Universidade de Aveiro
2014

Departamento de Eletrónica,
Telecomunicações e Informática

Manuel David Fonseca Montenegro

Previsão de Capacidade para Redes de Acesso Rádio

Capacity Forecasting for Radio Access Networks

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

o júri

presidente / president

Prof. Dr. José Carlos da Silva Neves
Professor Catedrático da Universidade de Aveiro

Prof. Dra. Ana Cristina Costa Aguiar
Professora Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Prof. Dr. Aníbal Manuel de Oliveira Duarte
Professor Catedrático da Universidade de Aveiro

agradecimentos

À minha família, em especial aos meus pais pelo esforço e apoio que tornaram possível alcançar este objetivo.

Ao meu orientador, professor doutor A. Manuel de Oliveira Duarte, pela disponibilidade e essencial contributo para este trabalho.

À equipa da Nokia Networks, por tornar possível este estudo e pelos válidos conselhos para o mesmo.

palavras-chave UMTS, Planeamento, Desempenho, Monitorização, KPI, Previsão

resumo

A estratégia comum dos operadores no mercado das redes móveis (área onde este trabalho se debruça) passa por uma consolidação da sua rede base já instalada e pela otimização dos recursos já existentes.

A crescente competitividade e agressividade deste mercado obrigam a que os operadores mantenham a sua rede atualizada e com o menor número de falhas possível, com a finalidade de oferecer a melhor experiência aos seus utilizadores. Neste contexto, esta dissertação apresenta um estudo que auxilia os operadores a aperfeiçoar futuras alterações na sua rede.

De um modo geral, esta dissertação compara alguns métodos de previsão (baseados maioritariamente na análise de séries temporais) capazes de assistir os operadores no planeamento da sua rede e ainda apresenta alguns indicadores de rede onde as limitações de desempenho são mais frequentes.

keywords UMTS, TMN, Planning, Performance, Monitoring, KPI, Forecasting

abstract

The mobile networks market (focus of this work) strategy is based on the consolidation of the installed structure and the optimization of the already existent resources.

The increasingly competition and aggression of this market requires, to the mobile operators, a continuous maintenance and update of the networks in order to obtain the minimum number of fails and provide the best experience for its subscribers. In this context, this dissertation presents a study aiming to assist the mobile operators improving future network modifications.

In overview, this dissertation compares several forecasting methods (mostly based on time series analysis) capable of support mobile operators with their network planning. Moreover, it presents several network indicators about the more common bottlenecks.

Index

Index i
Tables Index v
Figures Index vi
Acronyms ix
1. Introduction 1
1.1. Motivation 2
1.2. Objectives 2
1.3. Organization of the Dissertation 3
2. Mobile Network 4
2.1. Mobile Network Evolution 4
2.1.1. GSM Architecture 6
2.1.2. UMTS Architecture 10
2.1.2.1. Release 4 10
2.1.2.2. Release 5 11
2.1.2.2.1. IP Multimedia Subsystem 11
2.1.2.2.2. HSDPA 13
2.1.2.3. Release 6 14
2.1.2.3.1. HSUPA 14
2.1.2.4. Release 7 14
2.1.2.4.1. HSPA & HSPA+ 15
2.1.3. LTE Architecture 15
2.1.3.1. E-UTRAN 15
2.1.3.2. EPC 16
2.1.3.2.1. MME 16
2.1.3.2.2. S-GW 17
2.1.3.2.3. PDN-GW 18
2.1.3.2.4. PCRF 19
2.2. Network Management 19
2.2.1. Network Management Models 20
2.2.1.1. OSI NMM 20
2.2.1.2. Telecommunication Management Network 21
2.2.1.2.1. Functional architecture 22

2.2.1.2.2. Physical Architecture 27
2.2.2. Performance Management 29
2.2.3. KPIs 30
2.2.4. Configuration Management 32
2.2.4.1. Why CM? 32
2.2.4.1.1. Administrative Aspects 33
2.2.4.2. CM systems and methods 34
2.2.4.2.1. CM Operation 35
3. Forecasting Methodology 36
3.1. Introductory Concepts 36
3.1.1. Methods 36
3.1.2. Time Series Components 38
3.1.3. Graphics 39
3.1.4. Forecast Evaluation 40
3.1.4.1. Sum of Squared Errors 40
3.1.4.2. Mean Absolute Percentage Error 41
3.1.4.3. Durbin-Watson 41
3.1.4.4. Akaike Information Criterion 42
3.1.4.5. Coefficient of Determination 42
3.1.5. Outliers and influential observations 43
3.2. Linear Regression and Correlation 45
3.2.1. Simple Linear model 46
3.2.1.1. Forecasting results 47
3.3. Time Series 50
3.3.1. Moving Average Smoothing 50
3.3.1.1. Moving Average Filter 50
3.3.1.2. Centered Moving Averages 51
3.3.1.3. Examples of MA smoothing 52
3.3.2. Classical decomposition 53
3.3.2.1. Additive model 53
3.3.2.2. Multiplicative model 54
3.3.3. Holt-Winters Exponential Smoothing 55
3.3.3.1. Multiplicative model 56
3.3.3.1.1. Excel Implementation 57

3.3.3.2. Result analysis 59
3.3.4. AutoRegressive Moving Average 63
3.3.4.1. Auto-Regressive model 64
3.3.4.2. Moving Average model 65
3.3.4.3. Non-Seasonal ARMA 66
3.3.4.3.1. Result analysis 71
3.3.4.4. Seasonal ARMA 74
3.3.4.4.1. Forecast 75
3.4. General Purpose 77
3.5. Add-ons and Enhancements 78
3.5.1. Manual Growth 78
3.5.2. Level Offset 79
3.6. CS Throughput Analysis 79
3.7. PS Throughput Analysis 82
3.8. Conclusion 85
4. Forecasting Tool 86
5. RNC Capacity Planning 94
5.1. RNC user plane fill factor 95
5.1.1. Iub PS throughput 97
5.1.2. Iub CS throughput 98
5.1.3. Outcome analysis 98
5.2. Unit Load monitoring 100
5.2.1. ICSU 101
5.2.2. DMCU 105
5.2.2.1. DMPG 106
5.2.2.1.1. DSP 108
5.2.2.2. RSMU 109
5.2.2.3. GTPU 109
5.3. RRC connected subscribers 110
5.4. Busy Hour Call Attempt 111
5.4.1. CS BHCA 111
5.4.2. PS BHCA 112
6. Conclusions 114
7. Future Work 115

8. References 116

Tables Index

Table 1: 3GPP Releases 6
Table 2: Mapping function blocks into building blocks 28
Table 3: KPI categories 32
Table 4: Methods Summary 79
Table 5: Case Study #1 80
Table 6: Daily average PS throughput forecast results summary 84
Table 7: Models properties 85
Table 8: RNC user plane fill factor 95
Table 9: Measurements required to compute the Iub PS throughput 97
Table 10: AMR throughput counters 98
Table 11: M592 measurements 101
Table 12: ICSU load 102
Table 13: Counters used for ICSU overload inspection 103
Table 14: DMPG load 106
Table 15: DSP load 108
Table 16: RSMU load 109
Table 17: GTPU load 110
Table 18: RNC connected subscribers. 111
Table 19: BHCA counters 111
Table 20: PS BHCA counters 112

Figures Index

Figure 1: Global mobile traffic 1
Figure 2: Final Goal: 2
Figure 3: UMTS (Rel. 5) 3
Figure 4: Mobile Generations 5
Figure 5: GSM Architecture 6
Figure 6: GSM Architecture (GPRS) 9
Figure 7: UMTS (Rel-5) Architecture 11
Figure 8: IMS Architecture 12
Figure 9: Release 7 (HSPA and HSPA+) Architecture 14
Figure 10: LTE (Release 8) Architecture 15
Figure 11: MME logical connections and main Control Functions 17
Figure 12: S-GW logical connections and functions 18
Figure 13: PDN-GW logical connection and main functions 18
Figure 14: PCRF main functions 19
Figure 15: TMN and a telecommunication network 22
Figure 16: Function blocks 23
Figure 17: Reference Points 24
Figure 18: TMN logical layers 25
Figure 19: TMN related architectures 27
Figure 20: Relation between architectures 27
Figure 21: Mapping reference points upon interfaces 28
Figure 22: Network Management Level 30
Figure 23: Report creation 31
Figure 24: CM Architecture (example) 34
Figure 25: Forecasting methods 37
Figure 26: Daily average voice throughput in Iub interface 38
Figure 27: Daily average voice throughput 39
Figure 28: RNC user plane fill rate vs. ICSU CPU load 40
Figure 29: Mean and Standard Deviation Sliding Window 44
Figure 30: AMR daily throughput 45
Figure 31: Scatter plot investigation 46
Figure 32: Daily average CS throughput in Iub interface 48
Figure 33: Real and forecasted values 49
Figure 34: Outliers and influential values 49
Figure 35: Block diagram of a 6th order/7-tap FIR filter 50
Figure 36: Moving average of order k=7 52
Figure 37: Moving average of order k=31 52
Figure 38: Constant seasonal amplitude (Additive model) 53
Figure 39: Additive decomposition steps. 54
Figure 40: Seasonal amplitude proportional to the trend (Multiplicative model) 54
Figure 41: Multiplicative decomposition steps 55
Figure 42: Holt Winters Initialization 58
Figure 43: Holt-Winters Parameters Estimation 59

Figure 44: Solver tool 59
Figure 45: CS throughput 60
Figure 46: Holt-Winters forecast 60
Figure 47: Real values vs. forecasted values 61
Figure 48: Holt-Winters case study #1 61
Figure 49: Case Study #1 (w/ Level Offset) 62
Figure 50: Holt-Winters outliers' case study #2 62
Figure 51: Holt-Winters outliers' case study #3 63
Figure 52: Case study #3 without outliers. 63
Figure 53: Non-Stationary behaviors 64
Figure 54: p-order IIR filter implementing an AutoRegressive model 65
Figure 55: q-order FIR filter implementing an Moving Average model 66
Figure 56: FIR/IIR filter implementing an ARMA(p,q) model 66
Figure 57: ARMA cycle construction 67
Figure 58: AR(p) part of ARMA 69
Figure 59: MA(q), Error and Predictions 69
Figure 60: Parameters and Constraints 70
Figure 61: Solver Settings 70
Figure 62: Reliability Indicators 71
Figure 63: ARMA forecast 71
Figure 64: AMR daily traffic 72
Figure 65: ARMA case study #1 72
Figure 66: Case study #1 (Level Offset and Manual Growth) 73
Figure 67: AMR case study #3 73
Figure 68: AMR case study 3 without outliers 74
Figure 69: SARMA results 75
Figure 70: AMR daily traffic 75
Figure 71: SARMA case study #1 76
Figure 72: General model plus manual growth 78
Figure 73: Outliers Influence 81
Figure 74: Daily average PS throughput 82
Figure 75: PS throughput forecast 83
Figure 76: Forecasting tool 89
Figure 77: RNC Properties 90
Figure 78: Forecasting Section 90
Figure 79: Shortcuts 91
Figure 80: Methods 91
Figure 81: Methods Information 92
Figure 82: Results Section 92
Figure 83: RNC capacity 94
Figure 84: RNC capacity (PS vs. CS) 96
Figure 85: Forecast vs. Reality 99
Figure 86: User plane fill factor forecast 99
Figure 87: Fill rates relationship 100
Figure 88: RNC Architecture 101
Figure 89: Evolution of technologies 103

Figure 90: ICSU load forecast 104
Figure 91: Relationship between fill factor and ICSU load 105
Figure 92: DMCU architecture 106
Figure 93: Average DMPG CPU load forecast 107
Figure 94: Relationship between RNC Fill Factor and DMPG CPU Load 107

Acronyms

3GPP	3rd Generation Partnership
AAL2	ATM Adaptation Layer 2
ACF	Autocorrelation Function
AIC	Akaike Information Criterion
AMC	Adaptive Modulation Coding
AMR	Adaptive Multi-Rate
AR	AutoRegressive
ARIMA	Auto Regressive Integrated Moving Average
ARMA	Auto Regressive Moving Average
ATM	Asynchronous Transfer Mode
AuC	Authentication Center
BLC	Buffer Limit Control
BML	Business Management Layer
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CA	Carrier Aggregation
CAPEX	CAPital EXpenditure
CDMA	Code Division Multiple Access
CM	Configuration Management
CN	Core Network
CP	Control Plane
CPM	Communications Processor Module
CPU	Central Processor Unit
CS	Circuit Switch
CSCF	Call Session Control Function
CSD	Circuit Switched Data
DCH	Dedicated Channel

DCN	Data Communication Network
DHCP	Dynamic Host Configuration Protocol
DMPG	Data and Macro Diversity Processing Groups
DSP	Data Signaling Processor
E-DCH	Enhanced Dedicated Channel
EDGE	Enhanced GPRS
eICIC	enhanced Inter-Cell Interference Coordination
EIR	Equipment Identity Register
eMBMS	Multimedia Broadcast Multicast Services (MBMS)
EML	Element Management Layer
EPC	Evolve Packet Core
E-TACS	Expanded Total Access Communication System
E-UTRAN	Evolved UTRAN
FACH	Forward Access Channel
FCAPS	Functional, Configuration, Accounting, Performance and Security
FIR	Finite Impulse Response
FP	Framing Protocol
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switching Center
GPRS	General Packet Radio Service
GRG	Generalized Reduced Gradient
GSM	Global System for Mobile Communications
GTP	GPRS Tunneling Protocol
GUTI	Globally Unique Temporary Identity
H-ARQ	Hybrid Automatic Repeat Request
HeNB	Home NodeB
HLR	Home Location Register
HSCSD	High Speed Circuit Switched Data
HSDPA	High Speed Downlink Packet Access
HS-DSCH	High Speed Downlink Shared Channel

HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High Speed Uplink Packet Access
HW	Hardware
I-CSCF	Interrogating-Call Session Control Function
ICSU	Interface Control and Signaling Unit
IIR	Infinite Impulse Response
IMEI	International Mobile Equipment Identity
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IoT	Internet of Things
IP	Internet Protocol
ISDN	Integrated Services for Digital Network
ISO	International Standardization Organization
ISUP	ISDN User Part
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LGMANA	Leg Management Program Block
LLA	Logical Layered Architecture
LS	Least Squares Estimation
LTE	Long Term Evolution
MA	Moving Average
MAC	Medium Access Stratum
MAF	Management Applications Function
MAPE	Mean Absolute Percentage Error
MBMS	Multimedia Broadcast Multimedia Service
MGW	Media Gateway
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MRFC	Media Resource Function Controller
MRFP	Media Resource Function Processor

MS	Mobile Station
MSC	Mobile Switching Center
MSS	MSC Server
NE	Network Element
NEF	Network Element Function
NML	Network Management Layer
NMM	Network Management Model
OA&M	Operation, Administration and Maintenance
OFDMA	Orthogonal Frequency Division Multiple Access
OMS	Operation and Management Server
OS	Operation System
OSF	Operation System Function
OSI	Open Systems Interconnection
OSS	Operation Support System
PACF	Partial Autocorrelation Function
PCH	Paging CHannel
PCRF	Policy and Charging Rules Function
P-CSCF	Proxy-Call Control Function
PCU	Packet Unit Control
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDN-GW	Packet Data Network Gateway
PI	Performance Indicator
PLMN	Public Land Mobile Network
PM	Performance Management
PMIP	Proxy Mobile IPv6
PPC	PowerPC
PS	Packet Switch
PSTN	Public Switched Telephone Network
PTO	Public Telecommunication Operators
QAF	Q Adaption Function

QoE	Quality of Experience
QoS	Quality of Service
RA	Routing Area
RAB	Radio Access Bearer
RAN	Radio Access Network
RLC	Radio Link Control
RNC	Radio Network Controller
RRC	Radio Resource Connection
RRM	Radio Resource Management
RSMU	Resource and Switch Management Unit
SAE	System Architecture Evolution
SARMA	Seasonal ARMA
S-CSCF	Serving-Call Session Control Function
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SHO	Soft Handover
SIP	Session Initiated Protocol
SLF	Subscription Locator Function
SML	Service Management Layer
SON	Self-Organizing Networks
SSE	Sum of Squared Errors
TA	Tracking Area
TD-CDMA	Time Division CDMA
TDMA	Time Division Multiple Access
TF	Transformation Function
TMN	Telecommunication Management Network
TTI	Transmission Time Interval
UE	User Equipment
UI	User Interface
UMTS	Universal Mobile Telecommunication System
UP	User Plane

URA	UTRAN Registration Area
UTRAN	Universal Terrestrial Radio Access Network
VLR	Visitor Location Register
WAC	Windows Admission Control
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network
WSF	Workstation Function

1. Introduction

The Universal Mobile Telecommunication System (UMTS) brought the possibility to access data services in a practical and viable way. The networks evolved from a voice oriented service to an enormous diversity of multimedia data services (IP based). Although most of the mobile networks revenues still coming from voice traffic, traffic trend analysis shows that this pattern is significantly changing. The data traffic is exploding due to the increasingly use of smartphones, tablets, dongles, flat tariffs plans, etc. *Figure 1* by Eriksson clearly illustrates this trend.

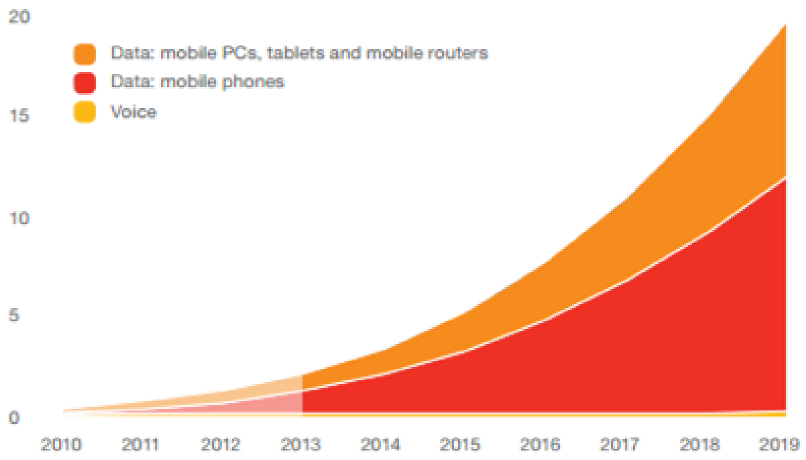


Figure 1: Global mobile traffic
Source: "On the Pulse of the Networked Society", Ericsson Mobility Report, June 2014

The diversity of services with multiple requirements is revealing itself a big challenge to mobile operators that are struggling to manage the performance of their networks and reduce their capital expenditure (CAPEX) and operational expenditure (OPEX). It is being difficult for them to follow the traffic's trend.

Facing this situation, the mobile operators have on one hand the investment in network expansion only when it is really needed, which means CAPEX increase. On the other hand the need to optimize the network performance by maximizing the usage of existing resources to prevent bottlenecks and loss of revenue.

The current performance tools are based on measure and display several performance indicators. These indicators can tell, among several other metrics, the capacity state of the Network Elements (NEs) in the network. Although the Key Performance Indicators (KPIs) [1] are a powerful tool capable of tell if there are any current issues in the network, these cannot forecast when the network starts experiencing these capacity issues.

[1] KPIs, as the name suggests, are network performance indicators. This topic is study later in *2.2.3 KPIs.*

1.1. Motivation

Facing this problem, network management entities have created mechanisms trying to prevent network capacity issues. These mechanisms are mainly based on the creation of reports and analysis of network indicators.

My motivation as a novice in the telecommunication management area is to develop forecast mechanisms capable of prevent network capacity issues.

1.2. Objectives

The main goal of this study is to develop a forecast tool, based in mathematical models, suitable for operators to prevent potential problems related with network capacity.

To do so, it is necessary to study the most common forecast techniques, be able to manage and apply them to this particular case. It is also necessary to study the KPIs related to this matter and select the most significant ones to include in the forecast tool.

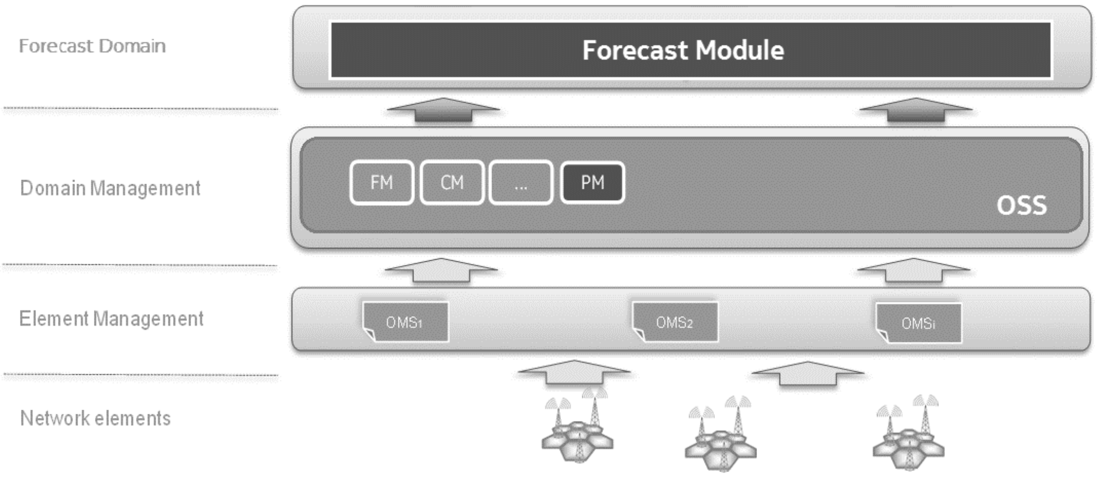


Figure 2: Final Goal

The objective is to make the forecast module take advantage of the existing resources, i.e. the forecast tool will use data (collected and filtered) from a real telecommunication vendor Operational Support System (OSS) and use it in its methods for forecasting.

This study is aimed at UMTS networks, specifically the Radio Access Network (RAN), UTRAN in *Figure 3*.

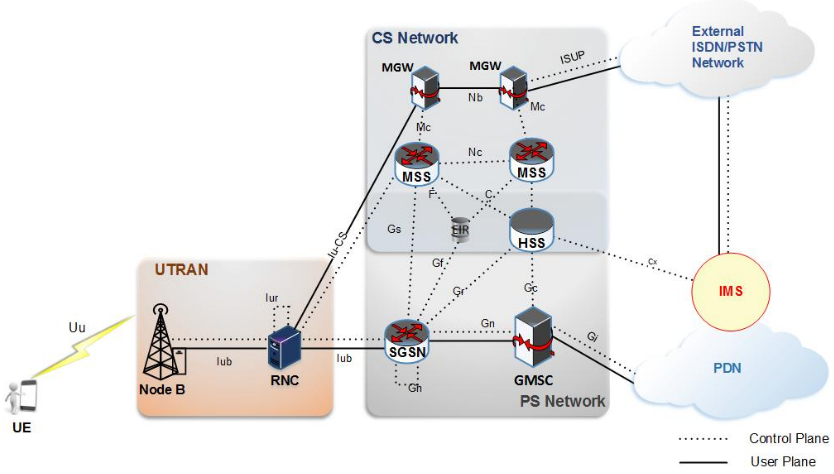


Figure 3: UMTS (Rel. 5)

1.3. Organization of the Dissertation

Besides this introductory chapter, this dissertation is organized in four main chapters:

- Chapter 2 introduces the UMTS network architecture, network elements and interfaces. After this, introduces the standardized network management models and discusses the importance of KPIs and performance management. Finally, it describes the role of configuration management in a network management environment.

- Chapter 3 studies classical methods used for forecasting. Starts to introduce some important concepts for a better understand of the chapter and then follow a structure of the type, theoretical study followed by a practical analysis, for each model. Besides that, it also presents several concepts and techniques used to evaluate and support the methods.

- Chapter 4 briefly introduces the forecast tool developed during the research and writing of this study. It is an important chapter that helps to recognize the origin of all the results analyzed in this document.

- Chapter 5 presents a detailed analysis of the KPIs chosen to use in the prediction models. Furthermore, gives examples of the potential problems, how to analyze them and suggests procedures to prevent the problems.

2. Mobile Network

During this chapter the reader will be introduced to the UMTS architecture and its evolution, as well as the architecture of the most used network management standard models. Furthermore, it will be discussed the concept and importance of KPIs, Performance Management (PM) and Configuration Management (CM).

2.1. Mobile Network Evolution

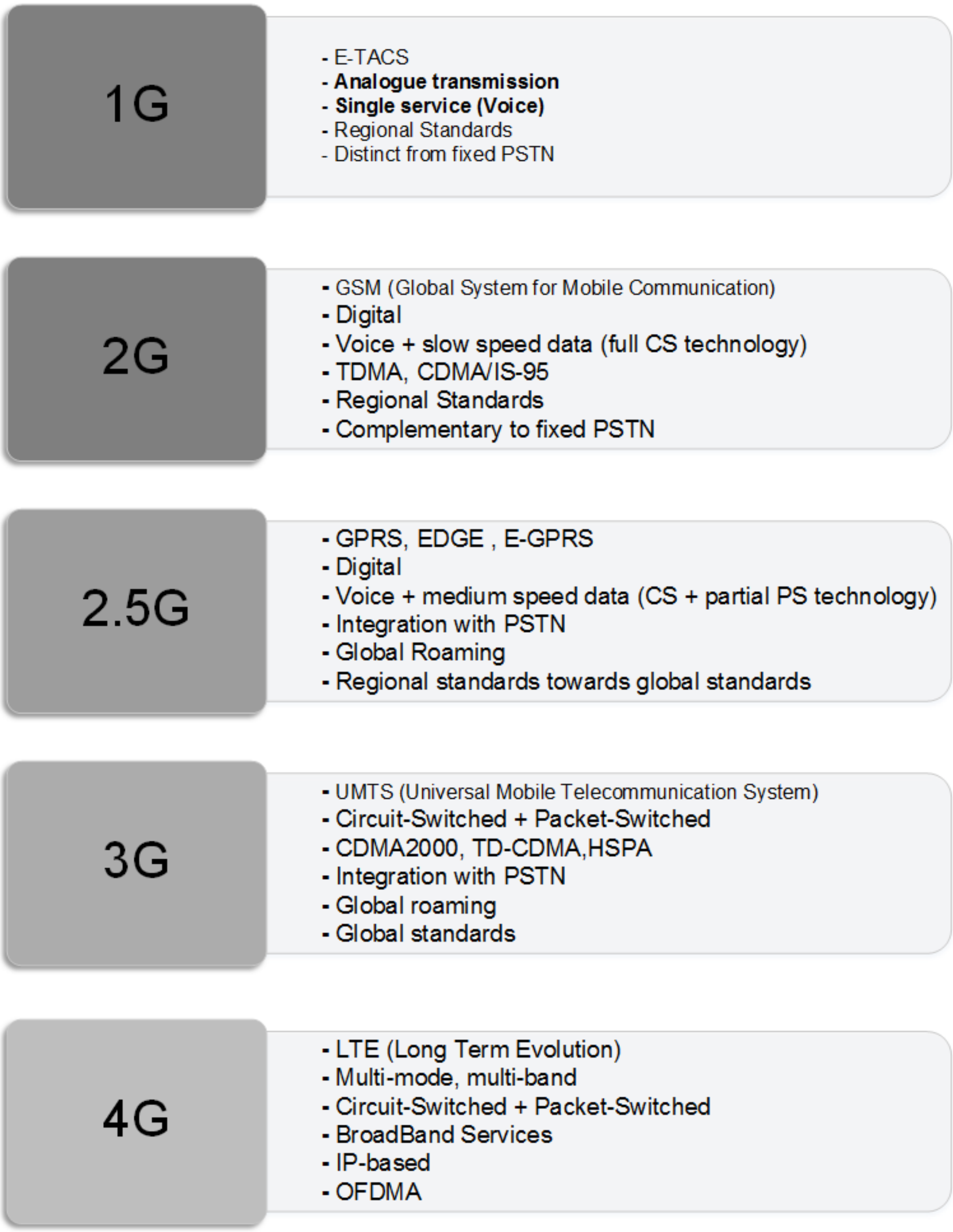
Somewhere by the mid-90s [3] the words third generation were used to designate the mobile network under research by that time and quickly became a “buzzword” in this field.

In order to better understand the current state of the mobile networks, especially the UMTS architecture (aim of this dissertation), it is important to make a step-back in the history of the mobile networks and track from closely the evolution of them.

This is the purpose of this section, to give an overview on the evolution of the mobile networks generations (Gs) from the first mobile network generation (1G) in the late 70s [1], through the GSM or second generation, until the UMTS or third generation (3G).

Also, the LTE architecture (actual state of the mobile networks) is study here, to compare the differences and similarities with the older technologies, especially with third generation, focus of this work.

A discontinuity or change in the paradigm sets the change of generations. *Figure 4* shows some of the important changes between generations.



1G

- E-TACS
- **Analogue transmission**
- **Single service (Voice)**
- Regional Standards
- Distinct from fixed PSTN

2G

- GSM (Global System for Mobile Communication)
- Digital
- Voice + slow speed data (full CS technology)
- TDMA, CDMA/IS-95
- Regional Standards
- Complementary to fixed PSTN

2.5G

- GPRS, EDGE , E-GPRS
- Digital
- Voice + medium speed data (CS + partial PS technology)
- Integration with PSTN
- Global Roaming
- Regional standards towards global standards

3G

- UMTS (Universal Mobile Telecommunication System)
- Circuit-Switched + Packet-Switched
- CDMA2000, TD-CDMA,HSPA
- Integration with PSTN
- Global roaming
- Global standards

4G

- LTE (Long Term Evolution)
- Multi-mode, multi-band
- Circuit-Switched + Packet-Switched
- BroadBand Services
- IP-based
- OFDMA

Figure 4: Mobile Generations

Digging a little bit in the architectures, the 3GPP organization has defined several hundred specifications categorized in releases. *Table 1* lists some of the 3GPP standards defining the 3G operation and beyond [2], [5].

The last release that is considered to belonging to 3G is the release 7 (HSPA+), some literature, refers to this release as the 3.9G architecture. Further superior releases are considered LTE.

The last architecture release of 3GPP is the release 11 with functional freeze date including stable protocols in June 2013 [4] (last checkup: May 2014). More detailed information on the 3GPP releases can be found in [4].

Release	Freeze Date	Features
99	Dec-99	WCDMA, TD-CDMA
4	Mar-01	All-IP core network, TD-SCDMA
5	Jun-02	HSDPA, IMS phase 1
6	Mar-05	HSUPA, IMS phase 2, MBMS
7	Dec-07	HSPA+ with MIMO, EDGE Evolution, IMS VCC
8	Dec-2008	OFDMA, MIMO, LTE, SAE, eNodeB, All-IP Network
9	Dec-2009	HeNB, SON, eMBMS
10	June-2011	Carrier aggregation(CA), enhanced inter-cell interference coordination (eICIC) , "HetNets",etc
11	June-2013	Enhancements to CA, MIMO and eICIC, etc.

Table 1: 3GPP Releases
Adapted from "Wireless Mobile Evolution to 4G Network", Mohammed Jaloun, Zouhair Guennoun

2.1.1. GSM Architecture

The mobile networks exploded by the beginning of the 90s with the GSM/2G networks and reach their cruise speed during the second half of the 90s. By the end of the 90s the mobile networks stopped to be a luxury and start to become a major daily need 116[1].

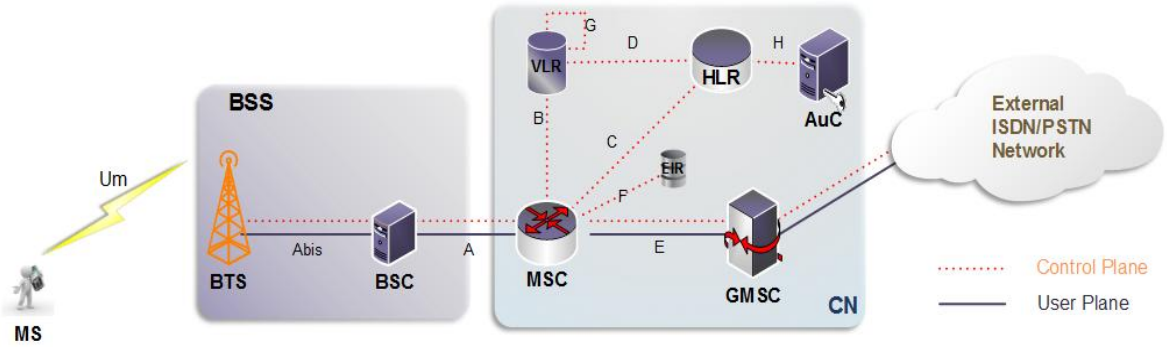


Figure 5: GSM Architecture

Figure 5 represents the GSM architecture where continuous dark line between Network Elements (NEs) represents the data plane and the orange dotted line represents the control plane.

The GSM architecture can be subdivided in three basic subsystems as *Figure 5* illustrates:

A. Mobile Station

Generally, the Mobile Station (MS) is the mobile equipment used by a subscriber to communicate with the GSM network**.** Some functions of the MS are:

- Speech to digital conversion to transmit to the BTS and digital to speech when receiving from the BTS.
- Radio Signal encryption.
- Monitoring the signal quality from BTS and reporting this back.

B. Base Station Subsystem

Base Transceiver Station (BTS) – Radio equipment (transceivers and antennas) needed to provide the best coverage to a designated area and communicate with the MS through the radio interface. Some of the most important functions are:

- Conversion of radio waves into a digital signal to send to the BSC and digital signals into radio waves to send to the MS.
- Monitoring quality and levels of the radio waves and reports to the BSC.

Base Station Control (BSC) – The BSC has the responsibility of control all the BTS attached to it. Some of the main functions are:

- Manage of radio resource allocation.
- Handover control taking into account the values from the BTS.
- Power control.

C. Core Network (CN)

The CN in GSM is in fact a Circuit Switched (CS) Core Network (CN), because even having slow speed data services these services are secondary due to the fact that it uses voice service resources. This fact will become clear later.

Mobile Switching Center (MSC) – The central element in the CS CN and its main task is to control the services provided by the system. The major difference between a MSC and a normal switching center in a fixed network is that the MSC has additional functions such as:

- Call control and mobility management.
- Gathering billing information.
- Location registration of MS.
- Authentication.

- Handover procedures (e.g. change to other MSC).
- Managing of radio resource allocation.

Visitor Location Register (VLR) – The VLR is a database containing information about the current MSs in its area. A single VLR can be associated to one or more MSCs.

The information in VLR, unlike in the HLR, is dynamic. When a MS enters a MSC area and try to connect with the network, the MSC is informed of its arrival and its associated VLR is update. After this, a message is sent to the HLR informing that the VLR contains the location of the UE. Some of the data registered in the VLR are:

- Subscriber location.
- Identification numbers of the mobile subscribers.
- Services that the subscriber can use.
- Security information for authentication purposes.

Gateway Mobile Switching Center (GMSC) – GMSC is an interface between the Home Public Land Mobile Network and the external ones (e.g. PSTN), i.e. it works like a support to the incoming calls from an external network, e.g. when the external network is not cable to interrogate the HLR because of the different protocols.

The GMSC receives all the incoming calls (like a default gateway) and asks the HLR for the address the user whereabouts and then routes the call to the MSC that is currently serving that user. The GMSC can be seen as a variation of the MSC with more capabilities.

Home location Register (HLR) – It is a central database that keeps information about all the subscribers within its network. The HLR holds:

- User identification, addressing and security information, i.e. network access control information for authentication and authorization.
- User location information at inter-system level.
- Handles the user registration, and stores inter-system location information.
- User profile (ex. services).

Authentication Center (AuC) – The AuC cooperates closely with the HLR database, it is responsible for:

- Encryption keys generation.
- Identify the MS and the network.
- Regulate the integrity of the transmission.

Equipment Identity Register (EIR) – The EIR is database with a list of numbers identifying a given mobile station for security purposes, e.g. to identify a stolen mobile or if it is causing problems in the network. The International Mobile Equipment Identity (IMEI) is used for this purpose.

Note that all the interfaces between the Network Elements (NEs) of the GSM architecture are represented in *Figure 5*. *T*hese interfaces determine certain rules for the communication between these NEs.

The major problem of GSM is the transmission of data by the circuit switched technology which provides very low transmission rates. The protocols used in this type of transmission were the Circuit Switched Data (CSD) and High-Speed Circuit Switched Data (HSCSD) [1]. The charging scheme were independent of the amount of data transmitted thus becoming a major constraint to the success of data transmission.

Therefore, in the year 2000 the introduction of the General Packet Radio Service (GPRS) added the packet switched technology and "kick started" the delivery of Internet on mobile equipment. Later on, an advance in GSM radio access technology brought us the EDGE (Enhanced Data rates for Global Evolution) or Enhanced GPRS. The EDGE technology can be considered as the "almost 3G".

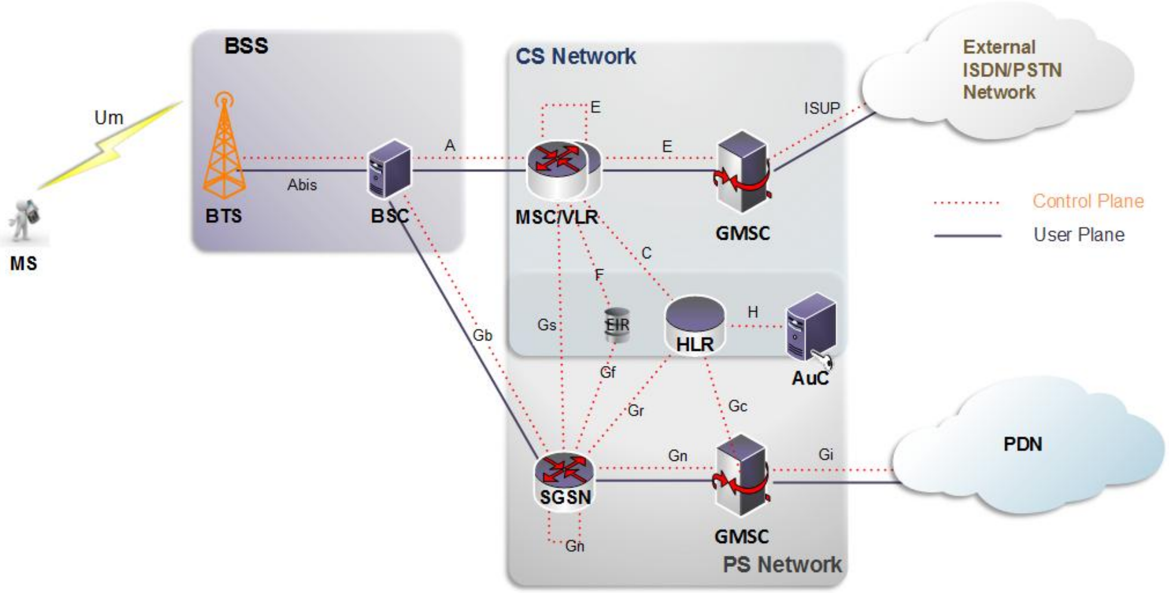


Figure 6: GSM Architecture (GPRS)

The GPRS brought some modifications to the old GSM architecture. A new item called Packet Control Unit (PCU) was added to the BSC with the function to separate the CS traffic from the PS traffic. However the major modification were in the core network, where two new NEs were added (refer to *Figure 6*).

The new SGSN and GGSN form now the PS CN and handle the packet data transmission.

Serving GPRS Support Node (SGSN) – It is the connection point between Radio Access Network (RAN) and PS CN and has a dynamic database which stores information about the current mobile station that is serving:

- Device location.

- Security information.
- QoS.

Also responsible for:

- Authentication.
- Mobility Management context (e.g. when a MS is PS attached, the SGSN keeps track of it to a Routing Area (RA) or specific cell).

Gateway GPRS Support Node (GGSN) – Provides the interface between the PLMN and external packet switched networks:

- Packets arriving from an external network are routed to the respective SGSN by the GGSN.
- Packets from the mobile subscriber are routed to external networks by the GGSN.

Like SGSN the GGSN has a data base that stores information about the mobile subscriber:

- QoS negotiated.
- Charging information.
- Address of the SGSN serving the user.

2.1.2. UMTS Architecture

This chapter is the continuity of a history about the evolution of the mobile networks. Thus, the proper understand of this chapter is strongly dependent of the previous chapter (*2.1.1 GSM Architecture*).

Back to the mobile networks evolution, the next big modifications were done at a radio access level, when by mid-2001 [1] the first commercial radio access technology based on Wideband Code Multiple Access (WCDMA) start. The first UMTS/3G releases (**release 99**) has been commercial initiated.

2.1.2.1. Release 4

Release 4 brought the packet transmission technology to the PS CN making the UMTS get close to its initial “dream”, reach an all-IP approach.

More efficient transport options and higher switching efficiency allowed new and lower cost services. Now the calls started to be transmitted through the MGW (refer to *Figure 7*) without need to be routed to the MSC Server site.

The separation of the control plane from the user plane facilitates the allocation of more efficient bearer services and with different characteristics. This architecture also simplifies the convergence with the packet switched domain.

As a result of these modification the old MSC, VLR and GMSC NEs were replaced by the new MSC Server (MSS), GMSC Server and MGW NEs.

