UNIVERSITY OF PIRAEUS **ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

Διδακτορική Διατριβή

ΔΙΑΧΕΙΡΙΣΗ ΕΤΕΡΟΓΕΝΩΝ ΕΥΡΥΖΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΜΕ ΧΡΗΣΗ ΜΗΧΑΝΙΣΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Αριστοτέλης Γ. Μάργαρης

Πτυχιούχος Τμήματος Ψηφιακών Συστημάτων
Πανεπιστημίου Πειραιώς

Κάτοχος Μεταπτυχιακού τίτλου σπουδών Τμήματος Ψηφιακών Συστημάτων

Πανεπιστημίου Πειραιώς

Πειραιάς, 2021

UNIVERSITY OF PIRAEUS **ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

PhD Dissertation

MANAGEMENT OF CELLULAR BROADBAND NETWORKS BY MEANS OF MACHINE LEARNING TECHNIQUES

Aristotelis G. Margaris

B.Sc. Department of Digital Systems
University of Piraeus

M.Sc. Department of Digital Systems

University of Piraeus

Piraeus, 2021

UNIVERSITY OF PIRAEUS ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αριστοτέλης Γ. Μάργαρης

Πτυχιούχος Τμήματος Ψηφιακών Συστημάτων
Πανεπιστημίου Πειραιώς
Κάτοχος Μεταπτυχιακού τίτλου σπουδών Τμήματος Ψηφιακών Συστημάτων
Πανεπιστημίου Πειραιώς

ΔΙΑΧΕΙΡΙΣΗ ΕΤΕΡΟΓΕΝΩΝ ΕΥΡΥΖΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΜΕ ΧΡΗΣΗ ΜΗΧΑΝΙΣΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Συμβουλευτική Επιτροπή: *Π. Δεμέστιχας, Καθηγητής Παν. Πειραιώς*
Α. Κανάτας, Καθηγητής Παν. Πειραιώς
Α. Ρούσκας, Καθηγητής Παν. Πειραιώς

Εγκρίθηκε από την επταμελή Εξεταστική Επιτροπή την 5η Φεβρουαρίου 2021,

Π. Δεμέστιχας
Καθηγητής
Πανεπιστημίου Πειραιώς

Κανάτα
Α. Κανάτας
Καθηγητής
Πανεπιστημίου Πειραιώς

Α. Ρούσκας
Καθηγητής
Πανεπιστημίου Πειραιώς

Μ. Λούτα
Αναπληρώτρια Καθηγήτρια ΠΔΜ

Ν. Μήτρου
Καθηγητής ΕΜΠ

Σ. Παπαβασιλείου
Καθηγητής ΕΜΠ

Γ. Βούρος
Καθηγητής
Πανεπιστημίου Πειραιώς

....

Δρ. Αριστοτέλης Γ. Μάργαρης

Πτυχιούχος Τμήματος Ψηφιακών Συστημάτων
Πανεπιστημίου Πειραιώς

Κάτοχος Μεταπτυχιακού τίτλου σπουδών Τμήματος Ψηφιακών Συστημάτων

Πανεπιστημίου Πειραιώς

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τη συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημές θέσεις του Πανεπιστημίου Πειραιώς.

....

Dr. Aristotelis G. Margaris

B.Sc. Department of Digital Systems
University of Piraeus

M.Sc. Department of Digital Systems

University of Piraeus

The concepts and conclusions included in this work express the author's personal opinion and should not be interpreted that they represent University of Piraeus official concepts

Αφιερωμένο σε όλα όσα φεύγουν, κάτι που θα μείνει για πάντα

Abstract

Cellular networks are one of the most impactful technologies of today's ICT industry. They provide wireless access to internet and services with very high availability and effectiveness. The evolution of this technology comes with the maturity of the 3GPP-based network and their upcoming releases that promise to deliver even higher quality of service, additional capabilities, and solutions to previous drawbacks. To achieve this, vendors of these technologies must analyze the complexity of these networks and their different deployment options and provide intelligent management software. Variations of cellular networks can be found in literature as Heterogeneous Cellular Networks (HetNets) or Ultra-Dense networks which are improved design flavors of the same system with increased complexity and configurations. The added capabilities of these networks must be used as a toolbox to improve various operational aspects of the networks such as energy efficiency, network performance and system fault prevention. The scope of this Doctorate Thesis is to analyze different approaches of optimizing HetNets in order to suggest plausible suggestions for extensions that will optimize all high-level objectives. Static management and configuration will be used in conjunction with knowledge-building to improve the energy efficiency of key simulation scenarios of 3GPP networks. Dynamic Resource allocation schemes will be used as a real time management algorithm to improve quality of service in a micro-scale. Predictive models based on acquired historical data will be used to predict network operational KPIs, evaluate the probability of network congestion and identification of unknown network element groups based on their behavior. These generated insights will help the infrastructure providers to impose countermeasures to prevent quality deterioration and enforce the technological standards. They will also lead to the reduction of the OPEX and the energy footprint of the system making technology investments sustainable and profitable for network operators. The framework for developing and testing these algorithms is a custom-designed software platform for HetNet simulations and algorithm experimentation. This system is designed according to standards and specifications in order to provide realistic results that will establish the suggested algorithms as strong candidates to be included in future 3GPP-based wireless networks.

Keywords: 3GPP, HetNet, Ultra-Dense Networks, Optimization, Management Knowledge-Building, High-Level Objectives, Energy Efficiency, EMF reduction, Quality of Service, Fault Prevention, KPI Forecasting, Network Element Clustering

