

[53] J. Caiado, *Métodos de previsão em gestão : com aplicações em excel*, 2nd ed. Lisboa: Sílabo, 2016.

[54] Statistics Canada, “Seasonal adjustment and trend-cycle estimation,” 2015. [Online]. Available: http://www.statcan.gc.ca/pub/12-539-x/2009001/seasonal-saisonnal-eng.htm. [Accessed: 25-Apr-2017].

[55] R. M. Kunst, “Evaluating predictive accuracy,” vol. 1, no. 1, Department of Economics of the University of Vienna, 2004, pp. 55–63.

[56] R. J. Hyndman, “Errors on percentage errors,” *Hyndsight*, 2014. [Online]. Available: https://robjhyndman.com/hyndsight/smape/.

[57] S. Makridakis and M. Hibon, “The M3-Competition: results, conclusions and implications,” *Int. J. Forecast.*, vol. 16, pp. 451–476, 2000.

[58] W. Bank and I. T. Union, “Internet users by world region.” [Online]. Available:

https://ourworldindata.org/grapher/internet-users-by-world-region?overlay=sources. [Accessed: 22-Oct-2017].

[59] R. F. Nau, “Statistical forecasting: notes on regression and time series analysis.” [Online]. Available: https://people.duke.edu/~rnau/411avg.htm#SMA. [Accessed: 03-Apr-2017].

[60] J. F. Ehlers, *Rocket science for traders : digital signal processing applications*. Wiley, 2001.

[61] W. J. Stevenson, *Operations management*, 11th ed. Rochester Institute of Technology, 2012.

[62] Nist, “Single Exponential Smoothing.” [Online]. Available: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc431.htm. [Accessed: 11-Oct-2017].

[63] H. V Ravinder, “Forecasting With Exponential Smoothing – What’s The Right Smoothing Constant?,” *Rev. Bus. Inf. Syst. – Third Quart.*, vol. 17, no. 3, 2013.

[64] P. Gavin, “The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems.” Duke University, p. 18, 2016.

[65] The Pennsylvania State University, “5.2 Smoothing Time Series | STAT 510.” [Online]. Available: https://onlinecourses.science.psu.edu/stat510/node/70. [Accessed: 15-Mar-2017].

[66] R. J. Hyndman and Y. Khandakar, “Automatic Time Series Forecasting: The forecast Package for R,” *J. Stat. Softw.*, vol. 27, no. 3, pp. 1–22, 2008.

[67] R. Upadhyay, “Forecasting & Time Series Analysis – Manufacturing Case Study,” 2015. [Online]. Available: http://ucanalytics.com/blogs/forecasting-time-series-analysis-manufacturing-case-study-example-part-1/. [Accessed: 23-Feb-2017].

[68] “Extract Seasonal & Trend: using decomposition in R - Anomaly,” *anomaly*, 2015. [Online]. Available: https://anomaly.io/seasonal-trend-decomposition-in-r/. [Accessed: 01-Mar-2017].

[69] The Pennsylvania State University, “5.1 Decomposition Models | STAT 510.” [Online]. Available: https://onlinecourses.science.psu.edu/stat510/node/69. [Accessed: 26-Feb-2017].

[70] “Time Series Decomposition - MATLAB & Simulink.” [Online]. Available: https://www.mathworks.com/help/econ/detrending.html. [Accessed: 27-Feb-2017].

[71] R. B. Cleveland, W. S. Cleveland, and at all, “STL:A Seasonal-Trend Decomposition Procedure Based on Loess,” *J. Off. Stat.*, vol. 6, pp. 3–73, 1990.

[72] J. H. Stock and M. W. Watson, *Introduction to econometrics*, Third edit. Addison -Wesley.

[73] The Pennsylvania State University, “14.1 - Autoregressive Models | STAT 501,” 2016. [Online]. Available: https://onlinecourses.science.psu.edu/stat501/node/358.

[74] E. Engineering, “IIR filters what does infinite mean?” [Online]. Available: https://electronics.stackexchange.com/questions/15206/iir-filters-what-does-infinite-mean.