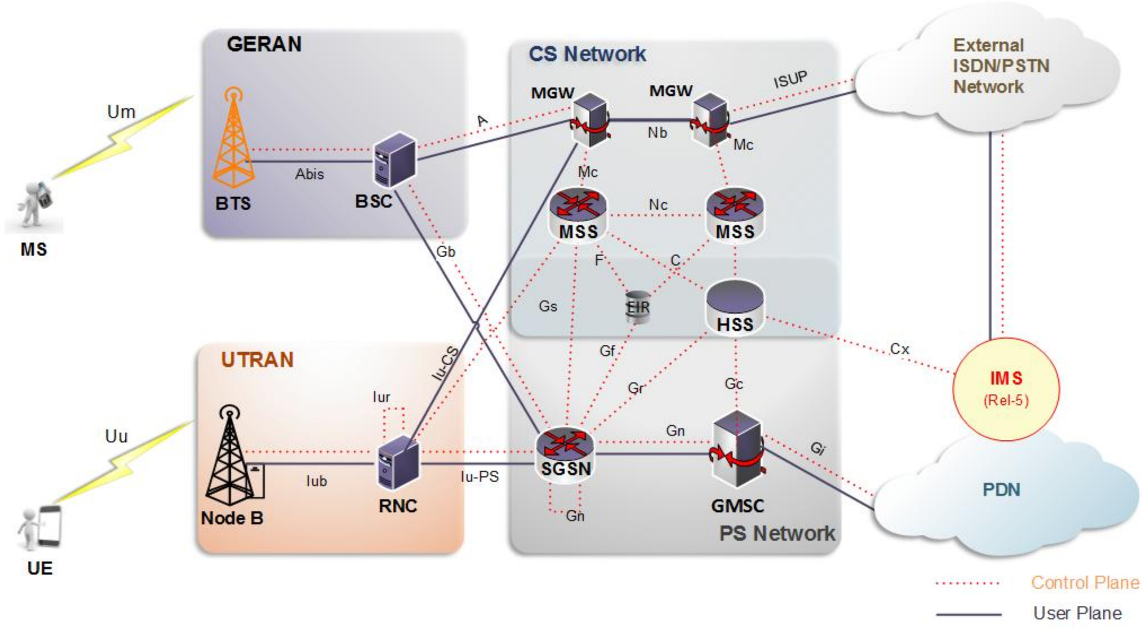


Figure 7: UMTS (Rel-5) Architecture

2.1.2.2. Release 5

Release 5 (*Figure 7*) has brought the IP Multimedia Subsystem (IMS) to the CN side and the High Speed Downlink Packet Access (HSDPA) protocol to the radio access side. Besides these modifications, the HLR and AuC were merged into one master database, the Home Subscriber Server (HSS).

2.1.2.2.1. IP Multimedia Subsystem

The IP Multimedia Subsystem (IMS) provides the possibility to set up multimedia sessions by means of the Session Initiated Protocol (SIP) for provision of multimedia services.

The IP Multimedia Core Network (IM CN) subsystem should enable the convergence, and access to, voice, video, messaging and web-based technologies for the wireless user, and combine the growth of the internet with the growth in mobile communications.

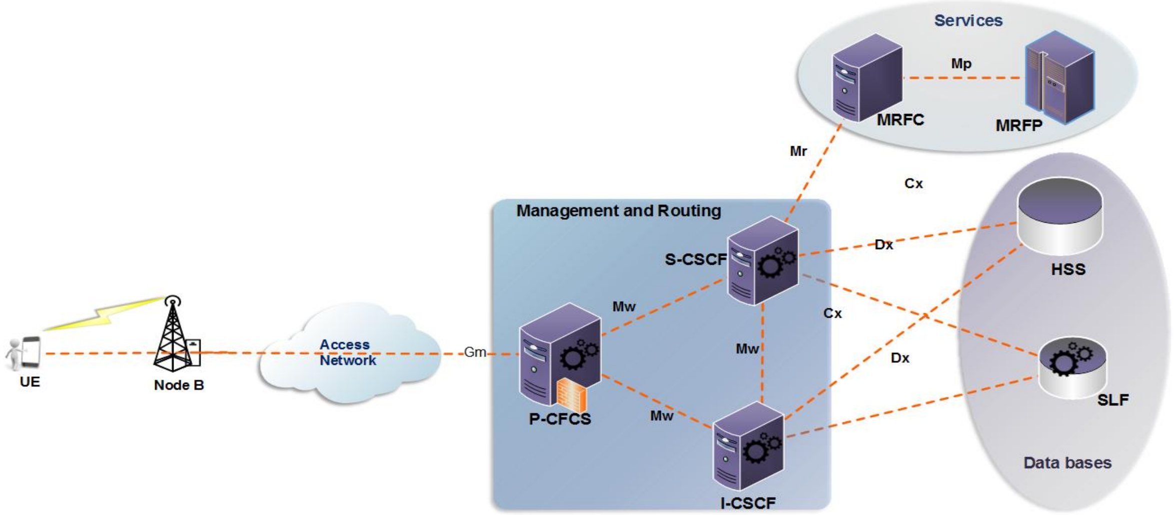


Figure 8: IMS Architecture

Figure 8 represents a simple overview of IMS architecture (3GPP standard). This overview represents the most relevant entities, functions and interfaces. The objective of this work is not a detailed study about the IMS. For more information, refer to 3GPP TS 23.002.

Most vendors follow the standardized IMS architecture although it is possible to find nodes implementing more than one function and functions distributed over more than one node [7].

A. Call Session Control Function (CSCF)

Three important CSCFs will be discussed: Proxy-CSCF, Serving-CSCF and Interrogating-CSCF. Each of the CSCFs have their special tasks, the most important are described next.

P-CSCF – The Proxy-Call Session Control Function (P-CSCF) is the first point of contact between the IMS UE and the IMS network. This means that all SIP signaling traffic from UE will be sent through the P-CSCF and vice-versa.

S-CSCF – Serving-Call Session Control Function (S-CSCF) is the central node of the signaling plane and it is responsible for registration processes, routing decision and service profile storing.

I-CSCF – The Interrogating-Call Session Control Function (I-CSCF) is a contact point between a subscriber and its operator's network.

B. Databases

Two databases are part of the IMS subsystem: the Home Subscriber Server (HSS) and the Subscription Locator Function (SLF).

HSS – Has mentioned earlier in *2.1.2.2 Release 5* the Home Subscriber Server (HSS) is the merging between the Home Location Register (HLR) and Authentication Center (AuC) in a single database. The functions of the HLR and AuC are described in *2.1.1 GSM Architecture*.

In addition to the functions required by PS and CS domains, the HSS stores data related with all IMS subscribers and services. The most important data includes [8]:

- User's identities.
- Registration Information.
- Access Parameters.

SLF – The Subscription Locator Function (SLF) is a support function that allows the I-CSCF and S-CSCF to find the address of the HSS that holds information data for a given subscriber when architecture of multiple and separately HSSs have been deployed by the network operator.

C. MRFC & MRFP (Service Functions)

The Service functions, also known as Media Resource Functions (MRF), provides the home network with a source of media and the ability for mix media streams (e.g. centralized conference bridge), media analysis, etc. These service functions are divided into a signaling plane – the Media Resource Function Controller (MRFC) and a media plane – the Media Resource Function Processor (MRFP).

2.1.2.2.2. HSDPA

As stated before, release 5 has introduced the HSDPA. HSDPA is a packet based data service which includes techniques such as:

- **Hybrid Automatic Repeat Request (H-ARQ)**: Having a fast error correction make the transmission delay, due to errors, decrease.

- **Improvements in the resource management**: Addition of High Speed Downlink Shared CHannel (HS-DSCH) with Adaptive Modulation Coding (AMC) and multi-code operation and short Transmission Time Interval (TTI)

- **Fast scheduling**: Moving the scheduling to the Node B (MAC-HS) enables a more efficient implementation of scheduling by allowing the scheduler to work with the most recent channel information. The scheduler can adapt the modulation to better match the current channel conditions and fading environment.

The goal of the implementation of these techniques is to increase throughput, reduce delay and achieve higher peak rates. A good aspect of the implementation of HSDPA on existing 3GPP systems is that it was a low-cost investment once in some cases was just a software upgrade. For more details on HSDPA refer to [9], [10] and [11].

2.1.2.3. Release 6

Later on, release 6 brought improvements to the IMS (CN side), the capability of interworking with WLANs (CN and radio access sides) and the High Speed Uplink Packet Access (HSUPA) technology in the radio access side.

2.1.2.3.1. HSUPA

The High Speed Uplink Packet Access (HSUPA) system is in fact a new uplink transport channel, the Enhanced Dedicated CHannel (E-DCH), which brought some of the same features to the uplink as the HSDPA with its new transport channel (HS-DSCH).

The E-DCH supports fast Node B based scheduling, fast physical H-ARQ and shorter TTI. Though, unlike in HSDPA the E-DCH is not a shared channel, but a dedicated one (i.e. each UE has its own dedicated E-DCH data path to the Node B that is continuous and independent from the others UEs [11].

Similarly to HSDPA technology, the HSUPA is inexpensive because it is based in software updates. For more details on this matter refer to [11].

2.1.2.4. Release 7

These last developments from the last releases (Rel-5 and Rel-6) brought improvements to the radio network side, and so did the emerging of release 7 that introduced the HSPA and HSPA+ technologies to the radio access network side.

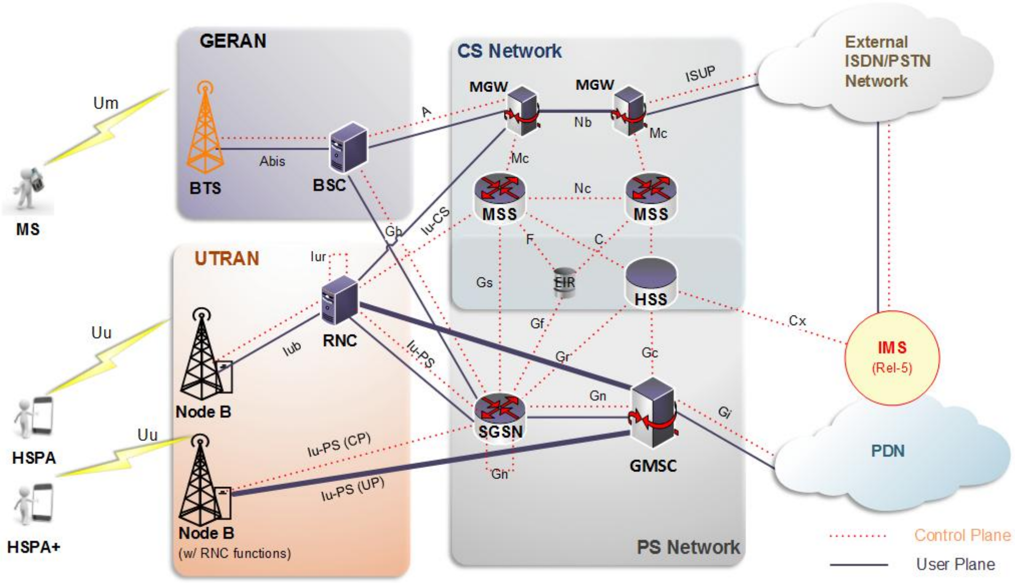


Figure 9: Release 7 (HSPA and HSPA+) Architecture

2.1.2.4.1. HSPA & HSPA+

The High Speed Packet Access (HSPA) is an upgrade to the WCDMA networks that increases the packet data transmission performance. The HSPA is the result of HSDPA and HSUPA combination. HSPA is a radio access network improvement which is associated with the PS CN enhancement (RNC – GGSN direct tunneling) [1].

The HSPA+ brought by release 7 is an evolution of the HSPA. New functions were added, such as: higher order modulation as well as Multiple Input Multiple Output (MIMO), used only in downlink. To learn more about HSPA and HSPA+ refer to [11] and [7].

Thus, release 7 becomes the launching pad to a new generation of mobile networks, the Long Term Evolution or 4G networks.

2.1.3. LTE Architecture

The Evolved Universal Terrestrial Radio Access Network (E-UTRAN) is the radio access part to an Evolved Packet Core (EPC).

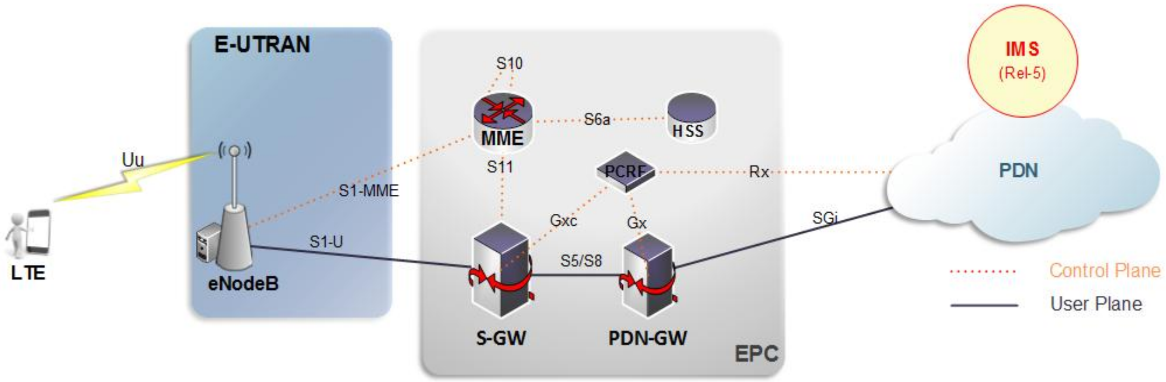


Figure 10: LTE (Release 8) Architecture

2.1.3.1. E-UTRAN

In the radio access part it is possible to see that, like in the HSPA+, the olds (BTS - BSC) and (Node-B - RNC) functions have been merged into one NE – the eNodeB. Essentially, the E-UTRAN is a network created only by eNodeBs, with no centralized controllers interconnected between them. The eNodeBs are connected to the core network over the S1 interface as you can see in *Figure 10*.

The new access solution is based in Orthogonal Frequency-Division Multiplexing (OFDM) that combined with spatial multiplexing (i.e. multiple antennas), high order modulation (up to 64QAM) [7] and large bandwidths, can achieve high data rates.

The MAC protocol layer level which is responsible for scheduling is now represented only in the UE and eNodeB leading to fast communication and decisions between them. In UMTS the MAC protocol and scheduling, is located in the RNC and when HSDPA was introduced an additional MAC sub-layer, responsible for HSPA scheduling, was added in Node-B.

The eNodeB is responsible for the follow Control Plane (CP) functions [14]:

- **Radio Bearer Management** – Includes radio bearer setup and release and also involves Radio Resource Management (RRM) functionalities for initial admission control and bearer allocation. These functions are controlled by the MME through the S1 interface during session setup, release and modification phases.

- **Radio interface transmission and reception** – Includes radio channel modulation/demodulation as well as radio channel coding/decoding.

- **Uplink and downlink Dynamic RRM and data scheduling** – The most serious function, because the eNodeB as to be able to multiplex different data flows over the radio interface using the available resources and in the most efficient way.

- **Mobility Management** – UE mobility handling while the terminal is in the active state. Handover algorithms for mobility decision and target cell determination.

- **User data IP header compression and encryption** – Used to maintain privacy over the radio interface and transmit IP packets in the most efficient way.

2.1.3.2. EPC

The Evolved Packet Core (EPC) is purely IP based, i.e. both real time and data services will be carried by IP protocol. The IP address is allocated when the mobile switches on and released when the mobile switches off [13].

2.1.3.2.1. MME

The Mobility Management Entity (MME) is the NE responsible for all CP functions related to subscriber and session management. MME establishes a direct CP path between itself and the UE which is used as primary control channel between network and UE (see *Figure 11).*

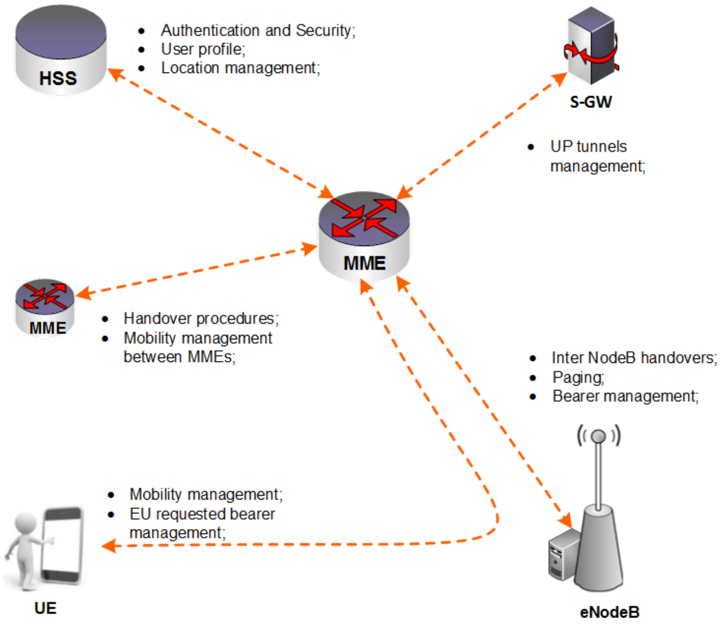


Figure 11: MME logical connections and main Control Functions
Adapted from "A tutorial on LTE Evolved UTRAN (EUTRAN) and LTE Self Organizing Networks", Dhruv Sunil Shah

The main functions of the MME are: the following:

- **Resource allocation and authentication** – Provide additional security to the UE, assigning to each UE a temporary identity, Globally Unique Temporary Identity (GUTI) [14].
- **Tracking area list management** – MME tracks the UEs presented in its area. If the UE is connected it tracks its location to an eNodeB level. If it is in idle mode tracks its location to a Tracking Area (TA) level.
- **Handover procedures** – Takes part on the control signaling for handover of an active mode UE between MMEs, S-GWs and eNodeBs.
- **Bearer management**

2.1.3.2.2. S-GW

The Serving Gateway (S-GW) it is not too involved with the CP operations but more concerned with the User Plane (UP) tunnel management. S-GW can use or GPRS Tunneling Protocol (GTP) or Proxy Mobile IPv6 (PMIP) tunnels for data flow depending on the data bearer setup (refer to *Figure 12)*.

The S-GW only controls its own resources and based on the requests from PDN-GW, MME or PCRF.

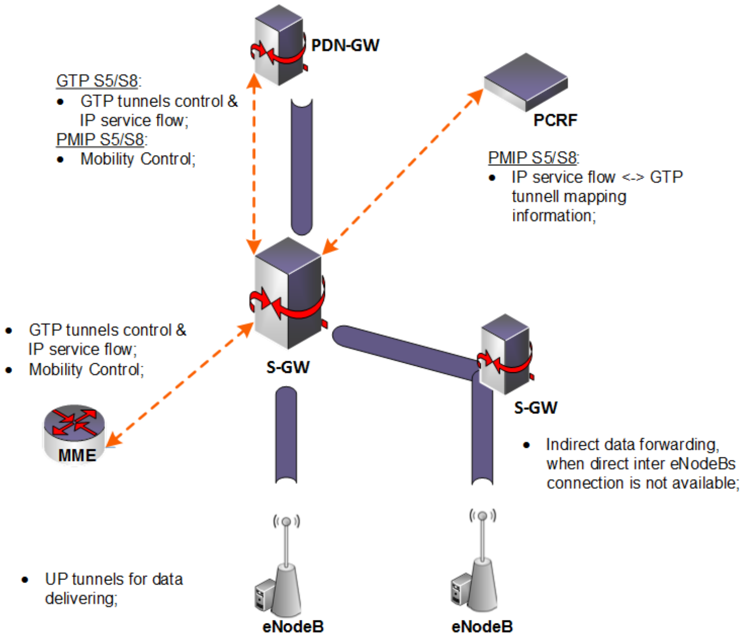

PDN-GW
GTP S5/S8:
GTP tunnels control & IP service flow;
PMIP S5/S8:
Mobility Control;
PCRF
PMIP S5/S8:
IP service flow <-> GTP tunnell mapping information;
GTP tunnels control & IP service flow;
Mobility Control;
S-GW
MME
S-GW
Indirect data forwarding, when direct inter eNodeBs connection is not available;
UP tunnels for data delivering;
eNodeB
eNodeB

Figure 12: S-GW logical connections and functions
Adapted from "A tutorial on LTE Evolved UTRAN (EUTRAN) and LTE Self Organizing Networks", Dhruv Sunil Shah

2.1.3.2.3. PDN-GW

The PDN GW is the gateway which terminates the SGi interface towards the Packet Data Network (PDN). The PDN-GW assigns an IP address to the UE which uses it to communicate with the external network and also performs the Dynamic Host Configuration Protocol (DHCP) functionality [14].

Each PDN-GW may be connected to several PCRFs, S-GW and external networks. If a UE is accessing more than one PDN, there may be more than one PDN-GW for that UE.

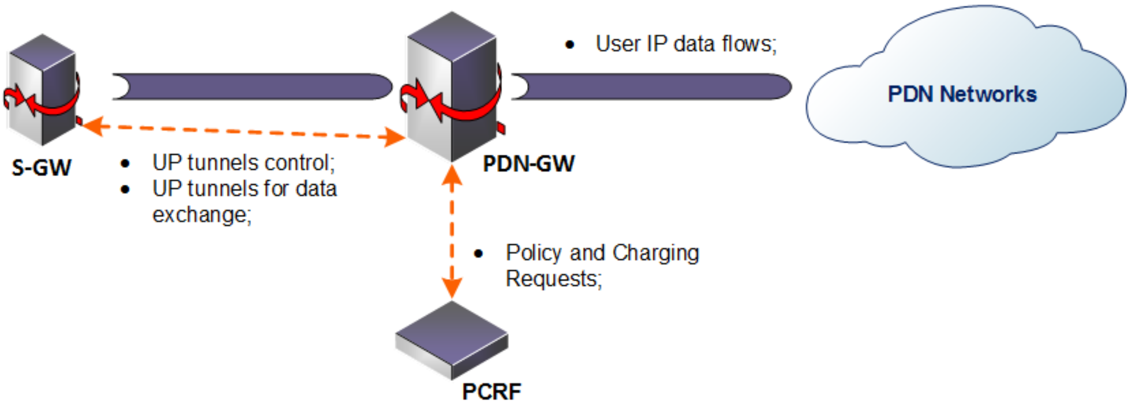
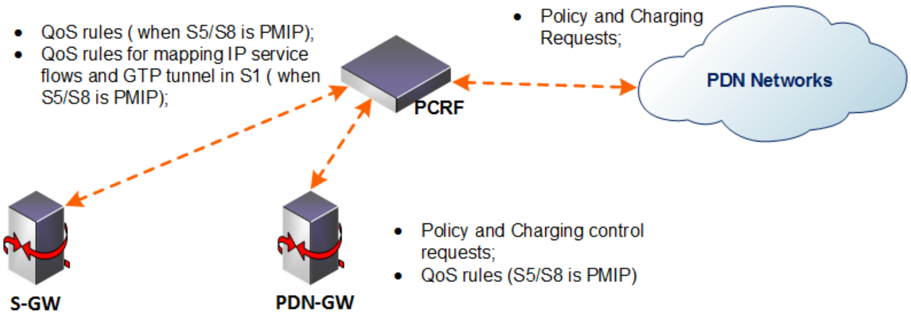


Figure 13: PDN-GW logical connection and main functions
Adapted from "A tutorial on LTE Evolved UTRAN (EUTRAN) and LTE Self Organizing Networks", Dhruv Sunil Shah

2.1.3.2.4. PCRF

The Policy and Charging Rules Function (PCRF) is the policy and charging control element of service data flows and IP bearer services.

Its decisions on how to handle the services are made based on the Quality of Service (QoS) and then send information to the PDN-GW and, if applicable, the S-GW can setup the appropriate bearers and policing (refer to *Figure 14*). Then the EPC bearers are set up based on those.



- QoS rules (when S5/S8 is PMIP);
- QoS rules for mapping IP service flows and GTP tunnel in S1 (when S5/S8 is PMIP);
- Policy and Charging Requests;
PCRF
PDN Networks
S-GW
PDN-GW
- Policy and Charging control requests;
- QoS rules (S5/S8 is PMIP)

Figure 14: PCRF main functions
Adapted from "A tutorial on LTE Evolved UTRAN (EUTRAN) and LTE Self Organizing Networks", Dhruv Sunil Shah

LTE advanced describes future LTE releases and its focus is on higher order capacity, using various techniques such as carrier aggregation, relay nodes, etc. But once the LTE advanced is not the aim architecture of this work no more will be discussed on this topic. For more detail on the LTE advanced refer to [15] and [7].

2.2. Network Management

The usage of computer based Operation Support System (OSS) started long before the emergence of the mobile network, beginning in 1960's and 1970's in the Bell System [17]. In those early days there were few applications that worth enough to justify the high costs of software and hardware, so the OSS were mainly to support paper based operations.

The reduction costs of the OSSs and high performance of operations made the OSS become a decisive NE in the network business platform.

Later on during the 90's, the massive usage of personal computers and the beginning of the mobile technologies were decisive in the evolution and reaching of the OSSs current state.

Network management is concerned with operation, maintenance, provision and administration of network systems. Operation is concerned with keeping the network and its services running with the minimum impact on the user experience. Maintenance is concerned with the state of the NEs (need for upgrades or repair). Provision is related to facilitate service requirements by means of network configuration. Finally, administration is concerned with network control.

2.2.1. Network Management Models

The network developments coupled with the introduction of new technologies and services in a possible multi-vendor environment required improvements regarding to network management. This, as consequence, made that different standardization organizations tried to develop a common network management model.

2.2.1.1. OSI NMM

One possible solution, and the most well-known, is the Open Systems Interconnection (OSI) Network Management Model (NMM) introduced by the International Standardization Organization (ISO).

There are three main components that need to be present in the elements of the management architecture in order to support a successful implementation of the OSI NMM [21], [1]:

- A functional component concerned with the activities performed in support of network management.
- A communication component focused in the matter of how the information is exchanged between the managed systems.
- An information component concerned with the management of the five major IT functional areas – Fault, Configuration, Accounting, Performance and Security management (FCAPS) – which provide rapid and consistent progress within each individual areas.

Although the ITU-T initially developed the concept of FCAPS it was in fact the ISO who applied the concept to data networks [33]. So the ISO model introduces the FCAPS functional areas as [20], [21]:

Fault Management (F)

Recognizing a problem in the telecommunication network is the first step in Fault Management. To have an effective fault management it is require detection, recognition, isolation and correction of faults that occurs in the network.

Configuration Management (C)

Monitor and configure NEs. Responsible of various functions such as: identify, exercise control over and collect data from NEs and provide configuration data to the various hardware and software versions of network elements (more detailed functions, refer to [20]).

Accounting Management (A)

This functional area is concerned with the collect of user usage statistics. These statistics can be controlled and improve the fairness of the network access and thus minimize network problems. Some important functions are [20]:

- **Usage measurement**.
- **Tariff and pricing** – Decision on the amount of payment for service use.

- **Collection and finance** – Receive payments, informs the user of payment dates, etc.
- **Enterprise control** – Supports the enterprise financial responsibilities.

Performance Management (P)

Evaluate and report about the behavior of telecommunication equipment and the effectiveness of the network or network element (NE). Gathers and analyze statistical data with the purpose of monitor and correct the behavior and effectiveness of the network, NE and/or other equipment. Moreover, assists in planning, provision and maintenance and quality measurement. The next functions are included in the performance management functional area [20]:

- **Performance quality assurance** – Includes quality measurements such as performance goals.

- **Performance monitoring** – Responsible for continuous collection of data concerning with NE performance.

- **Performance management control** – Includes setting thresholds and data analysis algorithms. It has no effect on the managed network.

- **Performance analysis** – Processing and analysis of collected performance records.

Security management (S)

Security management consists of two different areas:

- Security services for communications such as authentication, access control, data confidentiality, data integrity, etc. In addition a set of security mechanisms (e.g. event detection, security audit-trail management and security recovery) applicable to any communication are defined.

- Security event detection and reporting activities that may resides in a security violation (unauthorized user, physical tampering with equipment).

2.2.1.2. Telecommunication Management Network

The Telecommunication Management Network (TMN) has been defined by the ITU-T (International Telecommunication Union-Telecom Standardization) and has become the most widely used Network Management model and the main reference to the network management solutions providers.

The common requirements of the TMN is to support the network operator to manage (planning, maintenance, etc.) its network and services. Conceptually a TMN network interfaces different points of a telecommunication network to exchange information and control their operation. Thus, the TMN concept is to provide an organized architecture, with standardized interfaces, capable of interconnect various Operations Systems (OSs) and/or Network Elements (NEs) for exchange management information and to provide the network operator with access to the management information. The TMN includes a Data Communication Network (DCN) in charge of management traffic and so freeing the network from the management traffic.

Figure 15 illustrates in a very elementary way the relationship between the TMN network and telecommunication network managed by it.

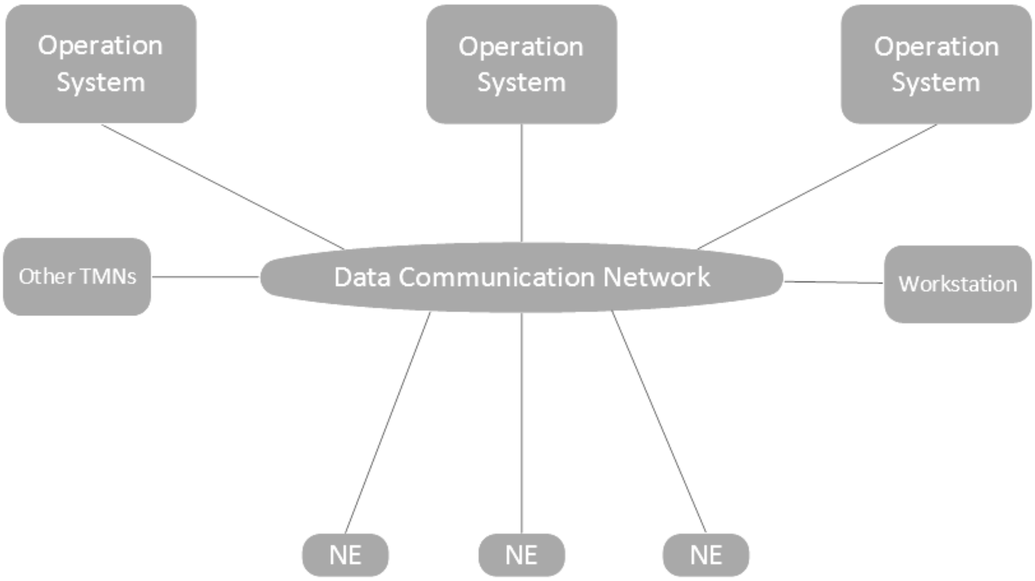


Figure 15: TMN and a telecommunication network
Source: "Principles for a telecommunications management network", ITU-T Recommendation M.3010 (2000)

2.2.1.2.1. Functional architecture

The TMN functional architecture is a framework of the management functionality. The functional architecture is constituted by the next elements:

- Function Blocks.
- Management Applications Functions (MAFs).
- TMN Management Function Sets and TMN Management Functions.
- Reference points.

Although all the above elements belong to the functional architecture, this document only discusses the function blocks and reference points. For more detailed information in all these elements please refer to [18].

2.2.1.2.1.1. Function blocks

Figure 16 illustrates the different types of TMN function blocks. Some of this function blocks can be partially in and partially out of a TMN as indicated in the figure.

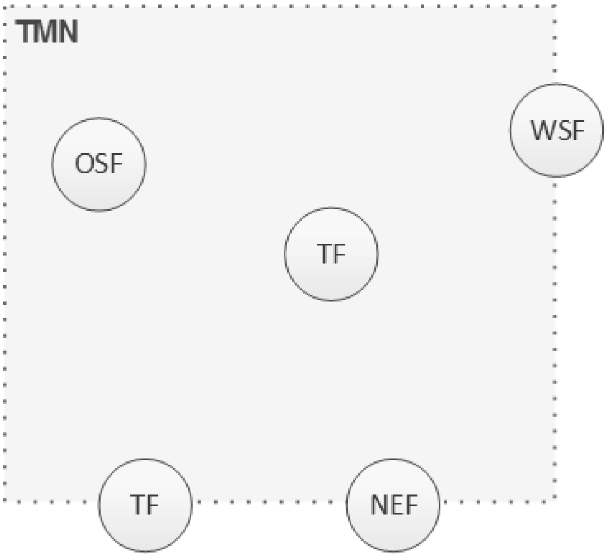


Figure 16: Function blocks
Source: "Principles for a telecommunications management network", ITU-T Recommendation M.3010 (2000)

Operation System Function (OSF)

The OSF block can be seen as the telecommunication manager function, i.e. processes the management information with the purpose of monitoring and control the telecommunication functions.

Network Element Function (NEF)

Communicates with the TMN with the purpose of be monitored, i.e. the NEF includes the telecommunication functions that need to be managed. Generally, the NEF function is the communication between Network Elements (NEs). These NEs constitute the network being monitored.

The NEF function has two parts, the one that perform telecommunication functions that doesn't count as part of the TMN (i.e. the NE) and the one that provides this representation in support of the TMN that is part of the TMN itself.

Workstation Function (WSF)

The WSF block translates between a TMN reference point and a non-TMN reference point (e.g. user), i.e. it provides the interface with the human user. As a result, a portion of this block is outside the TMN boundaries.

Transformation Function (TF)

The purpose of the TF block is to connect two entities with incompatible mechanisms of communication (e.g. different protocols). There are several circumstances in which the TF block is necessary [18]:

- Connect functional blocks with standardized, but different communication mechanisms within a TMN.

- Connect different TMN and non-TMN environments (at the boundaries).

2.2.1.2.1.2. Reference points

Reference point is a very important concept because it represents all the abilities that a determined function block is seeking from another function blocks. It also represents all the operation and/or notification that a function block can provide to a requesting function block. Generally speaking, a reference point defines one of several external views of functionality of a function block.

A TMN reference point only corresponds to a physical interface when the function blocks are implemented in different physical blocks.

Classes of reference points

- **q class**: between OSF, TF and NEF.
- **f class**: between OSF and WSF.
- **x class**: between OSFs of two different TMNs or between the OSF of a TMN and an entity equivalent to an OSF functionality of another network.
- **g class**: between WSF and users (non-TMN reference point).

Figure 17 illustrates all the possible pairs of function blocks associated via reference points [18].

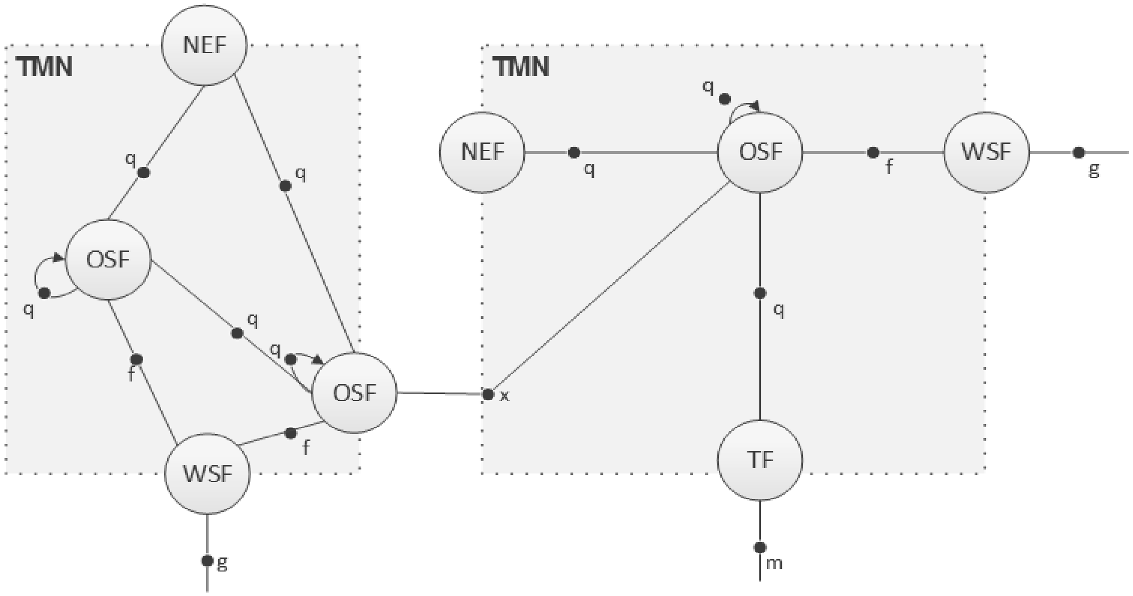


Figure 17: Reference Points

2.2.1.2.1.3. TMN logical layered architecture

A relevant aspect of the TMN architecture is the concept of layers. The TMN architecture has a strong relationship to the OSI standards and frameworks, thus the network management is grouped into functional areas such as FCAPS. In addition a Logical Layered Architecture (LLA) consists of five management layers.

Figure 18 illustrates the widely accepted layering, and the relationship between them. Each layer is responsible to provide the appropriate FCAPS functionality according to the layer definition. Each layer communicates with the layer above and below it.

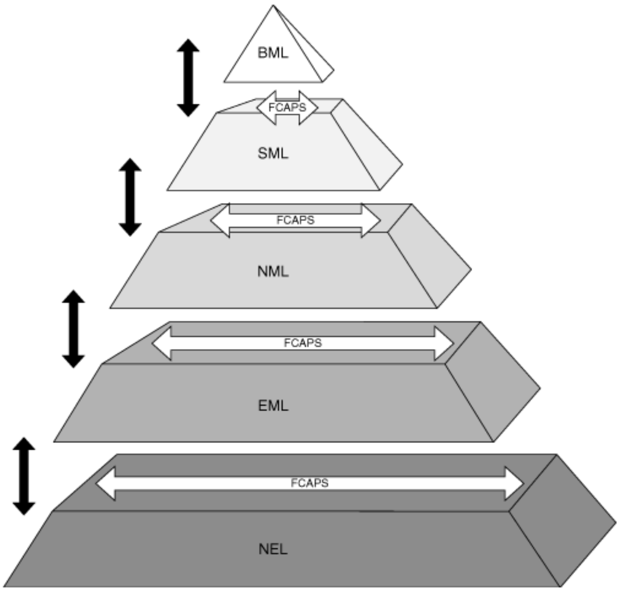


Figure 18: TMN logical layers
Source: "Network Management: Accounting and Performance Strategies", Benoit Claise, Cisco, June 2007

Element management layer

The Element Management Layer (EML) manages each NE on an individual or group basis. The EML can be constituted by one or more OSF elements that are individually responsible for a subset of network element functions.

The EML is responsible for the following functions [18]:

- Control and coordinate NEs in an individual NEF basis. Here, the OSFs support the interaction between NML and EML by processing the information to be exchanged between network OSFs and individual NEFs. OSF elements should provide full access to NE functionality.

- Maintain important data (e.g. statistical) about the elements within its scope of control.

Network management layer

The Network Management Layer (NML) has the responsibility to manage the network with the support of the EML. In this layer a more wide-ranging view and management of the network is considered, i.e., a wide geographical area is managed. Complete visibility of the whole network is typical and a technological independent view will be provided by the SML.

The following roles are the main duty of the NML [18]:

- Control and coordinate the NEs within its scope.
- Maintain the network capabilities.
- Provision, ceasing or modification of network capabilities to provide a service to the costumer.
- Maintain important data (e.g. statistical) about the network within its scope and work with the SML on performance, usage, availability, etc.

The NML provides functionality to manage and control and support the network demands made by the SML. The NML knows the available resources, how to control them, how these are related and geographical allocated. Furthermore, this layer is responsible for the network performance and must control the network capabilities and capacity to achieve the best results of accessibility and QoS.

Service management layer

The Service Management Layer (SML) is concerned with the contractual aspects of services provided to the customers.

Principal roles of SML [18]:

- Customer facing[2] and interfacing with other Public Telecommunication Operators (PTO).
- Interaction with service providers.
- Interaction between services.

Business management layer

The Business Management Layer (BML) is part of the overall enterprise management. While the service and network management layers are the optimal utilization of the network resources, the business management layer is the optimal investment and use of new resources.

The main roles of the BML are [18]:

- Support of the decision process for the optimal investment and use of new resources.
- Support the management of Operation, Administration and Maintenance (OA&M) budget.
- Support the supply and demand of OA&M related manpower.
- Maintain aggregate data about the total enterprise.

[2] Facing is the point of contact with the customers for all service transaction including service provision/cessation, accounts, QOS, fault reporting, etc.

2.2.1.2.2. Physical Architecture

So far, this document has discussed the TMN functional architecture. However TMN also defines a physical architecture. The TMN physical architecture shows how the TMN functions described in the functional architecture can be implemented into physical equipment.