Περίληψη

Τα κυψελωτά δίκτυα κινητών επικοινωνιών είναι μία από τις τεχνολογίες με την μεγαλύτερη επίδραση στην σημερινή βιομηχανία των τεχνολογιών επικοινωνιών και πληροφορικής. Παρέχουν ασύρματη πρόσβαση στο διαδίκτυο αλλά και μια πληθώρα άλλων υπηρεσιών με πάρα πολύ υψηλή διαθεσιμότητα και αποτελεσματικότητα. Η εξέλιξη αυτής της τεχνολογίας έρχεται με την ωρίμανση των δικτύων προδιαγραφών 3GPP, η πρόοδος των οποίων υπόσχεται να παρέχει ακόμα υψηλότερη ποιότητα υπηρεσιών, περισσότερες δυνατότητες αλλά και λύσεις σε προβλήματα των παλαιότερων γενεών. Για την επίτευξη αυτών των στόχων, οι κατασκευαστές αυτής της τεχνολογίας πρέπει να αναλύσουν προσεκτικά την πολυπλοκότητα αυτών των δικτύων αλλά και των δυνατοτήτων εγκατάστασης τους και να παρέχουν ευφυές λογισμικό διαχείρισης τους. Διαφοροποιήσεις σε κυψελωτά δίκτυα τύπου 3GPP όπως τα ετερογενή κυψελωτά δίκτυα αλλά και τα «υπερ-πυκνά» δίκτυα είναι εξελίξεις αυτών των δικτύων με αυξημένη πολυπλοκότητα και δυνατότητες που βρίσκεται στη βιβλιογραφία. Μέσω αυτών των δυνατοτήτων μπορούμε να βελτιώσουμε τους διάφορους λειτουργικούς στόχους της υποδομής όπως ενεργειακή αποδοτικότητα, δικτυακές επιδόσεις και αποφυγή σφαλμάτων. Αυτή η διδακτορική διατριβή έχει ως σκοπό να αναλύσει διαφορετικές προσεγγίσεις βελτιστοποίησης ετερογενών δικτύων καταλήγοντας έτσι σε προτάσεις για επέκταση τους επηρεάζοντας όσο το δυνατών περισσότερους στόχους-κλειδιά. Στατική διαχείριση και ρύθμιση σε συνδυασμό με συλλογή γνώσης θα χρησιμοποιηθεί για την βελτίωση ενεργειακή επίδοσης σεναρίων-κλειδιών για την 4η γενιάς κινητής τηλεφωνίας. Αλγόριθμοι δυναμικού διαμοιρασμού πόρων θα χρησιμοποιηθούν σαν μία μέθοδος διαχείρισης πραγματικού χρόνου με σκοπό την βελτίωση ποιότητας υπηρεσιών σε μικρό-κλίμακα. Τέλος, μοντέλα μηχανικής μάθησης θα εκπαιδευτούν σε ιστορικά δεδομένα με σκοπό την πρόβλεψη των λειτουργικών δεικτών του δικτύου, εκτίμηση της πιθανότητας δικτυακής υπερφόρτωσης και αναγνώριση άγνωστων ομάδων δικτυακών στοιχείων βασισμένα στην συμπεριφορά τους. Αυτές οι προβλέψεις θα βοηθήσουν την διαχείριση της υποδομής στην ενεργοποίηση αντίμετρων για την αποφυγή της υποβάθμισης της ποιότητας υπηρεσιών αλλά και την επικράτηση των προδιαγραφών της τεχνολογίας. Αυτές θα οδηγήσουν επίσης στην ελάττωση του OPEX αλλά και του ενεργειακού αποτυπώματος του συστήματος οδηγώντας έτσι σε βιώσιμες και επιτυχημένες επενδύσεις για τους παρόχους. Το πλαίσιο ανάπτυξης για αυτούς τους αλγορίθμους είναι ένα αυτοσχέδιο πρωτότυπο προσομοιωτή ετερογενών δικτύων και πλατφόρμα εκτέλεσης πειραματικών αλγορίθμων. Αυτό το σύστημα σχεδιάστηκε με βάση τις πρότυπες προδιαγραφές προσομοίωσης τέτοιων συστημάτων και τα ρεαλιστικά δεδομένα που θα εξαχθούν από αυτό θα βοηθήσουν στην υπεράσπιση της αποτελεσματικότητας των προτεινόμενων αλγορίθμων για την ένταξη τους στην τεχνολογία κυψελωτών επικοινωνιών 3GPP.

Λέξεις-Κλειδιά: 3GPP , Ετερογενή Κυψελωτά Δίκτυα, «υπερ-πυκνά» δίκτυα, βελτιστοποίηση, συλλογή γνώσης, υψηλοί στόχοι-κλειδιά, ενεργειακή αποδοτικότητα, μείωση ΗΜ ρίπων, ποιότητα υπηρεσιών, αποφυγή σφαλμάτων, πρόβλεψη μετρικών, ομαδοποίηση δικτυακών στοιχείων

Foreword

A PhD dissertation is the detailed diary of an academic's journey through the vast and overwhelming realm of scientific research. The completion of such a journey is a proof that you are an evolved individual, rich with novel tools and knowledge that allow you to move forward and make a difference in modern society.

In the beginning of this research, the goal was to fully understand the current technological achievements of modern 3GPP-based cellular communications. There we would identify problematic situations that would allow for novel approaches and algorithms. After that it became necessary to develop an accurate simulation environment that would be used to test our new approaches, compare them with existing proposed solutions and validate that our contribution was a successful improvement. The process of creating such an accurate software was long and required help from colleagues, foreign institutions, and rich literature. All this labor however was rewarded with a state-of-the-art engine for knowledge building and algorithm development. The next steps were to provide one technical solution for each of the three categories of network problems, energy efficiency, quality of service and network stability. These three problems required three different altogether approaches which proved that to move forward in the next generation of telecommunications, we need to embrace as much as possible new, radical advanced techniques.

This journey would never be accomplished if it weren't for all the fellow academics of University of Piraeus. First, I would like to personally thank my teacher and mentor Prof. Panagiotis Demestixas as well as the Prof. Athanasios Kanatas and Prof. Angelos Rouskas. Their contribution, guidance and support helped me chart an important course in these scientific fields leading me to the completion of this dissertation. The important members of the TNS lab Dr. Kwstas Tsagkaris, Dr. Andreas Georgakopoulos, Dr. Yiouli Kritikou, Dr. Dimitris Karvounas, Dr. Dimitris Kelaidonis, Dr Aimilia Bantouna played a key role in my research, providing constant input, wise suggestions and continuous inspiration. My colleagues and fellow PhD students Ioannis Belikaidis and Mixalis Mixaloliakos gave me the strength of a real fellowship as well as Dr Panagiotis Vlaxeas and Dr Vassilis Foteinos. In addition, I would like to thank my colleague George Poulios for being both a role model scientist, a dependable coworker, and a true rival. Finally, I would like to thank my family and my closest friends who supported me in pursuing this esteemed academic reward. Finishing this has allowed me to move one step closer to the completion of my academic ambitions.

Πρόλογος

Μία διδακτορική διατριβή μπορεί να τη φανταστεί κανείς ως το λεπτομερές ημερολόγιο ενός ακαδημαϊκού ταξιδιού στον αχανή και γεμάτο δέος κόσμο της επιστημονικής έρευνας. Η ολοκλήρωση ενός τέτοιου ταξιδιού είναι μία απόδειξη ότι έχεις μετατραπεί σε ένα εξελιγμένο άτομο, πλούσιο με νεοφυή εργαλεία και γνώση που σου επιτρέπουν να κάνεις ένα βήμα μπροστά και να φέρεις την αλλαγή σε μία μοντέρνα κοινωνία.