[75] N. S. Nalawade and M. M. Pawar, “Forecasting telecommunications data with Autoregressive Integrated Moving Average models,” in *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)*, 2015, pp. 1–6.

[76] A. Kasyoki, “Simple Steps for Fitting Arima Model to Time Series Data for Forecasting Using R,” *Int. J. Sci. Res. ISSN (Online Index Copernicus Value Impact Factor*, vol. 14, no. 3, pp. 2319–7064, 2013.

[77] The-MathWorks, “Residual Diagnostics.” [Online]. Available: https://www.mathworks.com/help/econ/residual-diagnostics.html#btb6mje. [Accessed: 03-Nov-2017].

[78] The Pennsylvania State University, “Applied Time Series Analysis.” [Online]. Available: https://onlinecourses.science.psu.edu/stat510/node/65. [Accessed: 03-Nov-2017].

[79] E. Mahdi and A. I. Mcleod, “Portmanteau Test Statistics,” *unknown*. Western University, 2014.

[80] Cisco, “Network Configuration Management.” [Online]. Available: https://www.cisco.com/en/US/technologies/tk869/tk769/technologies_white_paper0900aecd806c0d88.html.

[81] J. Ding, *Advances in network management*. CRC Press, 2010.

[82] C. Nuangjamnong, S. P. Maj, and D. Veal, “The OSI network management model- capacity and performance management,” in *2008 4th IEEE International Conference on Management of Innovation and Technology*, 2008, pp. 1266–1270.

[83] J. Sathyan, *Fundamentals of EMS, NMS, and OSS/BSS*. CRC Press, 2010.

[84] J. de R. S. Ramos, “Operator Metrics that Matter: From Network Metrics to Business Metrics!” [Online]. Available: http://www.metanoia-inc.com/blog/2012/05/14/operator-metrics-that-matter-from-network-metrics-to-business-metrics/.

[85] International Telecommunication Union, “Principles for a telecommunications management network-Recommendation M.3010,” unknow, 2000.

[86] Wikipedia, “Operations support system - Wikipedia,” 2017. [Online]. Available: https://en.wikipedia.org/wiki/Operations_support_system. [Accessed: 03-Oct-2017].

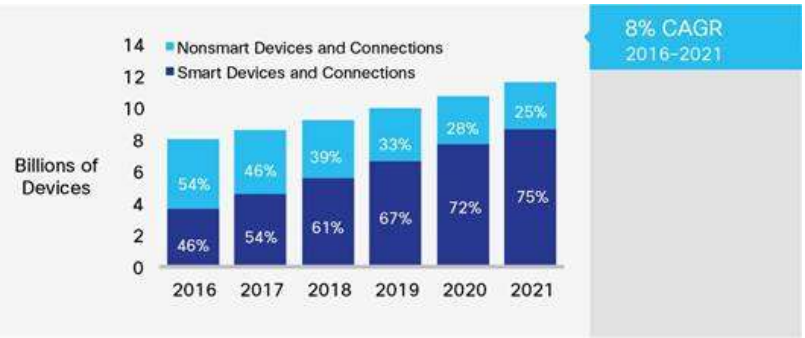
[87] K. Berquist and A. Berquist, *Managing information highways : the PRISM book : principles, methods, and case studies for designing telecommunications management systems*. Springer, 1996.

[88] S. Kasera, N. Narang, and S. Narang, *Communication networks : principles and practice*. McGraw-Hill, 2007.