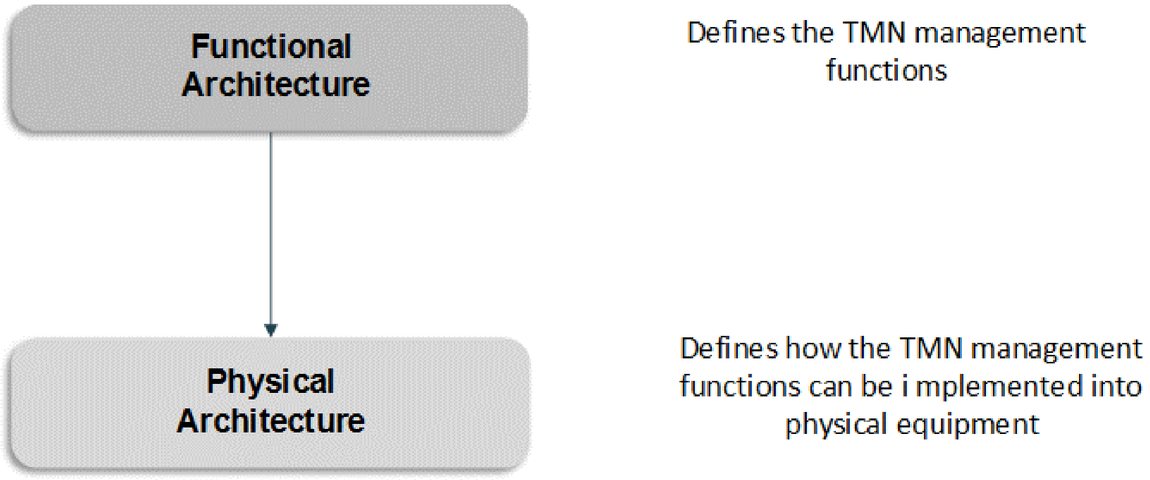

Functional Architecture
Defines the TMN management functions
Physical Architecture
Defines how the TMN management functions can be i mplemented into physical equipment

Figure 19: TMN related architectures

The physical architecture shows how the function blocks should be mapped upon building blocks (physical equipment) and reference points upon interfaces (refer to *Figure 20*). Note that a building block can map multiple function blocks and a function block may contain multiple functional components.

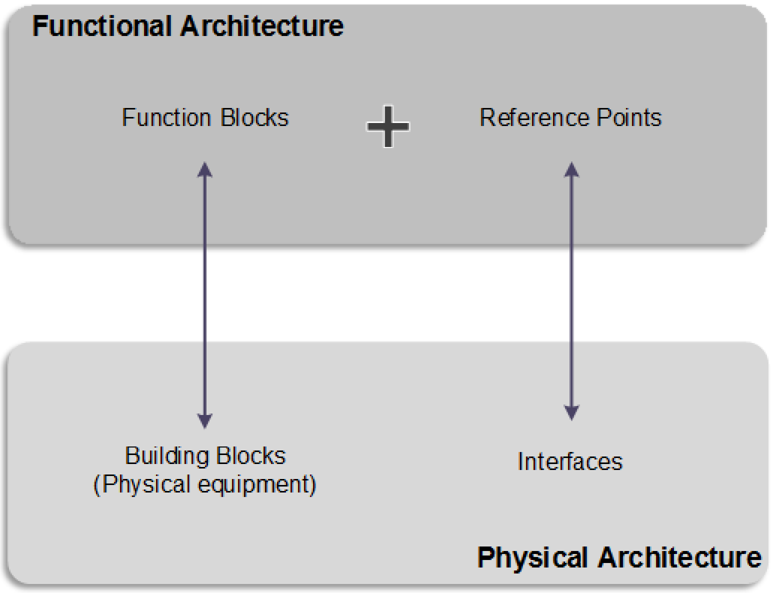


Figure 20: Relation between architectures

Building blocks always implement the function blocks of the same name, e.g. Network Element (NE) implements the Network Element Function (NEF).

As already mentioned, it is possible to implement multiple function blocks into a single building block, e.g. the Operations System (OS) building block may implement multiple OSFs, but may also be used to implement an OSF, MF, and WSF. The name of the building block is determined by the mandatory function block. *Table 2* shows which function blocks can be mapped to which building blocks [18][19][19].

	NEF	MF	QAF	OSF	WSF
NE	M	O	O	O	O*
MD		M	O	O	O
QA			M		
OS		O	O	M	O
WS					M
DCN					

Table 2: Mapping function blocks into building blocks

M- mandatory
O- Optional
O*- may only be present if OSF or MF is also present.

A special building block is the Data Communication Network (DCN) that doesn't implement a function block as the others building blocks. In fact, the DCN is used by the other building blocks to exchange information, i.e. it acts as a transport network.

Interfaces

Interfaces are related to the physical implementation of the reference points. Normally reference points and interfaces have one to one mapping and the name of the interfaces are the same as the related reference point but in capital letters. Some reference points do not have associated interfaces, such as:

- Reference points that interconnect function blocks inside a single block.

- Reference points lying outside the TMN (e.g. m and g, refer to
- *Figure* **21**). Implementation of these reference points are outside of the TMN scope.

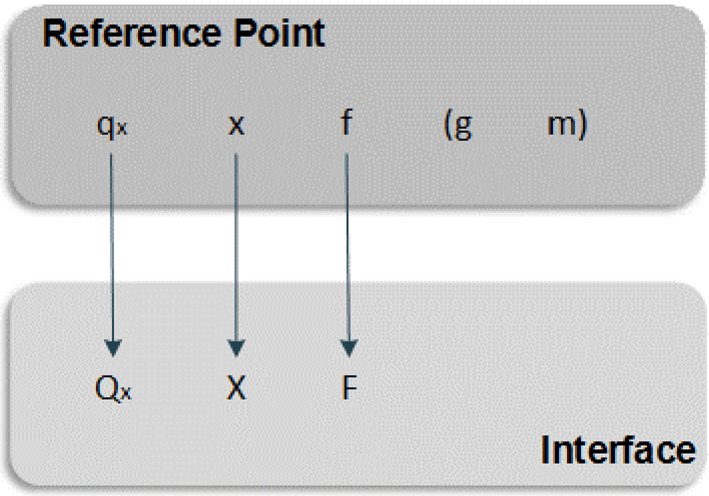


Figure 21: Mapping reference points upon interfaces

2.2.2. Performance Management

The main purpose of the Performance Management (PM) “field” is to collect network data information to support several activities, such as:

- Detection and identification of problems in the network as soon as possible.
- Continuously monitoring the network state.
- Assure optimum services and best QoE delivering to the subscribers.
- Verify the network configuration both logical and physical.
- Monitor the user behavior.

There are two types of performance management depending on how the information is collected from the network and used by the PM entities [23]:

- **Performance Monitoring**: The performance monitoring is used to detect severe problems that demand fast intervention. This short time intervention requires data collection with short time intervals because the speed is a crucial aspect on this type of performance.
- **Performance reporting**: The performance reporting is a long-term type of performance that provides information about the network over a large period of time. This type of performance is commonly used for example to support network planning, troubleshooting, etc.

The NEs produce a lot of PM data information which commonly covers different aspects of the network, e.g.:

- Network Configuration
- QoS
- Resource availability
- Traffic levels within the network

Due to the large amount of data, it is impossible for the management team to process all the information into practical reports. Thus, the PM applications have the work of data filtering and production of reports that suit the needs of the user (e.g. marketing area). Depending on the target audience, these reports can be tables, graphical reports, etc.

Figure 22 presents different combinations of reports. Each chunk of the pyramid corresponds to a report group that the PM applications can provide to the different groups.

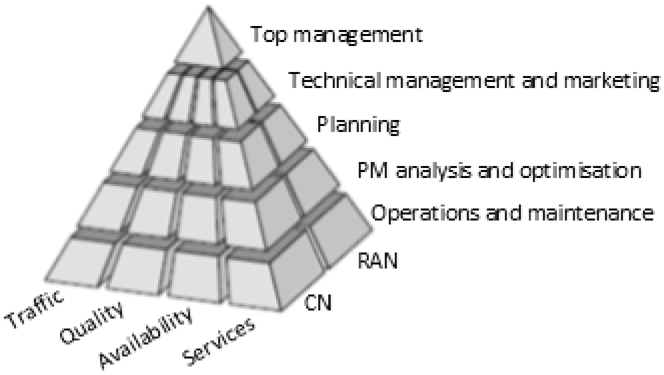

Top management
Technical management and marketing
Planning
PM analysis and optimisation
Operations and maintenance
RAN
CN
Traffic
Quality
Availability
Services

Figure 22: Network Management Level
Source: "End User Behaviour and Performance Analysis in 3G Networks", Igor Pais, Universidade de Aveiro, 2009

This different type of information can be used in several ways, depending on what the manager operator wants to evaluate.

2.2.3. KPIs

The Key Performance Indicators (KPIs) are the most important indicators expressing the state of the network. These can be used not only to detect problems but also for a big diversity of subjects, e.g. analyzing performance trends, marketing, etc.

The different NEs produce a large amount of counters related to their performance. These counters often provide data on a very specific aspect of the network, which make that the analysis of these counters separately (i.e. one by one) becomes practically impossible. The best approach to study the state of the network is to create reports. Reports normally contain more than one KPI based on several Performance Indicators (PIs). The PIs are simple counters that belong to a measurement in the network and a measurement can include several counters. Depending on the target audience and goals of the network operator, different types of reports can be created, with different KPIs and different formats such as: diagram, map, tabular, etc.

Figure 23 represents the process of creating a report. Note that the KPIs are generated from different PIs and the reports are generated from different KPIs depending on the needs of the operator.

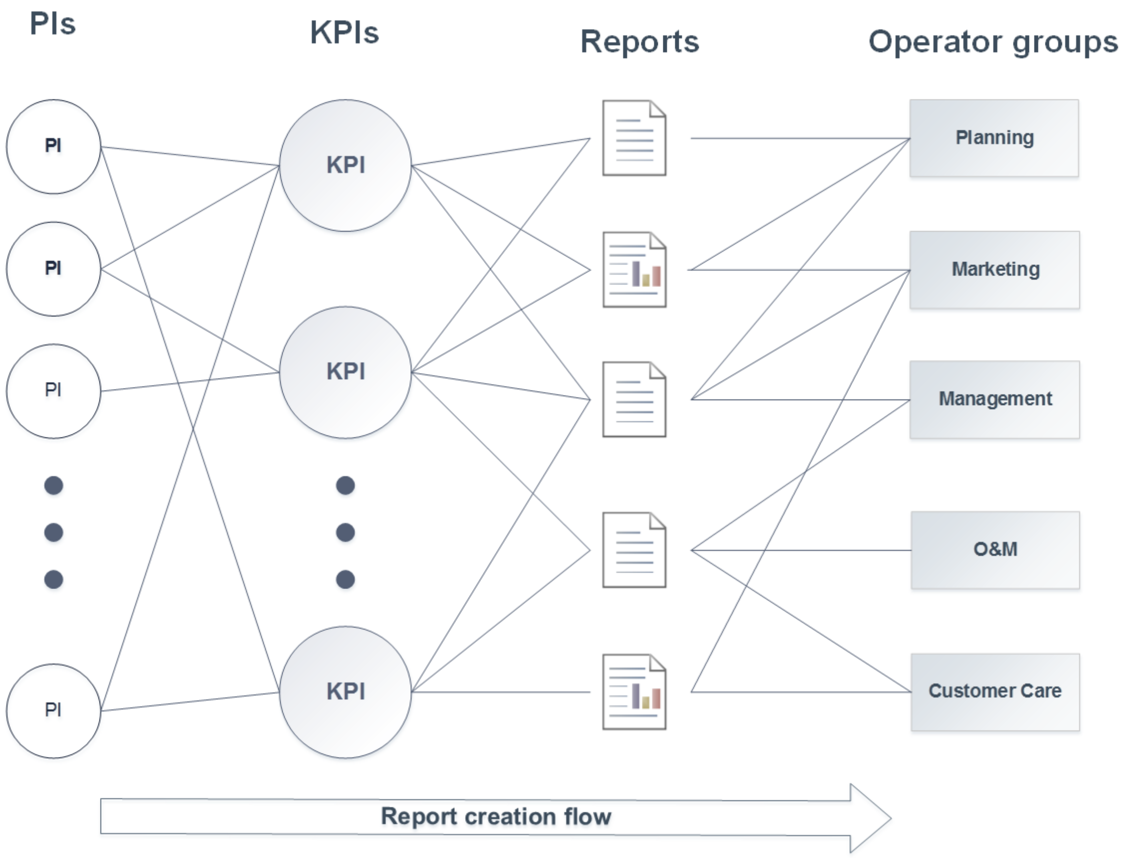


Figure 23: Report creation
Source: "End User Behaviour and Performance Analysis in 3G Networks", Igor Pais, Universidade de Aveiro, 2009

KPIs use counters from one or more measurements or use a formula based on several counters. As it is evident these counters in which the KPIs are based must be activated.

The 3GPP specifications allow different implementation that lead to non-uniform counter updates or definitions between vendors. For some KPIs, even if the KPI definitions could be common across equipment vendors and network operators, the related calculation formula and/or NEs counters may not be. Therefore, there are differences between vendors and changes over time and technologies. Concluding, these high level KPIs should be well described, but the details about which measurements (counters) or formulas to use should not be standardized [24].

When defining a KPI some rules should be followed. To do this, the 3GPP had created a template [25]. An important field of this template is the KPI category/group. When KPIs are defined, these should be classified into a category, *Table 3* illustrates some of these categories. The practice of classify the KPIs into groups supports the creation of reports by simplifying the search for the KPIs needed for the report.

KPI category	KPIs
Accessibility	KPIs related with the access phase: • RAB Establishment Success Rate • RRC Connection Establishment Success Rate • Call Setup Success Ratio
Retainability	KPIs related with cases of calls failing after the access phase: • RAB Abnormal Release Rate • Combined 2G 3G Call Drop Ratio
Integrity	KPIs related with transmission and retransmission errors and quality in the cells: • Average BLER (Block Error Rate) • HSDPA congestion Rate in Iub
Utilization	KPIs related with the capacity of the network: • Percentage of Established RABs • HSDPA throughput • Bit Rate Utilization
Mobility	KPIs related with all handover procedures and cell changes: • Soft Handover Success Rate • Intra-System Hard Handover Failure Rate

Table 3: KPI categories

Later in this work, KPIs belonging to utilization and mobility categories will be used to accomplish the final goal (RAN capacity forecasting).

2.2.4. Configuration Management

Configuration Management (CM) is concerned with the monitoring of the network configuration and any modification that take place. This is a crucial area since many issues arises as a direct consequence of network modifications. A decent CM strategy has to be concerned with all modification to the network software and/or hardware.

The possibility of change the network configuration is a crucial aspect in a network. It is essential for network operators to be able to make changes and correct problems very quickly with a minimum of effort and essentially, if possible, without affecting the user services.

This chapter starts with a brief explanation of the reasons why a network needs its CM strategy and the methods used by it. After that it presents some administrative aspects that must be followed. Last, presents a general idea about the techniques and methods used in CM.

2.2.4.1. Why CM?

A network development can be described in three phases and once the first one is complete, the network will cycle between the second and third phase. This is known as the network life-cycle [26][26]:

1. Install the 3G network and put into service.

2. The network can be modified to satisfy some short terms requirements, these modifications disturb the stability of the network and optimization actions are required again.

3. Regard to performance, capacity and user experience, the network is adjusted to meet long term requirements.

Short term requirements (related to phase 2) – The operator needs to be able to solve short terms incidents, e.g. network elements (NEs) that need a reconfiguration in its parameters and adapt to the day-to-day requirements.

Network enhancements (related to phase 3) – The demand for new services, subscriber requirements and the constant traffic growth in the network compels the operators to add new features and improve its infrastructure. Therefore, the configuration management goal is to ensure the best network configuration to meet its needs. These procedures are used for long term strategy.

Relatively to the two last phases of the network life-cycle, there are two possible techniques to optimize the network [26]:

Network Update – Software (SW) or equipment will be updated and/or replaced without adding new facilities or resources to the network. This procedure shall not disturb the network. Thus, techniques such as SW data downloading in parallel with on-going traffic can be used [26].

Network Upgrade – New features and/or facilities are developed, i.e. creation, deletion and/or modification of NEs and/or network resources.

2.2.4.1.1. Administrative Aspects

During the NE(s) creation, deletion or modification the operator should make sure that, there is not an uncontrollable impact in the network. To prevent this to happen, the network management uses some preventive methods [26]:

- **Security**: Network configuration changes should be performed only by authorized personal. The operator should employ mechanisms to control the access to the network configuration.

- **Data validation**: The NEs and the CM mechanisms have the responsibility of validate the data transferred in the system.

- **Consistency check**: Some information has to be in consistency between the local resources to be managed and configured, e.g. which information will be exchanged, how the information is transferred, etc.

2.2.4.2. CM systems and methods

In order to achieve short terms and long terms requirements the CM use different techniques (depending on the requirement) to configure the NEs.

The possibility to manage the NEs configuration is provided by computer systems used for this purpose, typically called **Operation Support System (OSS)**. The OSS holds the main CM role, allowing the execution of operations in the Radio Network, such as small day-to-day parameters changes, expansion procedures, site creations, etc.

In some cases, the radio network configuration can be modified using a graphical interface provided by an **Operation and Management Server (OMS)**, also known as the OMS GUI. This OMS GUI give to the user an overview of the network managed objects. All the modifications performed by the OMS GUI are updated into the OSS DBs so the OSS has the correct information. The OSS is the "master" of all configuration information.

Figure 24 shows a very simple illustration of a CM architecture using the RNC and BTS as the target NEs for the configuration data.

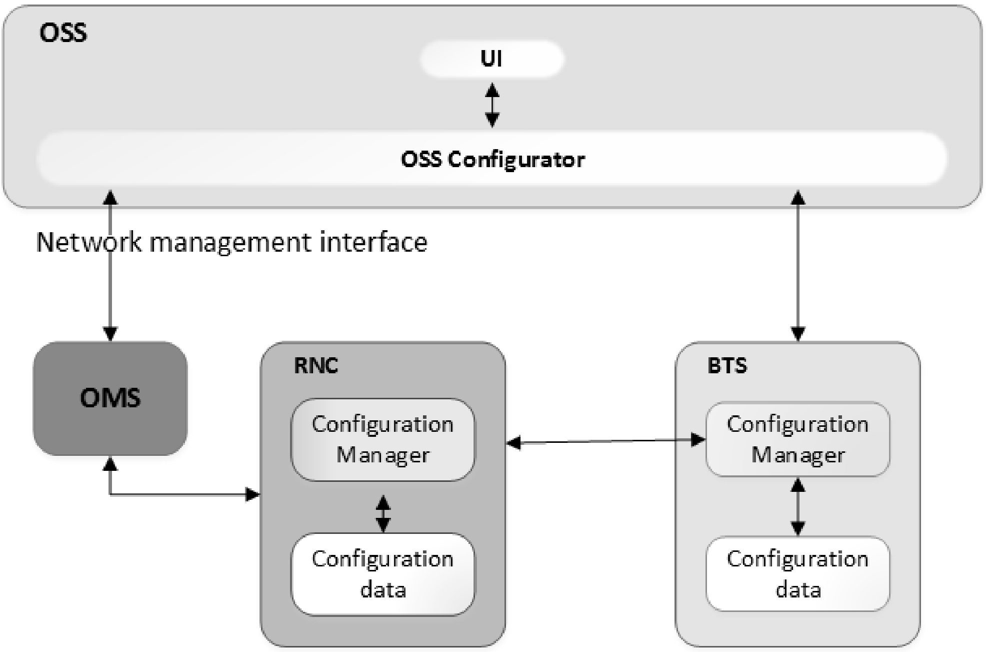


Figure 24: CM Architecture (example)

As it is possible to see in the figure the CM operation on the RAN involves some essential elements. The roles of these elements are:

User – Responsible to start the operation using the OSS or OMS depending on the operation.

OSS – It is the most important element in a CM operation. Responsible for initiate the operations and store in its databases, the configuration data of all NEs belonging to its network (safety feature).

OMS – Depending on the type of operation, it can work like a mediator between the OSS and the NEs or can be used to trigger the operations locally or remotely.

RNC – Target of configuration data upload, download and activation operations. Usually the RNCs have a local data storage intended for the configurations.

BTS – Target of configuration upload, download and activation operations. Normally, also provides a local storage for the configurations.

2.2.4.2.1. CM Operation

There are several stages in a CM operation that can be identified, some of that are next presented in a general view.

Download – A file is transferred to the respective network element. Once in the NE, these file are validated, if some errors occur these are reported to the OSS.

Activation – Once the file is downloaded to the NE, it is activated.

Upload – The configurations can be uploaded to the OSS and stored in it without any disturbs in the network performance. This is a safety feature in case of something goes wrong with new activations. So, an upload should be done before modifying the configuration of a network element.

Several other stages can exist on a CM operation of network vendor. The definitions above are only a small enlightenment about the CM operation. The CM operation strategy has differences from vendor to vendor and the deep analysis of these operations will not bring added value to this work.

3. Forecasting Methodology

Forecasting is used to estimate the future values of some variable of interest. The process of forecasting is required in several situations: climate changes, electricity demand, network traffic volume, etc. Regardless of data type (percentage of humidity, erlangs, etc.) and prediction horizon, forecasting is a very important aid to a good and strategic plan. Bad forecasts can be dangerous for an organization, consider the following famous predictions about computing [27]:

- *"Computers in the future may weigh no more than 1.5 tons."* (Popular Mechanics, 1949)
- *"There is no reason anyone would want a computer in their home."* (President, DEC, 1977)

The last one of these statements was made three years before the release of the first personal computer by IBM. Obviously, forecasting is a difficult activity and the businesses that do it well have a big advantage over those who forecasts fail.

This work aims to help the mobile operators to prevent capacity issues in their networks. The idea is to forecast when a NE (in this study, the RNC) and its objects (physical and logical) reach their maximum capacity, allowing the operators to prevent this to happen and avoid revenue losses.

With this problem in mind, this chapter studies some of the most widely used forecasting techniques and compares their performance when applied to data related with the RAN capacity (e.g. voice and packet traffic). In order to achieve some conclusions, it will be study which method gives the better results, under which circumstances and if the complexity justifies the results.

This chapter starts for introducing some important concepts related with the problematic of forecasting. Later on, it adopts a structure of type theoretical analysis followed by results analysis.

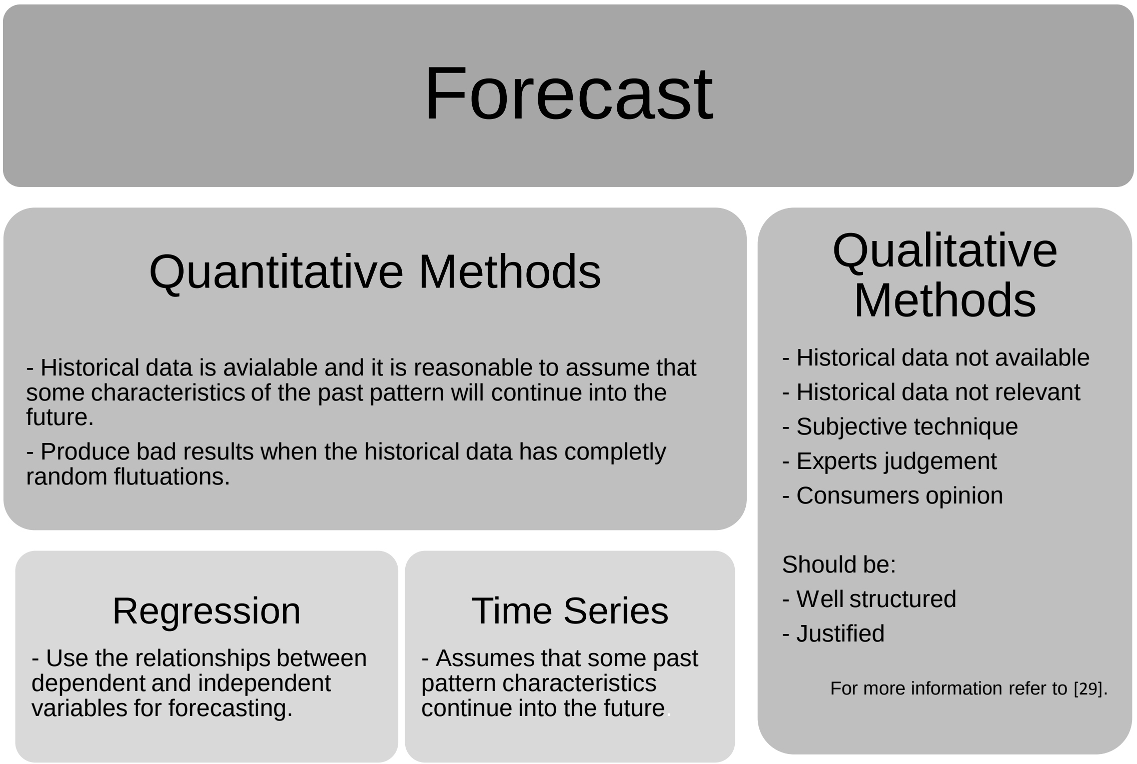
The data used to test and study the different methods in this work was taken from a real network operator.

3.1. Introductory Concepts

Before any further discussion on the forecasting methods, it is necessary to provide the reader with some concepts and techniques that will be useful in subsequent chapters.

3.1.1. Methods

More and more, the organizations start to store their data to create profiles, patterns and use it for several purposes, including forecasting. A very important step is to identify if the use of a forecast model will produce an accurately prediction. *Figure 25* shows a brief overview on the main forecast types:



Forecast

Quantitative Methods

- Historical data is avialable and it is reasonable to assume that some characteristics of the past pattern will continue into the future.

- Produce bad results when the historical data has completly random flutuations.

Regression

- Use the relationships between dependent and independent variables for forecasting.

Time Series

- Assumes that some past pattern characteristics continue into the future.

Qualitative Methods

- Historical data not available
- Historical data not relevant
- Subjective technique
- Experts judgement
- Consumers opinion

Should be:

- Well structured
- Justified

For more information refer to [29].

Figure 25: Forecasting methods

Since we have access to the databases of real mobile operators, this work will focus on the quantitative methods.

There's a wide range of quantitative forecasting methods, each one with its own complexity, accuracies and costs that should be taken into account when choosing which method to use.

The next three very simple methods are a good introduction and benchmark for more complex methods discussed later.

A. Naïve method

This method relies on set the forecasts of all future values to be equal to the last observation.

B. Seasonal naïve method

The seasonal naïve is very similar to the naïve method, but instead of all forecast be equal to last observation, the forecast values are equal to the last observation from the same season. For example, for daily data the forecast for all Monday values are equal to the last observed Monday value.

C. Average method

The average method states that all forecast values are equal to the mean of the historical data. Let and, then:

$$\hat{y} = \bar{y} = \frac{(y_1 + \cdots + y_T)}{T} \qquad (3.1)$$

- $\hat{y}$ are the forecast values;
- T is the total number of observations;

Figure 26 shows the three methods applied to daily average voice throughput.

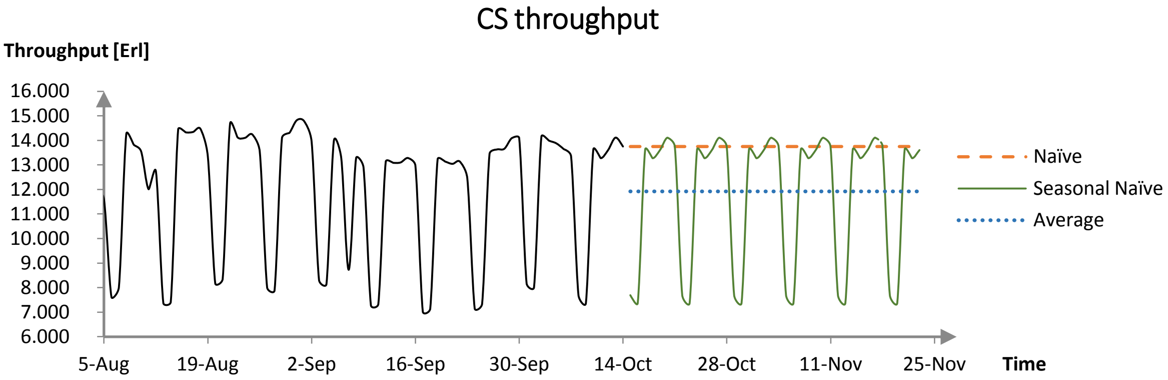


Figure 26: Daily average voice throughput in Iub interface

3.1.2. Time Series Components

A time series can exhibit a big variety of patterns, and it is helpful to classify some of the patterns and behaviors that can be observed. The classical decomposition helps to classify different behaviors of a time series. This method of decomposition states that each time series consists of the following components:

- **Trend (T)** – A trend exists when there is a long-term increase or decrease in the data.
- **Cyclical variation (C)** – Rises and falls that are not of a fixed time period.
- **Seasonality (S)** – Patterns of change influenced by seasonal factors. Seasonality is of a fix and known time period.
- **Irregularity (I)** – What is left over when the other components are extracted from the data, i.e. it is the residue of the data.

Seasonality can be confused with cycle, however they are quite different. While the seasonality has a fixed and known period of the changing patterns, the cycles do not have fixed time periods. Normally the length of the cycles is longer than the seasonal pattern.

The classical decomposition is the base for almost all the methods study in this work and can describe almost every type of data. Furthermore, it helps to understand time series and also improve forecasts.

3.1.3. Graphics

The first thing to do in any data analysis is to plot the data. The analysis of the graphs allows identifying several characteristics of the data, such as: unusual observations, patterns, relationships between observations, etc.

A. Time plots

For a time series analysis, the most simply and obvious plot to start is a time plot, where the consecutive observations are plotted as function of the consecutive time steps. Link these points by a straight line helps in the analysis of the data.

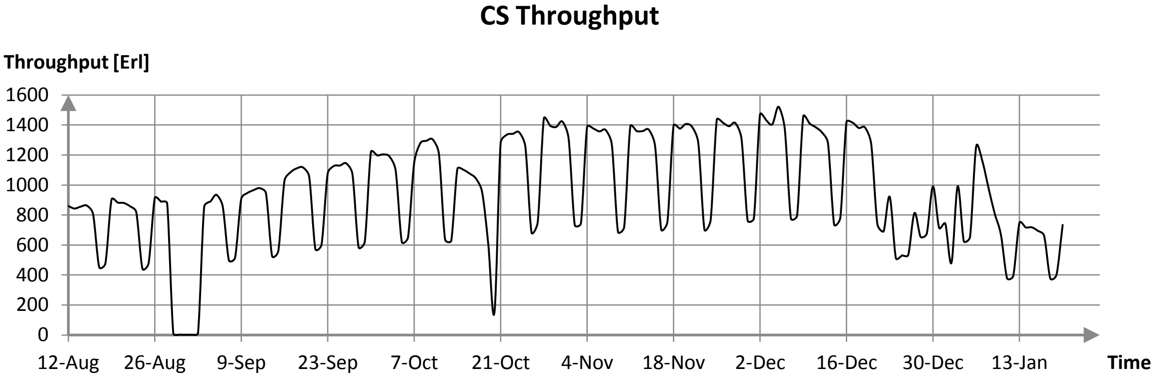


Figure 27: Daily average voice throughput

Figure 27 represents the time plot for the daily average of CS traffic in the Iub interface. A quick analysis instantly reveals some interesting properties in the data:

- The data exhibits a clear pattern.
- It is possible to notice an increase trend until around 20 December and then a fall down after that, probably due to the holiday season.
- There are unusual values around August 30 and October 20.

B. Scatter plots

A scatter plot is simply a graphical representation using Cartesian coordinates to display values for two variables of a data set. Scatter plots are very useful to study the relationship between variables. *Figure 28* shows the relationship between the RNC user plane fill factor and ICSU CPU load.

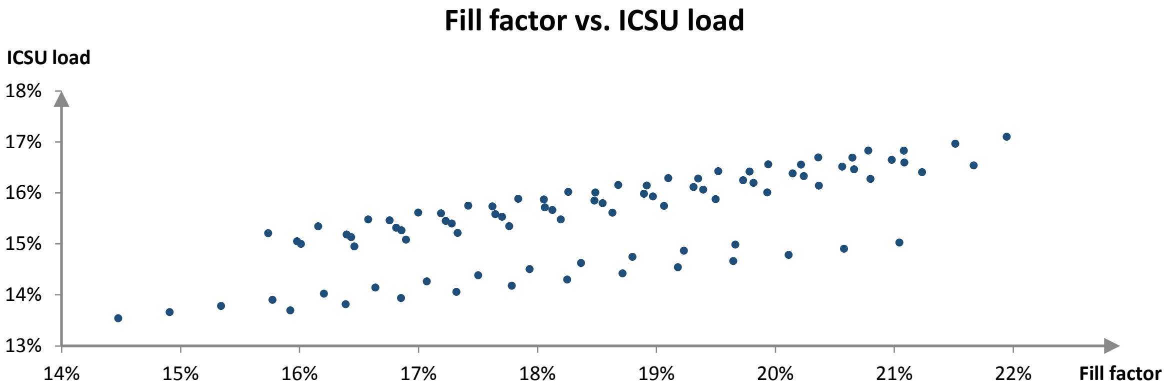


Figure 28: RNC user plane fill rate vs. ICSU CPU load

This is one of the most important relationships and it will be studied later on. For now and for example purposes, the scatter plot shows that these two variables have a linear relationship, as the fill factor increases the CPU load also increases. If this trend is kept, the fill factor will reach its maximum more rapidly than the CPU load.

3.1.4. Forecast Evaluation

A very important stage of using a quantitative forecasting method is to learn about the method performance in that type of data. This chapter presents some of the most commonly measures used to evaluate forecasting methods accuracy and goodness of fit.

3.1.4.1. Sum of Squared Errors

The forecast error is simply the subtraction between the real value and the forecast value, i.e. $e_t = y_t - \hat{y}_t$. A common measure based on forecast errors is the Sum of Squared Errors (SSE):

$$\mathrm{SSE} = \sum_{t=1}^{n} e^2(t) \qquad (3.2)$$

Some forecast methods discussed in this work are fitted based on the minimization of the SSE value. There are other criteria to fit the models, such as the Akaike Information Criterion (AIC) discussed in *3.1.4.4 Akaike Information Criterion.*

3.1.4.2. Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is the most commonly used measure to compare performance between different data sets. In a simple way, the MAPE is the average of the relative errors (in percentage). It is described by the following expression:

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(\left|\frac{\mathrm{y_t} - \hat{\mathrm{y}}_\mathrm{t}}{\mathrm{y_t}}\right| \times 100\right) \qquad (3.3)$$

- y_t are the actual values;
- $\hat{y}_t$ are the predicted values;
- n is the number of observations;

The forecasting method that minimizes the MAPE is considered the preferred method for forecasting.

The MAPE concept although being scale-independent and very simple, it has two main drawbacks in practical applications:

- If there are zero values, there will be a division by zero.
- A null MAPE means a perfect fit of the model. But there is no upper limit for MAPE values.

3.1.4.3. Durbin-Watson

The Durbin-Watson test computes a number that tests for autocorrelation in the residuals from a statistical regression analysis. It is defined by the following expression:

$$w = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \qquad (3.4)$$

- e_t is the residual associated to the observation at time t;
- T is the total number of observations;

The value of w is always between 0 and 4. If the value is equal to 2 it means that there is no correlation [29]. Low values of w indicates that the values for the successive error terms are closely to one another (in value) or positively correlated and values approaching $w = 4$ indicates that the successive error terms are very different, i.e. negatively correlated.

3.1.4.4. Akaike Information Criterion

The AIC criterion is a measure of relative quality of a model, e.g. it tells if an ARMA(3,1) fits best the data than an ARMA(1,1). Thus, this criterion does not tell nothing about the quality of the model in an absolute sense, i.e. if the models to be compared fit poorly the data, AIC doesn't informs about that. AIC calculates a trade-off between the goodness of fit and complexity of the model, thereby providing a means for model selection. In the equation form:

$$\mathrm{AIC} = \ln\left(\hat{\sigma}_{\varepsilon}^{2}\right) + \frac{2\mathrm{k}}{\mathrm{T}} \tag{3.5}$$

- k is the number of the estimated ARMA parameters $(p+q)$;
- T is the number of observations;
- $\hat{\sigma}_{\varepsilon}^{2}$ is the residuals variance;

Essentially, the model that presents the lower AIC is the one that best fits the time series in question. Nevertheless, there is the possibility that the model that produces the best AIC, doesn't produce the best forecasting results.

3.1.4.5. Coefficient of Determination

The coefficient of determination, usually referred as R^2 is a measure of the model adjustment to a set of observations. It is commonly used to evaluate linear regression, however in this work it will be calculated also for the other models. It is defined by the following expression:

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_{\mathrm{i}} - \bar{\mathrm{y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_{\mathrm{i}} - \bar{\mathrm{y}})^2} \tag{3.6}$$

- y_{i} are the observations;
- $\hat{\mathrm{y}}_{\mathrm{i}}$ are the estimates (predictions);
- $\bar{\mathrm{y}}$ is the mean of the observations;
- n is the number of observations.

The coefficient of determination varies between 0 and 1 indicating how much, in percentage, the model can explain the observations. The bigger the coefficient, the better the fit.

Note however, that a high R^2 doesn't mean a good model for forecasting and the inverse is also true. This type of evaluation has to be made depending on the type of data being analyzed.

3.1.4.6. Outliers and influential observations

When an observation takes an extreme value compared with the majority of the observations it's called an outlier. An outlier may be due to a measurement error in the NE and must be removed from the data set.

The bad behavior that the methods studied in this work can exhibit in the presence of outliers requires the use of a technique to manage outliers. There are several techniques to handle outliers, e.g. use the scatter plot to detect unusual observations. The use of the scatter plot however, is not the best option because it is not an automated technique and in some type of data is difficult to identify outliers.

3.1.4.6.1. A technique to deal with outliers

Thus, the technique used to remove outliers from the data (CS and PS throughput, unit load, etc.) is based on the mean ($\bar{x}$) and standard deviations (σ) of the data:

Standard deviation (σ) – The standard deviation indicates the amount of deviation of the data. A low standard deviation means that the data points are close to the mean and a high standard deviation indicates that the data points are spread out over a large interval.

$$\sigma = \sqrt{\frac{1}{N}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2]} \qquad (3.7)$$

$$\bar{x} = \tfrac{1}{N}(x_1 + x_2 + \cdots + x_N);$$

- $x_1,\ x_2, \ldots ,\ x_N$ are the data set;
- N is the number of observations;

Mean – The central value of a set of values. Specifically is the sum of the values divided by the number of values.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} \qquad (3.8)$$

- $x_1,\ x_2, \ldots ,\ x_N$ are the data set;
- N is the number of observations;

So, the proposal of this technique is to calculate lower and upper limits for each data point, compare the value with these limits and if the data point exceed one of those limits it is replaced by the value of the respective day from the week before (weekly seasonality).

$$x_L = \bar{x} - 2 \times \sigma \qquad (3.9)$$

$$x_U = \bar{x} + 2 \times \sigma \qquad (3.10)$$

- x_L denotes the lower limit;
- x_U denotes the upper limit;

In this particular work where the data set can be enormous and exhibit big seasonal variations (e.g. daily data during two years present variations due to the Christmas season and others), if the mean ($\bar{x}$) and standard deviation (σ) are calculated using the two year daily data, the seasonal variations can be judge as outliers by this technique.

To avoid the situation described in the last paragraph, instead of compute the mean and standard deviation for all the data, they are computed for each data point using a sliding window. Thus, let k be the size of the sliding widow, the equations (*3.7*), (*3.8*), (*3.9*) and (*3.10*) respectively become:

$$\sigma(t) = \sqrt{\frac{1}{k}[(x_{t-k} - \bar{x}\,)^2 + (x_{t-k+1} - \bar{x}\,)^2 + \cdots + (x_{t-1} - \bar{x}\,)^2]}$$

$$\bar{x}\,(t) = \frac{x_{t-k} + x_{t-k+1} + \cdots + x_{t-1}}{k}$$

$$x_L(t) = \bar{x}(t) - 2 \times \sigma(t)$$

$$x_U(t) = \bar{x}(t) + 2 \times \sigma(t)$$

In the limits, the deviation established here is two standard deviations (σ) from the mean ($\bar{x}$). This number of standard deviations was obtained iteratively.

Figure 29 illustrates a sliding window with a size of three data points. Here it is easy to understand how to calculate the mean and standard deviation for each data point using the sliding window. This figure is used only for example purposes.

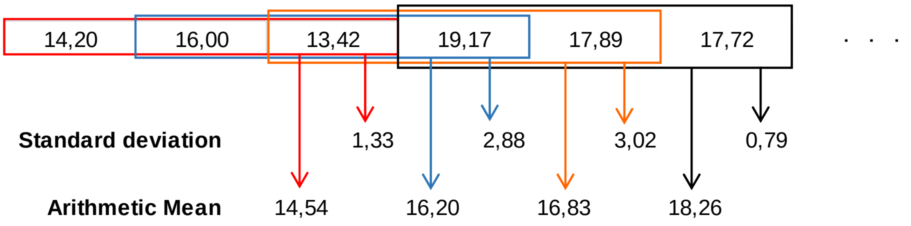


Figure 29: Mean and Standard Deviation Sliding Window

This technique has a problem, because the first $k-1$ observations don't have a mean and standard deviation associated and it is impossible to say if there are outliers or not in the first $k-1$ observations. A simple way, but efficient to solve this problem is to calculate the mean and standard deviation based on the observations ahead.