Στις απαρχές αυτής της έρευνας, ο στόχος μας ήταν η κατανόηση των σημερινών τεχνολογικών επιτευγμάτων των μοντέρνων συστημάτων κινητής τηλεφωνίας της 3GPP. Μέσω αυτού μπορέσαμε να εντοπίσουμε προβληματικές καταστάσεις οι οποίες θα χρήζουν νέες προσεγγίσεις και αλγορίθμους για την αντιμετώπιση τους. Αυτό μας οδήγησε στο συμπέρασμα ότι είναι απαραίτητο να αναπτυχθεί ένα κατάλληλο εργαλείο προσομοίωσης το οποίο θα χρησιμοποιούταν για να δοκιμαστούν οι νέες προσεγγίσεις και να συγκριθούν με τις υπάρχουσες λύσεις , επιβεβαιώνοντας την βελτίωση που μπορούν να επιφέρουν. Η διαδικασία της κατασκευής αυτού του λογισμικού ήταν μακριά και δύσκολη και απαίτησε την συνολική συνεισφορά συναδέλφων, ξένων οργανισμών / ιδρυμάτων αλλά και μελέτη της βιβλιογραφίας. Στο τέλος ο κόπος απέδωσε καρπούς, καθώς μας έδωσε το κατάλληλο εργαλείο τεχνολογικής αιχμής το οποίο θα χρησιμοποιούταν ως η βάση για την ανάπτυξη όλων των αλγορίθμων. Τα επόμενα βήματα που ακολούθησαν ήταν να προτείνουμε μία διαφορετική τεχνική λύση για κάθε έναν από τους κύριους άξονες διαχείρισης τέτοιων υποδομών , ονομαστικά την ενεργειακή αποδοτικότητα, την ποιότητα υπηρεσιών και την σταθερότητα του δικτύου. Τα προβλήματα που πηγάζουν από τις διαφορετικές αυτές ενότητες, χρήζουν διαφορετικής προσέγγισης επίλυσης τους κάτι το οποίο αποδεικνύει ότι για να οδηγηθούμε στη νέα τεχνολογική γενιά θα πρέπει να «αγκαλιάσουμε» όσο το δυνατόν περισσότερο ριζοσπαστικές και σύνθετες τεχνολογίες.

Αυτό το ταξίδι δε θα μπορούσε ποτέ να έχει ολοκληρωθεί χωρίς τους συνάδελφους ακαδημαϊκούς του Πανεπιστημίου Πειραιώς. Πρωτίστως θα ήθελα να ευχαριστήσω προσωπικά τον καθηγητή και μέντορα μου Παναγιώτη Δεμέστιχα, καθώς και τα υπόλοιπα μέλη της τριμελής μου καθηγητές Αθανάσιο Κανάτα και Άγγελο Ρούσκα. Η συνεισφορά, καθοδήγηση και υποστήριξη τους ήταν μία σημαντική βοήθεια στην ολοκλήρωση αυτής της διατριβής. Τα μέλη του εργαστηρίου Τηλεπικοινωνιών, Δικτύων και Υπηρεσιών Δρ. Κωνσταντίνος Τσαγκάρης, Ανδρέας Γεωργακόπουλος, Γιούλη Κρητικού, Δημήτρης Καρβουνάς, Δημήτρης Κελαηδόνης, Αιμιλία Μπαντούνα έπαιξαν σημαντικό ρόλο στην έρευνα μου, παρέχοντας μου συνεισφορά, ευφυείς παραπομπές και συνεχή έμπνευση. Οι συνάδελφοι και Υποψήφιοι διδάκτορες Ιωάννης Μπελικαίδης και Μιχάλης Μιχαλολιάκος μου παρείχαν τη δύναμη μίας πραγματικής ακαδημαϊκής συντροφιάς καθώς επίσης και οι Δρ. Παναγιώτης Βλαχέας και Βασίλης Φωτεινός. Επιπλέον θα ήθελα να ευχαριστήσω τον συνάδελφο Γιώργο Πούλιο ο οποίος αποτέλεσε για εμένα πρότυπο επιστήμονα, ακαδημαϊκό αντίζηλο αλλά και αξιόπιστο συνεργάτη. Τέλος , θα ήθελα να ευχαριστήσω την οικογένεια μου και τους κοντινότερους φίλους μου για την υποστήριξη τους στην διεκδίκηση αυτού του πολυπόθητου μου στόχου. Η ολοκλήρωση του είναι ισοδύναμη με ένα ακόμα σταθερό βήμα στην ολοκλήρωση των ακαδημαϊκών μου φιλοδοξιών.

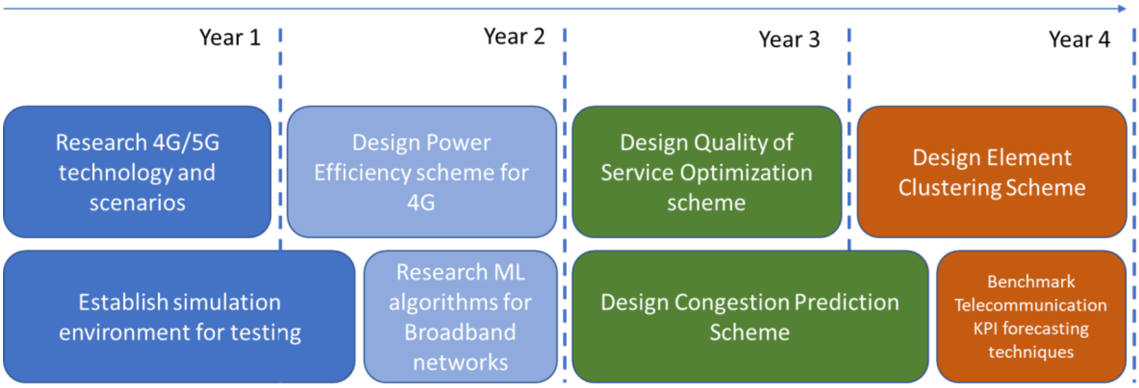
Table of Contents

List of Figures

List of Tables

Thesis Timeline

The initial planning for this research project was included in the doctorate proposal and it included 4 yearly stages that would result in the final doctorate thesis document. It is of key importance that research follows accurate scientific steps in order to understand a problem statement, study the existing status of the scientific community, implement the SOTA, design an improvement and implement it in order to extract feedback on the benefits it can bring.