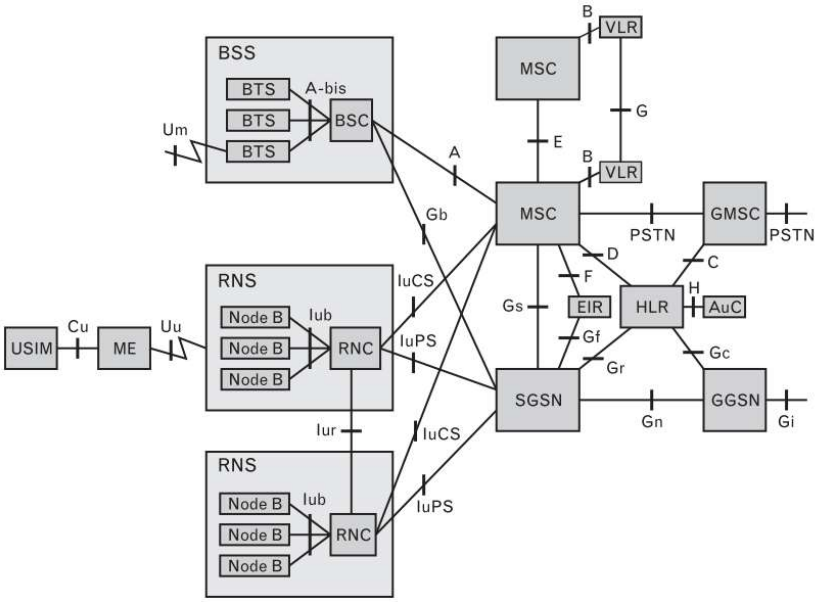
Appendix 1.1: Comparison of data speeds offered by various generations of mobile technology

Generation	Standard Variation Name	Maximum Theoretical Download Speed	Maximum Theoretical Upload Speed
2G	GSM	14.4 Kbits/s	14.4 Kbits/s
	GPRS	53.6 Kbits/s	26.8 Kbits/s
	EDGE	217.6 Kbits/s	108.8 Kbits/s
3G	UMTS	384 Kbits/s	128 Kbits/s
	HSPA	7.2 Mbits/s	3.6 Mbits/s
	HSPA+ (Evolved HSPA- Release 6)	14.4 Mbits/s	5.76 Mbits/s
	HSPA+ (Evolved HSPA- Release 7)	21.1 Mbits/s or 28.0 Mbits/s	11.5 Mbits/s
	HSPA+ (Evolved HSPA- Release 8)	42.2 Mbits/s	11.5 Mbits/s
	HSPA (Evolved HSPA- Release 9)	84.4 Mbits/s	11.5 Mbits/s
	HSPA+ (Evolved HSPA- Release 10)	168.8 Mbits/s	23.0 Mbits/s
4G	LTE	100 Mbits/s	50 Mbits/s
	LTE-A (LTE Advanced)	1 Gbits/s	500 Mbits/s

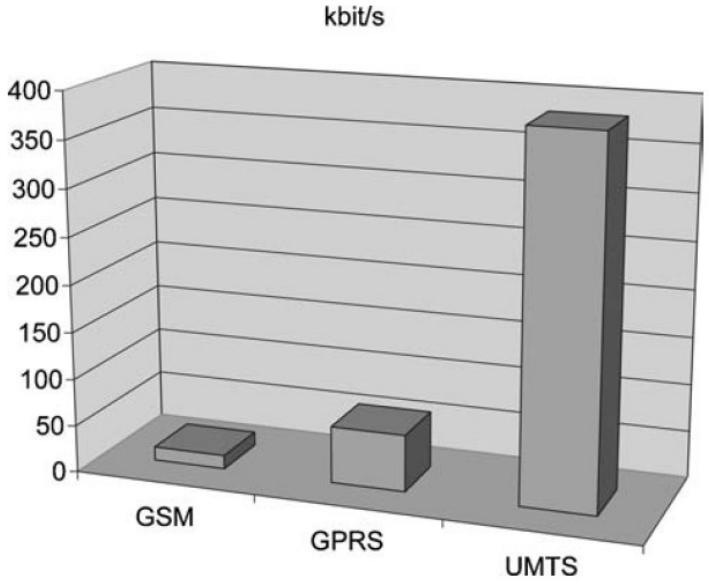
Appendix 1.2: Global Growth of Smart Mobile Devices and Connections



Appendix 1.3: UMTS network elements and interfaces [7].



Appendix 1.4: Mobile Generations Speed Evolution [16]



Appendix 1.5: Network Management Models

1.5.1 FCAPS Network Management

The ISO is the entity responsible for the creation of the Telecommunications Management Network model and framework for network management called FCAPS. It stands for an acronym for fault, configuration, accounting, performance, security which is used to define networks management responsibilities. The FCAPS network management model is also called OSI/ISO network management model.