Figure 30 presents the daily average CS throughput in the Iub interface of a real operator. This is a good example of the good results and improvements that this technique can bring for the matter of forecasting.

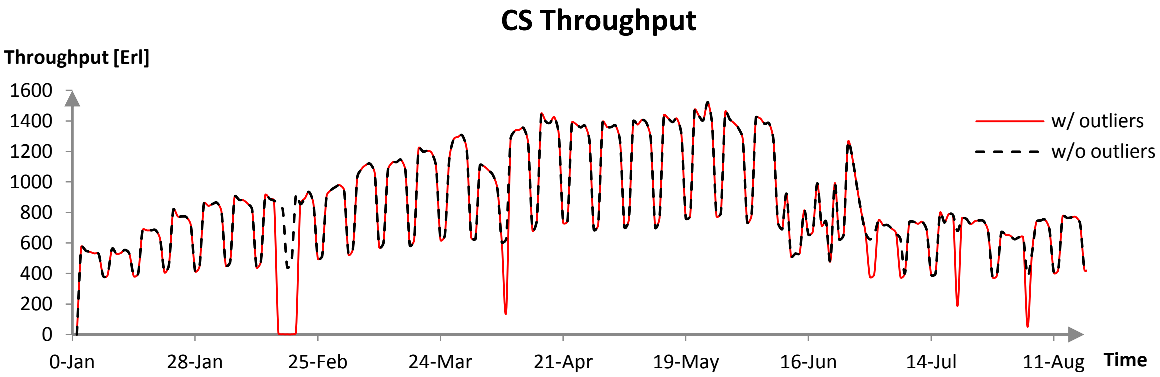


Figure 30: AMR daily throughput

Around Jun-16 there is an unusual pattern (or variations) of the data that are not mended by the technique because all the values are inside of the limits calculated. This technique repair low or high values but not variations on the data pattern.

3.2. Linear Regression and Correlation

The regression analysis studies the relationship between two variables, the dependent variable (Y) and independent variable (X), i.e. it consists in obtaining a mathematical model that represents this relationship.

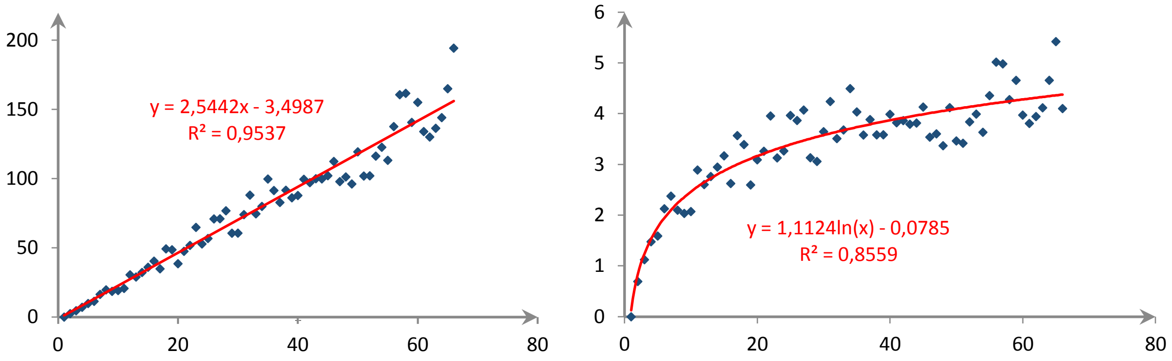
An initial analysis to the behavior and relationship between the two variables is possible through a scatter plot. This type of diagram helps to study the correlation[3] between two variables. Three main situations can occur:

- Positive correlation – when a variable grows the other one is also growing.
- Negative correlation – when a variable is growing the other one is decreasing.
- No correlation – when the points are dispersed with no notion of direction.

Still in the scatter plot analysis, the relationship between X and Y can exhibit different types of behavior: linear, logarithmic, exponential, etc. (refer to *Figure 31*). In order to create the

[3] Correlation coefficient measures the association degree between two variables.

mathematical model that best describes the data set, it should be inspected which are the type of curve and mathematical equation that most closely represent the values in the scatter plot.



a) Linear behavior data set

b) Logarithmic behavior data set

Figure 31: Scatter plot investigation

It is possible to see in *Figure 31* that the points of the scatter plot are not perfectly adjusted to the mathematical model identified. A distance exists between the majority of the points of the scatter plot and the curve of the mathematical model.

The most widely used method to obtain the relationship between X and Y is to estimate an equation that minimizes the distance between the data points and the mathematical model. This method is known as the Least Squares Estimation (LSE). In a simple way, the LSE minimizes the sum of the square distances between the data points and the regression line, finding so a functional relation between X and Y with the minimum possible error.

The coefficient of determination, usually referred as R^2 is an approach to evaluate how well the linear regression fits the data (goodness of fit). If predictions are close to the actual values, we would expect a R^2 close to 100%, in the other hand, if the predictions are dissimilar from the actual values we can expect a R^2 near zero. The coefficient of determination is explained in more detail in *3.1.4.5 Coefficient of Determination*.

3.2.1. Simple Linear model

This chapter introduces the simple linear regression. The objective here is to forecast a variable Y assuming that it has a linear relationship with variable X. The simple linear model is defined by the following equation:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + e_i \qquad (3.11)$$

- X_i is the ith point of the independent variable X;

- Y_i is the observed value of Y at the ith point of the independent variable X;
- β_0 is the intercept of the tendency line (red line referring to *Figure 31*), i.e. the value of Y when $X = 0$, which in this case is around -3.5;
- β_1 is the is the tendency (slope) of the line, i.e. it represents the variation of Y as function of one unit variation in X. Referring to *Figure 31, β_1 is* around 2.54 (positive slope, positive correlation);
- e_i is the error associated with the distance between the observed value and the correspondent value of the purposed model;

The least square principle provides a way to find the β parameters minimizing the sum of the squared errors. Thus, we have:

$$e_i = Y_i - \beta_0 - \beta_1 \cdot X_i \tag{3.12}$$

square both sides of the equation,

$$\sum_{i=1}^{n} e_i^2 = \sum\nolimits_{i=1}^{n} [Y_i - \beta_0 - \beta_1 \cdot X_i]^2 \tag{3.13}$$

The purpose of LSE method is to achieve estimators for β_0 and β_1 that minimizes *(3.13)*, with some mathematical calculus (refer to [28]) it can be demonstrated that the least squares estimators are:

$$\beta_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{3.14}$$

and

$$\beta_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \tag{3.15}$$

- $\bar{x}$ is the average of the x observations;
- $\bar{y}$ is the average of the y observations;
- N is the number of observations;

3.2.1.1. Forecasting results

The pattern of the data that will be analyzed in this work follows closely the pattern present in the daily average CS throughput (Iub interface) represented in *Figure 32*.

Once the linear regression only captures the intercept and the slope, we can conclude a priori that the linear regression is not a suitable method to fit this type of data. However, the analysis with the linear regression is a good start point to understand some concepts related with forecast and a good reference for the next methods.

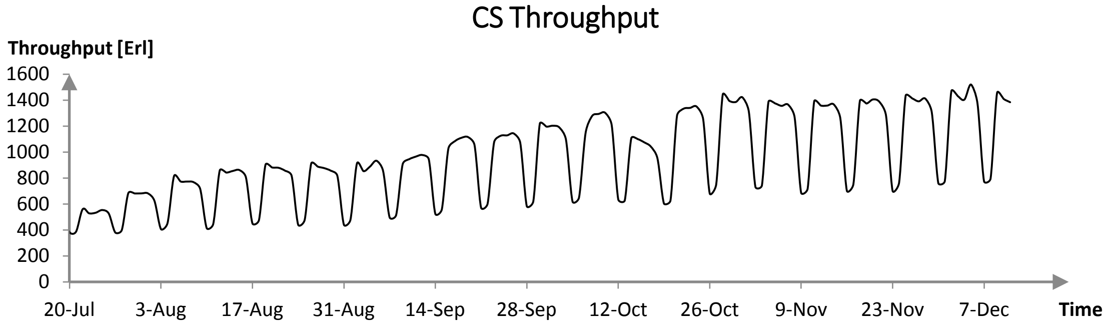


Figure 32: Daily average CS throughput in Iub interface

The ability to forecast using a linear regression is possible through the next equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x \qquad (3.16)$$

where x (or predicator) is the value for which it's required a forecast. When the value of x is an observation from the data set, the resulting $\hat{y}$ is called a "fitted value". This fitted value is not a genuine forecast because the actual value of y for that predictor was used to create the model so the value of $\hat{y}$ is affected by the true value of y. If x is a new value (i.e. it was not used to estimate the model), the predicted value $\hat{y}$ is a genuine forecast.

The linear regression is one of the simplest techniques used for forecasting, but this simplicity comes with costs. As mentioned before, the linear regression only captures the slope (trend) of the time series and predicts the values based on this trend discarding all the other components that the time series may have (e.g. seasonal cycle).

Figure 33 represents the CS throughput and the forecast values obtained with the linear regression. At the left of the black line is represented the real values and fitted values and at the right of the black line the forecasted values and the real values.

As it is possible to see the linear regression captures the level of the past values, which continues to the forecast values. However, around Dec-24 the level of the CS throughput falls off for a lower level (probably due to the holiday season) and the linear regression predicts values higher than the real ones.

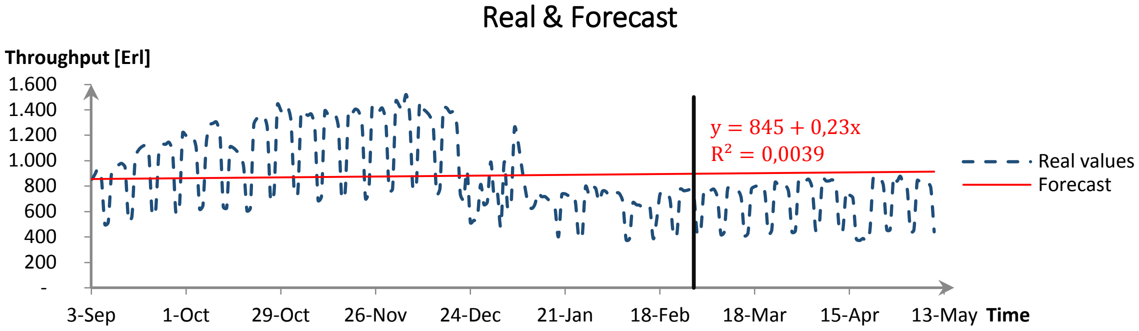


Figure 33: Real and forecasted values

The fact that the linear regression doesn't capture seasonality is reflected in a poor coefficient of determination (R^2) equals to 0.0039. Again, the pattern observed in the figure is a common pattern (weekly seasonality) for this kind of data. Thus, the linear regression is not the best method for this type of data.

The disparity between the real and predicted values is reflected in a 46% of mean of the relative errors (not so good forecast), which was expected since the linear regression does not capture historical patterns (e.g. seasonality).

Case study

Figure 34 represents a simple example that clearly exposes a drawback of the linear regression (and other methods too as we will see) – the presence of outliers in the data.

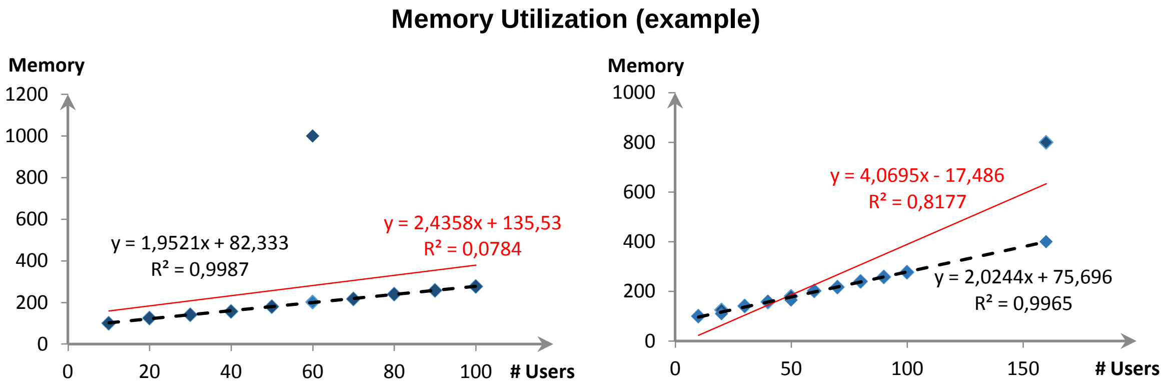


Figure 34: Outliers and influential values

In the left chart an outlier is intentionally introduced, and in the right side chart an influential observation. The black dashed lines in the graphs are the linear regression of the dataset without the outliers influence and the red lines are the linear regression with the outliers. This allows us to observe the big influence that a single outliers has on the path of the linear regression. The presence of outliers can lead to terrible results, depending on their nature.

3.3. Time Series

A time series is a sequence of successive observations in time and spaced with uniform time intervals. Some examples of a time series are: daily temperature, car sales monthly values, throughput in an interface, hourly CPU capacity usage, etc.

The analysis of time series has particular interest to this work. Usually, time series assumes the past pattern to continue into the future and the analysis of the data history can be helpful to support network planning based on forecasting.

3.3.1. Moving Average Smoothing

The moving average technique provides a simple manner to smooth a time series, replacing an observation by the average of this observation and the observations in its vicinity. This is a technique used in the classical decomposition to estimate the trend-cycle component.

3.3.1.1. Moving Average Filter

The moving average process is an alternative representation of a Finite Impulse Response (FIR) filter. The simple moving average that will be presented here is the simplest form of a FIR filter, with all coefficients being equal.

As the name implies a moving average filter operates by averaging a number of points from the input data to produce a smooth output version of the data. Let x be the input data and y the output, the moving average filter is written as follows:

$$y(t) = \frac{1}{k}\sum_{j=0}^{k-1} x_{t-j} \qquad (3.17)$$

Figure 35 represents an example of a FIR implementing a 7-MA.

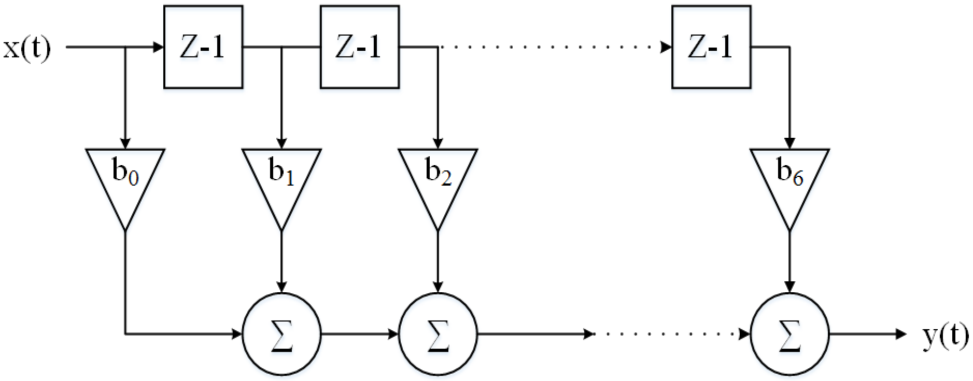


Figure 35: Block diagram of a 6th order/7-tap FIR filter

To implement an equal weighted 7-MA, the coefficients must have the same weight:

$b_0 = b_0 = \cdots = b_6 = \frac{1}{N+1} = \frac{1}{7}$

As an alternative, the group of points from the input signal can be chosen symmetrically around the output point, i.e.:

$$y(t) = \frac{1}{k} \sum_{-m}^{j=m} y_{t+j} \qquad (3.18)$$

where,

$$m = \frac{k-1}{2}$$

The MA technique eliminates some randomness in the data leaving a smoothing trend-cycle component. Thus, the trend cycle in period t is obtained by averaging m values on the vicinity of t.

3.3.1.2. Centered Moving Averages

The simple moving average is used for odd numbers of k, though it is less suitable for even number of k. Imagine a 4-MA moving average, we have two possibilities for the averages:

$$MA = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1}}{4}$$

$$MA = \frac{y_{t-1} + y_t + y_{t+1} + y_{t+2}}{4}$$

Which one should we choose? The best decision is to average the two, making an even order moving average symmetric. So the answer is to apply a moving average of order 2 to the results of the first moving average of even order, for example for a 4-MA the notation should be 2 × 4-MA and can be written as follows:

$$2 \times 4MA = \frac{1}{2} \cdot \left[\frac{y_{t-2} + y_{t-1} + y_t + y_{t+1}}{4} + \frac{y_{t-1} + y_t + y_{t+1} + y_{t+2}}{4} \right]$$

The centered moving average is usually used to estimate the trend-cycle from seasonal data. Consider the 2 × 4-MA:

$$2 \times 4MA = \frac{1}{8} y_{t-2} + \frac{1}{4} y_{t-1} + \frac{1}{4} y_t + \frac{1}{4} y_{t+1} + \frac{1}{8} y_{t+2}$$

When applied to quarterly data, each quarter of the year is given the same weight, note that the first and last term is applied to the same quarter in consecutive years.

Concluding, if the seasonal period is an even number, it should be used a $2 \times k$ MA. If the seasonal pattern is of an odd number use a k-MA moving average. Some common moving averages are: 2×12 -MA to estimate the trend-cycle of monthly data and 7-MA to estimate the trend-cycle of daily data.

3.3.1.3. Examples of MA smoothing

Figure 36 and *Figure 37* shows how the trend-cycle (red line) looks like when obtained with a 7-MA and 31-MA respectively. Note how the application of a moving average smooth the data and captures the main movement of the time series without the small fluctuations.

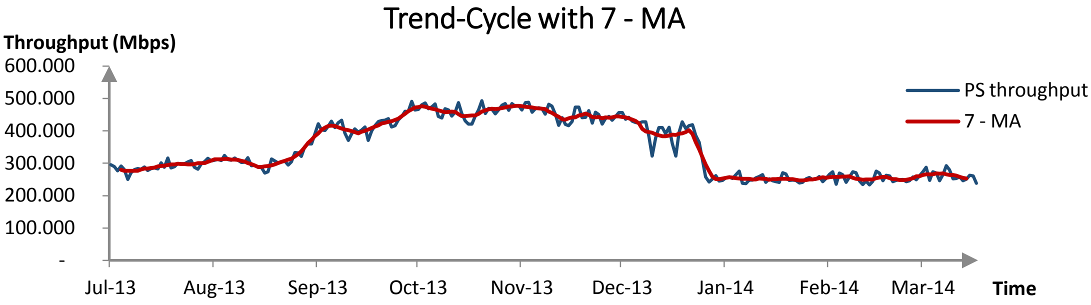


Figure 36: Moving average of order k=7

The greater the order of the moving average the smoother the trend-cycle line is. This is possible to notice when comparing the *Figure 36* and *Figure 37*. This fact happens because the number of values used to calculate the average for each period of time is bigger, hence the 31-MA gets a smoother trend-cycle line than the 7 – MA one.

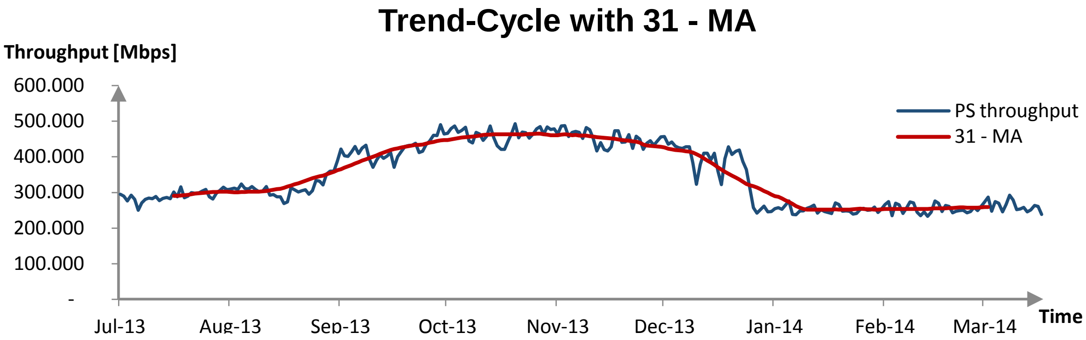


Figure 37: Moving average of order k=31

This method of smoothing the data in order to obtain the trend-line component is very useful to start the classical time series decomposition and forecasts.

3.3.2. Classical decomposition

The classical decomposition created in the 1920s [27], is a simple useful method to understand the basis of most time series methods decomposition. There are two types of decomposition that can be identified, the additive and multiplicative models.

3.3.2.1. Additive model

The additive model, assumes that the data is the sum of the time series components mentioned *3.1.2 Time Series Components*. Here, the seasonal component is independent of the trend, and thus the seasonal movement is constant over time as shown in *Figure 38*.

$$Y = T + C + S + I \qquad (3.19)$$

If the data doesn't contain one of the components, the value for that component is equal to zero.

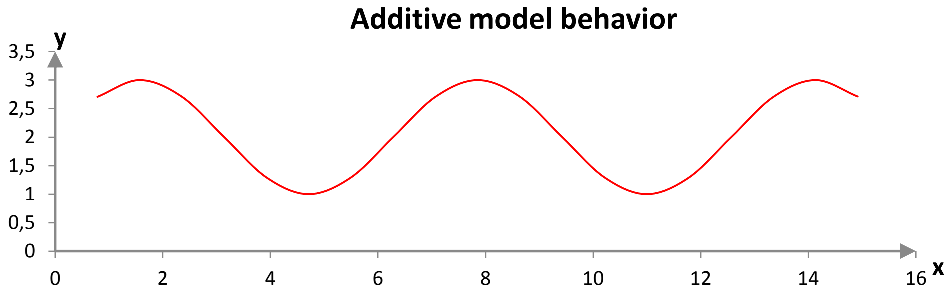
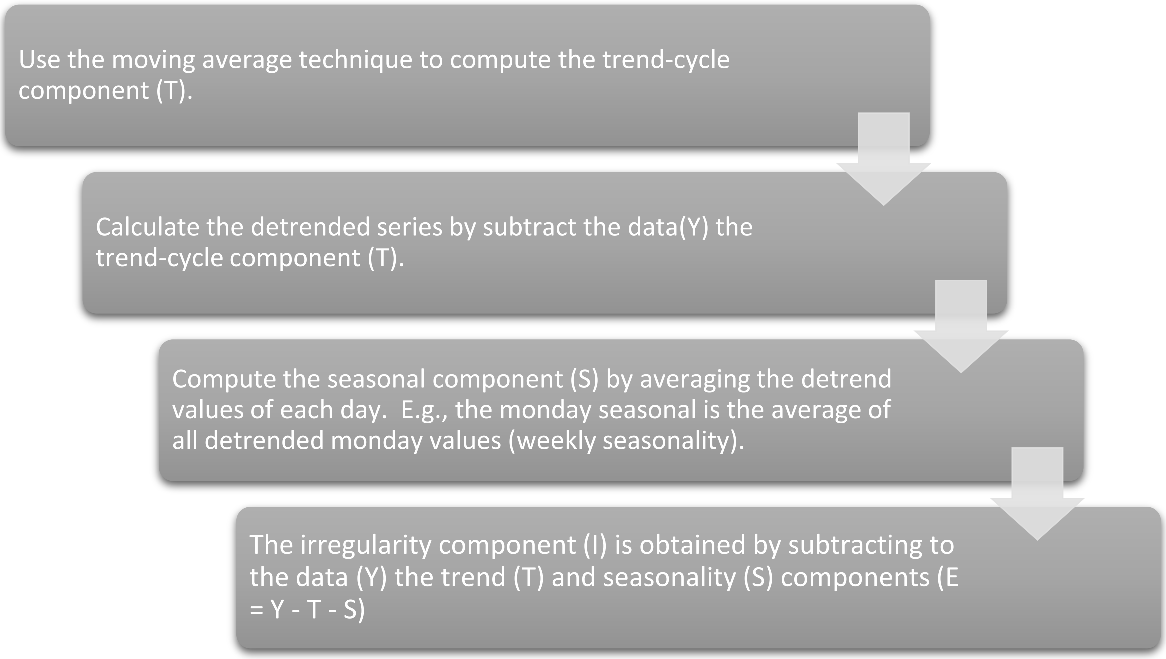


Figure 38: Constant seasonal amplitude (Additive model)

The time series components can be obtained by follow the steps presented in *Figure 39*.



Use the moving average technique to compute the trend-cycle component (T).

Calculate the detrended series by subtract the data(Y) the trend-cycle component (T).

Compute the seasonal component (S) by averaging the detrend values of each day. E.g., the monday seasonal is the average of all detrended monday values (weekly seasonality).

The irregularity component (I) is obtained by subtracting to the data (Y) the trend (T) and seasonality (S) components (E = Y - T - S)

Figure 39: Additive decomposition steps.

3.3.2.2. Multiplicative model

Here the data is assumed to be the product of the different components of a classical decomposition. In this case, the seasonal component is proportional to the trend, i.e. the swing of the seasonality increase or decreases according to the behavior of the trend, see *Figure 40*.

$$Y = T \times C \times S \times I \tag{3.20}$$

If the data doesn't contain one of the components, the value for that component is equal to one.

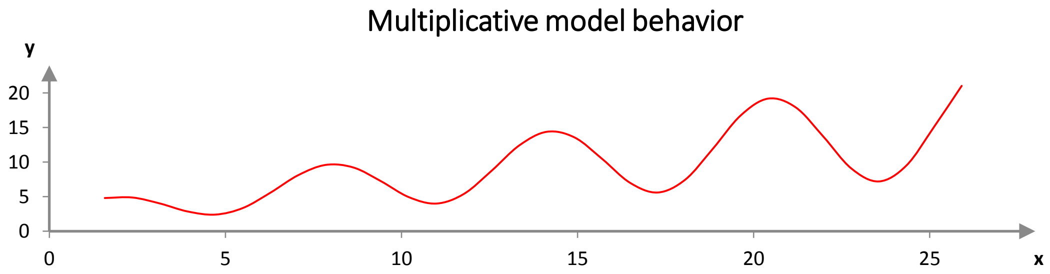
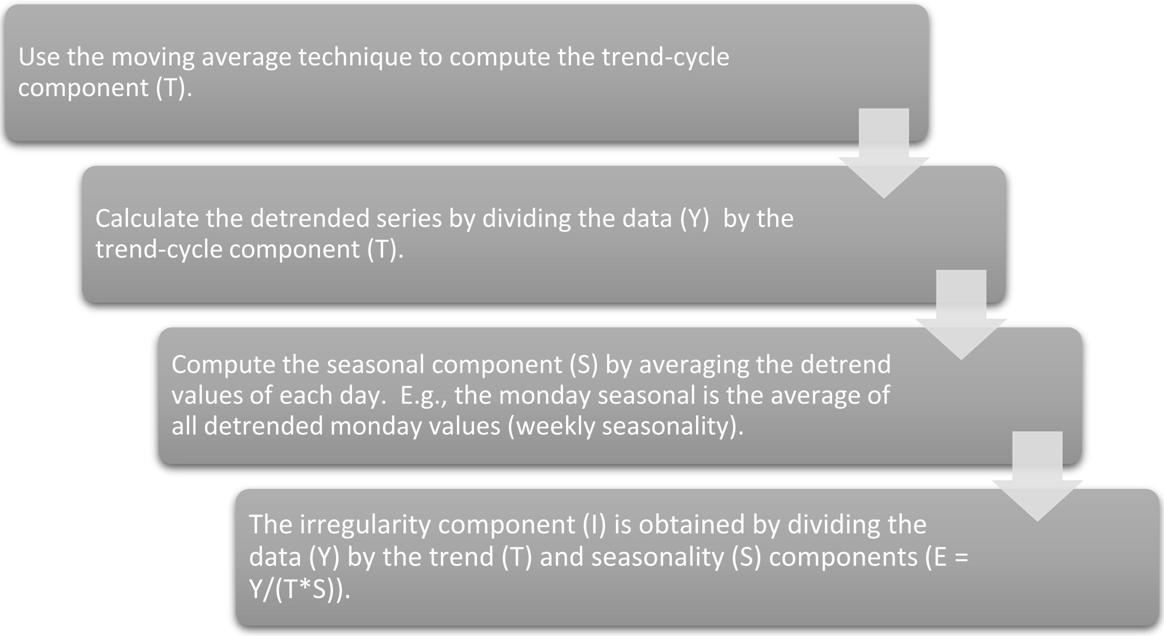


Figure 40: Seasonal amplitude proportional to the trend (Multiplicative model)

The multiplicative model is very similar to the additive model except the additions are replaced by multiplications and the subtractions by divisions, refer to *Figure 41*.



Use the moving average technique to compute the trend-cycle component (T).

Calculate the detrended series by dividing the data (Y) by the trend-cycle component (T).

Compute the seasonal component (S) by averaging the detrend values of each day. E.g., the monday seasonal is the average of all detrended monday values (weekly seasonality).

The irregularity component (I) is obtained by dividing the data (Y) by the trend (T) and seasonality (S) components (E = Y/(T*S)).

Figure 41: Multiplicative decomposition steps

Although, sometimes is not easy to classify precisely a dataset as additive or multiplicative. A simple way solve this problem is to look at the forecasts obtain with both models and choose the one that minimizes the Sum of Squared Errors (SSE) and seems appropriate for the data in question.

The classical decomposition is widely used in several work areas, however its vulnerabilities require a carefully employment. Some problems of the classical decompositions are:

- Sometimes the data contains outliers, so when computing the seasonal component by simply using a moving average, the outliers are included in the calculations. There are some methods to identify outliers and remove them; this could be an improvement to this problem.
- The classical decomposition assumes that the seasonal component repeats over time which in some series it is not reasonable to assume. Classical decomposition is unable to capture changes over time.
- The trend component is not available in the first and last observations. This reality is possible to see in *Figure 36* and *Figure 37*.

3.3.3. Holt-Winters Exponential Smoothing

The Holt-Winters model is one of the time series classical methods. This method is widely used when the time series to be analyzed presents seasonality (S) and trend (T). As already mentioned before, time series with seasonality is usually characterized by repetitive cyclic patterns normally with constant time intervals.

Once the type of data being analyzed by this work presents a multiplicative seasonality and trend, the properties of this method are ideal.

This method contains three smoothing equations: one for the level L_t, one for trend, T_t and one for the seasonal component S_t, with smoothing parameters α, β and γ. The m here still denoting the period of the seasonality (e.g. quarterly data $m = 4$, monthly data $m = 12$ and daily data $m = 7$).

The Holt-Winters method has two variants: the additive model and the multiplicative model. This division is based in the idea of the additive and multiplicative models of the *Classical decomposition* (*3.3.2*). In the additive model the seasonal variation amplitude is constant over time and in the multiplicative model increases or decreases over time.

3.3.3.1. Multiplicative model

The multiplicative model is the correct model to series with trend and multiplicative seasonality. The following equations describe the Holt-Winters multiplicative model.

$$L_t = \alpha \frac{y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad (3.21)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \qquad (3.22)$$

$$S_t = \gamma \frac{y_t}{L_t} + (1-\gamma)S_{t-m} \qquad (3.23)$$

$$\hat{y}_{t+k} = (L_t + kT_t)S_{t+k-m} \qquad (3.24)$$

- y_t is the observation at time t;
- L_t is the level component;
- T_t is the trend component;
- S_t is the seasonal component;
- m is the seasonal period;
- $k = 1, 2, \dots, h$;
- h is the prediction horizon;
- $\hat{y}$ is the prediction value at time $t + k$;

The Holt-Winters multiplicative method works by applying recursively its equations to the time series. Thus, this kind of application has to start some point in the past where the values of L_t, T_t and S_t must be estimated. This can be achieved by initializing the level (L_t) and trend (T_t) in the same period t.

Despite of different literatures present alternative ways to initialize the level and the trend, in this work the next initializations will be used:

- **Level** – Let n be the total number of observations, the initial level is the average of all observations.

$$L_t = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \tag{3.25}$$

- **Trend** – To initialize the trend is recommended the use of two seasonal periods, i.e. $2 \times m$:

$$T_t = \frac{1}{m}\left(\frac{y_{m+1} - y_1}{m} + \frac{y_{m+2} - y_2}{m} + \cdots + \frac{y_{m+m} - y_m}{m}\right) \tag{3.26}$$

- **Seasonal Indexes** – The seasonal indexes have to be initialized to the first season (weekly seasonality in this case). These are obtained using a 7-MA (simple moving average). Refer to *3.3.1 Moving Average Smoothing.*

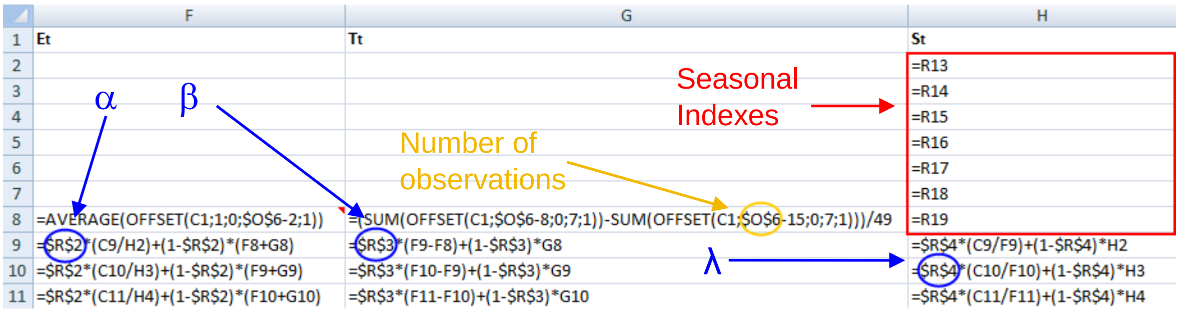
3.3.3.1.1. Excel Implementation

The implementation of the Holt-Winters method can be divided in three main parts.

- **Method Initialization**
  - Initialize level, trend and seasonal values.
  - Perform the calculus of the fitting values.
- **Parameters Estimation**
  - Use the Solver to calculate the smoothing parameters that minimize the errors (i.e. best fitting smoothing parameters)
- **Forecast**
  - Recursively use the forecast equation (3.24).

A. Method Initialization

As already mentioned before, different literatures present different forms of initialization of the methods components (Level, trend and seasonality). The aim of this section is to demonstrate how to implement on Excel these components initializations.



	F	G	H
1	Et	Tt	St
2			=R13
3			=R14
4			=R15
5			=R16
6			=R17
7			=R18
8	=AVERAGE(OFFSET(C1;1;0;$O$6-2;1))	=(SUM(OFFSET(C1;$O$6-8;0;7;1))-SUM(OFFSET(C1;$O$6-15;0;7;1)))/49	=R19
9	=$R$2*(C9/H2)+(1-$R$2)*(F8+G8)	=$R$3*(F9-F8)+(1-$R$3)*G8	=$R$4*(C9/F9)+(1-$R$4)*H2
10	=$R$2*(C10/H3)+(1-$R$2)*(F9+G9)	=$R$3*(F10-F9)+(1-$R$3)*G9	=$R$4*(C10/F10)+(1-$R$4)*H3
11	=$R$2*(C11/H4)+(1-$R$2)*(F10+G10)	=$R$3*(F11-F10)+(1-$R$3)*G10	=$R$4*(C11/F11)+(1-$R$4)*H4

Figure 42: Holt Winters Initialization

Figure 42 illustrates the initialization process in a very general way. This figure simplifies the identification of the initialization equations as well as the methods equations. The seasonal indexes are obtained using the moving average as already stated.

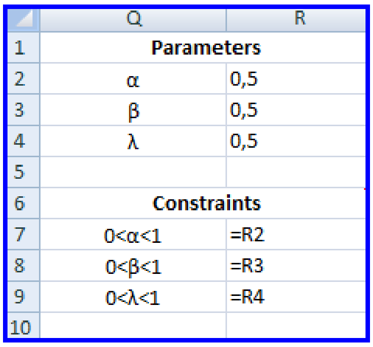
Note however that this example serves only for illustration, because the automation of the method requires several conditions on the calculus.

B. Parameters Estimation

The estimation of the smoothing parameters (α, β and γ) is done using an *Excel* add-in called *Solver* with the aim of minimize the Mean Squared Errors (MSE). This section pretends to show how to use the *Solver*.

Figure 43 represents the smoothing coefficients of the Holt-Winters equations and the constraints that they must obey.

Note that the initial guess for all the coefficients is 0.5. The initial guess can be any value. However, it must obey to the constraints. The reason of these constraints is the stability of the method, and it is assumed that older observations have less weight in the predictions.

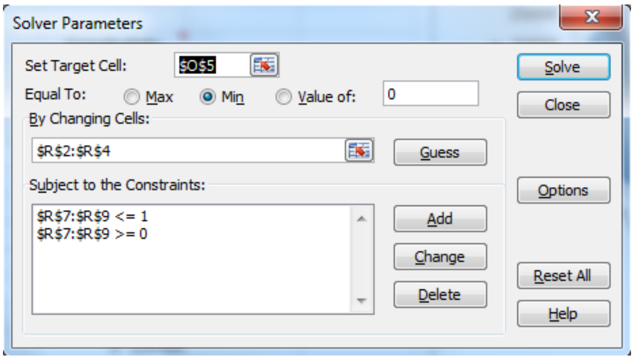


	Q	R
1	**Parameters**	
2	α	0,5
3	β	0,5
4	λ	0,5
5		
6	**Constraints**	
7	0<α<1	=R2
8	0<β<1	=R3
9	0<λ<1	=R4
10		

Figure 43: Holt-Winters Parameters Estimation

The *Solver* takes as inputs a target cell, the values to change (the smooth parameters in this case) and the constraints to these. *Figure 44* shows the setting of the solver tool in accordance to *Figure 43*.

Note that, in *Figure 44*, it is chosen to minimize the cell $O$5 that corresponds to the MSE by changing the cells $R$2 to $R$4 (the smooth parameters, see *Figure 43*) subjected to the constraints shown in *Figure 43*, i.e. the smooth parameters must be between 0 and 1.



Solver Parameters

Set Target Cell: $O$5

Equal To: ○ Max ● Min ○ Value of: 0

By Changing Cells:

$R$2:$R$4

Subject to the Constraints:

$R$7:$R$9 <= 1
$R$7:$R$9 >= 0

Figure 44: Solver tool

The *Solver* uses a method called Generalized Reduced Gradient (GRG) to compute the parameters. This is a very useful tool concerning the matter of forecasting (fitting the models to the observations), and it is also used to fit the ARMA and SARMA methods later discussed in this document.

3.3.3.2. Result analysis

The examples that will be analyzed in this chapter were obtained using AMR traffic (refer to 5. *RNC Capacity Planning*) acquired from a real network operator. Also, these examples were obtained using the forecast tool presented in 4. *Forecasting Tool*.

The method is evaluated based on values studied in *3.1.4 Forecast Evaluation* and the results are also compared with real values. In addition, it is also discussed some particularities in the data (e.g. outliers, external conditions, etc.) which can influence the forecasting results.

Before any further discussions, it is necessary to identify the seasonality of the data, plot the time series is the best way to do it.

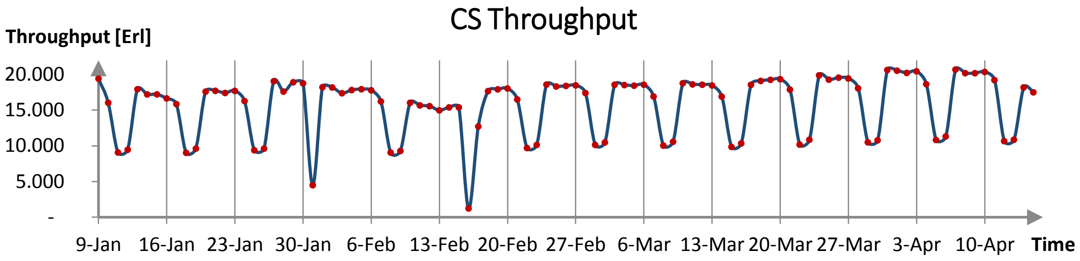


Figure 45: CS throughput

Figure 45 represents the daily average AMR throughput during the time from January 9 until April 10 (year 2014). This figure only represents a small subset of data to facilitate the recognition of the pattern in the data. Thus, it is possible to observe that the throughput takes the same shape every week. The red dots in the figure help this recognition, for example the level of throughput in the first day of the first week is identical to the first days of the next weeks. Thus, the seasonal period (m) that must be used to compute the Holt-Winters model is $m = 7$.

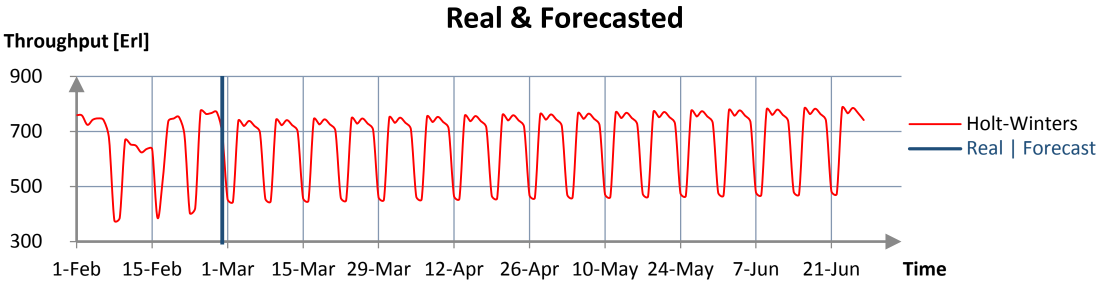


Figure 46: Holt-Winters forecast

The blue line in *Figure 46* marks the beginning of the predictions. This figure illustrates the forecasting results for a period of 4 months using the Holt-Winters method.