Year 1
Year 2
Year 3
Year 4
Research 4G/5G technology and scenarios
Design Power Efficiency scheme for 4G
Design Quality of Service Optimization scheme
Design Element Clustering Scheme
Establish simulation environment for testing
Research ML algorithms for Broadband networks
Design Congestion Prediction Scheme
Benchmark Telecommunication KPI forecasting techniques

Figure 1 - Timeline of the doctorate thesis

In the 1st year we focused our research in the next generation of cellular networks, analyzing the modern literature, 3GPP and 5GPPP standards and reference scenarios. This included the technological features of cellular technologies and their key differences with other wireless network communications such as the 802.11 standard. We identified the design principles behind the modern deployment versions of such networks from homogeneous placements of GSM network to HetNet and Ultra-Dense networks. Our Participation in Green-Touch consortium, conferences and publications allowed us to formulate the requirements for an accurate simulation environment that would work as the testbed for all our future optimization attempts. We also studied the S.O.T.A in other simulation environments such as Omnet++, NS2 , NS3 and OpNet in order to finalize the design process of the software. 3GPP reports contained crucial key scenarios of the LTE standard that we used as baseline to design our future optimization / improvements. We also identified the KPIs for the various management high level objectives to include them in the calculations.

The development of the simulation software was finalized in the 2nd year while simultaneously we focused on our 1st important optimization use case, energy efficiency in LTE HetNets. In order to achieve a sustainable energy consumption status for the reference scenarios, we used cross-operator infrastructure sharing (leveraging the stochastic nature of the traffic demand) and supplemented the quality of service gap by introducing strategically planned Pico cells in traffic hotspots. The simulation software was then used to validate the results and it resulted into the first journal publication on IEEE vehicular technologies magazine. The way forward led into research for application of other types of algorithms that are more data-driven and more flexible than network redesign. Machine learning is a promising field for network optimization and rich literature can be found on different directions. This literature was split into 3 different subcategories, semi-supervised classification, clustering and forecasting. The last 2 were pushed to be studied in the final (4) year of the thesis while the 1st was the next item to be included in the study.

In the 3rd year of this doctorate thesis, we moved further into additional HetNet optimization scenarios for various important situations that are foreseen as problematic. We studied the literature for the best approach on tackling the quality of service optimization problem in dense urban deployments by the means of applying intelligent dynamic resource allocation in the radio link control module of LTE. Appropriate simulation scenarios were selected to showcase the importance of real time management by the means of SON functions is crucial for the runtime of HetNet infrastructure. Also, we visited different approaches such as class-based resource allocation and policy enforcement. The simulation results showed promising benefits in the selected scenarios, and this resulted in our 2nd journal publication is spring wireless communication journal. We also studied a different use case in which network was being congested due to a change in the underlying state of the active user equipment terminal devices. Predictive modeling that used semi-supervised learning and Self-Organizing maps was used to identify the congestion and appropriate counter measures in the handover algorithm were activated in order to optimize the network performance. The analysis of the outcome was split into two parts: a) a performance analysis on the machine learning model fit, that would encourage us to include is as a component for the optimization and b) the network KPI improvement due to the application of the predictive countermeasures. The results of the ensemble scheme were a dramatic prevention of the network congestion for various network load conditions and situations.

In the 4th and final year, we focused on the rest of the machine learning predictive models to identify novel use cases for their application. Literature for unsupervised learning and dimensionality reduction techniques was studied in order to solve the management complexity of Ultra-Dense networks. In detail, we used applied clustering to identify network elements that belong to the same behavioral categories. Elements that serve the same amount of traffic get grouped together and can then be effectively managed. The same methodology was used in a different scenario in which different classes of users were identified and grouped together, in order to apply different radio resource allocation schemes. In both cases the various clustering algorithms showed promising accuracy in the identification of the hidden groups. The second half of the final year was dedicated to network KPI forecasting algorithms. The importance of KPI forecasting lies in the value of the incident prevention. A timely prediction of congestion or a high throughput spike can be used in conjunction with countermeasures to provide network robustness and stability. A large set of forecasting models were benchmarked on measured network KPIs produced by the simulator for key scenarios to test the limitations of their forecasts. The results show us that forecasting is a useful tool for short-to-mid-term predictions and can also be a robust tool for the future of cellular networks.

Chapter 1 – 3GPP Cellular Networks of 4th and 5th generation

1.1 Introduction

In this introductory chapter, we will analyze the characteristics that differentiate cellular networks of the 4th / 5th generation from their predecessor (2G/3G) in terms of architecture[1], technologies and management methodologies. In detail, we will focus on a) the usage of heterogeneous network coverage elements that consist a type of network sometimes referred to as HetNet, b) The density and geospatial diversity of the placement for such network elements that lead to a new definition of mobile networks, Ultra-Dense Networks, c) The usage of multiple simultaneous carrier frequency groups in bands that are not traditionally used for mobile telecommunications such as microwave and mm-wave bands in order to deliver higher capacity and meet the user demand. Finally, we will analyze the necessity and difficulty of intelligent management for these systems which will be the centerpiece of this doctorate thesis

1.2 Heterogeneous Cellular Networks (HetNets)

Heterogeneous cellular networks[2][3][4] are up and coming network architectures for the cellular network providers and have dominated the architectural models in the last decade, beginning with the release 8 of the LTE (4G network). In contrast with the homogeneous cellular networks (which consist of the repetitive placement of homogeneous network elements with respect to the coverage area, capacity specifications and capabilities), HetNet architecture is utilizing radio coverage elements of different specifications and capabilities in terms of transmit power, antenna type and functionalities (coverage capabilities, inter-communication with other elements, backhauling capabilities etc.) creating a more capable and also a more complex access network that is better suited for the present environments.