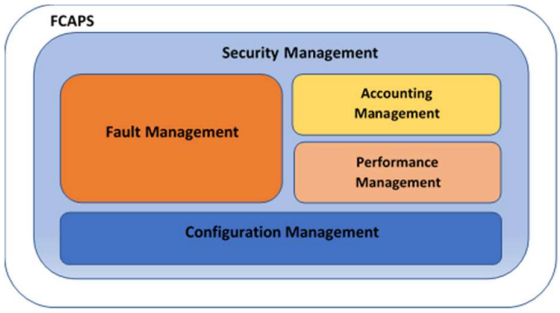


Figure 119: FCAPS architecture
Source: [80]

FCAPS model sections has five functional areas which are described as follows [81][82]:

A. Fault management

The fault management is used to identify, separate, fix and record the identified faults which has occurred. The fault management accomplish this by applying a trend analysis on the network to prevent errors of occurring, so the network stays always available. Being responsible for identifying issues within the network by constant and continues monitoring, the fault management has the following life cycle [83] :

- Fault and problem detection: captures indications of disorder or system mal function and determines the cause of failure in the network.
- Fault isolation: after the detection of a fault it is very important to find the element which causes the issue in the network than isolate it. Fault isolation can be divide into the following functions:
  - Handling and acknowledging alarms sent by devices;
  - Fault and problem isolation using a filtration and correlation process;
- Fault correction and recovery: repairs the damaged caused by a fault and also filter all redundant resources.

B. Configuration management

Consists on the process of initial configuration and adjustment to requirements changes of a network system. It continuously tracks critical network attributes from the network elements.

Configuration management is well known to facilitate the control of any system configuration both on the hardware and software side and the ability to track and log changes to device configurations. Some tasks of configuration management include the following items:

- Creation of controls;
- Monitoring and enforcement of hardware and software;

- Backup and restore of data configurations and creation of both log and report about configurations changes;

C. Accounting management

This area is mainly concerned about system users features such as charging and billing of service and also regulation of used service. It keeps track of usage statistics related to accounting such as link utilization, device resource, CPU and circuit utilization. Some of the main accounting management includes:

- Service tracking and resource usage underlying: used to determine charges of a given customer.
- Tariff: applies the correct charging rate on a given customer for the use of data and may also apply discounts.
- Usage restrictions: involves the creation of restriction measures to be applied to a particular user. Those restrictions include network bandwidth, disk space, and so forth.

D. Performance management

Performance management consists on guaranteeing acceptable network performance levels to which a network manager can dwell to determine the efficiency of the network in relation to investments applied. It gathers information about the network throughput, response time, percentage utilization reporting the behaviour and effectiveness of these features, so the manager can be prepared in the future to analyse the network and be able to prevent issues such as bottlenecks and near reaching capacity by altering the system modes of operation.

The information data that constitute performance management are collect on:

- NE bulk download: here the performance management data are collect at regular intervals of time and stored in folders on the NE.
- Send by NE on generation: a predefined protocol is used to send data to the management applications which is always listening for NE data.
- NE database queried: in here the management application collect from the Management information base all the performance and dynamic updated performance attributes stored data.

E. Security management

Security management consists on the process of supervising all assets access in a network and gather all necessary security related information for regular analysis. It is primarily concerned with controlling network devices access by using management network authentication, authorization, and auditing among other. And also includes the configuration and management of network firewalls, intrusion detection systems, and security policies.