The AMR traffic pattern being analyzed here is almost always the same, i.e. doesn’t diverge of the normal pattern. *Figure 46* helps us to notice that the Holt-winter method captures very well the historical pattern.

Figure 47 represents both real and forecasted values. This figure helps to observe how much the forecast deviates from the reality.

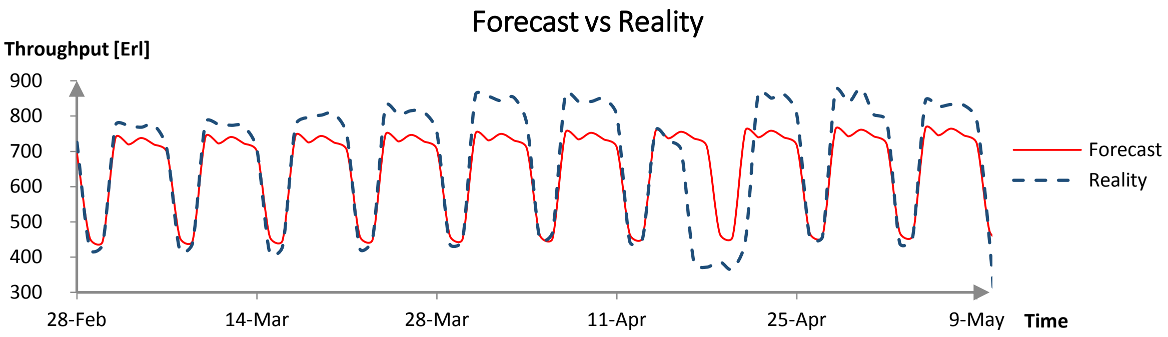


Figure 47: Real values vs. forecasted values

Observe that the real values keep the historical pattern as expected. The forecasting shape doesn't follow exactly the reality, but the mean for the relative error is 10.98%, which is 31.48% less than the relative error achieved with the linear regression in the same data and with the same prediction horizon.

This model is a good alternative to the linear regression technique. However, the use of Holt-Winters method and the analysis of its results must be done very carefully. *Figure 48* presents a case study of the Holt-Winters method.

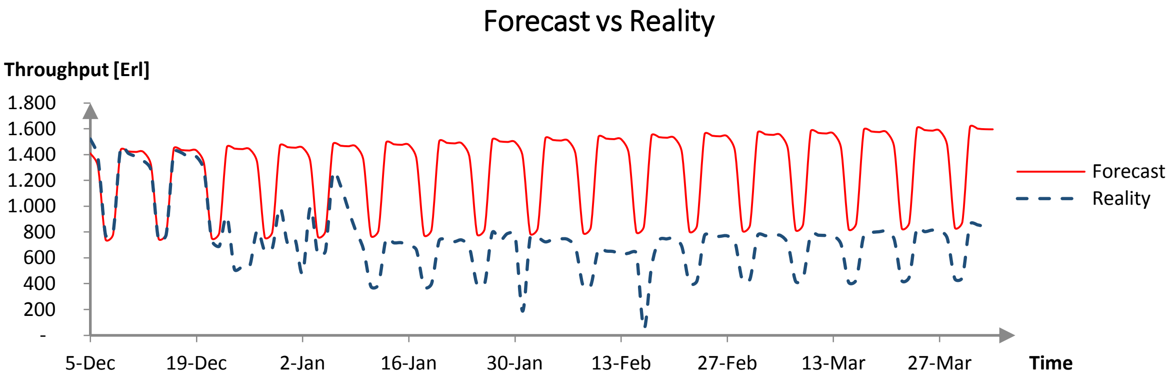


Figure 48: Holt-Winters case study #1

The data used to produce the forecast in *Figure 48* has been rising until around December 19 (probably due to Christmas season). The Holt-Winters property of capturing the trend of the data becomes a problem in this situation because suddenly around December 24 the AMR throughput falls to a very low level while the Holt-Winter methods continues growing.

Clearly, the Holt-Winters doesn't predict these type of situations and produce bad results. This document suggests a technique to attenuate these situations, the *Level Offset* discussed later in *3.5.2 Level Offset.* If the user foresees that, for some reason, the throughput will decrease around 50% around December 23 this can be input in the forecast tool (refer to *Forecasting Tool)*.

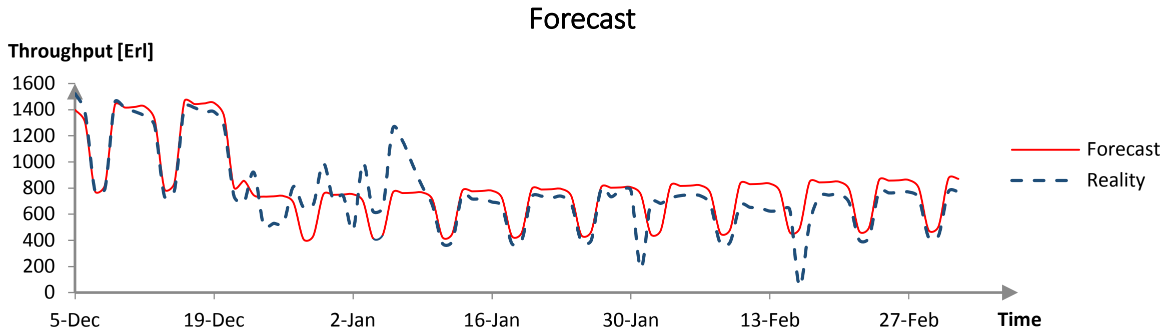


Figure 49: Case Study #1 (w/ Level Offset)

Figure 49 represents the results with the “Level Offset” input of -50%. The mean of the relative errors fell from 92.24% to 15.87% (values calculated without outliers on the real values and for a 3 months forecast period).

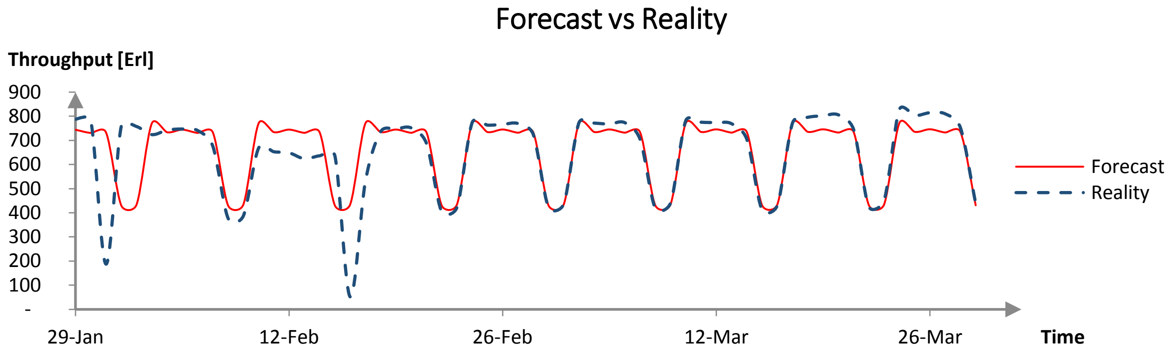


Figure 50: Holt-Winters outliers’ case study #2

The case study in Figure 50 pretends to show that the use of the mean of relative errors between the forecast results and the reality can mislead the results evaluation. The existence of two clear outliers at January 31 and February 16 reflects in the mean of relative errors that rise to a value of 20.48 %, but if you remove these two outliers the value falls to 10.47 %, a much better value. The greater the number of outliers, worst the results.

This case study shows that when doing this type of analysis (using the mean of relative errors) the user should always observe the chart to avoid a bad reading of the results. The user can always use the technique studied in *3.1.4.6 Outliers and influential observations* to deal with outliers, however it is always wise to visualize the results.

Note that this case only applies when the user wants to analysis the response of the method to a specific data type ad compare the forecast with the reality. This case study is common for all the methods studied in this document and it will not be discussed again.

Referring to equations *(3.21), (3.22), (3.23), (3.24),* all calculations for the prediction values $\hat{y}_{t+k}$ have a strong dependence with the time series last values, in particular with the last one, because the calculation for all the predicted values are directly related with it. This can be a problem if there is not a precautious analysis before using the Holt-Winters method to forecast.

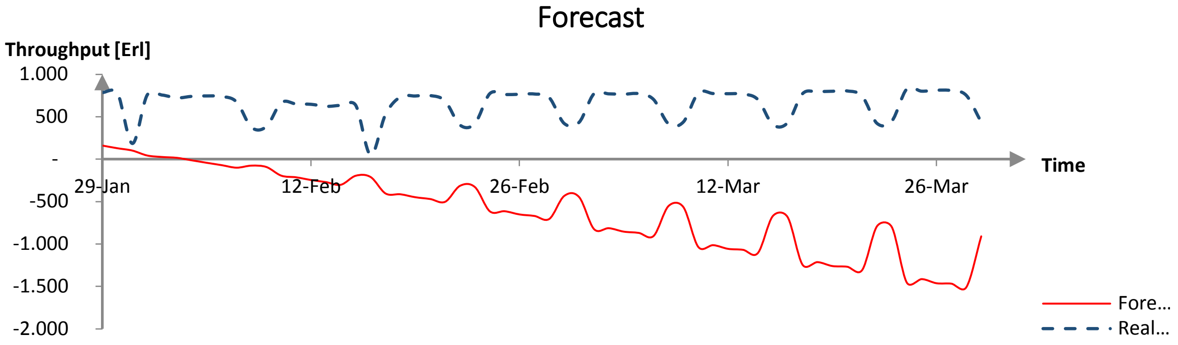


Figure 51: Holt-Winters outliers' case study #3

Figure 51 represents a case study where the last observation is an outlier and no outliers' technique is used. It is possible to see that only one value (last value) has a strong influence in the entire result. The mean of relative errors in this example is 277.52 % and predict negative values which are impossible to occur.

Figure 52 represent the forecasting results obtained for the same period of *Figure 51* after the technique to deal with the outliers had been applied (refer to *3.1.4.6 Outliers and influential observations)*.

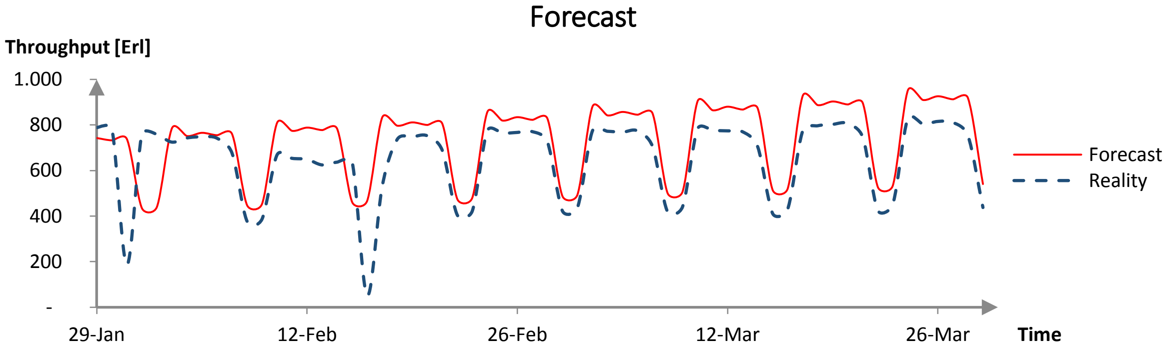


Figure 52: Case study #3 without outliers.

Note that the results are much more close to the reality. The mean of relative errors in this case is 19.98 % which is much less than 257.54 %.

Concluding, the Holt-Winters methodology doesn't deal with outliers and produce bad results in its presence. Therefore, before using this methodology it must be applied some technique to deal with outliers.

3.3.4. AutoRegressive Moving Average

The Box & Jenkins (1970) methodology is widely used in time series analysis. This methodology consists in the adjustment of Auto-Regressive (AR) Integrated Moving Average Models – ARIMA (p, d, q) – to a set of data.

Three basic models can be identified in the ARIMA methodology: The AR (p) (autoregressive), MA (q) (moving average) and their combination, the ARMA (p, q). When differentiation is performed, the combination AR (p) and MA (q) is called ARIMA (p, d, q), the letter *I* refers to the differentiation procedure. For example an ARIMA(1,0,0) corresponds to an AR(1), an ARIMA(0,0,1) to a MA(1), an ARIMA(1,0,2) to an ARMA(1,2), etc.

Before introduce the ARIMA methodology the concept of stationarity must be introduced.

Stationarity

A time series is considered stationary when its properties such as mean, autocorrelation, etc. are constant over time. Trend, seasonality and their combination are examples of non-stationary behaviors (see *Figure 53*).

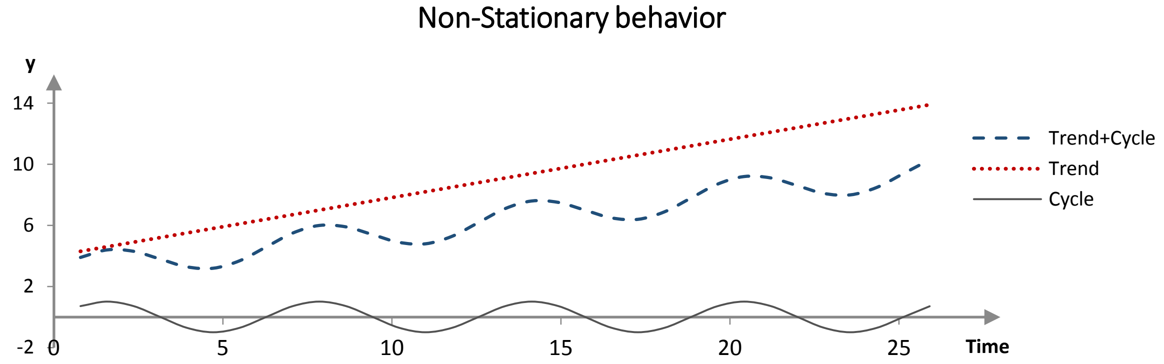


Figure 53: Non-Stationary behaviors

The ARIMA model is capable of fit stationary and non-stationary time series since it doesn't exhibit an explosive behavior, i.e. homogeneous stationarity [30].

The first requirement when applying an ARIMA model is to be facing a stationary time series or else the predicted values can result in a bad forecast. A non-stationary time series can be transformed into a stationary one by differencing consecutive observations ($y_t^{'} = y_t - y_{t-1}$). There are different order and types of differencing, such as: first order differentiation, seasonal differentiation, etc.

Taking into account the type of data that this model will fit (AMR throughput) the differentiation component will not be covered by this work. Refer to [27] and [31] if you are interested in this topic.

The decision of leaving the differentiation topic aside is due to the stationarity (or almost) nature of the AMR throughput for most RNCs and also simplicity. However, take into account that this method should be applied to more stationary markets.

3.3.4.1. Auto-Regressive model

The AR model is a subset of the ARMA models. It is called autoregressive (AR) because the output values are calculated based on the regression of the previous output values, it works like a multiple regression where the predictors are the lagged values of y_t. The AR processes can be represented by a simple Infinite Response Filter (IIR) block diagram (*Figure 54*).

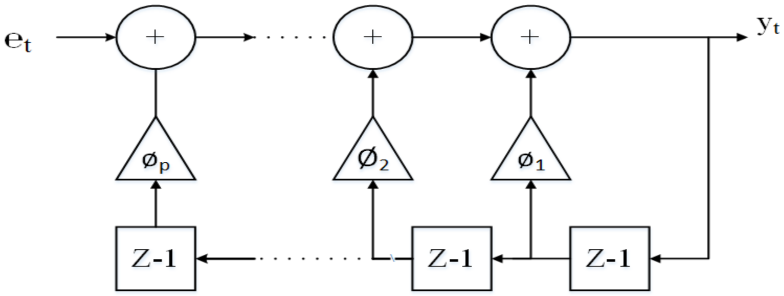


Figure 54: p-order IIR filter implementing an AutoRegressive model

This model (filter) is a linear function of past values:

$$y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + e_t \tag{3.27}$$

ϕ_k → Autoregressive coefficients

p → Order of the filter (or AR process)

e_t →Residuals (assumed to be Gaussian white noise)

3.3.4.2. Moving Average model

Another subset off the ARMA model is the Moving Average (MA). Here instead of using the past values of y_t in a regression, the moving average model uses past forecast errors in a regression-like model.

Thus an moving average model of order q – MA(q) – can be denoted as:

$$y_t = \sum_{k=1}^{q} \theta_k e_{t-k} + e_t \tag{3.28}$$

where e_{t-q} are the residuals in the lag q and $\theta_k$are the moving average coefficients. This moving average coefficients (θ_k) must respect the stationarity constraints $-1 < \theta_k < 1$ [27].

The moving average model is essentially a FIR filter, but instead of using past observations (x_t) as input, uses the residuals e_t. *Figure 55* shows a block diagram representation of a residuals moving average filter.

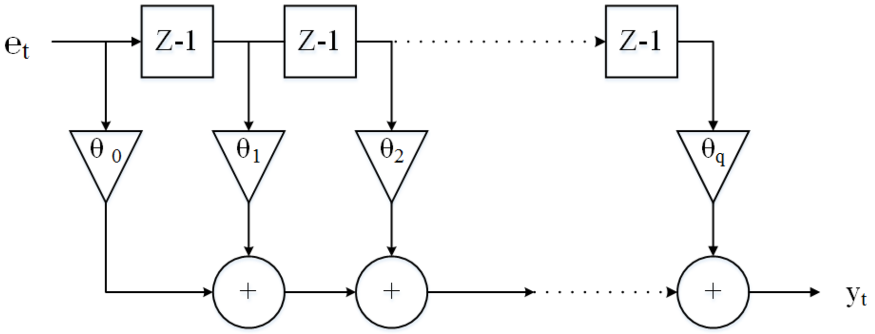


Figure 55: q-order FIR filter implementing a Moving Average model

Note that this moving average model should not be confused with the *Moving Average Smoothing* discussed in *3.3.1.* The *Moving Average Smoothing* is used to estimate the trend-cycle of past values ($\mathrm{x_t}$) while the moving average model is used for forecasting future values.

3.3.4.3. Non-Seasonal ARMA

The combination of the AutoRegressive (AR) and Moving Average Model (MA) discussed above forms the AutoRegressive Moving Average (ARMA). The ARMA model can be written as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t \tag{3.29}$$

Here the model uses both lagged values of y_t and lagged values of the residual e_t. In the notation ARMA (p, q) the *p* letter refers to the AutoRegressive order and *q* to the Moving Average order. The ARMA model can be represented by the combination of a FIR (with the input equal to the residuals) and IIR filters, in fact the ARMA filter representation is the combination of the AR (*Figure 54)* and MA (*Figure 55*) filters discussed before.

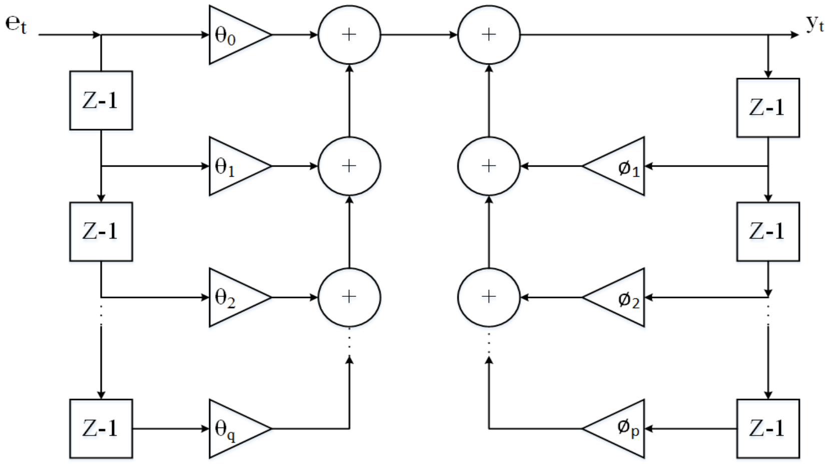


Figure 56: FIR/IIR filter implementing an ARMA(p,q) model

The construction of these models is based in an interactive cycle, where the model structure is influenced by the data in question. *Figure 57* represents a general view of this cycle in a diagram form.

The diagram illustrates the stages to "manually" construct the method, but one of the main objectives of this work is automation. Thus, next is explained how to automate the main stages of the cycle.

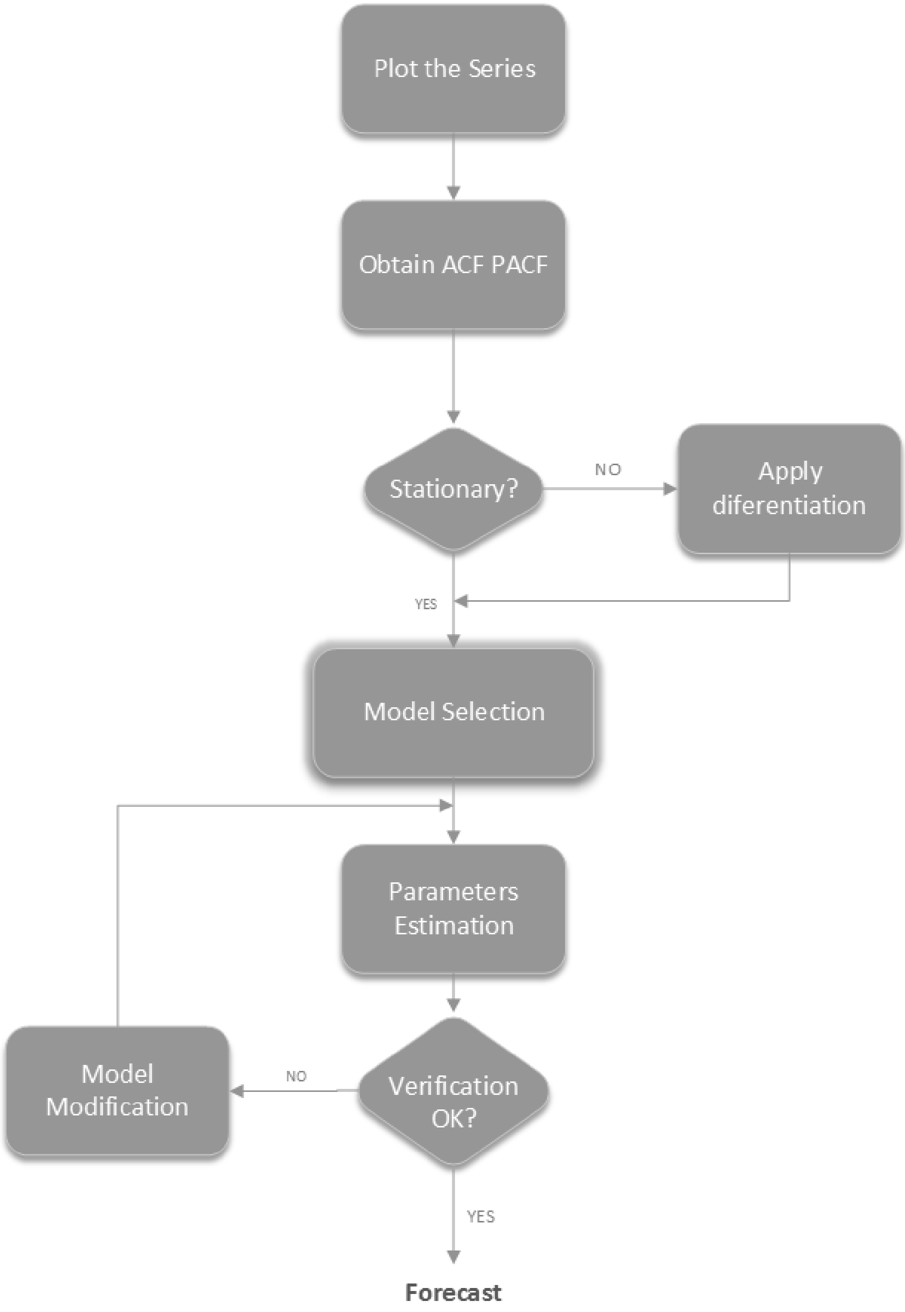


Figure 57: ARMA cycle construction

Model Selection

The model estimation is basically the identification of the model order, for a certain set of data. This identification is “manually” done by using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

The inspection of the ACF and PACF to identify ARMA models is somewhat of an art rather than a science, which makes this technique of model identification a very subjective task and not an easy one, even for time series analysts.

Trying to solve this problem, time series analysts have been pursuing alternative objective methods for the identification of ARMA models. One of the most widely used criteria is the Akaike Information Criterion (AIC) [27]. Refer to *3.1.4.4 Akaike Information Criterion.*

Alternative to AIC and widely used too, is the minimization of the SSE discussed in *3.1.4.1 Sum of Squared Errors*. The difference is, while AIC assume a trade-off between the complexity/order for the model and the minimization of the errors, the SSE minimization only takes into account the minimization of the errors.

Conscientious of who the audience for this work is (i.e. mobile operators), the implementation of objective methods is essential since there are no trained experts for the construction of the forecast methods. The model selection is done jointly with the parameters estimation, next explained.

Parameters Estimation

Here the aim is to find the autoregressive ($\phi_1,\ \phi_2, \dots\ \phi_p$) and moving average ($\theta_1, \theta_2, \dots, \theta_q$) parameters that achieve the best fitting. The computation of the best fitting parameters is done by minimize the Sum of Square Errors (SSE) studied in *3.1.4.1 Sum of Squared Errors*:

$$\boldsymbol{SSE}(\phi, \theta) = \sum_{t=1}^{n} e^2(t)$$

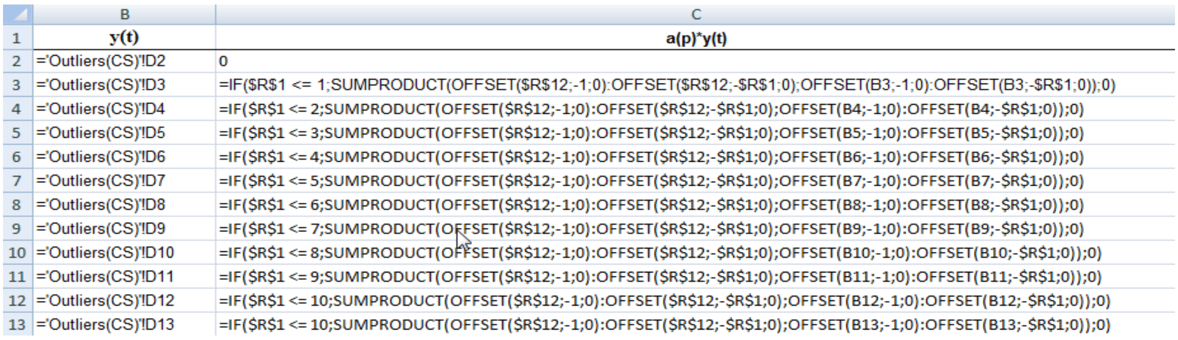
with,

$$e_t = y_t - \left[\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}\right]$$

Note that, to calculate the first e_t, we need to know the values of $e_{t-1}, e_{t-2}, \dots, e_{t-q}$, which we do not know. The convention is to assign zero to all the unknown values previous to e_t.

The model selection and parameters estimation are the main important stages in the construction of the model. Thus, the next figures demonstrate by steps how to compute the model using Excel.

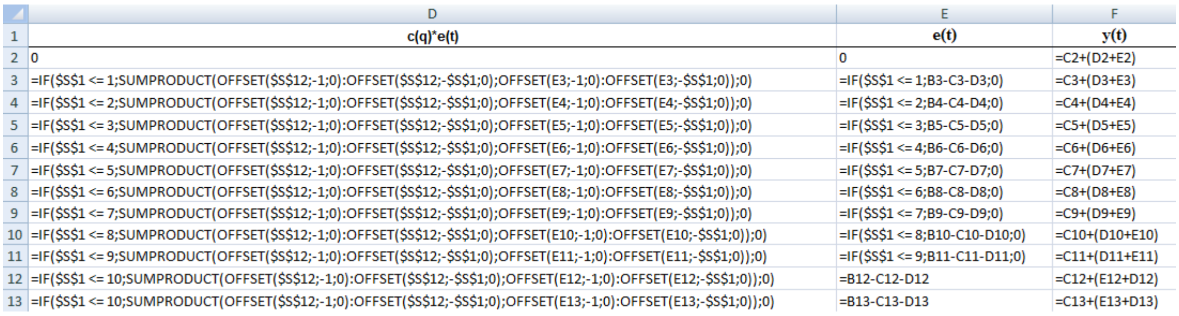
Figure 58 shows how the AR(p) part of the ARMA model is performed. Note that the column B has the observations after passing through the outliers' technique discussed in *3.1.4.6 Outliers and influential observations.*



	B	C
1	y(t)	a(p)*y(t)
2	='Outliers(CS)'!D2	0
3	='Outliers(CS)'!D3	=IF($R$1 <= 1;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B3;-1;0):OFFSET(B3;-$R$1;0));0)
4	='Outliers(CS)'!D4	=IF($R$1 <= 2;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B4;-1;0):OFFSET(B4;-$R$1;0));0)
5	='Outliers(CS)'!D5	=IF($R$1 <= 3;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B5;-1;0):OFFSET(B5;-$R$1;0));0)
6	='Outliers(CS)'!D6	=IF($R$1 <= 4;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B6;-1;0):OFFSET(B6;-$R$1;0));0)
7	='Outliers(CS)'!D7	=IF($R$1 <= 5;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B7;-1;0):OFFSET(B7;-$R$1;0));0)
8	='Outliers(CS)'!D8	=IF($R$1 <= 6;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B8;-1;0):OFFSET(B8;-$R$1;0));0)
9	='Outliers(CS)'!D9	=IF($R$1 <= 7;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B9;-1;0):OFFSET(B9;-$R$1;0));0)
10	='Outliers(CS)'!D10	=IF($R$1 <= 8;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B10;-1;0):OFFSET(B10;-$R$1;0));0)
11	='Outliers(CS)'!D11	=IF($R$1 <= 9;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B11;-1;0):OFFSET(B11;-$R$1;0));0)
12	='Outliers(CS)'!D12	=IF($R$1 <= 10;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B12;-1;0):OFFSET(B12;-$R$1;0));0)
13	='Outliers(CS)'!D13	=IF($R$1 <= 10;SUMPRODUCT(OFFSET($R$12;-1;0):OFFSET($R$12;-$R$1;0);OFFSET(B13;-1;0):OFFSET(B13;-$R$1;0));0)

Figure 58: AR(p) part of ARMA

Figure 59 shows the computation of the MA(q) part of the ARMA model, the errors and the predictions.



	D	E	F
1	c(q)*e(t)	e(t)	y(t)
2	0	0	=C2+(D2+E2)
3	=IF($S$1 <= 1;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E3;-1;0):OFFSET(E3;-$S$1;0));0)	=IF($S$1 <= 1;B3-C3-D3;0)	=C3+(D3+E3)
4	=IF($S$1 <= 2;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E4;-1;0):OFFSET(E4;-$S$1;0));0)	=IF($S$1 <= 2;B4-C4-D4;0)	=C4+(D4+E4)
5	=IF($S$1 <= 3;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E5;-1;0):OFFSET(E5;-$S$1;0));0)	=IF($S$1 <= 3;B5-C5-D5;0)	=C5+(D5+E5)
6	=IF($S$1 <= 4;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E6;-1;0):OFFSET(E6;-$S$1;0));0)	=IF($S$1 <= 4;B6-C6-D6;0)	=C6+(D6+E6)
7	=IF($S$1 <= 5;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E7;-1;0):OFFSET(E7;-$S$1;0));0)	=IF($S$1 <= 5;B7-C7-D7;0)	=C7+(D7+E7)
8	=IF($S$1 <= 6;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E8;-1;0):OFFSET(E8;-$S$1;0));0)	=IF($S$1 <= 6;B8-C8-D8;0)	=C8+(D8+E8)
9	=IF($S$1 <= 7;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E9;-1;0):OFFSET(E9;-$S$1;0));0)	=IF($S$1 <= 7;B9-C9-D9;0)	=C9+(D9+E9)
10	=IF($S$1 <= 8;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E10;-1;0):OFFSET(E10;-$S$1;0));0)	=IF($S$1 <= 8;B10-C10-D10;0)	=C10+(D10+E10)
11	=IF($S$1 <= 9;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E11;-1;0):OFFSET(E11;-$S$1;0));0)	=IF($S$1 <= 9;B11-C11-D11;0)	=C11+(D11+E11)
12	=IF($S$1 <= 10;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E12;-1;0):OFFSET(E12;-$S$1;0));0)	=B12-C12-D12	=C12+(E12+D12)
13	=IF($S$1 <= 10;SUMPRODUCT(OFFSET($S$12;-1;0):OFFSET($S$12;-$S$1;0);OFFSET(E13;-1;0):OFFSET(E13;-$S$1;0));0)	=B13-C13-D13	=C13+(E13+D13)

Figure 59: MA(q), Error and Predictions

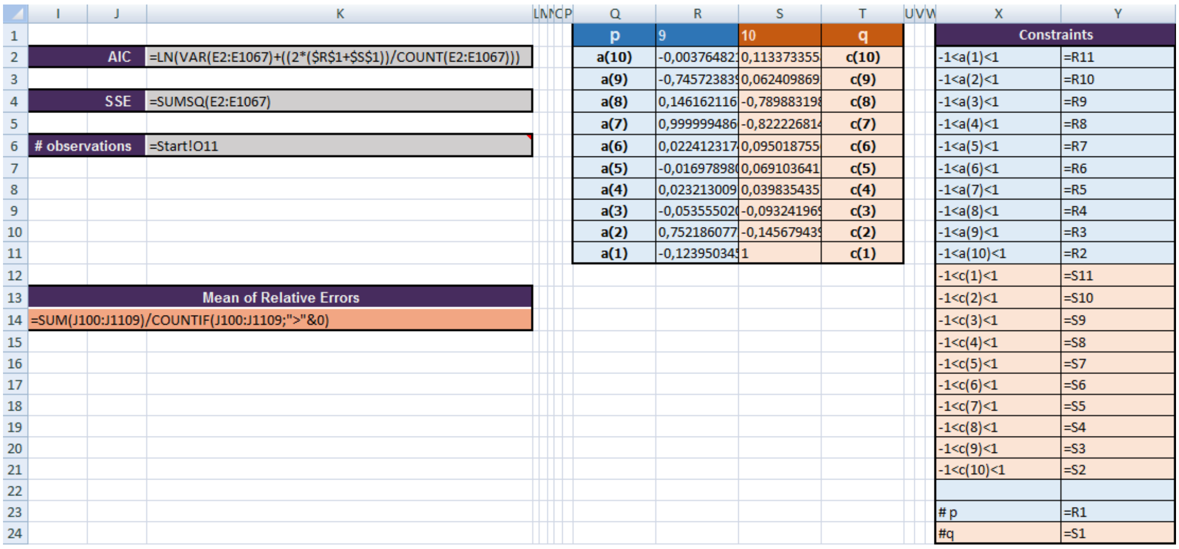
The errors (column E), are computed by subtract to the real observation, the prediction value:

$$e_t = y_t - \left[\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}\right]$$

The prediction values (column F) are the sum of the AutoRegressive (column C), Moving Average (column D) and the errors (column E):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

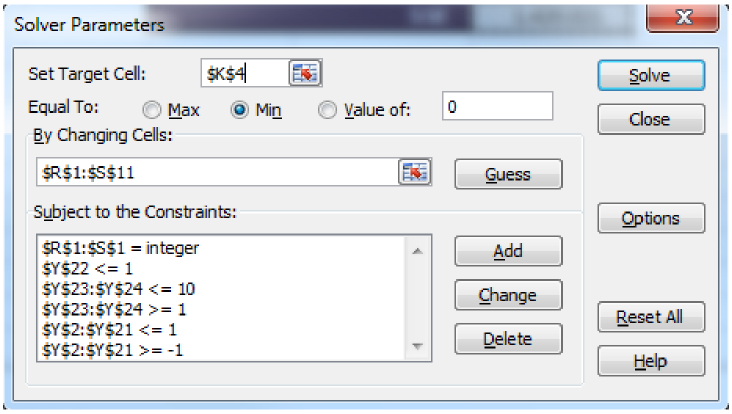
Note, in *Figure 60*, that the SSE is easily calculated. The figure also shows the method parameters as well as the constraints that must be fulfilled. These are the important fields because they are inputs to the *Solver* add-in used to fit the model.



	I	J	K	Q	R	S	T	X	Y
1				p	9	10	q	Constraints	
2	AIC		=LN(VAR(E2:E1067)+((2*($R$1+$S$1))/COUNT(E2:E1067)))	a(10)	-0,00376482	0,11337335	c(10)	-1<a(1)<1	=R11
3				a(9)	-0,74572383	0,062409869	c(9)	-1<a(2)<1	=R10
4	SSE		=SUMSQ(E2:E1067)	a(8)	0,146162116	-0,78988319	c(8)	-1<a(3)<1	=R9
5				a(7)	0,999999486	-0,82222681	c(7)	-1<a(4)<1	=R8
6	# observations		=Start!O11	a(6)	0,022412317	0,095018755	c(6)	-1<a(5)<1	=R7
7				a(5)	-0,01697898	0,069103641	c(5)	-1<a(6)<1	=R6
8				a(4)	0,023213009	0,039835435	c(4)	-1<a(7)<1	=R5
9				a(3)	-0,05355502	-0,09324196	c(3)	-1<a(8)<1	=R4
10				a(2)	0,752186077	-0,14567943	c(2)	-1<a(9)<1	=R3
11				a(1)	-0,12395034	1	c(1)	-1<a(10)<1	=R2
12								-1<c(1)<1	=S11
13	Mean of Relative Errors							-1<c(2)<1	=S10
14	=SUM(J100:J1109)/COUNTIF(J100:J1109;">"&0)							-1<c(3)<1	=S9
15								-1<c(4)<1	=S8
16								-1<c(5)<1	=S7
17								-1<c(6)<1	=S6
18								-1<c(7)<1	=S5
19								-1<c(8)<1	=S4
20								-1<c(9)<1	=S3
21								-1<c(10)<1	=S2
22									
23								# p	=R1
24								#q	=S1

Figure 60: Parameters and Constraints

The utilization of *Solver* was already explained in *3.3.3.1.1 Excel Implementation* to compute the smooth parameters in the Holt-Winters method. However, the use of it is reminded here.



Solver Parameters

Set Target Cell: $K$4

Equal To: ○ Max ● Min ○ Value of: 0

By Changing Cells: $R$1:$S$11

Subject to the Constraints:

```
$R$1:$S$1 = integer
$Y$22 <= 1
$Y$23:$Y$24 <= 10
$Y$23:$Y$24 >= 1
$Y$2:$Y$21 <= 1
$Y$2:$Y$21 >= -1
```

Figure 61: Solver Settings

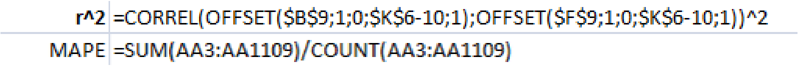
Figure 61 shows the setting of *Solver*. The target cell is the SSE (K4), the objective is to minimize the cell (Equal To: Min) by changing the ϕ and θ parameters (matrix R1:S11) subjected to the constraints in interval Y2:Y24.

Note that the order (p,q) of the model is entered as a constraint between 1 and 10, this is the stage of "Model Selection" done automatically instead of using the inspection of ACF and PACF.

Verification diagnosis

There are different indicators concerning the reliability of the method obtained. Some of the most commonly used are presented in *3.1.4 Forecast Evaluation*. *Figure 62* shows the *Excel* calculations of these techniques. The Durbin-Watson is also an indicator widely used in this type of verification.

The techniques represented in the figure are focus on the verification of the model goodness of fit, i.e. how well the model fits the data.



r^2	=CORREL(OFFSET($B$9;1;0;$K$6-10;1);OFFSET($F$9;1;0;$K$6-10;1))^2
MAPE	=SUM(AA3:AA1109)/COUNT(AA3:AA1109)

Figure 62: Reliability Indicators

Forecast

After all the latter steps are fulfilled, the conditions to produce predictions are reunited. Equation *(3.29)* can be applied to produce the prediction y_t. To predict h-steps ahead, it can be done by recurrent application of the equation.

3.3.4.3.1. Result analysis

Figure 63 illustrates the forecasting results for a period of time of around three months using an objective model based on the minimization of the SSE. The estimated model is an ARMA(8,9) with a MAPE of 9.02% and a mean of relative errors equal to 14.47%. The data used to evaluate the methods in this work is the same allowing a fair comparison between them.

The MAPE for the equivalent forecast using Holt-Winters is 0.23%, which is much less than the 9% obtained for the ARMA model. This is reflected in an increase of around 4% for the mean of relative errors.