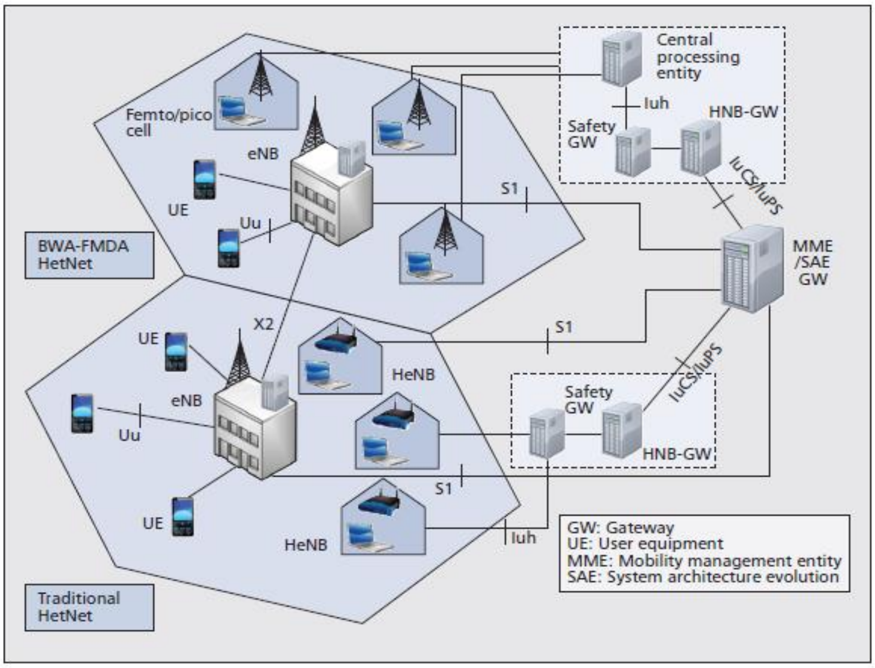


Figure 2 - Example Heterogeneous Cellular Network (HetNet) [1]

The benefits of this architecture aim at the geospatial imbalance between different urban areas (city areas, rural areas or industrial areas) in terms of traffics demands and/or human density which is a paradigm shift in relation to the older planning principles. It also benefits from the city layout and the impact that it has on the radio transmission environment (reflections, delay diversity, absorption and shadowing). For the design of such networks, detailed population density maps are utilized as the majority of commercial cellular modems are currently handheld devices such as smartphones, tablets and other 4G/5G devices. The antenna systems of a heterogeneous cellular network are mainly categorized by the size of their radio coverage (effective coverage as it results from the accompanying antenna it has). The most commonly seen HetNet elements are: a) Macro Cell (eNodeB , coverage of 250 to 1500m), b) Micro cell (250 to 100 m) , c) Pico cell (100 to 50 m) , Femto cell (<25m mostly indoor elements), Wi-Fi Access Points (<25m) and also Remote Radio Heads (RRH) which vary in coverage and are mostly used as repeaters / relays for a long range macro eNodeB transmission (i.e. it does not include its own Layer 2 and onwards network stack so it is referred to as a passive network element)

1.3 Ultra-Dense Networks

The term "Ultra-Dense networks" is mentioned in modern literature as telecommunication networks of the 4th cellular generation which consist of multiple overlapping layers of radio coverage technologies. These layers can consist of network elements with a) different radio specifications (e.g. transmit power, carrier frequency, bandwidth), b) non-symmetrically placed radio elements, c) different element categories (e.g. Macro / Pico / Femto cells) and d) multiple simultaneous generations of radio technologies (e.g. combination of 2G, 3G ,4G along with Wi-Fi access points controlled by operators). This architecture is also contradicting the traditional symmetrical and homogeneous design of the previous generations which started with GSM and was followed by the UMTS network.

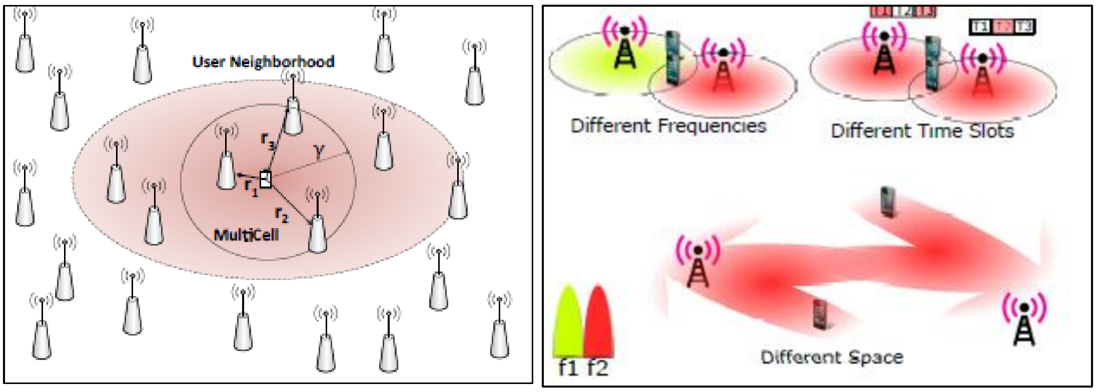


Figure 3 - Example of UltraDense Networks (UDN) [5][6]

Utilizing the methodology of multiple division of a coverage area to smaller and denser network elements, the network designers can achieve higher spatial performance in indices such as geospatial spectral density (Mbps / square meter) and average spatial interference or SINR. However, ultra-dense deployments are shown that can lead to various management problems and require advanced management algorithms and methodologies for their smooth operation.

1.4 Higher Frequency Cellular Networks

Historically[7][8], cellular networks utilize the initial 2nd and 3rd generation allocated radio frequency bands (800Mhz, 900Mhz, 1800 Mhz, 2100 Ghz etc.) which are each correlated to a different generation of networks. This results in a dedicated bandwidth for the service of the specified quality of service level that each technology promises. The increasing required capacity however that arrives with next generation of #GPP technologies require expansion on additional bands that lie above the 2.4Ghz band towards micro-wave and mm wave bands that were previously used for wireless backhauling and satellite links. In addition, carrier aggregation techniques allow for multiple carrier frequency combination that results in an even higher effective bandwidth for the base-band unit and consequently for the data link layer. These new frequencies (central frequencies of 6-13, 15-42, 80,100,150 Ghz) have their own shortcomings and challenges in order to be used as an access network frequency. Pathloss factors increase with the increase of frequency which causes problems in the propagation and refraction (which is an important principle for access networks as line of sight is almost always unavailable). However, they provide a very large and relatively "clean" bandwidth (in the order of Ghz) and the technological advancements now allow us to utilize them as well provided intelligent radio link control schemes are active. Ultimately this will lead to a tremendous increase in the networks capacity and capabilities that will lay strong foundation for the future evolution.