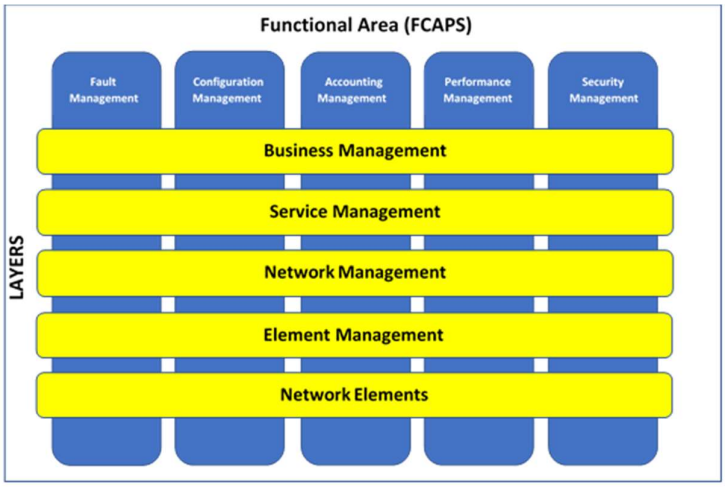


Figure 120: FCAPS model functional areas and TMN logical layers. Adapted: [84]

Some of security management tasks includes the following items:

- Identification of sensitive information and control of network physical and logical access resources;
- Enhance security by applying network intrusion detection systems;
- Respond to security breaches;
- Configuration of encryption policies;

The following figure illustrates the widely accepted relationship between the FCAPS (OSI/ISO network management model) and TMN where each layer communicates with the layer above or below it. The TMN is described in the following section.

1.5.2 Telecommunication Management Network

Telecommunication Management Network consists on a model for management of open systems in network communications defined by the ITU-T in its recommendation M.3000 series. ITU-T defines as TMN main objectives the provisioning of framework for telecommunications management by applying network models for the management of a variety of equipment network and services to support different management areas. These areas cover the planning, operations, administration, maintenance and provisioning of telecommunications networks and services. Moreover, TMN has the following two main objectives [81]:

- Multi-vendor environment functionality;
- Network functionality optimization;

There is a variety of TMN, which can start from a simple connection among an OS and a piece of telecommunications equipment to a more complex network that connects different types of telecommunications system. As it can be seen in Figure 121, a relationship between a telecommunications network and its Telecommunication Management Network, where the last interfaces the first at different point of operations to provide communications.

The basis for a communication network management is based on a TMN paradigm though is necessary that network management systems cover both management needs at different types of element levels such as BTD, RNC, CN and also network and service management levels.

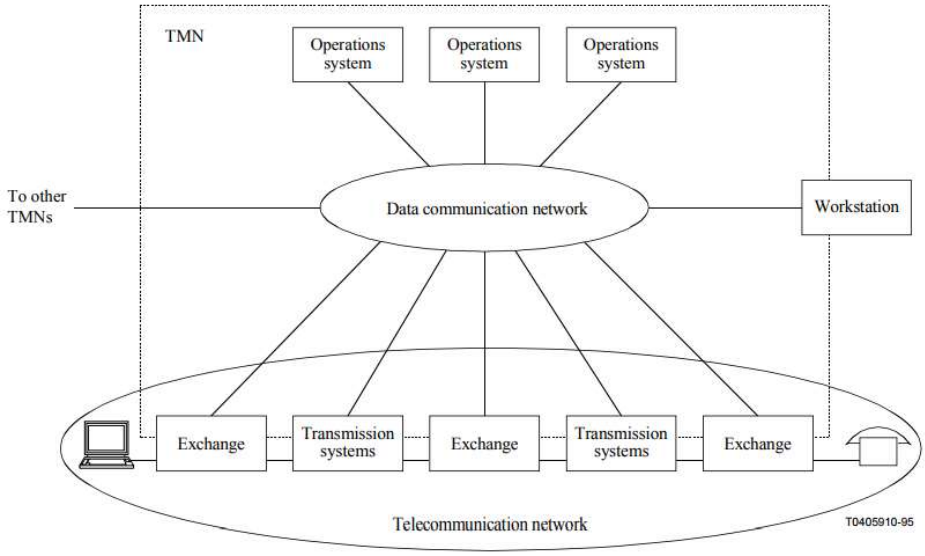


Figure 121: General relationship of a TMN to a telecommunication network. Source:[85]

1.5.2.1 Telecommunication-Management Architecture

Telecommunication-Management provides an interconnection between different types of Operations Systems (OS) and equipment by using an architecture which interfaces different protocols and messages for the exchange of management information.