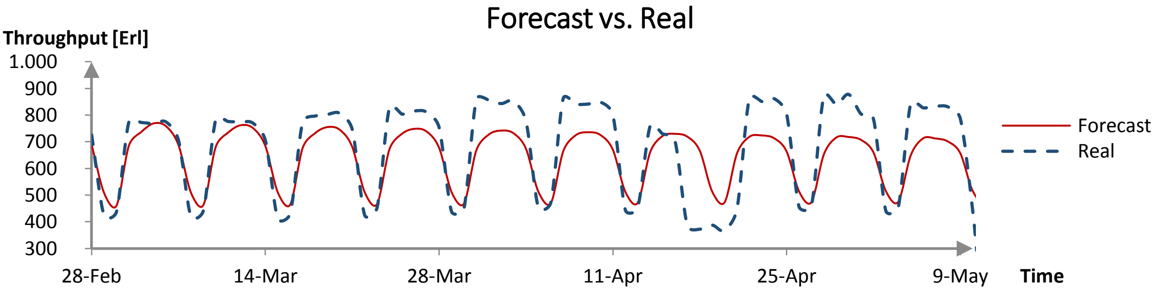


Figure 63: ARMA forecast

Comparing *Figure 63* with *Figure 47* (equivalent for the Holt-Winter method) is possible to observe that the Hot-Winters method captures the time series pattern much better. A potential reason for this to happen is that the ARMA adapts quickly to the data and the days before the forecasting were not good for fitting the model, refer to *Figure 64*.

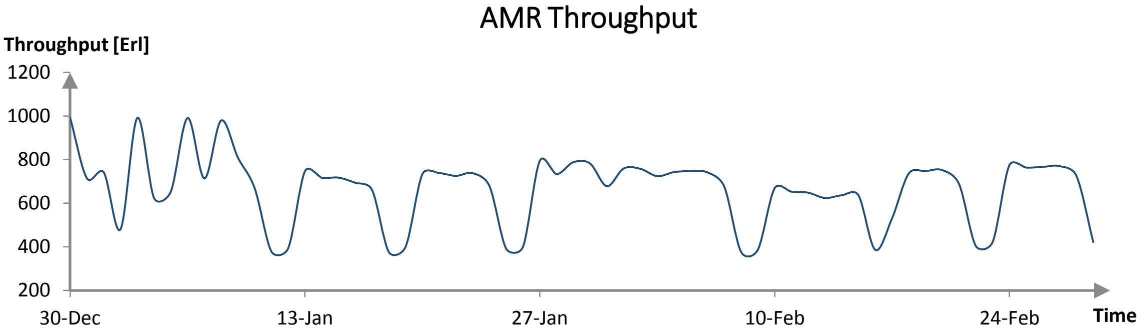


Figure 64: AMR daily traffic

The following case studies are equivalent to the ones analyzed in *3.3.3.2* for the Holt-Winters methods. This allows comparing the behavior of the model for the same problems.

As already mentioned before the ARMA model should be applied to more stable data (stationary property). However, *Figure 65* shows that the model presents a normal response (identical to the Holt-Winters method).

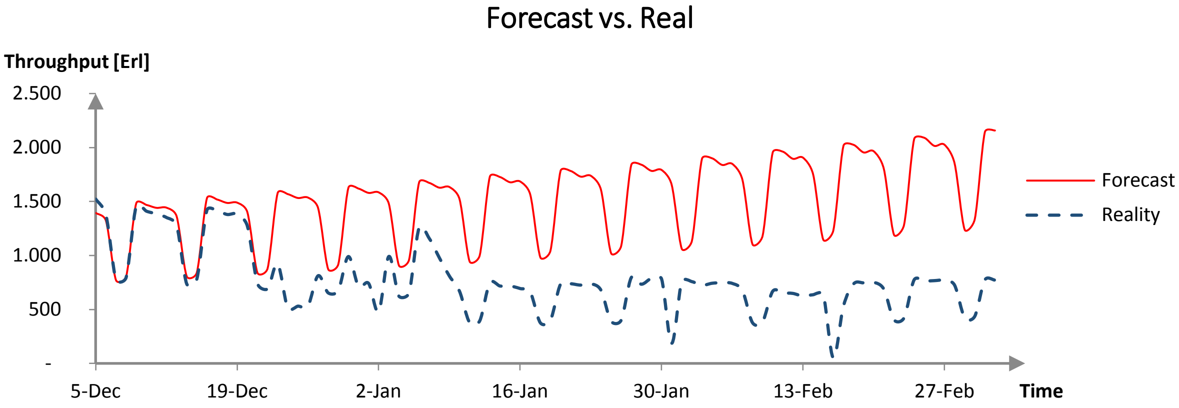


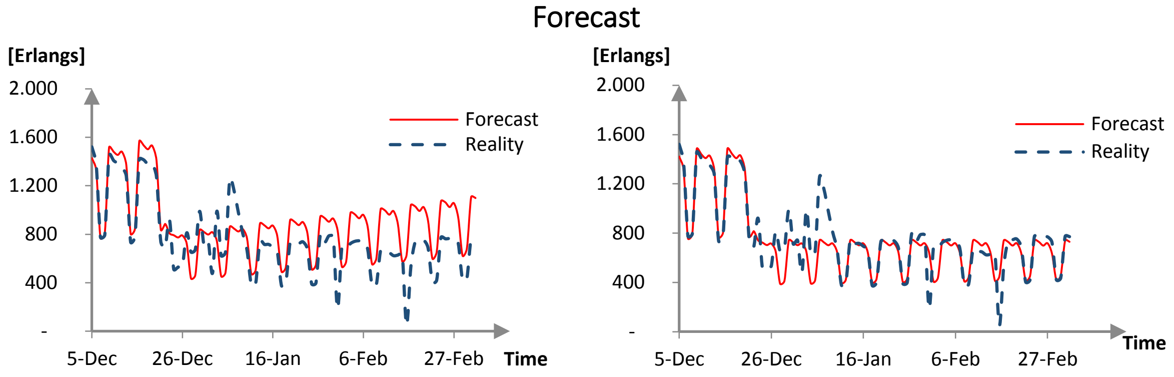
Figure 65: ARMA case study #1

Observe that the model has captured the increasing trend of the past values and continues with this trend to the forecast values. The problem here is that around Dec-20 the level of the daily average AMR throughput falls to a lower level while the forecast values continue its increment.

The MAPE for this example is 3.64% which suggest a very good fitting of the model. Note that the pattern is much better capture than for the latter example (*Figure 63*) where the MAPE was 8%.

However, this method presents a more explosive reaction than the Holt-Winters in this situation and the mean of relative errors is 188.60%, more 59.48% than the obtained for the Holt-Winters.

It is clear that this method cannot predict unusual situations. Again if these situations are somehow foreseen, they can be inputted in the *Forecasting Tool.*



a) “Level Offset” (only) implemented

b) “Level Offset” and “Manual Growth” implemented

Figure 66: Case study #1 (Level Offset and Manual Growth)

The chart a) is applied a level offset of -50%, but note that the forecast result continues with the trend captured from the previous observations. This trend is possible to remove using the “*Manual Growth”* explained in *3.5.1 Manual Growth.*

Using these two techniques the mean of relative errors fell to 70.40% and 13.70% respectively. These results were calculated without outliers on the real values.

Figure 67 is the equivalent to the 3rd case study analyzed for the *Holt-Winters Exponential Smoothing* method. The model used here is an ARMA(7,7). The blue dashed line represents the real values and the solid red line represents the forecast results.

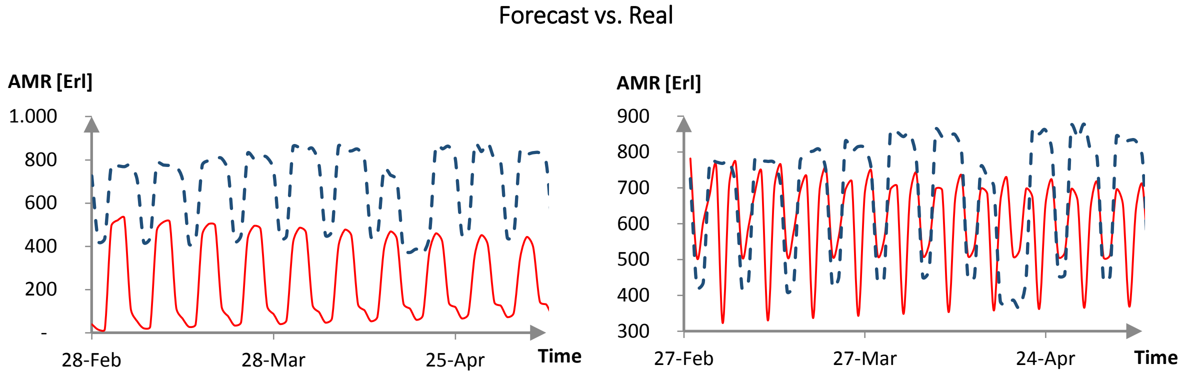


Figure 67: AMR case study #3

Equation *(3.29)* shows us that the calculus of the forecast values $\hat{y}_t$ is directly related with the last observations and errors, in this particular case using an ARMA(7,7), $\hat{y}_t$ is directly related with the seven last observations and the seven last residues.

Figure 67 represents two examples in which the data contains outliers in the last observations. The kind of effect on the forecast results depends on which observation(s) is (are) the outlier(s).

This case study is clear evidence that this method doesn't produce good results in the presence of outliers, especially in the last observations. Thus, before using this method the outliers must be removed.

Figure 68 represents the forecast results obtained with the same model and data used to produce the results in *Figure 67*, but without outliers. The outliers were handled using the technique explained in *3.1.4.6 Outliers and influential observations*.

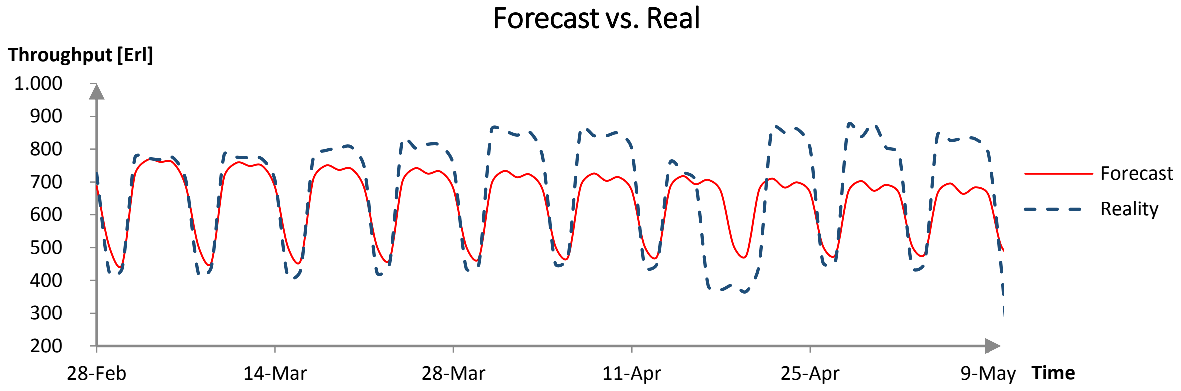


Figure 68: AMR case study 3 without outliers

While the mean of relative errors obtained in the forecasts of *Figure 67* were respectively 68.18% and 24.30%, the forecast results of *Figure 68* is 14.28%.

3.3.4.4. Seasonal ARMA

The fact of the data presents seasonality leads to introducing the seasonal ARMA, denoted by ARMA(p,q)$_S$, which can be written as follows:

$$y_t = \phi_1 y_{t-1s} + \phi_2 y_{t-2s} + \cdots + \phi_p y_{t-ps} + \theta_1 e_{t-1s} + \theta_2 e_{t-2s} + \cdots + \theta_q e_{t-qs} + e_t \qquad (3.1)$$

ϕ_p → Autoregressive coefficient for p-lag

θ_q → Moving average coefficient for q-lag

e → Residuals

y_{t-ps} → Observation at lag $p \times s$

$s \rightarrow$ Seasonal period length

As the non-seasonal ARMA, the Seasonal ARMA can also be represented by a block diagram similar to the presented in *Figure 56.* The only difference is the delay block instead of delaying one day it must delay s (seasonal period length) days. This because the seasonal ARMA presents the series in terms of its past values at lag equal to the length of the period s, while the non-seasonal ARMA does it in terms of its past values at lag 1. Seasonal ARMA incorporates the seasonality into the model.

3.3.4.4.1. Forecast

The data that is use to test the Seasonal ARMA (S-ARMA) is the same used to test the latter forecasting techniques (Linear Regression, Holt-Winters and ARMA).

Figure 69 shows very poor forecasting results obtained with a SARMA(10,10)$_7$. These results are consequential of a big instability on the data during the time Dec-22 until Jan-10 (refer to *Figure 70*).

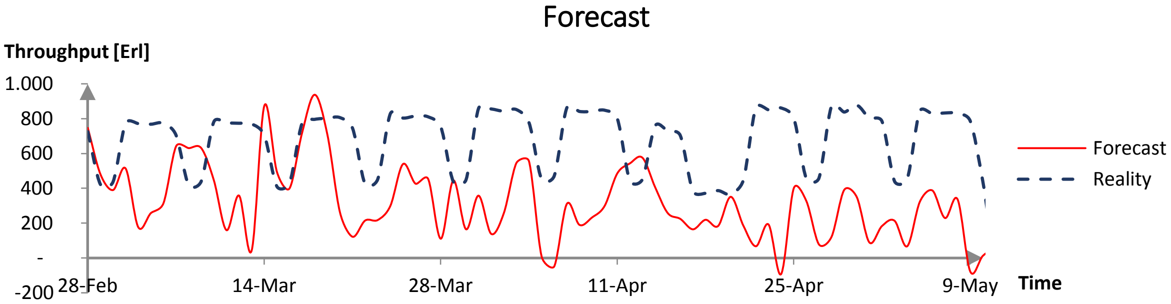


Figure 69: SARMA results

Referring to equation *(3.24)* and *Figure 70* we quickly conclude that the predicted values have a direct dependence on the instable data leading to poor forecast results.

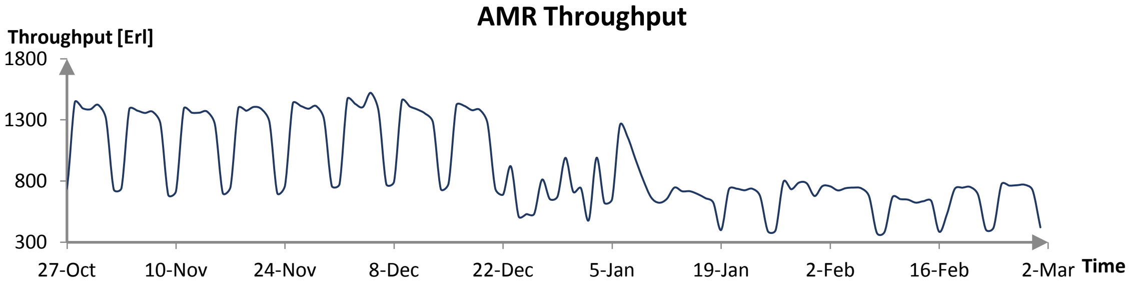


Figure 70: AMR daily traffic

The instability visible in *Figure 70* during the time from Dec-22 until Jan-10 is a possible consequence of the holiday season. This instability falls completely off the normal pattern of the data, which has negative repercussion on the forecast when including these values to predict into the future, which is the case of *Figure 69*.

The problem discussed above shows us that the Seasonal ARMA doesn't works very well when there are outliers (instability) in the observations used for forecasting, which was expected. Also the value of 12.21% for MAPE is the second higher after the linear regression (that doesn't capture seasonality), which indicates that the model has the second worst fitting model to the data set.

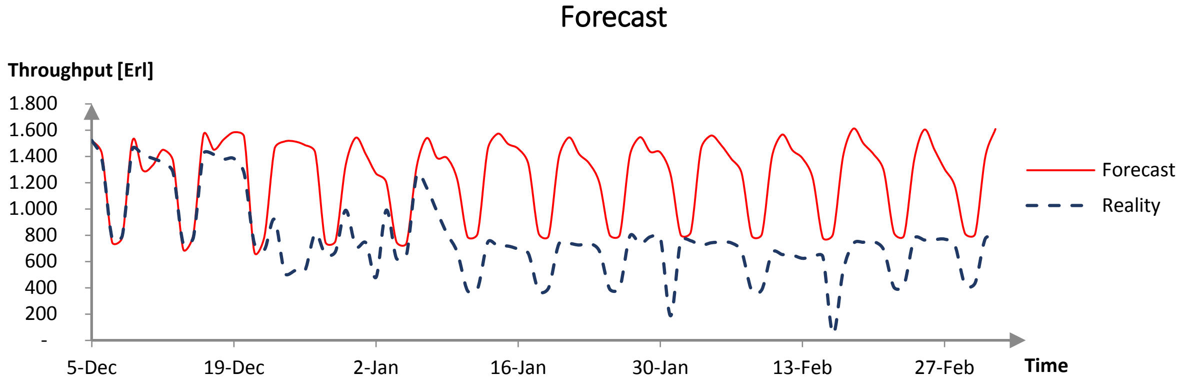


Figure 71: SARMA case study #1

Figure 71 illustrates the case study were the data rapidly increases in the last data observations and suddenly falls to a much lower level.

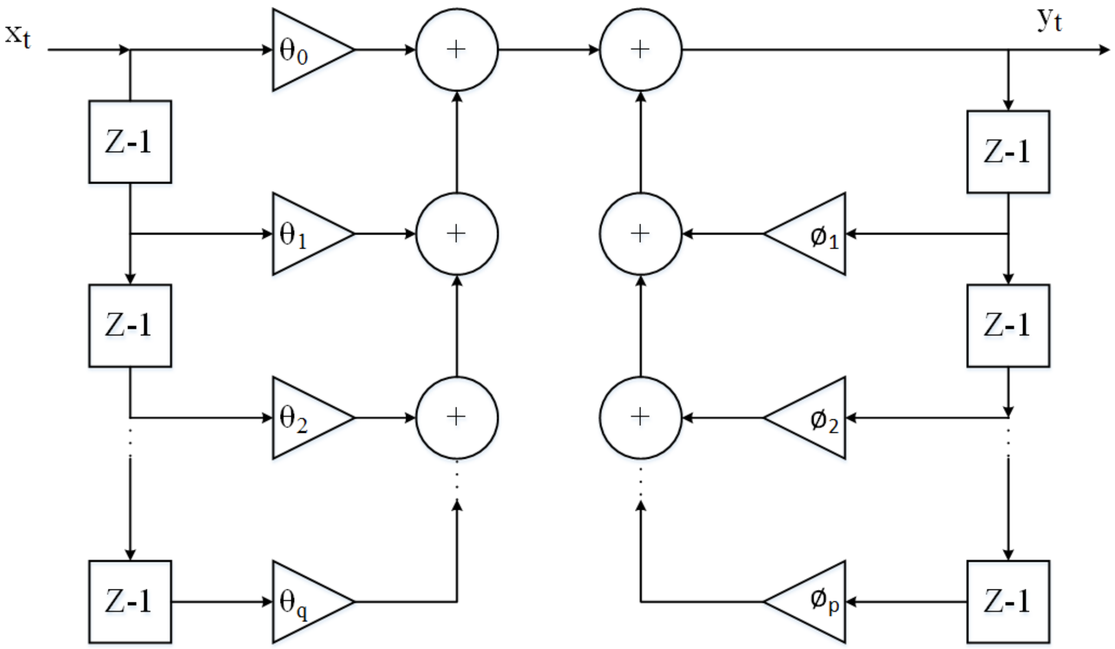
The value for the MAPE is 3.58% which means that the model fits the data very well. However, identically to the Holt-Winters and ARMA models, the SARMA also doesn't handle this phenomenon very well. Still, this method captures a longer trend in the data that's why the forecasted trend is softer than the obtained with the ARMA model.

The mean of relative errors obtained is 92% which still worse than the Holt-Winter method but an improvement relatively to the non-seasonal ARMA.

Note that the techniques "*Level Offset*" and "*Manual Growth"* (*3.5.1*) also can be applied here.

3.4. General Purpose

Aside of the Linear Regression model that is used more as a benchmark, all the other models (Holt-Winters, ARMA and SARMA) can be described using the combination FIR-IIR digital filters. However in this work, the Holt-Winters method uses this general purpose only to initiate the method, which can be replaced by other initiation technique.



The **ARMA** model can be described by this general purpose where x are the residuals, θ the MA coefficients, the $\emptyset$ the AR coefficients and y the prediction value. The **SARMA** model can be described in the same way that the ARMA model, the only difference is the delays.

Making all θs equal to zero the result is an IIR filter describing an **AutoRegressive** model. Making the $\emptyset s$ all equal to zero, the result is a FIR filter describing a **Moving Average** model.

The **Holt-Winters** use this general model (in a FIR/MA filter form) to perform some initializations, but this general model cannot perform the Holt-Winters method in any way.

Besides the methods described above, this general purpose architecture can be used in the classical decomposition described in *3.3.2 Classical decomposition (*the foundation of several forecast methods), more specifically to perform moving averages. This general model is then a powerful concept for the forecasting matter.

3.5. Add-ons and Enhancements

This chapter has the objective to present some ideas that emerged during the development of this work and had proven to be very useful.

3.5.1. Manual Growth

Unusual increases/decreases in the data due to external factors are for sure an occurrence that the methods cannot predict. Facing this problem, the manual growth feature pretends to attenuate it. Imagine that the management personal foresees that in the next months there will be an increase of 30% per day of the market; it would be good to add this increase to the forecast obtained.

Figure 72 shows the general purpose with the addition of the manual growth. As already mentioned the general purpose doesn't perform the Holt-Winters in any way, but in case of Holt-Winters the idea remains the same.

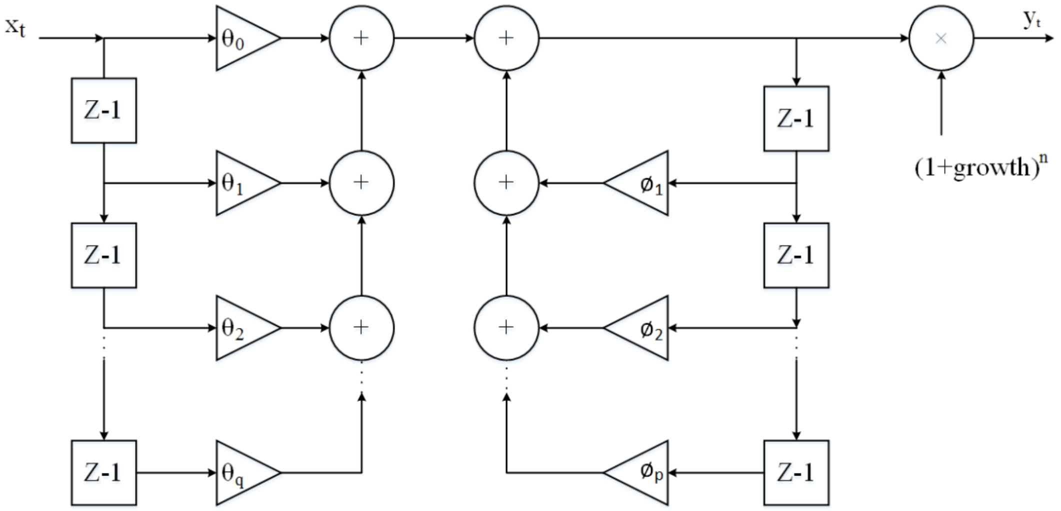


Figure 72: General model plus manual growth

In the image,

$growth$ → Value for the growth (e.g.0.3 = 30%)

n → Prediction number ($n = 1,2, \ldots, N$)

N →Number of predictions

Note that the growth is applied to the granularity period of the data, i.e. to daily data is applied a daily growth.

The same type of idea applies to add a trend to the forecast but based in a number of months defined by the user in charge for forecasting.

3.5.2. Level Offset

A problem studied before in this chapter (cases study #1) is that the methods are unable to predict unusual observations due to external factors, such as: conventions, sites shutdown, sites modifications, new application, etc.

The level offset aims to attenuate this problem by allowing the user to add/subtract an offset to the forecast results. This idea is very similar to the manual growth and also easily implemented.

These two ideas allow the user to input some of his knowledge about what will happen in the future, that the systems don't have. This has a big impact in the forecast results and when well-done can produce very good forecasts.

3.6. CS Throughput Analysis

This chapter has the objective to summarize the results obtained in *3.3Time Series* and take some conclusions based on it.

Table 4 summarizes the most relevant indicators about the forecast methods. The values in this table were obtained using daily average AMR (CS) throughput data.

	Linear Regression	Holt-Winters	ARMA	SARMA
MAPE [%]	35,71%	0,23%	9,02%	12,17%
Relative Error[4]	46,05%	10,98%	14,47%	58,94%
SSE	23.003.926	1.401	2.182.604	2.631.454
R^2	0,0024	0,9999	0,9999	1,0000
Durbin-Watson	-	-	2,00	1,94

Table 4: Methods Summary

Note that this is a summary for a particular case and before choosing a method all the aspects discussed in the latter chapters must be taken into account. For example, if we based on the MAPE value to choose a method, we would say that the SARMA is not a bad one. However, its predictions are very far from the reality, this is a perfect example of a bad option.

Reading the table:

- Rapidly concludes that the Holt-Winters method is the one that achieve the best results. The MAPE (as well as the SSE) says that the Holt-Winters is the method that best fits the data. The mean for the relative errors is also the lowest of the methods.

[4] This value is in fact the mean of the relative errors between the predictions and real values.

- Note that the MAPE for the SARMA is lower than the obtained with the linear regression. However, the mean for the relative errors is higher for SARMA. This is clear evidence that the method that achieves the minimum MAPE may not be the one that produces the best results. The reason for the bad SARMA results is explained in chapter *3.3.4.4.1*.
- Note that the difference of MAPEs between the ARMA and SARMA models is not so high, however the difference between the mean for the relative errors is huge. The reason for this is explained in chapter *3.3.4.4.1*.

In a general point of view all the indicators about the quality of the methods are in accordance with the results achieved.

Case Study #1

Table 5 shows the values obtained with different methods for the first case study. Recalling the first case study, it has to do with the rapidly increase of the data (possibly due to external factors) and a sharp fall to a lower level.

	Linear Regression	Holt-Winters	ARMA	SARMA
MAPE [%]	25,88%	0,29%	3,65%	3,58%
Relative Error	192,26%	129,12%	188,60%	92,36%
SSE	7.943.567	1.269	264.958	178.384
R^2	0,4822	0,9999	1,0000	1,0000
Durbin-Watson	-	-	1,94	1,72

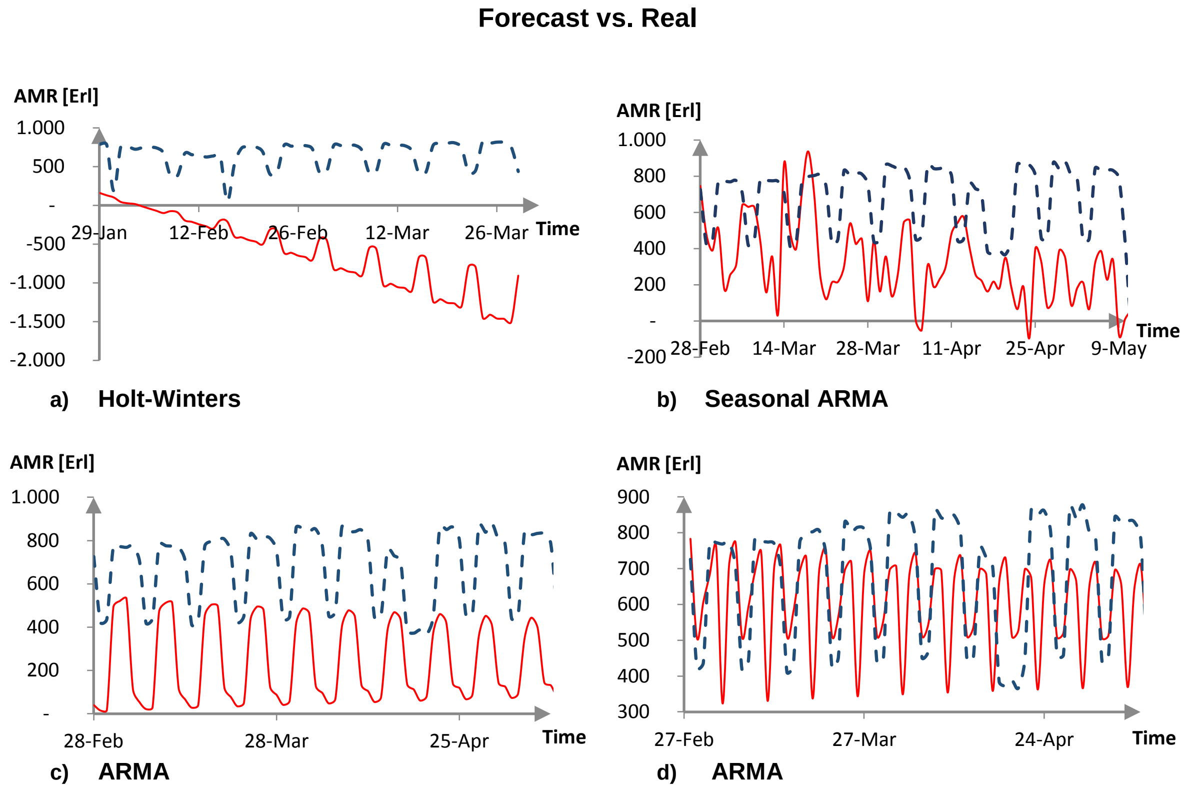
Table 5: Case Study #1

- The Linear Regression presents the worst MAPE, as expected since it doesn't capture seasonality. Accordingly to that it also has the worst mean for the relative errors.
- Looking to the MAPEs values says that the Holt-Winters is the method that best fits the data, and at a first sight, the obvious choice would be the Holt-Winters. However, this method is not the one that has the lower mean for the relative errors. In fact, SARMA holds the lowest value for the mean of relative errors, because the Holt-Winters method follows best the trend of the data while the SARMA stabilizes. This is possible to see if you relate to *Figure 48 and Figure 71.*
- The ARMA model holds the second worst MAPE values as well as the mean for the relative errors. This is due to the explosive behavior in this type of conditions. *Figure 65* shows this behavior and is possible to notice an increasing in the trend. This is good example why this method should be used in more stable markets (data stability).

The bad results produced in this case study can be attenuated using the "*Level Offset*" and "*Manual Growth"* (*3.5.1*). These improvements were already done in within the explanation of each method in the latter chapters.

Case study #3

The third case study is related with the presence of outliers in the data. Here, the use of numeric results is not the best approach to evaluate the methods. Thus, *Figure 73* presents the behaviors of the different methods in the presence of outliers in the last observations. The solid red line represents the forecast result and the dashed blue line represents the real values.



a) **Holt-Winters**

b) **Seasonal ARMA**

c) **ARMA**

d) **ARMA**

Figure 73: Outliers Influence

Holt-Winters (a) – Exhibit the worst behavior in the presence of outliers. Note that this behavior was capture with an outlier in the last observation, which is directly related with the predictions calculations.

SARMA (b) – In fact the case study #3 for the SARMA model is not particularly discussed in *3.3.4.4 Seasonal ARMA*. However it is possible to deduce some conclusions from chart (b), replica of *Figure 69.* The source of this bad result is the unusual variations in the data covered by the SARMA prediction calculations.

ARMA (c&d) – The charts c) and d) illustrate two cases where the outliers are in different observations (in time). The objective here is to show that the outliers can cause different behaviors of the model. The same happens for SARMA while the Holt-Winters is more affected only in the last observation. The causes for this are explained in the latter chapters.

Linear Regression – Although no figure is presented about the linear regression it must be discussed here too. The influence of outliers in the linear regression is explained in *3.2.1.1* with more detail. In general, the influence of outliers on the data depends on the length of the data used to calculate the slope and intercept, the position of the outliers and how influential the outlier is.

3.7. PS Throughput Analysis

The latter chapters had study several mathematical forecasting methods. Its properties, complexities and important cases study. That analysis used daily average CS throughput.

The choice of CS throughput data for the analysis of the methods was deliberated. The reason of this choice was the strong weekly seasonality in this type of data and typical variances (seasonality) during the seasons (i.e. Christmas, holidays, etc.), which provide a widely methods analysis.

This chapter pretends to study the characteristics of the PS throughput data type, as well as, compare the performance of the different models.

Figure 74 presents typical values for the daily average PS throughput. The black dashed line represents the actual values without any technique to manage outliers and the solid red line represents the values after passing through the outliers handling technique studied in *3.1.4.6 Outliers and influential observations.*

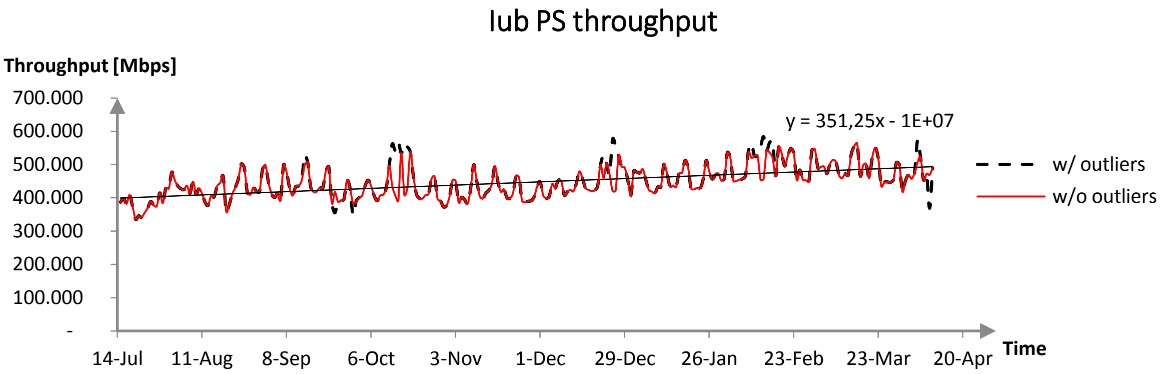


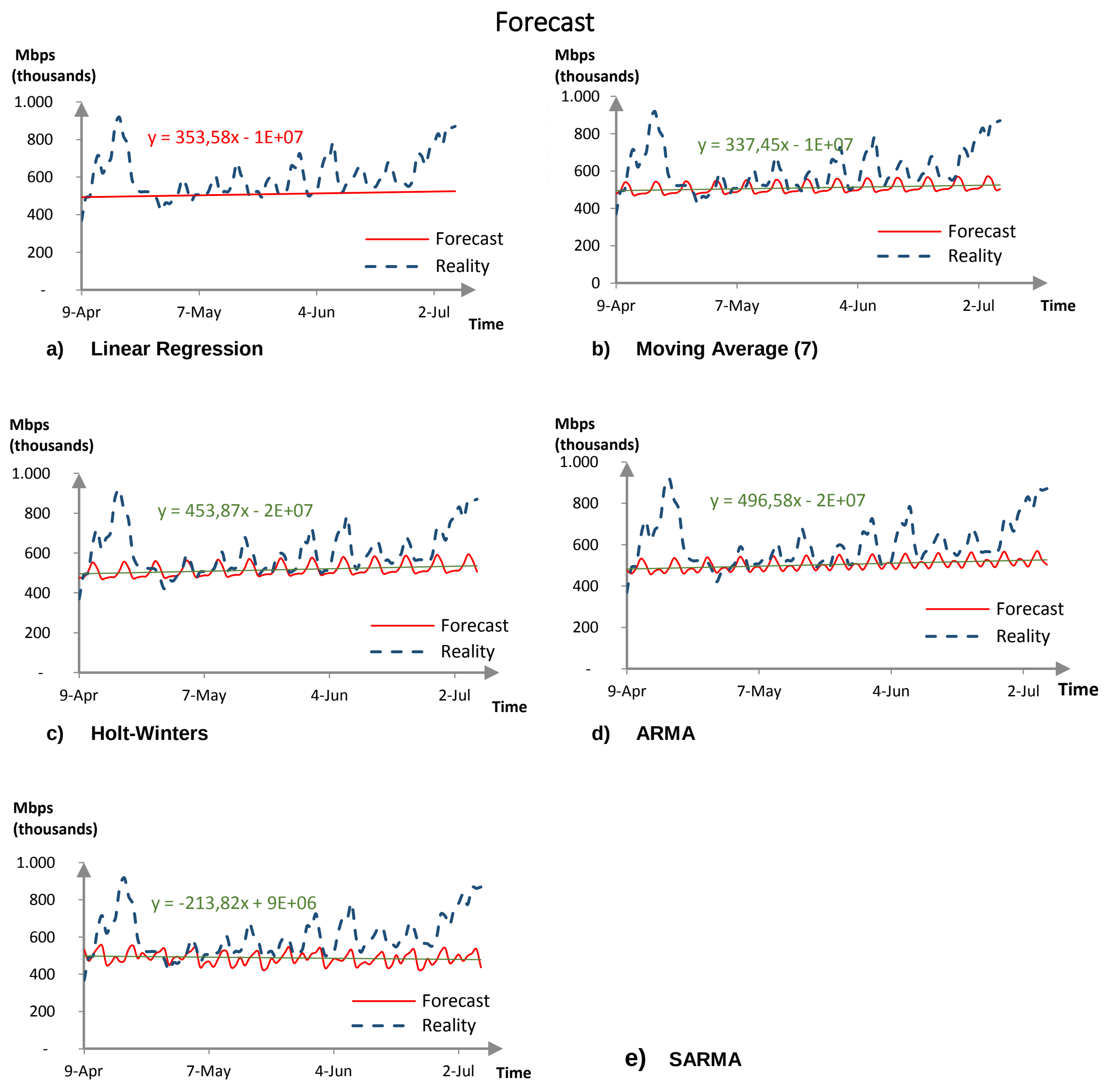
Figure 74: Daily average PS throughput

Although possible to see that the data presents weekly seasonality, this is not as evident as in the CS data. Another easy observation is, unlike CS data the PS throughput doesn't presents variances with the season of the year, e.g. in the Christmas season usually exists a decreasing in the CS throughput, here it is possible to see that the PS throughput continues its growth.

One of the motivations of this work is the changing in the character of the mobile networks that is making the data traffic explode. *Figure 74* also illustrates this growth*.*

Due to the non-stationary behavior of this type of data, the ARMA and SARMA methods should not be used. Thus, the prediction will be compute using the Linear Regression, Holt-Winters and a multiplicative classical decomposition as explained in

Multiplicative **model***.*



a) **Linear Regression**

b) **Moving Average (7)**

c) **Holt-Winters**

d) **ARMA**

e) **SARMA**

Figure 75: PS throughput forecast

Figure 75 shows the forecast result using all methods for the same time period. The black line represents real values, the red line represents the forecast results and the green line is the trend line of the forecast results. Note that all the methods, a part of the linear regression (as expected), capture the weekly seasonality in some way.

Linear regression – As expected the forecast results doesn't present any type of seasonality because this method is unable to capture seasonality, it only captures the trend. Note however, comparing with the reality, that in the last observations the real values start to have a sharper slope. The slope of the prediction depends on what observation where used to calculate it.

Moving Average – It was used a moving average filter of sixth order (or 7 taps), as represented in *Figure 35*. This makes possible to capture the week seasonality. To this seasonality it was add the trend of the last observations. Note that the slope of the forecast depends on what observation where used to calculate it. For example in *Figure 75* the slope was obtained with the maximum of last observations (270 days) and the result is 337.45 Mbps per day, if the number of last observations used is reduced to 200 days, the slope becomes 456.83 Mbps per day but if it is reduced to 60 days becomes -734.66 Mbps per day because decays in that period. Thus, this choice must be done carefully. The wisest choice is normally the maximum number of observations.

Holt-winters – The Holt-Winter can capture the seasonality and also the trend and produce a good forecast result. However the data passes through the outliers handling technique, the presence of an unusual value in the last observation can cause a bad forecast.

ARMA – Although not appropriated to data with trend, the forecast result doesn't present any instability. However, it presents the most "explosive" slope because it was used an ARMA(10,10), which means that the last ten observations are directly related with the forecast and the growth in these observations is huge. This is the reason for a so good forecast and also the reason why it should not be used.

SARMA – The seasonal ARMA is the one that presents the worst forecast once it predicts that the PS throughput will decrease. The fact that the predictions are decreasing is due to the past observations directly related with the predictions are decreasing. However, it is not recommended to use this method in data with trend.

As already known from previous analysis, the methods used cannot predict the occurrence of unusual events (e.g. the peak around April-18). Although, techniques such as the *Manual Growth* and *Level Offset* studied before help to attenuate this problem.

Table 6 summarizes the most important indicators of the forecast results in *Figure 75*.

	Linear Regression	MA(7)	Holt-Winters
MAPE [%]	6.88%	5.20%	0.15%
Relative Error[5]	15.76%	15.66%	14.67%
Relative Error w/o outliers	10.74%	10.62%	9.73%
$\mathbf{R}^2$	0.35	0.59	0.99997

Table 6: Daily average PS throughput forecast results summary

A quickly analysis to the table shows that all the values are in accordance, i.e. the method that presents the small MAPE also presents the better coefficient of determination and the small relative error.

Referring to the MAPE value, the linear regression is the method that poor fit the data but the moving average is not better. The Holt-Winters is by far the best fitting model. However, the relative error is not much better than the relative errors for the linear regression and moving average.