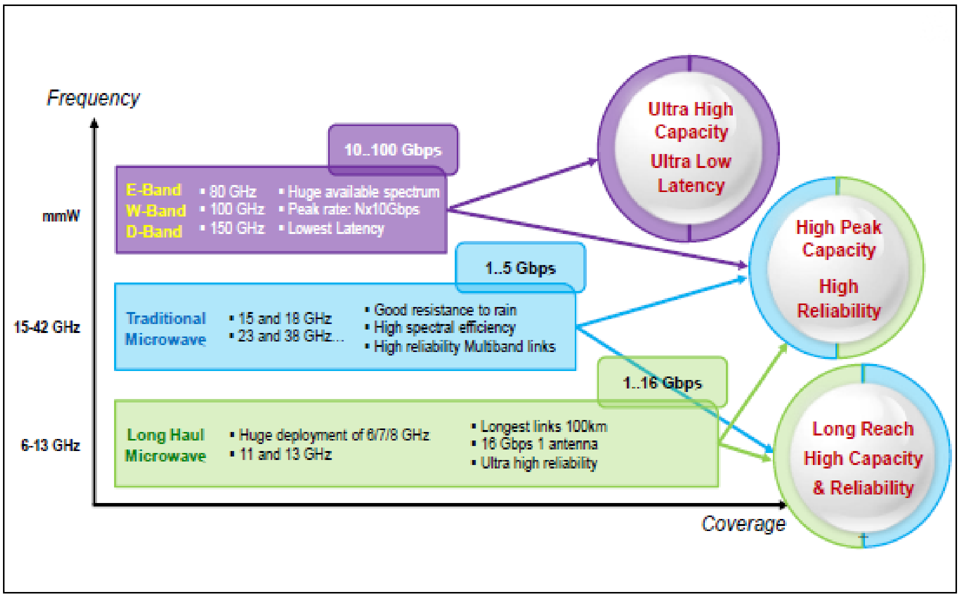


Figure 4 - New Access Frequency Bands (MicroWave and mmWave+) [7]

1.5 High Level Objectives, KPIs and hierarchical Cellular Network management

Management of complex intelligent network infrastructures[1][2][9] is a complex and divisible optimization problem which can be approached with different ways such as a top down approach (i.e. from a higher level of perspective, from network goals to element goals) or a bottom up approach (i.e. focusing on micro-optimization in a local level which in turn result into system-wide problem solving). In reality, mix of the two strategies are utilized which results into a hybrid solution for the best results. In following chapters, we analyze the term "infrastructure management" into its respective subcomponents and the means to achieve it along with the types of results it can derive. A complex telecommunication system can function in various operational modes by focusing the management and configuration capabilities towards a specific central KPI "axis". This axis is the central policy that will dictate the operation of the system and can sometimes be referred to as High Level Objective (HLO). The axes can have conflicting and reciprocated components therefore the optimization of their sub-objectives can result into the deterioration of the other. In literature this is referred to as a "tradeoff" (i.e. two different aspects of the system are trading their states from effective to ineffective) and every management action must be analyzed for its tradeoffs into different KPI axes. Our approach for the different HLOs of the telecommunication infrastructure and HetNets is split into the following, conflicting HLOs: A) Efficient Resource Utilization, B) Subscriber quality of service / quality of experience, C) Energy Efficiency / Power consumption of infrastructure. The contextual separation between these 3 HLOs is clear, however they cannot be satisfied simultaneously due to their correlation.

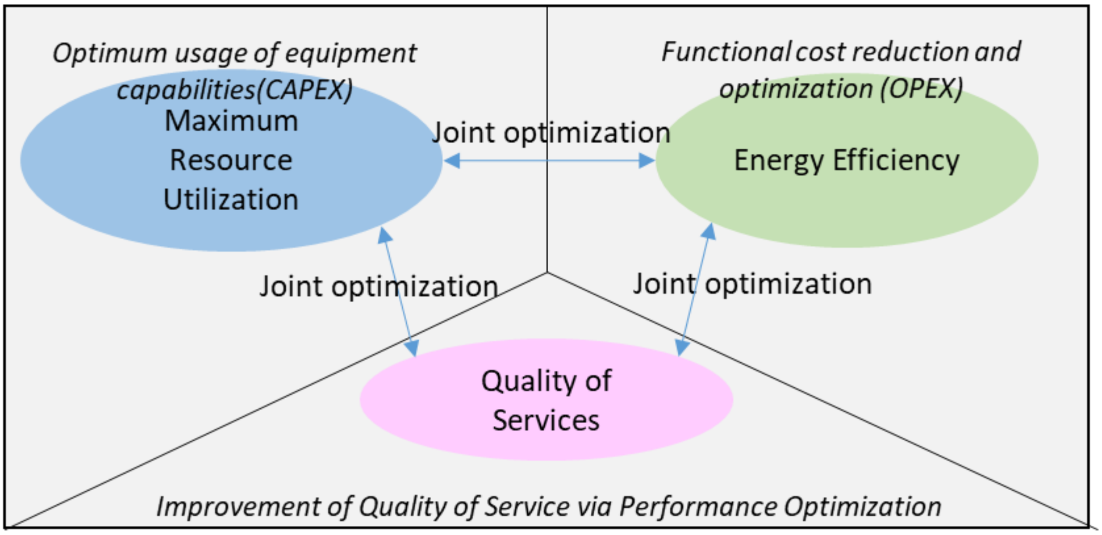

Optimum usage of equipment capabilities(CAPEX)
Maximum Resource Utilization
Functional cost reduction and optimization (OPEX)
Energy Efficiency
Joint optimization
Joint optimization
Joint optimization
Quality of Services
Improvement of Quality of Service via Performance Optimization

Figure 5 - Network Management HLO correlation analysis

For application purposes in real telecommunication environments, the selection of the proper high-level policy is a crucial and demanding decision that must take into consideration various factors (financial, technological, geographical etc.). In addition, the complexity of the correlations between the HLOs require a proper sub-objective analysis and planning in order to achieve the wanted goals. The sub-objectives are being generated following a tree-like dependency structure that follows the network hierarchy from the system-wide KPIs to the element-wide KPIs. In addition, hierarchy can be applied in the temporal scope of any management action which separates long-term and short-term lifecycles to observe the impact of any action.

Table 1 - Different Contextual hierarchies in system management

Generic	Time	Hierarchy	Example in 3GPP cellular technologies	Example of translating the policy into a specific KPI
	Long-Term	Management Platform Level	High level network management system (NMS)	Average monthly instantaneous network power consumption (Watt)
	Mid-Term	Management Agent Level	Mid level element management system (EMS)	Average per element power consumption (Watt/element)
Specific	Short-Term	Enforcing Entity Level	eNodeB (implementing the management actions)	Change Transmit Power or application of intelligent algorithm for reduction of power consumption (Watt /h)

The tree-like architecture introduces additional complexity in the infrastructure management algorithms (and schemes) but helps increase the precision and level of detail for different management solutions. This information increases the effectiveness of the operations and helps us predetermine the impact that it will have on the system. Additional diagram analysis can also give us a general "picture" of how the optimization / management methodology affects a telecommunication system (and its respective KPIs as a whole). Such algorithms can lead to relevant knowledge extraction from various control loops which can then be added in existing management schemes to fine-tune them with the new state of the system.