Both UMTS and LTE TMA (telecommunication management architecture) has a base on the telecommunication management network. Defined by the ITU-T the telecommunication management network was a part of series of recommendation and has its base also in the ITU-T recommendation of the OSI management.

Nowadays Telecommunication management network are used by communication service provider to supply automated management functions which run upon an operations support system (OSS) software. By the end of the 20th century ITU-T had done in TMN models some definition for the OSS[86]. OSS are software or hardware application used to support back-office activities which helps communication service provider to operate networks (maintaining services and providing services to customers).

As OSS is used to support back-office, the BSS comes as front-office set of separated applications used to support activities such as commercial revenues and customer-relationships. When combined, the use of both OSS and BSS can help a communication service provider to run all network and services.

The ITU-T defined in its standardization sector by the Recommendation M.3010 the following types of TMN architecture level of abstraction [85]:

A. Functional architecture

Used to describe several different management functions the TMN functional architecture can be defined as a structural and generic framework of management functionality. The functional architecture is very useful when:

- Analysing and describing existing operation system and process flows;
- New operation systems and process flows specifications;
- Operation plans analysis and development of new system operations;

The structure of the functional architecture is composed of the following main elements [85]:

- Function blocks;
- Management Application Functions (MAFs);
- TMN Management Function Sets and TMN Management Functions;
- Reference points;

From all the functional architecture elements, only the function blocks and the reference point will be explained given the fact of both being the main. For more detailed information about the other elements please refer to [85][81][83].

3. Function blocks

Function blocks are a set of functional entities that allow a TMN to perform its management functions. TMN defines four different types of functional blocks in its architecture in which not all of them have to be present in a TMN configuration. The main types of functional blocks are [81][83]:

- **OSF (Operations Systems Function) block**: has the function of collecting network elements data and process information about telecommunications management to use it to monitor and control telecommunications management functions. It represents a server and the network element the clients.
- **NEF (Network Element Function) block**: this block offers all the function support required by the network being managed and communicates to the TMN to be monitored and controlled. Despite no being part of the TMN, NEF functions includes telecommunication which is represented to the TMN by the NEF itself. Another function the NEF has is providing representation to support TMN.
- **WSF (Workstation Function) block**: used to provide the operators the capacity to interpret management related information and also can be used to interface a human user given the fact that these users are not considered a part of TMN.
- **QAF**: used to interconnect the TMN entities to non-TMN entities (such as NEF-like and OSF-like) which does not support his standard reference points. Nevertheless, the QAF serves as a translator between q (TMN reference point) and m (non-TMN reference point) reference points. Because the QAF interfaces with TMN and non-TMN it lies at the edge of the TMN boundary.
- **MF**: has the responsibility of acting on all the information (store, filter, conversion and condense the information) passing between the NEF and OSF or QAF and OSF. Also, the use of MF can be used to connect a single or multiple NEF and QAF to OSFs.

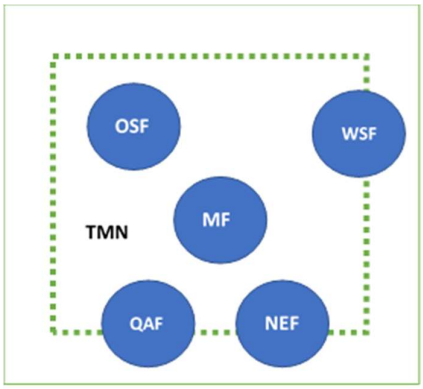


Figure 122: TMN Function Blocks. Source: [85]

As depicted in the previous figure about the TMN Function Blocks, is possible to see that some blocks a partly inside or outside of the TMN boundary (QAF, NEF and WSF) and the others are totally inside (MF and OSF) though only the function involved in management are part of a TMN in other hand these functions are specified by the TMN. Some literature uses the TF instead of MF and QAF though the TF is described as follows [85]:

TF (Transformation Function) block: connects two functional entities that have different communication mechanisms (protocols or information models). Some of the situation where the TF can be applied are described as follows [85]:

- If the TF is inside the TMN boundary it can be used to connect two standardized function blocks that have distinct mechanism of communication.
- May be used at the boundary when communication among two TMNs or between a TMN and a non-TMN environment.