If the real values to be compared with the forecast results are passed through the outliers' technique the relative error becomes around 5% lower.

In general, all the methods produce good results if there are no unusual observations and the difference between them is not so big. The best option here is to choose the less complex method.

[5] This value is in fact the mean of the relative errors between the predictions and real values.

3.8. Conclusion

The latter analyses show that the methods studied in this chapter can be a good asset for mobile networks to plan their networks. However, the exact prevision about the future is impossible and the use of these methods has to be cautious.

These methods produces their predictions based on past history and can capture trends and seasonality but if the future proves to be very different from the history, these methods are a waste of time. Extreme changes in the normal pattern of the data due to external factors (e.g. conventions, site shutdowns, site creations, new apps, etc.) are impossible to predict. Another problem is the presence of outliers in the data, which can destabilize the methods. Nevertheless, the *Manual Growth, Level Offset* and the outliers technique studied before are able to attenuate these problems.

Table 7 presents some of the most important factors and the response of the methods to them. Note this table is applicable to the analysis made in this document and to specific conditions and data type. Yet, this is a much summarized information and do not replace the read of the previous chapters.

	Linear Regression	**MA**	**Holt-Winters**	**ARMA**	**SARMA**
Seasonality	×	✓	✓	✓	✓
Trend	✓	✓	✓	~	~
Model Complexity	Small	Small	Medium	High	High
Influence of Outliers	Small	Small	Enormous	Big	Big
Response to External Factors	×	×	×	×	×
Manual Growth	✓	✓	✓	✓	✓
Level Offset	✓	✓	✓	✓	✓

Table 7: Models properties

The results between the methods were not so different in some cases and in these cases the best method to use is the simplest one. The simple methods are normally fast in producing the predictions and use fewer resources. However, the choice of the methods has to be done very carefully.

4. Forecasting Tool

Besides writing this document concerning the subject of RAN capacity forecasting, another objective of this work was to create a tool capable of assist mobile operators' network planning. Thus, in parallel with the research and writing of this document a forecasting tool was developed.

This tool is the final product of all the knowledge acquired during the research and development of this work. The bases for this tool are the forecasting methods studied earlier (*Forecasting Methodology)* and the KPIs to be study in *RNC Capacity Planning.*

This chapter pretends to present the developed tool used to obtain all the results presented in this work, some of them were already discussed in *Forecasting Methodology*.

The future objective for this tool is to be employed in a real mobile operator environment and make it capable of actually help mobile operators. For now, the Microsoft Excel software (by Microsoft Corporation) was used as a first problem approach but already shows very interesting results.

4.1. Tool Architecture

This section pretends to present the units that constitute the forecasting toll and their functions. Figure 76 illustrates the tool architecture and its elements.

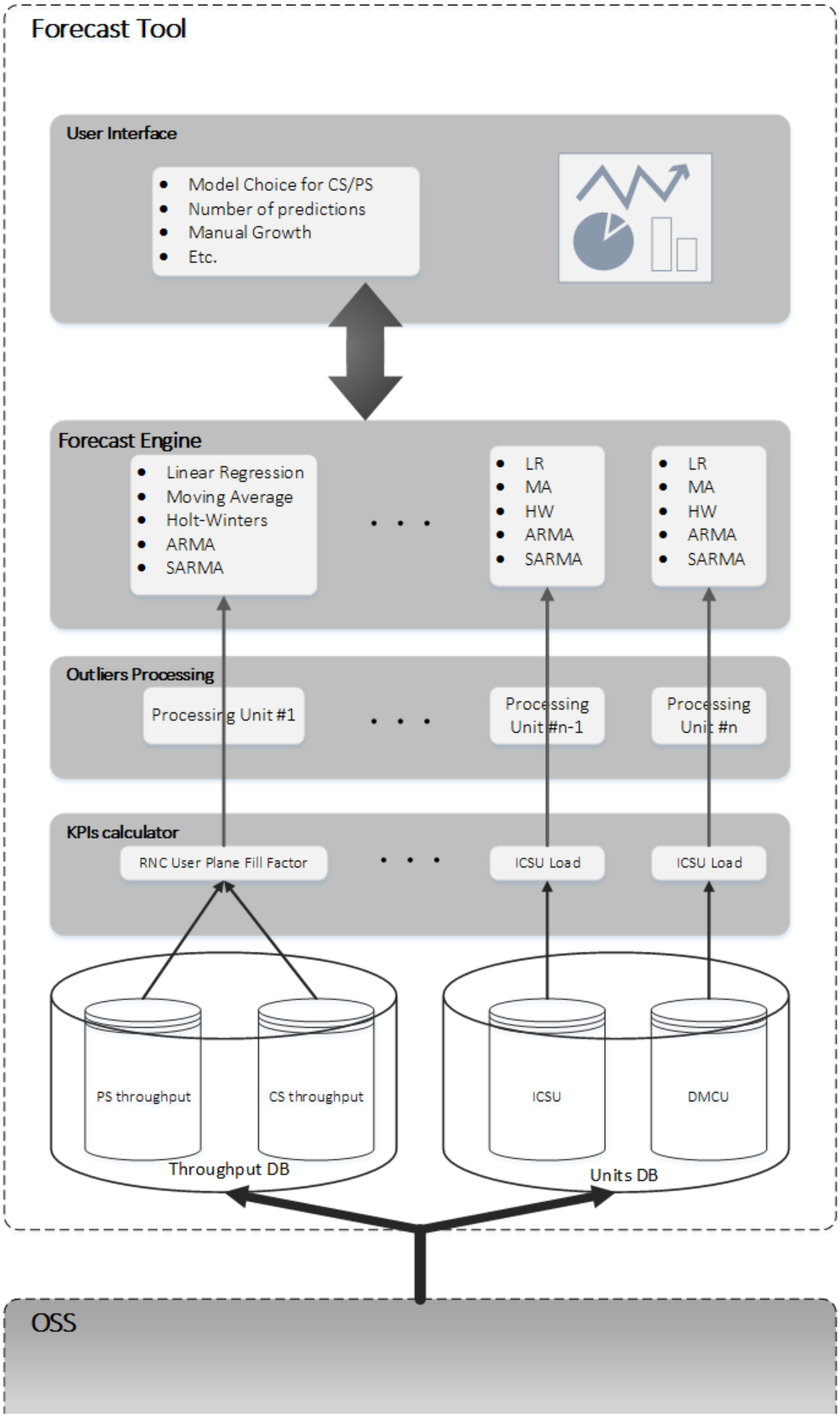


Figure 76: Tool Architecture

4.1.1. OSS

In fact, the OSS is not a part of the forecast tool but it is a very important element in the overall functionality of the tool. This because all the data needed in the tool is taken from an OSS from a real mobile operator.

4.1.2. Data Bases

There are two main data bases in which are stored data taken from an OSS of a real mobile operator. These databases are divided in two main databases:

- Throughput DB – Although the name, this database doesn't save only data related with the throughput. But it stores data that fed the KPIs calculator to calculate indicators such as the CS&PS throughput, fill rates, etc.
- Units DB – As the name indicates, this database stores data related with the load of the RNC units. It is important to keep the RNC units working well.

4.1.3. KPIs Calculator

As the name implies, the main function of the KPIs Calculator module is to perform all the calculations needed to obtain important network indicators. All the KPIs performed in this module are explained next in chapter *5. RNC Capacity Planning.* Note however, that this module to other KPIs, depending on the needs of the mobile operators.

4.1.4. Outliers Processing

Before the methods being fed with data, all data passes through the outliers processing. This module tries to eliminate the influence of outliers on the forecast results. As already seen the outliers have a big influence in the results, producing bad forecasts.

The technique used in this module is described in *3.1.4.6 Outliers and influential observations.*

4.1.5. Forecast Engine

After the data passes through the outliers processing it enters the Forecast Engine where all the methods are performed. Note that the models fitting are not automatic and it is controlled by the user in the User Interface (UI). The UI and Forecast Engine exchange information, to provide the results in accordance with the user requests.

Note that the different types of data (CS/PS throughput, ICSU load, etc.) are treated separately in all the modules. This is done in order to achieve more flexibility in the UI.

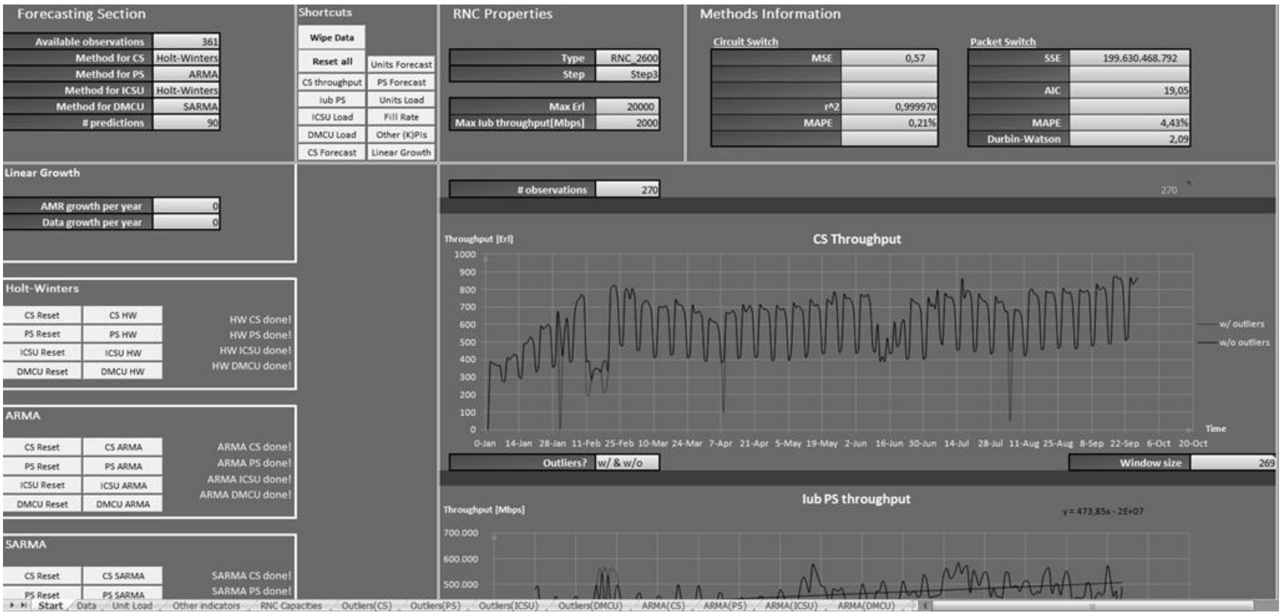
4.1.6. User Interface

The User Interface (UI) allows the user to have information about the state of the tool (Number of observations, methods performed, MAPE, etc.) and to control it. The information about the users' choices are passed to the Forecasting Engine to control the methods application to the data.

The UI also displays all the results accordingly to the users' choice.

4.2. Tool Description

The description of the tool will be supported with screenshots of it. *Figure 77* shows the general aspect of the forecasting tool.



Forecasting Section

Available observations 361
Method for CS Holt-Winters
Method for PS ARMA
Method for ICSU Holt-Winters
Method for DMCU SARMA
# predictions 90

Linear Growth

AMR growth per year 0
Data growth per year 0

RNC Properties

Type RNC_2600
Step Step3
Max Erl 20000
Max Iub throughput[Mbps] 2000

Methods Information

Circuit Switch

MSE 0,57
r^2 0,999970
MAPE 0,21%

Packet Switch

SSE 199.630.468.792
AIC 19,05
MAPE 4,43%
Durbin-Watson 2,09

observations 270

Outliers? w/ & w/o

Window size 269

Holt-Winters

HW CS done!
HW PS done!
HW ICSU done!
HW DMCU done!

ARMA

ARMA CS done!
ARMA PS done!
ARMA ICSU done!
ARMA DMCU done!

SARMA

SARMA CS done!
SARMA PS done!

Figure 77: Forecasting tool

The excel book is divided in several sheets with several purposes, such as:

Start – This is the main sheet where all the actions (e.g. methods choice, RNC type, etc.) can be taken and all the results are printed.

Data – The sheet where all the data needed for the tool is inserted and where the KPIs (studied in *RNC Capacity Planning)* are performed.

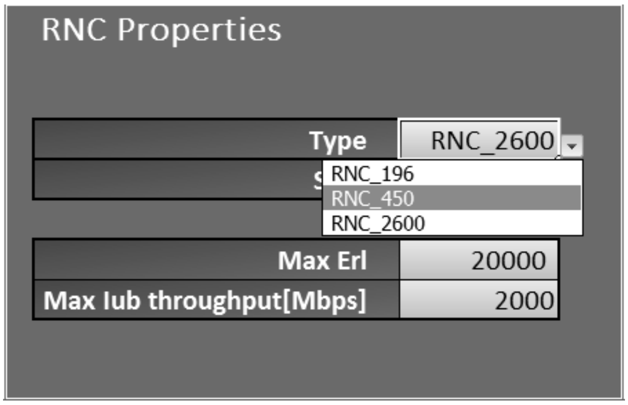
Unit Load – Contains KPIs and data related with the RNC units load.

Other indicators – Contains important KPIs that must be followed together with the forecast results.

Outliers – The outliers' technique, described in *3.1.4.6 Outliers and influential observations,* is applied to the data here. Before been inputted in the methods, all the data pass through this technique.

Linear Regression/Holt-winters/ARMA/SARMA – There are several sheets used to compute the methods in an automatic way. They hold the major complexity in the tool.

Figure 78 represents a subsection of the "Start" sheet called "RNC Properties". This subsection is composed of two drop-down lists where it is possible to choose the RNC type and the capacity step for that RNC type.



RNC Properties

Type RNC_2600

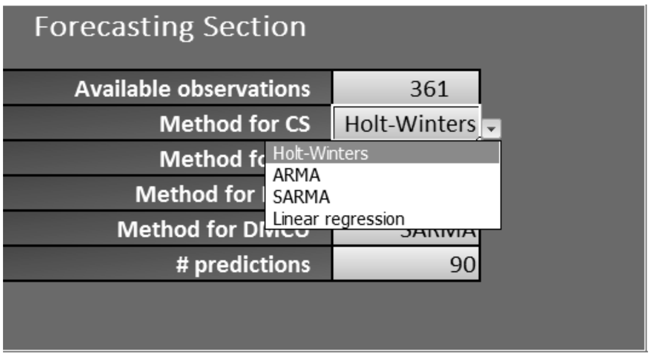
Max Erl 20000

Max Iub throughput[Mbps] 2000

Figure 78: RNC Properties

Besides the drop-down lists, it has two other fields, the "Max Erl" and "Max Iub throughput" that shows the Maximum capacity for the CS and PS services respectively and accordingly to the RNC type and step chosen. These values are used in the calculus of some KPIs in the "Data" sheet.

Figure 79 illustrates the "Forecasting Section", related with the methods choice.



Forecasting Section

Available observations 361

Method for CS Holt-Winters

Method fo

Method for

Method for DMCO SARMA

predictions 90

Figure 79: Forecasting Section

This subsection is majority composed by several drop-down lists used to choose which method is applied to which services (throughput/unit loads). Besides that, it presents the number of observations in the “Data” sheet and allows the user to choose the number of predictions he wants to perform.

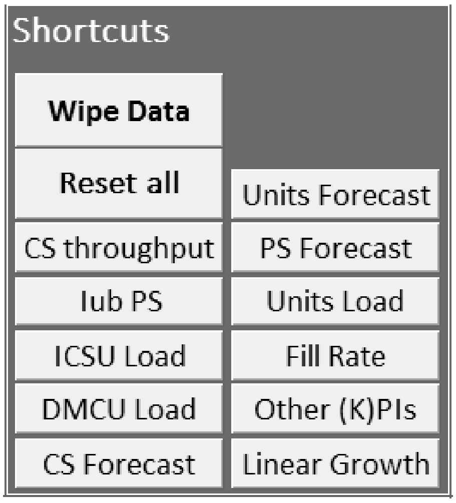


Figure 80: Shortcuts

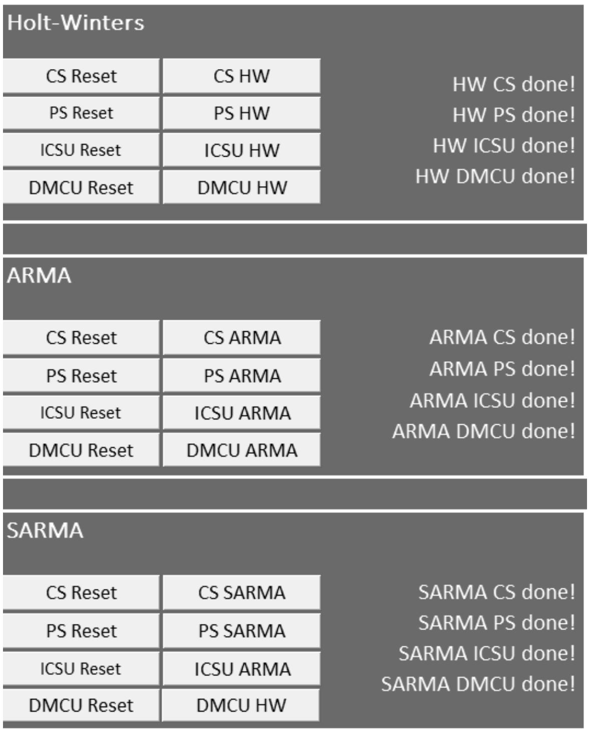
Figure 80 illustrate the “Shortcuts” subsection. As the name indicates, this section has several shortcuts that facilitate the use of the tool. The functions:

Wipe Data – Used to wipe all the data existent.

Reset all – The methods must be reset before use. This reset all the methods parameters.

The other buttons make the user go to the section named on them.

Figure 81 presents a section of the “Start” sheet with the methods options for the different indicators.



HW CS done!
HW PS done!
HW ICSU done!
HW DMCU done!

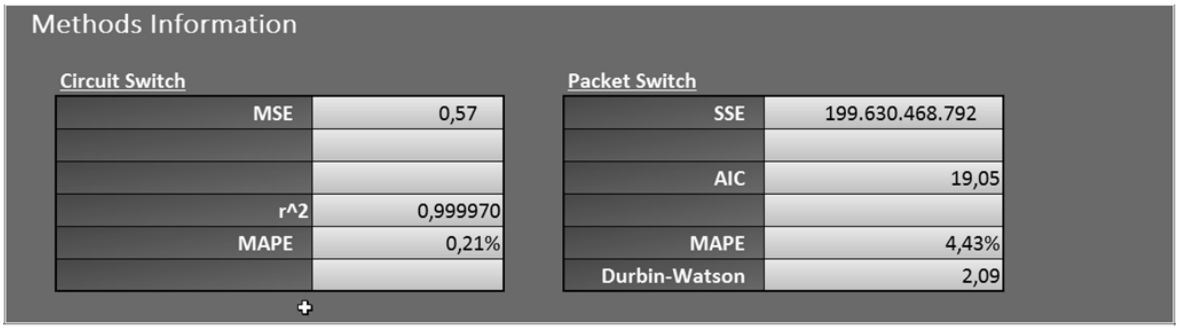
ARMA CS done!
ARMA PS done!
ARMA ICSU done!
ARMA DMCU done!

SARMA CS done!
SARMA PS done!
SARMA ICSU done!
SARMA DMCU done!

Figure 81: Methods

After all the data had been inserted on the “Data” sheet the tool is ready to perform the forecasting methods. This section allows the user to perform each method individually to each category being analyzed. Reset button are used to reset the parameters for the method and type of data in question.

Figure 82 illustrates the “Methods Information” section. This section is concerned with the quality of the methods and their quality to fit the data. It shows several indicators studied in this document, such as AIC, SSE, MAPE, etc.



Methods Information

Circuit Switch

MSE	0,57
r^2	0,999970
MAPE	0,21%

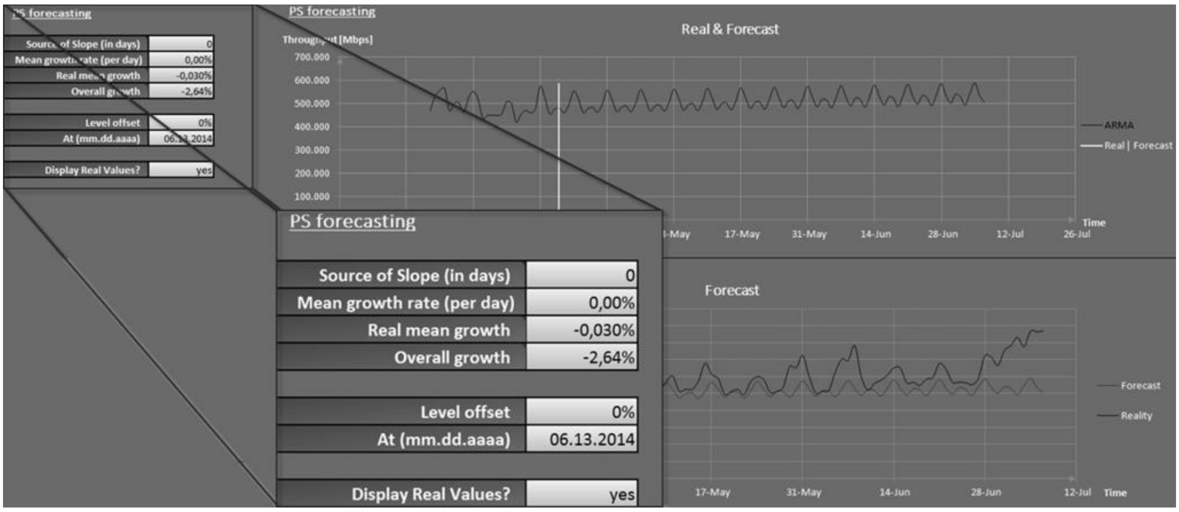
Packet Switch

SSE	199.630.468.792
AIC	19,05
MAPE	4,43%
Durbin-Watson	2,09

Figure 82: Methods Information

The type of information shown depends on which methods were chosen in the “Forecasting Section”.

Figure 83 illustrates an example from the results section in “Start” sheet. All the results are presented in this section.



PS forecasting

Source of Slope (in days)	0
Mean growth rate (per day)	0,00%
Real mean growth	-0,030%
Overall growth	-2,64%
Level offset	0%
At (mm.dd.aaaa)	06.13.2014
Display Real Values?	yes

Figure 83: Results Section

Associated with the results are usually a set of options and information (zoomed in the picture):

Source of slope – This field allows the user to add to the forecast results the slope of the inserted number of previous days.

Mean growth rate –Allows the user to add to the forecast result a growth per day. This is useful when the user forsees that, for some external reason, that type of data will increase/decrease.

Real mean growth – This is an indicator, that informs the user about the actual mean growth rate. For example if the user inserts some mean growth rate, this will be added to the forecast results that probably already have some kind of increase/decrease. This shows the sum of both method and manual growth.

$$growth\ rate = \left(\frac{last\ prediciton}{first\ prediction}\right)^{\frac{1}{n}} - 1$$

n → Number of predictions

Overall growth – Indicates the growth experienced from the first prediction to the last.

$$growth = \frac{last\ prediction - first\ prediction}{first\ prediction}$$

Level offset – Allows the user to insert a sharp level fall/rise to the data being analyzed. This is useful when the user predicts that it will be a sharp level fall/rise in the data due to external factors (e.g. site creation, conventions, etc.)

Display Real Values – The user has the ability to insert a certain amount of data and use only a set of it to forecast and the remainder to compare with the results. This can be useful to test the methods in that specific type of data. This field allows the user to show or hide the real values.

5. RNC Capacity Planning

This chapter covers some of the most important RNC capacity issues. These issues must be followed carefully by the personal in charge for the network performance management, in order to detect possible capacity bottlenecks and solve them quickly and/or prevent these bottlenecks to happen.

This dissertation was developed under a telecommunications vendor environment, therefore the methods and/or names specified in this chapter can differ from other vendors.

The user is able to track the RNC capacity problems by monitoring the follow capacity objects.

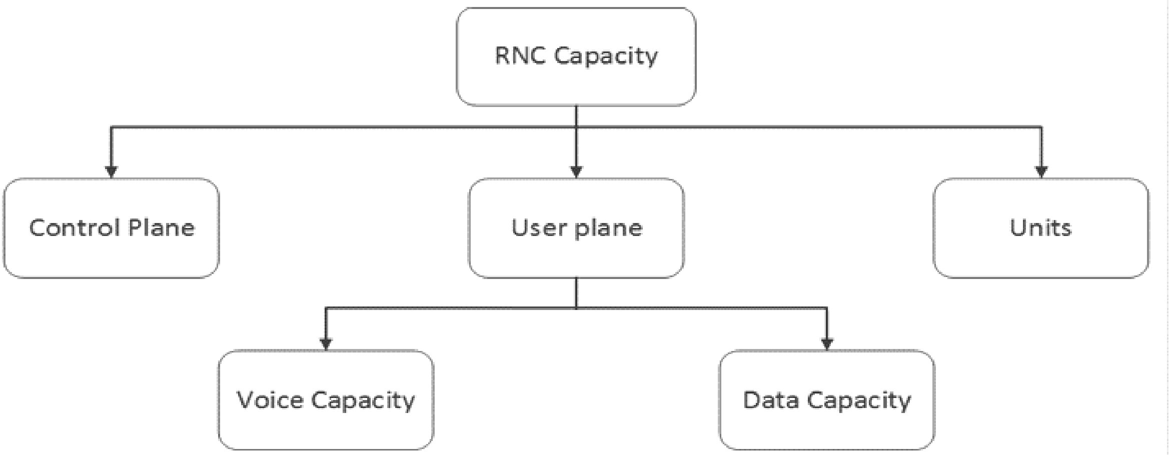


Figure 84: RNC capacity

Related to the user plane capacity in RNC there are four groups worthy to be mention [32][33]:

A. Voice capacity

The AMR defines the number of simultaneous AMR Radio Access Bearers (RABs) calculated periodically by the RNC. The AMR represents the real time speech. In fact, AMR stands for Adaptive Multi-Rate and is currently one of the most used speech codec.

As expected there is a maximum of capacity for CS services (Erlangs) in the product descriptions, this maximum defines the maximum number of simultaneous AMR RABs.

B. Data capacity

The RNC periodically calculates the amount of bytes in the GTP (GPRS Tunneling Protocol) received in each Iu-PS interface, these samples are averaged over a long period of time and the resulting throughput is converted to Mbps. The throughput on the Iub interface is factored by adding the FP header overhead factor (11%) [32] to the measured GTP throughput. Here the Iu-CS traffic, Iur traffic or soft handover overhead of the PS data throughput is not counted by the system.

Licensing is used to upgrade the RNC capacity. Once the limit of the RNC data capacity is reached changes have to be made to the network in order to reduce the load in the specific RNC or a new license need to be acquired.

C. User plane fill factor

The user plane fill factor represents the RNC capacity at a high level, for simplicity, it considers the two types of traffic in the RNC, i.e. the CS and PS data traffic.

The user plane fill factor is a very important indicator about the state of the RNC capacity and must be followed closely by the network managers.

D. RRC connected mode users

The number of Radio Resource Control (RRC) connected states is limited to values that depend on the RNC capacity. Other capacity bottlenecks might be limiting before the maximum of RRC-connected users is reached.

5.1. RNC user plane fill factor

1. Monitored Item	RNC user plane fill factor In short, the RNC user plane fill factor is the combination of PS and CS throughputs. $RNC\ user\ plane\ fil\ factor = \frac{IuPS\ throughput}{Max\ IuPS\ throughput} + \frac{CS\ Erl}{Max\ CS\ Erl}$ $Max\ IuPS\ throughput$ → Maximum IuPS UP capacity available in the Iu interface. $Max\ CS\ Erl$ → Maximum IuCS UP available in the Iu interface.		
2. Proactive monitoring	**Counter/KPI**	**Name [unit]**	**Description**
	RNC_5204a	RNC user plane fill factor [%]	Compares the IuPS throughput and AMR Erlangs with the state of the RNC capacity.
	M802C8	IU_PS_THR_AVERAGE	The average IuPS throughput in downlink direction from the core network to the RNC.
3. Reactive monitoring	**Counter/KPI**	**Name [unit]**	**Description**
	M609C3	DSP_SERVICE_FAIL_RES_ALLOC [#]	Number of DSP resource allocation failures.
	M609C4	DSP_SERVICE_FAIL_RES_MODIFY [#]	Number of DSP resource modification failures.
Packet session setup failures	RNC_1080b	PS setup failure rate due to lack of DMCU [%]	Packet call setup failure rate due to lack of DMCU resources for interactive, streaming and background traffic class.
4. Analysis	The user should follow the DSP resource allocation and modification failures with the proactive indicator (RNC user plane fill factor), because the RNC user plane fill factor can be hiding other capacity bottlenecks such as suboptimal DSP pool configuration.		
5. Overload	The RNC overload control, limits the incoming traffic. Nevertheless, it is recommended to start the capacity extension planning when the fill factor reaches 70%.		
6. Upgrade	The RNC user plane resources are partitioned into pools. Before upgrading, it is recommended to analyze the pools to see if it is possible to repartition them.		

Table 8: RNC user plane fill factor
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC user plane fill factor is a very important indicator for performance management since it considers the two types of traffic, CS and PS.

As already mentioned in *Table 8*, this indicator is the sum of the relative loads of both traffic types. The RNC user plane fill factor has to be less than one. The relative load can be obtained by dividing the measured traffic by the maximum allowed traffic for each type.

$$\frac{AMR(Erl)}{MaxAMR(Erl)} + \frac{IubPSthroughput(Mbps)}{IubMaxPSthroughput(Mbps)} \leq 1 \qquad (5.1)$$

When the CS and/or PS capacity is reaching its maximum licensed capacity, it is possible to upgrade the RNC capacity by purchase a new license with more capacity. Of course there is a hard limit – the hardware limit. Once this HW limit is reached, hardware changes need to be made in order to solve the RNC capacity problem.

When upgrading the license capacity for CS and/or PS services, a commitment between the two service capacities must be fulfilled. For example, it is possible to have the maximum HW capacity only for the CS services (e.g. 20000 Erl) but none for PS services or the opposite. Clearly, this is not a wise choice for networks that provide CS and PS services.

Figure 85 helps understanding the facts discussed in the latter paragraphs.

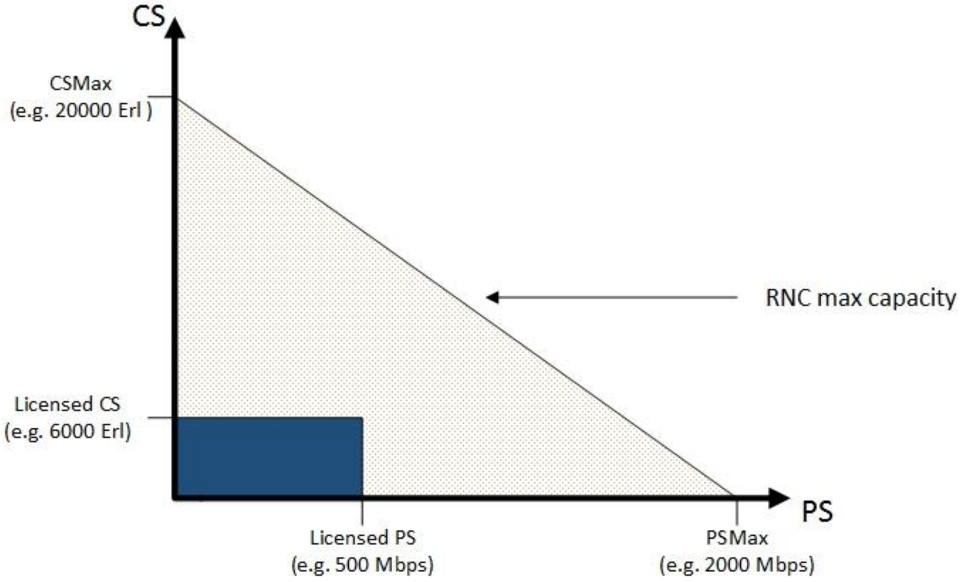


Figure 85: RNC capacity (PS vs. CS)
Adapted from "RNC Capacity Management", Raija Lilius, Nokia Solutions and Networks

To understand better the RNC user plane fill factor, we need to drill into the KPI (RNC user plane fill factor) and closely analyze the PIs, counters and calculations used to obtain this indicator. For a reason of simplicity this analysis will be divided in two parts, the PS fill factor (PS services related) and the CS fill factor (voice services related).

5.1.1. Iub PS throughput

The counters required for the Iub PS throughput calculations are presented in *Table 9*. First of all, there are some considerations about the Iub PS throughput:

- It is defined in the downlink direction.
- Take into account the FP head overhead (11%).
- Includes the SHO overhead (not included for HSDPA).

Calculations

$$Iub\ PS\ throughtput\ [Mbps] = ((IuPS\ throughput - HSDPA) \cdot SHO + HSDPA) \cdot 1.11 \qquad (5.2)$$

$$SHO = \frac{one\_cell\_in\_act\_set\_for\_nrt + two\_cells\_in\_act\_set\_for\_nrt + three\_cells\_in\_act\_set\_for\_nrt}{one\_cell\_in\_act\_set\_for\_nrt + \frac{two\_cell\_in\_act\_set\_for\_nrt}{2} + \frac{three\_cell\_in\_act\_set\_for\_nrt}{3}} \qquad (5.3)$$

$$HSDPA = \frac{RECEIVED\_HS\_MACD\_BITS}{1.05 * period\ time * 1000} \qquad (5.4)$$

$period\ time$ → This value depends on the granularity of the data being managed, e.g. if it is hourly data the period time would be 3600 (1h = 3600s), for daily data would be 24*3600 (1 day = 24*3600 seconds).

Counter/KPI	Counter name	Aggregation(time)	Unit	Description
M802C8	IU_PS_THR_AVERAGE	avg	Kbit/s	This is the average Iu-PS throughput in downlink direction from the core network to the RNC. It includes only user data, not the GTP header.
M5000C126	RECEIVED_HS_MACD_BITS	sum	Kbit	Amount of data received from the RNC in MAC-d PDUs. UPDATED: When a MAC-d PDU is received by the BTS.
M1007C19	one_cell_in_act_set_for_nrt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been one. This counter covers active sets related to NRT connections.
M1007C20	two_cells_in_act_set_for_nrt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been two. This counter covers active related to NRT connections.
M1007C21	three_cells_in_act_set_nrt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been three. This counter covers active sets related to NRT connections.

Table 9: Measurements required to compute the Iub PS throughput

5.1.2. Iub CS throughput

The CS throughput here referred as AMR (voice services) can be calculated using utilization KPIs. All the counters used in the calculations are described in *Table 10*.

$$AMR = \frac{dur\_for\_amr\_12\_2\_dl\_in\_srnc + dur\_for\_amr\_12\_2\_dl\_in\_drnc}{100 * period\ time * SHO\_rt} \quad (5.5)$$

$$SHO\_rt = \frac{one\_cell\_in\_act\_set\_for\_rt + two\_cells\_in\_act\_set\_for\_rt + three\_cells\_in\_act\_set\_for\_rt}{one\_cell\_in\_act\_set\_for\_rt + \frac{two\_cell\_in\_act\_set\_for\_rt}{2} + \frac{three\_cell\_in\_act\_set\_for\_rt}{3}} \quad (5.6)$$

Counter/KPI	Counter name	Aggregation(time)	Unit	Description
M1002C49	dur_for_amr_12_2_dl_in_srnc	sum	10 ms	The summary of RT DCH allocated durations for AMR 12.2 kbps in DL in the SRNC. UPDATED: This counter is updated with a 1-second sampling interval during the DCH allocation time.
M1002C285	dura_for_amr_12_2_dl_in_drnc	sum	10 ms	A summary of RT DCH allocated durations for AMR 12.2 kbps in DL in the DRNC. UPDATED: This counter is updated with a 1-second sampling interval during the DCH allocation time.
M1007C0	one_cell_in_act_set_for_rt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been one. This counter covers active sets which are related to RT connections.
M1007C1	two_cells_in_act_set_for_rt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been two. This counter covers active sets which are related to RT connections.
M1007C2	three_cells_in_act_set_rt	sum	100 ms	Sum of time periods during which this cell has belonged to the active set the size of which has been three. This counter covers active sets which are related to RT connections.

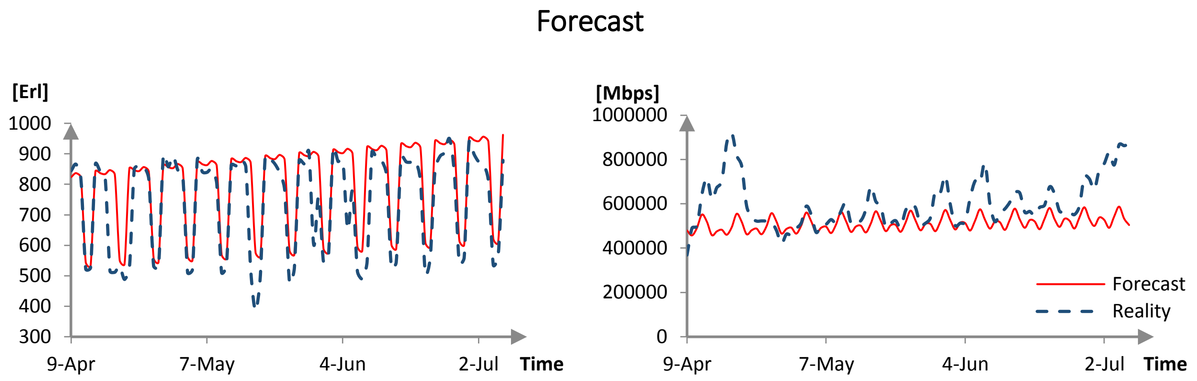
Table 10: AMR throughput counters

5.1.3. Outcome analysis

This chapter presents important output results from the forecast tool (refer to *Forecasting Tool)*.

Throughput Forecast

The first and simpler result is to plot the forecast results. *Figure 86* represent the forecast for CS and PS throughputs, respectively.



a) CS (AMR) daily average throughput

b) PS daily average throughput

Figure 86: Forecast vs. Reality

This type of charts, besides allowing to see how the throughput is evolving into the future also allows examining if the forecast is in accordance with the past, i.e. if the results are not completely off the expected.

The tool also allows comparing the performance of the method with real values as shown in *Figure 86*.

Fill Factor Forecast

Figure 87 shows both CS (AMR) and PS fill factors and the combination (sum) of both, the “Fill factor”.

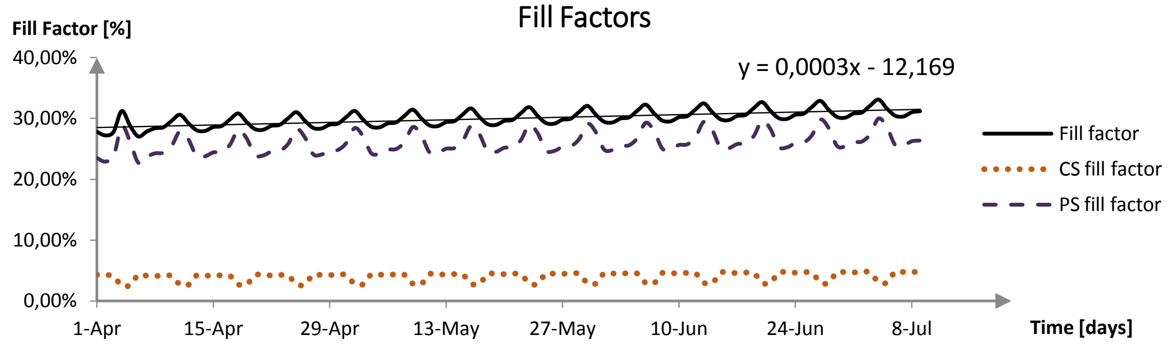


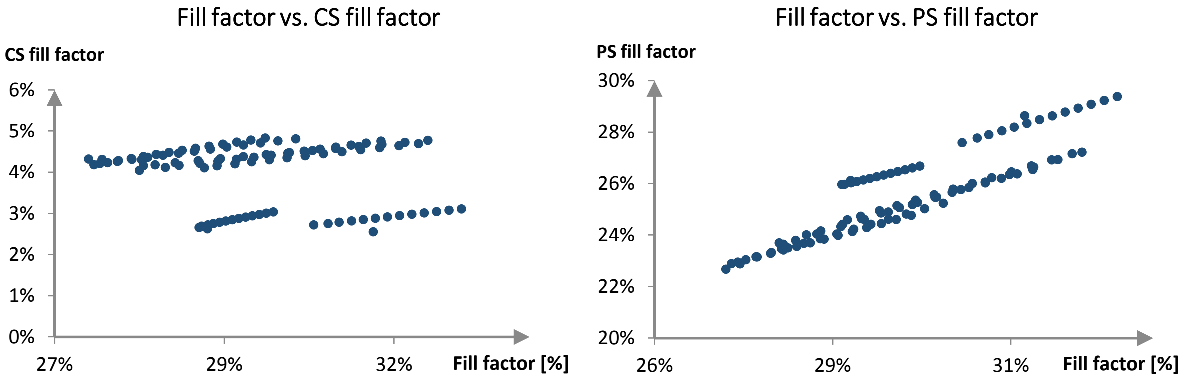
Figure 87: User plane fill factor forecast

This type of chart allows the user to see the filling percentage of both services and compare which is the dominant one. The chart represents forecast results and note that the fill factor is slightly increasing.