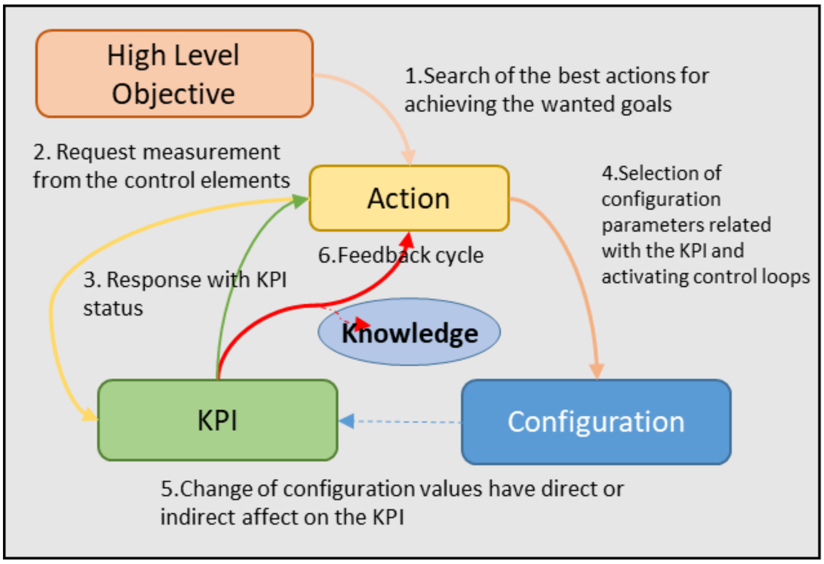

High Level Objective
1.Search of the best actions for achieving the wanted goals
2. Request measurement from the control elements
Action
4.Selection of configuration parameters related with the KPI and activating control loops
6.Feedback cycle
3. Response with KPI status
Knowledge
KPI
Configuration
5.Change of configuration values have direct or indirect affect on the KPI

Figure 6 - Knowledge-building during Intelligent Management

1.6 HetNet Management Schemes

In this chapter we are enumerating in a short description all the general categories of infrastructure management schemes for HetNets. As in all dynamic information systems that operate in real time, HetNets can be influenced by a large variety of problems in different operational stages. These stages begin from their initial setup / design phase and can continue along the operational function up until the termination phase. For every stage with different characteristics, different problem-solving methodology can be selected that will be more tailored to the nature of the problem.

1.6.1 Static Management and HetNet Design

In this static management category of telecommunication infrastructures, we can group all the actions that revolve around the placement / positioning and parameterization of the various network elements in the designated area. This, as the name of the scheme implies, occurs in a predetermined, manual or static' way based on calculations that have occurred and estimations of the operational environment. Initial design of a system is always a very detailed and complex decision point for such systems as it is further analyzed into multi-variable sub-problems that co-depend simultaneously. A lot of mathematical and mechanical modeling can be found in the literature for cellular network design based on optimization of various aspects (coverage, energy consumption, quality of service etc.). The mathematical models that are used in this stage vary from estimates to very detailed simulation models of the real environment in order to provide the algorithm with the best possible inputs and lead to the best results. In practice however, it is observed that these models are both the strengths and the weaknesses of static management scheme. While the simulation and calculation models provide a form of "concentrated knowledge" that is close to accurately predict the parameters of the system's environment, they suffer from the vulnerabilities and error of all the statistical estimation methods. Applied statistics can often miss out on "outlier" data points and lead to "average" estimations which, in great number of observations are accurate, but fail to accommodate for micro-management and more detailed phenomena. The impact that these statistical errors have on the output of the static management scheme can be very high resulting in serious KPI decline. Therefore, some problems are bound to have better suited solution schemes. Designers of HetNet infrastructures frequently use these models to acquire predictions results of new technologies based on trained models that used data from previous generations. These models cannot always be used in this generalized manner leading to mistakes in the calculations. Different technologies hide implicit differences that are very hard to quantify and include them as parametrization in models, especially when moving from one telecommunication generation to another. Design and static management cannot always accommodate for the geospatial evolution of urban environments in civilian areas. Urban evolution and population evolution models can be used but this will demand continuous re-parametrization of network elements to respond to the demands of the increase or decrease in telecom traffic (e.g. caused by the addition of a single train station in an area). Regardless of all the drawbacks mentioned, static models and planning cannot in any case be called obsolete as it can be used as a solid basis for optimized telecom infrastructures with some extensions and adjustments.

1.6.2 Dynamic Management with control loops and SON Functions

Dynamic management of HetNet is a real time operation that uses the predefined configuration points of the network as provided by the vendor of telecom equipment and infrastructure. In many references (either in literature or commercial software products) they can be referred to as SON functions, functions for the self-organization of networks. These functions are management control loops that are the first step towards AI-based models in commercial telecommunication networks. In principle SON-enabled systems have platforms that allow software to be plugged, installed or uninstalled and then activated when certain criteria are met to perform dynamic management. Different states of the network elements lead to different operational policies that ultimately impact the element KPIs. Many categories of such SON algorithms are currently used commercially and operating on cellular networks worldwide. In addition, the academic literature has numerous publications for SON functions because of the important paradigm shift from the previous, static management schemes. These management functions have many advantages versus the volatility of the ever-changing context of the network. With proper programming implementations, SON algorithms can solve a large variety of network problems in an automated manner. This automation leads to important benefits and improvements over the operator's KPIs and, ultimately, the OPEX of the telecommunication system. The negative drawbacks of SON functions are largely related with the additional complexity that is being introduced to the infrastructure to be able to execute such control loops. This can include increased signaling between network elements (i.e. more resources used and worse QoS) and increased complexity on the network operational environment that can result in reduced understanding of the causes between various incidents or network behavior. This further results into a decline in the prediction capability of future behavior of the system which can result into chaotic states. Another drawback of the SON functions that can be found in the public literature is their scope of application (with respect to the entirety of the cellular network). SON functions are sometimes designed to be applied locally or in a telecom site that includes 3-8 cells (radio units) performing an algorithm (e.g. dynamic power control) which uses as input information and measurements that are related only to the selected subset of (3-8) elements. This can often lead to an environment of competitive optimization which can lead to both successful (in cases of e.g. distributed Load-balancing SON) or to failed (in cases of handover optimization) applications. Centralized, global-scope SONs are shown to tackle this problem by taking in consideration all the relevant elements of the network but are very complex and require a very large amount of signaling between the managed nodes that leave a significant footprint on the network operation. Centralized SON's great workload can sometimes lead to requirements of dedicated computation hardware that is required for solving the management problem. The effectiveness of SON functions, regardless of the drawbacks, is gaining more and more ground in HetNet environments for their innovational perspective on the problems that rise.