4. Reference points

The concept of reference points was introduced by the TMN functional architecture as a matter to delineate service between two TMN function blocks and exchange information between all the function blocks.

Nevertheless, a group of five difference reference points were identified where from those three, the q, f and x, are described directly by the TMN recommendation M.3010. As for the other group, the g and m reference point, are described outside the TMN and are partially described. Despite that most TMN configuration being able to support multiple function blocks, note that from all existing types there is no need to have all of them implemented / present in a TMN configuration.

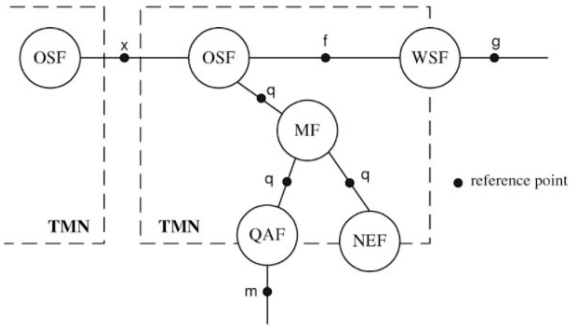


Figure 123: Reference Point between Function blocks. Source: [81]

As can be seen in the figure above, a group of reference points and functions blocks are provided, and is also possible to see that using the q reference is possible to reach the MF and to reach the QAF from the outside the reference m can be used.

B. Information architecture

Based on the OSI Management information model, the information architecture provides an object-oriented approach. Since it has an object-oriented approach all the management views from this model are visible at the boundary of the object being managed. The managed views are described in the following terms:

- Attributes with the characteristics of the object;
- Operations and activities performed upon the object;
- Behaviour to exhibit in response to operations;
- Notifications emitted by the object.

TMN information architecture has following fundamental elements structure:

- Reference points;
- Information elements and model of a reference point;
- Information models;
- interaction models;

C. Physical architecture

Located at the lowest level of abstraction of the TMN, the physical architecture shows how the different management function are implement into physical equipment. Also, it can be used to demonstrate how function blocks have to be mapped when building blocks (physical devices) and reference point to interfaces.

This architecture is composed of two fundamental elements, the physical blocks and physical interfaces. In Figure 124, a relation between functional architecture and physical architecture is depicted, where for each function block is possible to have various functional components and for each building block multiple function blocks can be implemented.

The physical architecture defines different types of building blocks which is composed of a group of computational objects operating in the same layer of the logical system. Besides management some of building blocks proprieties include security, independence and placement of the system. Though the building block defined by the physical architecture are listed as follows [87][88]:

- Operations systems (OS);
- Network element (NE);
- Work station (WS);
- Mediation device (MD);
- Q-adapter (QA);

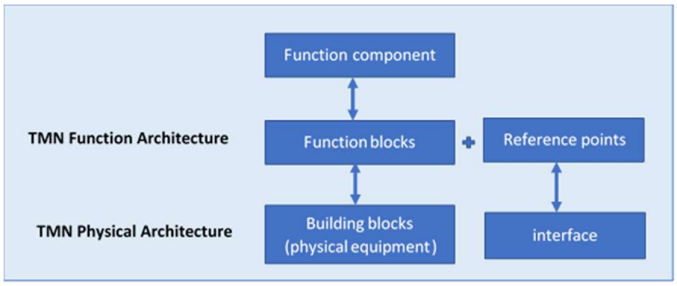


Figure 124: Relation between Functional architecture and physical architecture Source:[81]

Table 41: Example of function block and building blocks relationship Source:[83]

	NEF	MF	QAF	OSF	WSF
NE	M	O	O	O	O*
MD	-	M	O	O	O
QA	-	-	M	-	-
OS	-	O	O	-	O
WS	-	-	-	-	M
DCN	-	-	-	-	-

Legend:

O: Optional
M: Mandatory
O*: can only be present when OSF or MF are present

Nevertheless, the DCN (Data Communication Network), another building block can be added to the previous list. Despite not implementing any of the TMN function blocks, the DCN has a major application given the fact that it is used by other building blocks to exchange management information, though performing the network transport task.