It is important to have in mind *Figure 85*, because the CS and PS fill factors are calculated based on the maximum capacity for each service, respectively. But remind that there is a commitment between them, i.e. the sum of both CS and PS fill factors (which is the “Fill factor”) cannot exceed 100 %. For example, if one is 100% the other is 0% and the “Fill factor” is 100%, which by the way is a bad sign. The idea here is to clarify that the maximum “Fill factor” is 100%.

Fill Factors Relationship

Figure 88 illustrates the relationship between the CS and PS fill factors and the overall fill factor.



a) Fill factor vs. CS fill factor (forecast)

b) Fill factor vs. PS fill factor (forecast)

Figure 88: Fill rates relationship

This type of relationship allows observing the contribution of CS/PS services to the overall fill factor. For example, from *Figure 88* is possible to notice that the CS fill factor and the overall fill factor have a linear relationship but the overall fill factor is growing more rapidly. In the other hand, *Figure 88* shows that the PS fill factor is growing at almost the same speed than the overall fill factor. This is a normal situation, once the overall fill factor is the sum of both CS (AMR) and PS fill factors.

The figures represent simple relationships, but they can be very enlightening on how the services are growing and how they affect the user plane fill factor.

5.2. Unit Load monitoring

The RNC fill factor is a big asset to the network planning and performance. Unfortunately, the RNC complexity demands a performance monitoring that cannot be only based on the RNC fill factor. Obviously there are other “gamers” that can cause bottlenecks in the RNC besides the Iub interface.

Here it will be discussed the most critical units and their loads for Control Plane (CP) and User Plane (UP). The objective is to monitor and forecast the processor loads in computer units.

The M592 family has counters to measure the peak and average CPU load, these measurements are the basis for the unit load KPIs.

Counter	Counter Name	Aggregation (Time)	Unit	Counter Description
M592C0	AVERAGE_LOAD	avg	%	The Average Load for monitored computer unit. The value is the arithmetical average of samples taken from the processor load.
M592C1	PEAK_LOAD	max	%	The Peak Load of monitored computer unit. This is the highest recorded value of the processor load during a measurement period.

Table 11: M592 measurements

Figure 89 represents a RNC architecture mainly consisting of units (HW/SW) and interfaces (internal and external). This is an example of a specific RNC, different models and vendors may have different architectures. There are units and interfaces in this architecture that are not discussed in this chapter because their leaning to create bottlenecks is very small.

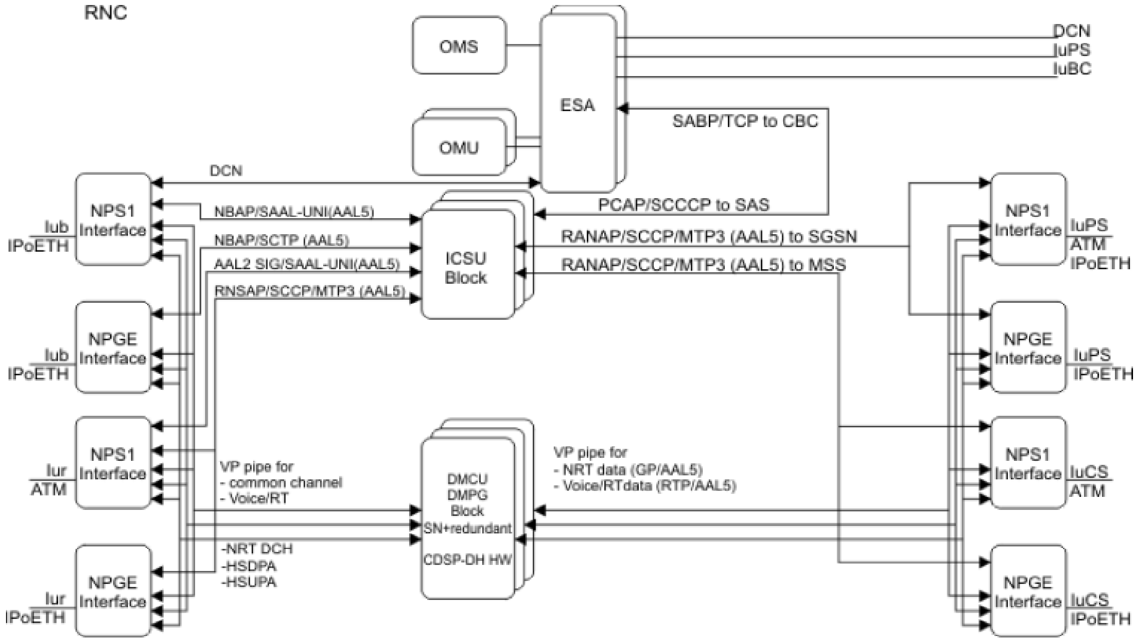


Figure 89: RNC Architecture
Source: "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

5.2.1. ICSU

The Interface Control and Signaling Unit (ICSU) handle signaling functions and the associated traffic control functions, including tasks such as:

- Admission control;
- Radio resource control;
- Resource management;
- Handover control;
- Packet scheduling;
- Signaling protocols to Iu and Iub interfaces, including: NBAP, RNSAP and RANAP;

To protect itself from overload, the unit uses an overload control system called the Windows Admission Control (WAC).

1. Monitored Item	ICSU load Reactive monitoring based on the WAC Overload Control (M594) can be used to see if the system has rejected RRC connection requests because of overload. M1006C206 counter measures if the BLC is causing RRC connection requests rejects and M1003C47 counter measures how many paging messages the system has deleted because of ICSU overload.		
2. Proactive Monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	RNC_5228a	Average CPU load of the most loaded ICSU unit [%]	The RNC ICSU unit load calculated as the average CPU load of the most loaded unit.
	RNC_1870a	Average ICSU CPU load [%]	RNC_ICSU unit CPU (unit _type = 329) load.
3. Reactive monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	RNC_5206a	Window Access Control (WAC) reject ratio [%]	Ratio of rejected WAC gate requests to all WAC gate requests on RNC ICSUs.
	M1006C206	RRC CONN REJECT DUE TO BUFFER LIMIT CTRL [#]	Updated when the RNC sends a RRC CONNECTION REJECT message to the UE due to the number of unhandled messages in a signaling unit has exceeded a specific threshold value. M1006C21 and M1001C618 are updated along with this counter.
	M1003C47	NBR OF DELETED PAGING MESSAGES DUE TO ICSU OVERLOAD	Number of panging messages deleted due to ICSU overload. Updated when a paging message is deleted because of ICSU overload.
4. Analysis	Note that the WAC is a general overload control mechanism and the reason for the reject might be on some other unit than the ICSU, or even beyond the RNC.		
5. Overload	The RNC overload control grants the CPUs to go up safely until 70% (80%). After that, the overload control should cause QoS degradation. The ICSU overload is caused by the subscriber or UE activity, the call setups, that is, BHCA. There are two thresholds for the overload control. Firstly only conversational and emergency calls are accepted. If the number of unhandled messages and the CPU load exceed the pre-defined higher values, only emergency calls are accepted.		
6. Upgrade	When ICSU CPU loads are close to the targets reduce the RNC C-plane load by optimizing the network: - Reduce the SHO area - Selectively remove 2G to 3G neighbor relationships - Other techniques refer to [32] If the optimization is not possible or enough, upgrade the RNC.		

Table 12: ICSU load
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC _5228a, RNC_1870 and RNC_5206a belong to the usage class and can be calculated as follows:

$$RNC\_5228a = MAX(M592C0) \qquad (5.7)$$

$$RNC\_1870a = \frac{SUM(M592C0) - MIN(M592C0)}{NUMBER\_OF\_UNITS - 1} \quad (5.8)$$

$$RNC\_5206a = \frac{\sum WAC\_GATE\_REQ\_TOTAL\_REJ}{\sum WAC\_GATE\_REQ\_TOTAL} \times 100 \quad (5.9)$$

Counter	Counter name	Aggregation(time)	Unit	Description
M592C0	AVERAGE_LOAD	avg	%	The Average Load for monitored computer unit.
M594C0	WAC_GATE_REQ_TOTAL	sum	Integer	Total number of WAC Gate requests that has been released and rejected. When an entity wants to establish some kind of a signaling connection it should firstly ask permission to WAC Gate. If the maximum number of accesses is reached, the request is rejected or put into a queue, waiting for resources to be released by other.
M594C1	WAC_GATE_REQ_TOTAL_REJ	sum	Integer	The total number of rejected WAC Gate requests. When an entity wants to establish some kind of a signaling connection it should first ask permission from WAC Gate. If the maximum number of accesses is reached, the request will be rejected or put into a queue, waiting for resources to be released.

Table 13: Counters used for ICSU overload inspection

A. Why the ICSU is so important to monitor?

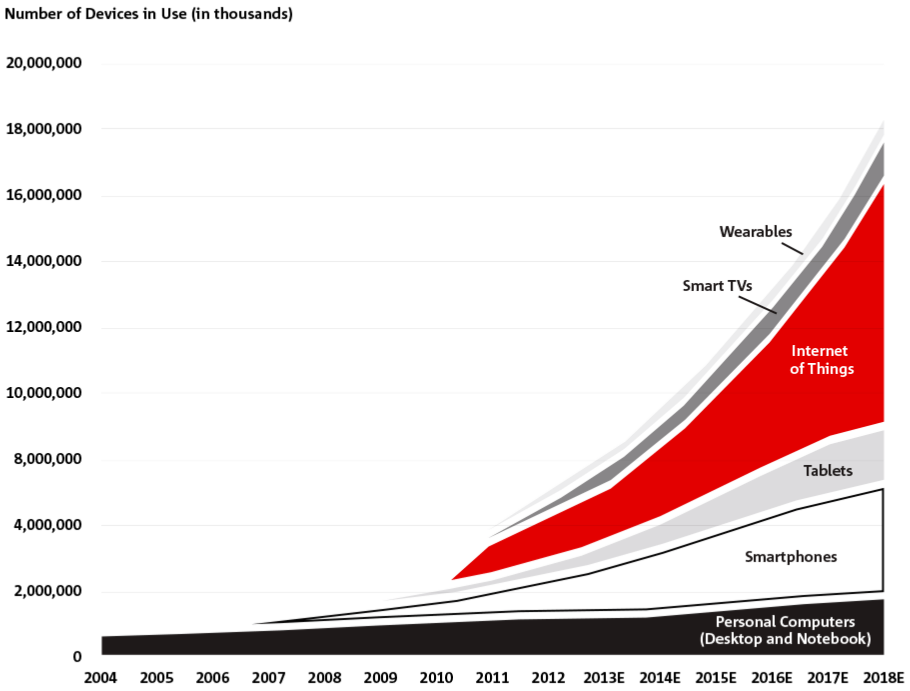


Figure 90: Evolution of technologies
Source: Gartner, IDC, Strategy Analytics, Machina Research, company filings, BN estimates

The smartphones are already widely used and becoming more and more common in our society. With it the development and use of new apps using high signaling level, which is creating a problem to the mobile operators that start to experiencing high signaling traffic which is causing problems in their networks.

Furthermore, the era of the Internet of Things (IoT) and "wearable" technologies are coming. These are two areas already existents, growing and with big investments planned. The signaling traffic generated by these technologies is huge threatening the good work of the mobile networks. Therefore, the mobile operators have to be aware of this problem and prepare themselves to confront it.

Thus, there is a threatening of ICSU overload caused by the circumstances discussed in the latest two paragraphs. This makes the ICSU a crucial unit to monitor.

B. ICSU overload problems

The overload of the ICSU generates a big problem. Once its limit is reached, the ICSU and its mechanisms of overload control starts to reject RRC connections even if the capacity for user data (data plane) is not in its limit (possibly even close) meaning that there are capacity not being used, reflecting bad management of the network resources.

C. Outcome Analysis

The principal forecasting tool objective is to predict the evolution of the ICSU load that is directly related with the signaling traffic. *Figure 91* represents the forecast results using an ARMA model.

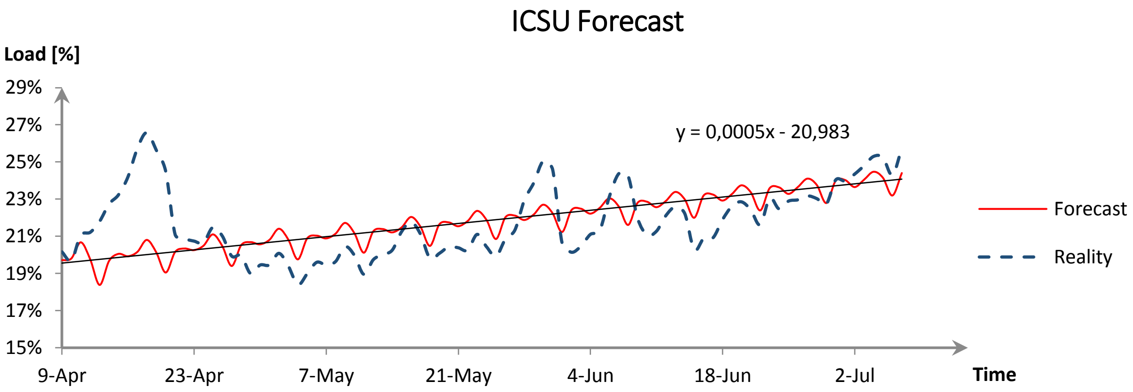


Figure 91: ICSU load forecast

This type of chart is useful to have an idea about which will be the CPU load of the ICSU unit in a few months ahead. Note that the slope of the forecast is 0.05%, i.e. the load is increasing over time. The forecasting tool also allows comparing forecast results with real values. This can be useful to see if the method produces good results in that type of data/market.

Figure 92 represents the relationship between the user plane fill factor (CS + PS fill factors) and ICSU CPU load (forecast results).

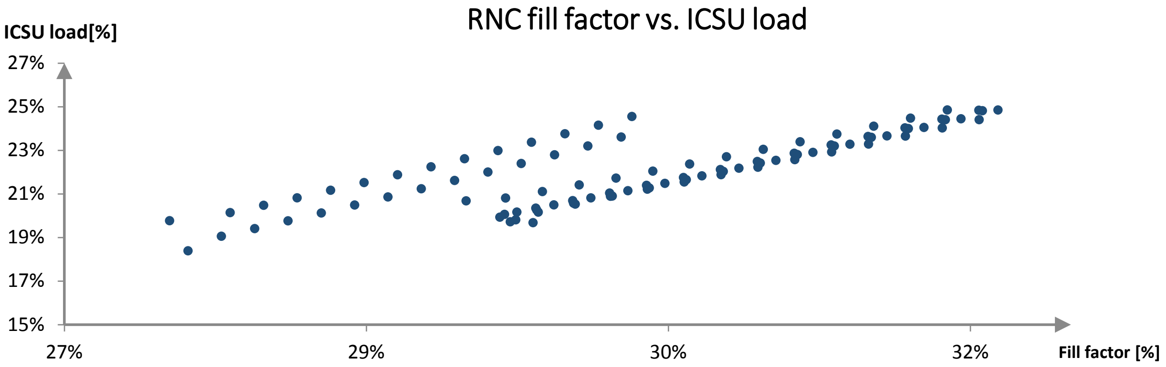


Figure 92: Relationship between fill factor and ICSU load

This is an important relation to observe because it can give us answers related with rejection of connections. Furthermore, it gives a perspective about the evolution of these two indicators.

It is easy to see, that if the trend in the figure continues, the ICSU CPU load will reach its maximum before the RNC fill factor.

The ICSU load reaching its maximum first than the RNC fill rate is a problem, because it means that there will be connections rejected by the ICSU unit due to overload in the Control Plane (CP) and the Iub throughput is not in its maximum or even close depending on the severity of the situation, i.e. there are User Plane (UP) capacity not being used, retracting a bad management of the network.

5.2.2. DMCU

The Data and Macro Diversity Combining Unit (DMCU) although considered a signal processing unit, it actually performs some control plane functions besides it signal processing tasks.

The DMCU has a hierarchical HW structure consisting of Data and Macro Diversity Processing Groups (DMPG) and Data Signaling Processors (DSPs), refer to *Figure 93*. The DSP performs signal-processing tasks.

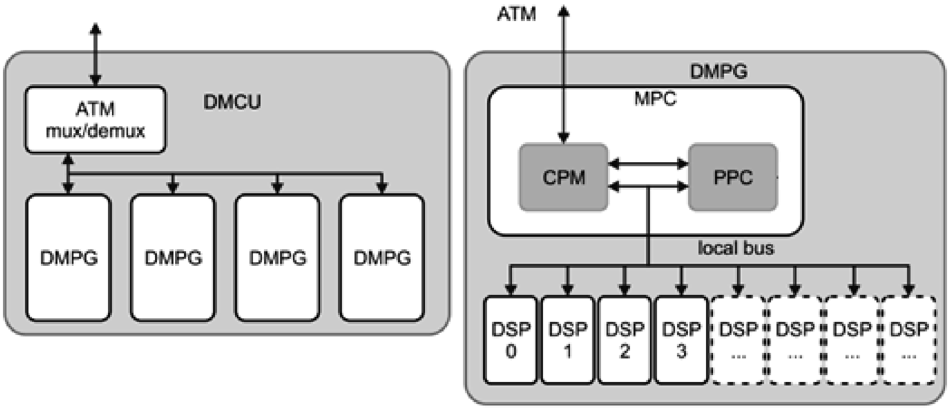


Figure 93: DMCU architecture
Source: “Managing WCDMA RAN and Flexi Direct””, Nokia Solutions and Networks, 17 June 2013

The DMPGs are independent from each other, and communicate through an Asynchronous Transfer Mode (ATM) switch. The DSPs communicate with the PowerPC (PPC) and Communications Processor Module (CPM) only. There are variants of the DMCU HW where the architecture might differ.

5.2.2.1. DMPG

The PowerPC capacity is related to the RNC Control Plane (CP). The CPM handles the low-level communication functions towards other units in the RNC.

1. Monitored Item	RNC DMPG (PPC) capacity The KPI for monitoring the PPC CPU load is based on the M592 measurement. The DMPG has mechanisms of overload control, whose task is to reduce the blocking probability of the more critical services in overload situations.		
2. Proactive monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	RNC_1981a	Average CPU load of the most loaded DMPG (PPC) unit [%]	Average CPU load of the most loaded RNC DMPG unit.
3. Reactive monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	-	-	-
4. Analysis	The KPI shows the maximum average PPC CPU loads on all DMPG units.		
5. Overload	The mechanisms of overload control have steps, depending on the severity of the overload. The order is that the more expendable work stops first: PS, CS and Emergency calls.		
6. Upgrade	This KPI allows us to get the average load of the most loaded DMPG (PPC) unit but doesn’t identify the most loaded DSP pool. Thus, before upgrading it is recommended an inspection to individual unit loads to see if the DSP pools can be repartitioned.		

Table 14: DMPG load
Adapted from “Managing WCDMA RAN and Flexi Direct”, Nokia Solutions and Networks, 17 June 2013

The RNC_1981a belongs to the KPI usage class and can be calculated as follows:

$$RNC\_1981a = MAX(M592C0) \qquad (5.10)$$

A. Why the DMPG is important to monitor?

The DMPG load is related with the data traffic and the Data Signaling Processor (DSP) units. In case of overload these start blocking services.

B. Outcome Analysis

Here, the forecasting tool ambition is to predict the evolution of the DMPG load accurately. *Figure 94* represents the average DMPG CPU load forecast results for a period of three months using the Holt-Winters method. The green line represents a linear regression of the forecast line.

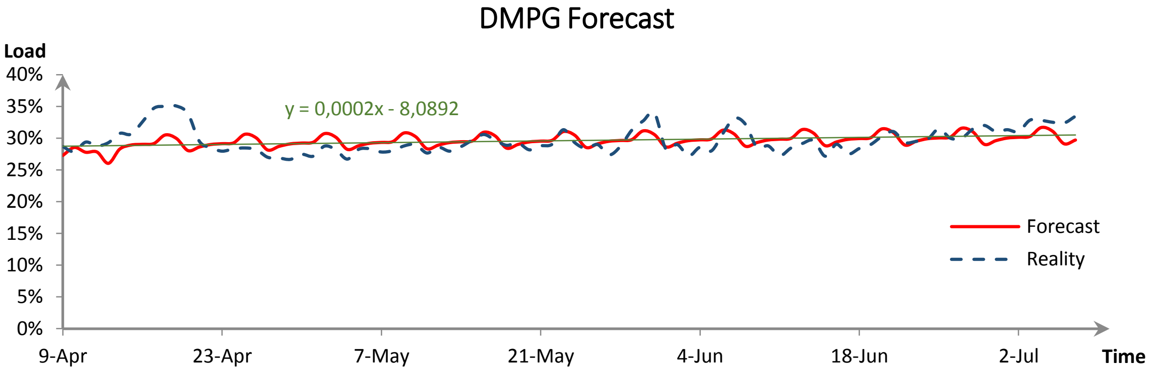


Figure 94: Average DMPG CPU load forecast

Besides representing the prediction values for the DMPG load, this type of chart shows the trend of the market giving an idea about a far future. Note that the DMCU load is increasing over time with a slope of 0.02%. It is also possible to compare the forecast results with the reality and inspect if the method used produces good results in that type of data/market.

Figure 95 represents the relationship between the RNC Fill Factor and the DMPG Unit Load. The values used in this relationship are the forecast results for the RNC Fill factor (*Figure 87*) and DMPG Unit Load (*Figure 94)*.

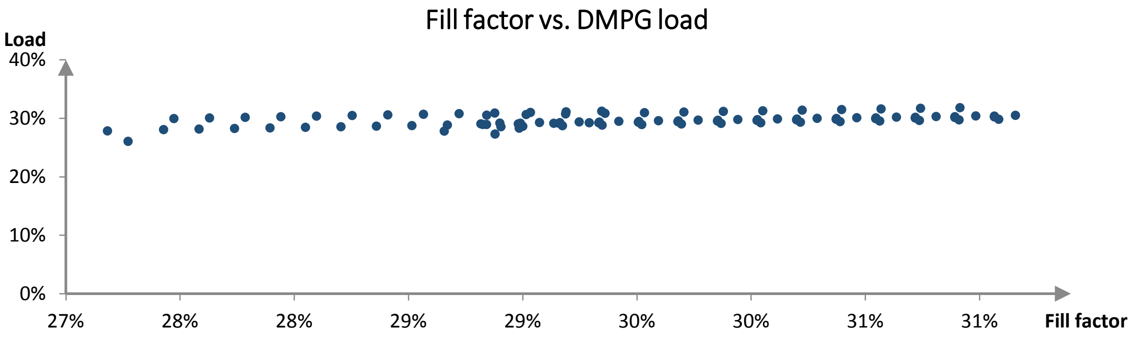


Figure 95: Relationship between RNC Fill Factor and DMPG CPU Load

Note that the DMPG load values are from the most load unit. This is an important relationship because it makes possible to identify if the DMPG unit will block services or the bottleneck will be the RNC user plane fill factor.

It is possible to notice a linear relationship between the two quantities, and both RNC fill factor and DMPG load are growing with a similar pace, which is good. Moreover, the forecast results in the figure show that in the next three months, there will be no problems due to DMPG CPU overload and/or RNC UP fill factor.

5.2.2.1.1. DSP

Recalling *Figure 93*, a DMPG unit is composed of several DSP units and the DMPG load studied earlier gives information about the most loaded unit. Thus, there is a need for a more detailed inspection, which is possible through the DSP load KPIs inspection.

The RNC user plane resources are partitioned into pools. The HS-DCH pool handles both HSPA and R99 DCH traffic. Overall, the R99 DCH and HSPA DSP capacity depends on the total number of DSP processors in a given RNC capacity step, and the number of processors configured for the pool types.

1. Monitored Item	RNC DSP load Measures the utilization of the RNC user plane processing hardware.		
2. Proactive monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	RNC_5209a	Average CPU load of the most loaded DSP unit [%]	The RNC DSP load, calculated as the average CPU load of the most loaded unit.
3. Reactive monitoring	-	-	-
4. Analysis	The DSP CPU load is one of the indicators that the system uses for load balancing between the DSPs. If the loads are high, the DSPs are overload. However, if the loads are low the RNC might be experience overload due to other factor, such as uneven distributions of the services to DSP pools.		
5. Overload	If the CPU load is above 80%, packet transmission from TCP/IP application slows down. If the CPU load is above 90%, the DSP starts to drop packets. Packet drops are applicable to the whole MAC-d flow, irrespective of the traffic class and priority.		
6. Upgrade	Before upgrade it is recommended an inspection to the pools, to see if it is possible to repartition them.		

Table 15: DSP load
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC_5209a belongs to the KPI usage class and can be calculated as follows:

$$RNC\_5209a = MAX(M617C1) \tag{5.11}$$

5.2.2.2. RSMU

The Resource and Switch Management Unit (RSMU) performs UE connection control, accordingly with the requests received from the ICSU. The ICSU is the source of most of the load on RSMU. The RSMU also perform centralized resource management tasks, such as DSP management resource tasks, etc.

The load of the RSMU is directly related with the number of leg setups. The legs are connections inside the RNC (between internal functions of the RNC). The Leg Management Program Block (LGMANA) on the RSMU provides the service to establish the leg connections. LGMANA monitors the size of requests and the CPU load. If the free load of the CPU is not sufficient to receive the new leg requests, the LGMAN frees low-priority requests so that requests with high priority can be served.

1. Monitored Item	RSMU load The unit CPU load formula for the RSMU unit is based on the M592C0 counter. The RSMU protects itself against overload caused by the ICSU and the ICSU unit reports the incident in the M1006C205 counter.		
2. Proactive monitoring	**Counter/KPI**	**Name [unit]**	**Description**
	RNC_1871a	Average CPU load of the active RSMU unit [%]	The RSMU unit load, calculated as the average CPU load of the most loaded unit.
3. Reactive monitoring	**Counter/KPI**	**Name [unit]**	**Description**
	M1006C205	RRC CONN REJECT DUE TO CENTRALIZED UNIT OVERLOAD [#]	Number of RRC connection rejection because of centralized signaling unit RSMU overload.
4. Analysis	RSMU CPU load The RSMU CPU load is directly related with leg establishments. The BHCA load and RRC state changes are causing leg setups and DSP resource allocations. RRC CONN REJECT DUE TO CENTRALIZED UNIT OVERLOAD This counter is updated when the RNC send the RRC CONNECTION REJECT message to the UE, due to the RSMU unit overload. In addition, the M1006C21 counter (total number of RRC Connection Request Reject messages) are updated along this counter.		
5. Overload	Because of overload, call setups and soft handover leg additions start to fail.		
6. Upgrade	The capacity extension planning should start before the RSMU load reaches 70%.		

Table 16: RSMU load
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC_1871a belongs to the KPI usage class and can be calculated as follows:

$$RNC\_1871a = MAX(M592C0) \qquad (5.12)$$

5.2.2.3. GTPU

The GPRS Tunneling Protocol Unit (GTPU) assists the connections between the RNC and the SGSN by performing user plane functions that are related to handling GTP in the Iu-PS interface. The GTPU is directly proportional to the frequency of the IP packets (UL+DL). The average size of the packets has no major impact on the throughput.

1. Monitored Item	RNC GTPU load The GTPU load can be measure using the M592C0 counter.		
2. Proactive Monitoring	**Counter/KPI**	**Name [Unit]**	**Description**
	RNC_1872a	Average CPU load of the most loaded GTPU Unit [%]	Average CPU load of the most loaded RNC GTPU unit.
3. Reactive monitoring	**-**	**-**	**-**
4. Analysis	The GTPU load is directly correlated with the packet ratio.		
5. Overload	The GTPU monitors the system's message queue and discards packets for the CPU overload. The system doesn't drop packet sessions.		
6. Upgrade	Activate the Dedicated GTPU for real-time IP support. Reconfigure the Iu-PS, in case of impossibility upgrade the units.		

Table 17: GTPU load
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC_1872a belongs to the KPI usage class and can be calculated as follows:

$$RNC\_1872a = MAX(M592C1) \qquad (5.13)$$

5.3. RRC connected subscribers

The possibility that the number of connected users exceeds the maximum supported by the RNC is another possible bottleneck in the radio access network. This point discusses the RRC connected subscribers indicator and the importance to follow this indicator closely.

The number of RRC connected subscribers has a limit, making of it an important indicator to follow. If the RRC connected subscribers reaches its limit, it starts to reject new connections even if the other resources are not in their limit.

The monitoring of the number of subscribers is possible using the M802 measurements. It is also possible to monitor the traffic profile using these counters. *Table 18* describes the RNC capacity usage (M802) counters.

1.Monitored capacity	Number of RRC connected subscribers Use M802 measurements (RNC capacity usage) to monitor the number of subscribers in the various RRC connected states.		
2. Proactive monitoring	**Counter**	**Name [Unit]**	**Description**
RRC connected mode users	RNC_2173a	Maximum number of RRC connected mode users [#]	Peak number of RRC connected mode users in RNC (all states) during the measurement period.
	M802C17	AVE_RRC_CONN_MODE_USERS [#]	The average number of RRC connected mode users in RNC (all states) during measurement period.
	M802C18	MAX_RRC_CONN_MODE_USERS [#]	Peak number of RRC connected mode users in RNC (all states) during measurement period.
	M802C19	AVE_USERS_CELL_DCH [#]	Average number of users in Cell-DCH state in RNC during measurement period.
	M802C20	AVE_USERS_CELL_FACH [#]	Average number of users in Cell-FACH state in RNC during measurement period.

	M802C21	AVE_USERS_CELL_PCH [#]	Average number of users in Cell-PCH state in RNC during measurement period.
	M802C22	AVE_USERS_URA_PCH [#]	Average number of users in URA-PCH state in RNC during measurement period.
3. Reactive monitoring	-	-	-
4. Analysis	The analysis of this measurements depends on the product (RNC) version, for example: - The maximum RRC connected mode users is greater than 80 % of the available capacity, it might require an upgrade in the RNC (The available capacity to use in the calculation is provided in the product description). - Different RRC-connected mode state is restricted by an overload control mechanism. The limits for the different states are provided in the product description.		
5. Overload	In case of RRC connected mode users overload, the RNC starts to reject RRC setups and/or move CELL-PCH users to idle mode.		
6. Upgrade	Select a higher capacity model if no other option exists.		

Table 18: RNC connected subscribers.
Adapted from "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

The RNC_1872a belongs to the KPI usage class and can be calculated as follows:

$$RNC\_2173a = \frac{\sum M802C18}{COUNT(DISTINCT\ RNC\_ID)} \qquad (5.14)$$

5.4. Busy Hour Call Attempt

The Busy Hour Call Attempt (BHCA) is the number of call attempts in the busy hour, which is an important parameter to decide about the processing capacity of the RNC. The higher the BHCA, the higher the network processors stress.

5.4.1. CS BHCA

The CS BHCA can be calculated using the M1001 family counters listed in *Table 19*.

Counter	Counter Name	Aggregation (Time)	Unit	Description
M1001C66	rab_act_att_cs_voice_rnc	sum	Integer	Number of RAB setup attempts for CS voice, conversational and streaming calls, respectively. UPDATED: When the RNC receives a RANAP: RAB ASSIGNMENT REQUEST message, the purpose of which is to initiate the establishment of a new CS voice, data conversational or data streaming RAB respectively.
M1001C67	rab_stp_att_cs_conv	sum	Integer	
M1001C68	rab_stp_att_cs_strea	sum	Integer	

Table 19: BHCA counters

The CS BHCA is simply the maximum of call attempts for CS services:

$$CS\ BHCA\ =\ MAX(M1001C66 + M1001C67 + M1001C68) \qquad (5.15)$$

5.4.2. PS BHCA

The PS BHCA can be calculated using the M1022 family counters listed in *Table 20*.

Counter	Counter Name	Aggregation (Time)	Unit	Description
M1022C3	HS-DSCH/E-DCH PACKET CALL ATT FOR INTERACTIVE	sum	Integer	Number of HS-DSCH/E-DCH packet call attempts for interactive traffic class. UPDATED: When the UE-specific packet scheduler attempts to allocate an HS-DSCH/E-DCH transport channel for the NRT RAB.
M1022C4	HS-DSCH/E-DCH PACKET CALL ATT FOR BACKGROUND	sum	Integer	Number of HS-DSCH/E-DCH packet call attempts for background traffic class. UPDATED: When the UE-specific packet scheduler attempts to allocate an HS-DSCH/E-DCH transport channel for the NRT RAB.
M1022C5	HS-DSCH/DCH PACKET CALL ATT FOR INTERACTIVE	sum	Integer	Number of HS-DSCH/DCH packet call attempts for interactive traffic class. UPDATED: When the UE-specific packet scheduler attempts to allocate an HS-DSCH/DCH transport channel for the NRT RAB.
M1022C6	HS-DSCH/DCH PACKET CALL ATT FOR BACKGROUND	sum	Integer	Number of HS-DSCH/DCH packet call attempts for background traffic class. UPDATED: When the UE-specific packet scheduler attempts to allocate an HS-DSCH/DCH transport channel for the NRT RAB.
M1022C7	DCH/DCH PACKET CALL ATT FOR INTERACTIVE	Sum	Integer	The number of DCH/DCH packet call attempts for interactive traffic class. UPDATED: When the UE specific packet scheduler attempts to allocate a DCH/DCH transport channels for the NRT RAB.
M1022C8	DCH/DCH PACKET CALL ATT FOR BACKGROUND	sum	Integer	The number of DCH/DCH packet call attempts for background traffic class. UPDATED: When the UE specific packet scheduler attempts to allocate a DCH/DCH transport channels for the NRT RAB.

Table 20: PS BHCA counters

The PS BHCA is the maximum of the sum of the attempts for all the Packet Switch services types:

$$PS\ BHCA =\ MAX(M1022C3 + M1022C4 + M1022C5 + M1022C6 + M1022C7 + M1022C8) \qquad (5.16)$$

The BHCA can be calculated with different granularities (e.g. hourly, daily, etc.). The examples in this work use daily values, thus is important that these values are taken in the BHCA moment. The reason for this is that in this way the methods for forecasting are fed in with the worst values of the day (for daily values) and will produce worst case scenario predictions.

If contrarily to this the methods are fed with the mean value of the day for example, the forecast have to be read as a mean value for the day.

6. Conclusions

The capacity to forecast with precision what will happen is a skill that everyone, in several distinct areas, would like to have. The mobile operators are not an exception, the ability of predict what will happen in their networks would help them to better planning modifications on the network and reduce their monetary investments. However, the task of obtain a good forecast is difficult.

This study is mostly based on classical methods. However the methods referred during this work are widely used in several areas, this study has an unusual approach to the forecast matter in which connects these classical methods with the subject of digital filters. Besides that, it has an evident approach to the forecasting applied to radio access networks. Thus, this work offers a different and systematic approach for forecasting.

This work is evidence that it is possible to generate predictions based on the historical data if the data maintain its typical characteristics. However, it is impossible to predict the future with precision, e.g. unusual occurrences are impossible to predict using only mathematical models. In these cases it must exist other inputs to the models, which mean that the use of these models must be done with caution.

The KPIs presented are vendor specific and were chosen very carefully. This doesn't mean that the study done here only is applicable to these specific KPIs. In fact, the tool developed allows the analysis of other type of metrics.

Until now the mobile operators planning were based in a very general view of traffic growth and qualitative predictions. This study and the forecast tool resultant from it are certainly a good asset to the performance and planning areas of a mobile operator, since it has a more specific inspection of network aspects. However, the result analysis must be done in a very critical and interpretative way.

7. Future Work

During the development of this work some issues were raised. The most significant one was the incapacity of the models to predict unusual observations. This problem was attenuated with the introduction of external inputs to the models, such as the user ability to add a manual growth/decrease factor and add/subtract a level to the forecast results. The truth about most of these unusual events (growths, level falls, etc.) is that they have explanations, such has: seasonal effects, conventions, site shutdown/creation, emerge of new apps, emerge of new technologies, etc. The challenge is to develop and/or add capabilities to the methods that take into account these changes.

The number of forecast methods is vast, it would be interesting to study other forecasting methods that were not covered by this work and compare with the ones studied here. Another important topic are the outliers, which was an area underexplored in this work and would be interesting to explore and test other techniques.

Related with the KPIs, this is a vast area that can be explored. Study of new KPIs, introduction of KPIs to the forecasting tool and most important, the expansion of this tool to another parts of the network and other technologies.

8. References

[1] José Silva, “Proposta 3GPP de Indicadores de Desempenho de Rede – R4 CS Core Network “, Universidade de Aveiro, 2011

[2] Mohammed Jaloun, Zouhair Guennoun, “Wireless Mobile Evolution to 4G Network”, February 24, 2010

[3] Atílio Gameiro, Adão Silva, “Introduction to Wireless Communication”, Universidade de Aveiro

[4] 3GPP- The Mobile Broadband Standard, “Releases”, Available at: <http://www.3gpp.org/specifications/67-releases>. Latest access: October 2014.

[5] Alastair Brydon, “Summary of 3GPP Standard Releases for LTE”, October 2012. Available at: <http://www.unwiredinsight.com/2012/3gpp-lte-releases>. Latest access: May 2014

[6] ETSI TS 123 228, “Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); IP Multimedia Subsystem (IMS); Stage 2”, Release 5

[7] Gonzalo Camarillo, Miguel A. García-Martín, “The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds”, John Wiley & Sons, 2006

[8] Miika Poikselkä, Georg Mayer, “The IMS: IP Multimedia Concepts and Services”, John Wiley & Sons, 2009

[9] Mervi Berner, “High-Speed Downlink Packet Access HSDPA – Improving the WCDMA downlink”, December 2005. Available at: <http://www.comlab.hut.fi/opetus/4210/presentations/6_hsdpa.pdf>. Latest access: May 2014

[10] 3GPP TR 25.855, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; High Speed Downlink Packet Access; Overall UTRAN Description (Release 5)”

[11] Harri Holma and Antti Toskala, “HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications”, John Wiley & Sons, 2006

[12] 3GPP- The Mobile Broadband Standard, “HSPA”, Available at: <http://www.3gpp.org/technologies/keywords-acronyms/99-hspa>. Latest Access: October 2014

[13] 3GPP- The Mobile Broadband Standard, “LTE”, Available at: <http://www.3gpp.org/technologies/keywords-acronyms/98-lte>. Latest Access: October 2014

[14] Dhruv Sunil Shah, “A tutorial on LTE Evolved UTRAN (EUTRAN) and LTE Self Organizing Networks”, The University of Texas at Arlington, December 2010

[15] 3GPP- The Mobile Broadband Standard, “LTE-Advanced”, Available at: <http://www.3gpp.org/technologies/keywords-acronyms/97-lte-advanced>. Latest Access: October 2014

[16] 3GPP TR 36.913, "LTE; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced)", November 2011

[17] Rand Edwards , "History and Status of Operations Support Systems", Springer Science+Business Media, LLC 2007, 30 October 2007

[18] ITU-T Recommendation M.3010 (2000), Principles for a telecommunications management network.

[19] Aiko Pras, Bert-Jan van Beijnum, Ron Sprenkels: "Introduction to TMN", University of Twente Netherlands, April 1999.

[20] Benoit Claise: "Network Management: Accounting and Performance Strategies", Cisco, June 2007

[21] Chompu Nuangjamnong, Stanislaw P. Maj, David Veal: "The OSI Network Management Model – Capacity and performance management", Edith Cowan University, 2008.

[22] Jeff Parker: "FCAPS, TMN & ITL, Three Key Ingredients to Effective IT Management", OpenWater Solutions, May 2005

[23] Igor Pais, "End User Behaviour and Performance Analysis in 3G Networks", Universidade de Aveiro, 2009

[24] 3GPP TR 25.814, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication Management; UTRAN and GERAN Key Performance Indicators (KPI); (Release 5)"

[25] ETSI TS 132 410, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Key Performance Indicators (KPI) for UMTS and GSM", 2012

[26] 3GPP TS 32.106 (V1.1.0): "Technical Specification Group Services and System Aspects; 3G Configuration Management", October 1999

[27] Rob J Hyndman and George Athanasopoulos, "Forecasting: principles and practice", October 2013, Available at: <https://www.otexts.org/book/fpp>. Latest Access: October 2014

[28] Luiz Alexandre Peternelli, "Capítulo 9 - Regressão Linear e Correlação", Available at: <http://www.dpi.ufv.br/~peternelli/inf162.www.16032004/materiais>. Latest Access: October 2014

[29] Investopedia, "Durbin Watson Statistic", Available at: <http://www.investopedia.com/terms/d/durbin-watson-statistic.asp>. Latest Access: November 2014

[30] Manoel Ivanildo Silvestre Bezerra, "Apostila de Análise de Séries Temporais", 2006

[31] Tzveta Iordanova, "Introduction to Stationary and Non-Stationary Processes", Available at: <http://www.investopedia.com/articles/trading/07/stationary.asp>. Latest Access: October 2014

[32] "Managing WCDMA RAN and Flexi Direct", Nokia Solutions and Networks, 17 June 2013

[33] Raija Lilius, "RNC Capacity Management", Nokia Solutions and Networks, May 2010