Usually a building block implements one function blocks of the same name and is also possible to implement multiple function blocks by using a single building block. For example, in order to the building block of MD to perform is mandatory of it to implement the MF. Thus, the implementation of a function block into a building block can be seen in the table below.

D. Logical layered architecture

An important aspect of M.3010 when talking about logical layered architecture is the concept of hierarchal layers where similarly to FCAPS network management tasks are grouped into management functional (please see Figure 125) area helping to focus on a specific aspect of management application. Though, despite new models were built by applying the TMN as a reference, the use of the logical layered architecture is still being used which helps to reduce all the complexity.

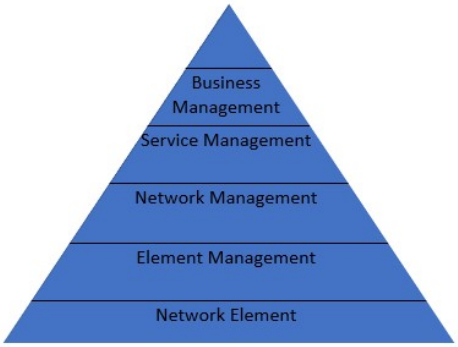


Figure 125:TMN pyramid logical layers
Source:[5]

Though to deal with the complexity of management the following logical layers are defined:

1. **NEL (Network Element Layer)**

The NEL is the lowest layer present in the model, often involved with the management of physical network elements functionalities, it defines interfaces for those elements by instantiating functions for device instrumentation. The use of this layer is very important for the effectiveness of management systems since it cover all FCAPS areas and is from where all the management information is collected.

All the different physical elements such as controllers and terminals are referred as Network Element (NE). In order to Q-adapter and the NE to be located in the NEL, the NEL needs to adapt the TMN and non-TMN information.

2. **EML (Element Management Layer)**

An Element Management Layer is the layer above the NEL and is used to provide to network elements the management functions on individual or group basis which may include the support of functions abstraction applied by the NEL. Since this layer contains all the functions and resources required for a TMN interconnection between the NE and the NML the functions provided by the EML includes view, change of network elements configurations and monitoring.

Also, includes collecting statistical data, logging event notifications and performance statistics. It has on it one or more OSFs element which are responsible for some network element functions. The EML work on base of the following three main roles [85][88]:

- Control and coordination of a subset of network elements on an individual NEF basis;
- Control and coordination of a subset of network elements on collective basis, thereby managing the relationship between NEF;
- Maintaining statistical, log and other data about elements;

3. NML (Network Management Layer)

The NML is used to manage network relationships and dependencies offering a holist view between various devices as supported by the EML. In this layer, different from the EML is concerned only with the working of the network elements dealing only with individual network elements interaction. It has all the functions and resources necessary for NE management. Some of NML main principle function roles are [88][83]:

- Control and coordination of the network view of all network elements within its layer domain;
- Provision and modification of network capabilities to support different service to customers;
- Maintain statistical information about the network and interact with the service manager;
- Network OSFs management of relationships (e.g. connectivity) between NEFs.

4. SML (Service Management Layer)

SML main concern and responsibility is with contractual aspects involved in services being provided to customers and guaranty that these services are running and functioning as intended. To supply this service the management layer makes use of the information provided by the network management layer. Some of SML main function are [85]:

- customer facing (Note) and interfacing with other PTOs/ROAs;
- interaction with service providers;
- maintaining statistical data (e.g. QOS);
- interaction between services.

5. BML (Business Management Layer)

The BML consist mainly of business related aspects such as: business planning, product planning financial issues which includes total enterprise and proprietary functionalities. It is included in the TMN architecture in order to simplify all the specification about capability. From the management layers the Business OSF can access all the information and functionalities where it can also interact with other OSF present in the same or other layers within the same TMN using the q reference point.

BML has the following main principal roles [85]:

- Decision-making process support to achieve an optimal investment and make use of new telecommunications resources;
- supporting the management of OA&M related budget;
- supporting the supply and demand of OA&M related manpower;
- maintaining all the data related to the total enterprise.