1.6.3 Management with Knowledge-based predictive models

As mentioned in the previous chapter, static management schemes of HetNets can use internal mathematical and/or computational models that are generated by statistical studies on historical data and also natural electromagnetic transmission phenomena during the operation of the network. Analyzing this approach led us into the conclusion that these models need to be continuously re-calibrated or updated in order to adapt to the rapid changes that can occur during the system operation. To meet this requirement, we need to implement machine learning methodologies to improve the complexity and accuracy of the models. Machine learning is a new methodology of solving complex problems that derives from a "data-driven" philosophy. In simple terms, machine learning is the parametrization of a mathematical model as an optimization problem of the prediction error calculated on historical data. These models aim on predicting and calculating the circumstances / future states that will come and apply tailored methodologies to avoid unwanted events or exploit them to the system's benefit. The advantages of these methods are based on the advantages of both management schemes (dynamic and static management) as they minimize the negative aspects of each schemes while maintaining all the benefits. The negative aspect of machine learning is the internal process of learning itself. This procedure is an internal optimization problem that comes in addition to the other problems that are part of the system. The complexity of solving this problem and the accuracy of the solution is mostly based on acquiring a large amount of historical data that are enough to lead to the best parameters of the model. However, since data can be found in abundance for various problems, academic literature is booming with references to machine learning-based solutions of various systems including HetNets. This research will ultimately lead to a stable, industrialized and accurate model-building machine learning implementation for commercial environments.

1.7 Conclusion

Having enumerated the various methodologies for management of heterogeneous cellular networks (i.e. static management and planning, dynamic management and SON functions and also management by utilizing machine learning and knowledge building), it is clear that not one single management scheme can be used to cover all aspects of the system. This leads to the necessity of constructing an experiment software platform that will perform simulations, apply tests and algorithm methodologies and evaluate the KPI results for various management schemes. The outputs of this system will lead to a high-level policy of which types of problems are best solved with which algorithms and methodologies. The synthesis of all these solutions will be the total output of this doctorate thesis and will lead to the best possible operation of HetNet systems. Future systems will provide interfaces for such intricate management schemes and it will be possible to apply them in real environments and therefore benefit from the improvements that will be brought.

1.8 Chapter References

[1] P. Demestichas, K. Tsagkaris, A. Georgakopoulos, "Knowledge-based Management for Wireless/Mobile Broadband Infrastructures", University of Piraeus, White paper, 2013

[2] A.Georgakopoulos, P. Demestichas, V. Stavroulaki, K. Tsagkaris, A. Bantouna, "Mechanisms for Information and Knowledge Sharing in Wireless Communication Systems", submitted to International Symposium on Wireless Communication Systems (ISWCS) 2012, Paris, France, August 28-31, 2012

[3] D. Karvounas, P. Vlacheas, A. Georgakopoulos, M. Logothetis, V. Stavroulaki, K. Tsagkaris, P. Demestichas, "Coverage and Capacity Optimization in Heterogeneous Networks (HetNets): A Green Approach", in Proc. International Symposium on Wireless Communication Systems (ISWCS) 2013, Ilmenau, Germany, 27-30 August 2013, International Conference Papers

[4] Qualcomm, "A Comparison of LTE Advanced HetNets and Wi-Fi", October 2011, white paper

[5] M. Kamel, W. Hamouda, A. Yussef, "Ultra-Dense Networks: A survey" IEEE communications, surveys and tutorials, Volume: 18, Issue: 4 May 2016

[6] W. Yu, H. Xu, A. Hematian, D. Griffith, and N. Golmie, "Towards Energy Efficiency in Ultra Dense Networks" 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), Dec. 2016

[7] Y. Wang, J. Li, L. Huang, J. Yao, A. Georgakopoulos, P. Demestichas, "5G Mobile: Spectrum Broadening to Higher-Frequency Bands to Support High Data Rates", IEEE Vehicular Technology Magazine, vol.9, no.3, pp.39-46, 2014, Journal Papers

[8] ETSI (R. Lombardi), "Microwave and Millimetre-wave for 5G Transport", White Paper No. 25, February 2018

[9] M. Lauridsen, L. Giménez, I. Rodriguez, T. Sørensen, and P. Mogensen "From LTE to 5G for Connected Mobility", IEEE Communications Magazine, March 2017

Chapter 2 - Simulation and Algorithm Application Platform

2.1 Introduction

For this doctorate thesis, it is necessary to develop appropriate simulation software with the capabilities of performing the operation of an actual HetNet infrastructure, applying various methods and policies for operating in an optimized manner and gathering measurements to evaluate the results. In order to achieve the level of calculation realism required, appropriate literature was studied and incorporated into the design analysis and implementation of this software. A software simulator is a large study item for various fields and this chapter will be dedicated to analyzing the various modular aspects of the tool. The simulator developed is designed to perform realistic network operation for HetNet radio and network components "serving" in addition to the user equipment terminals that are the "clients" of the network. The serving elements as mentioned in previous chapters are comprised of sub-components of the various OSI layers, mapped to their corresponding implementation from the 3GPP standards. The types of nodes that are being used as components for simulations are sector antenna multi-sized 3-node cells called eNodeB and also smallest range components called Micro or Pico cells with various ranges of circular coverage. The software implementation for these components includes the physical layer calculation for link budget for each link, the data link layer namely the RRC (Radio Resource Control) and RLC (Radio Link Control) units which involves dynamic modulation and coding schemes and handover operations. The environment of a simulation plays a key role into scenarios that provide useful insights for the various studies. For this reason, the software includes a lot of radio environment aspects and geospatial or mobility capabilities. The most important aspects that affect the operation of a HetNet are:

- Topology, geography, and the position of the stationary elements of the simulation.
- Mobility and the relative position of the user equipment devices that affect the line of sight and shadowing of the link
- The telecommunication channel with its various properties (radio frequency, bandwidth and all the positive and negative aspects of noise and interference)

- The generated traffic model that emulates the user equipment device usage with different traffic patterns for different services

The studied bibliography, standards and white paper were selected to provide the best accuracy to all these aspects of the simulation software. In addition, new models were designed and implemented to provide further extension to the existing public literature.

2.2 Software Specifications

The simulation / validation software created is based on thorough state-of-the-art research[1][2] of other network simulators that are used for publications and standardization studies. The specifications for these simulators are aligned with large standardization organisms for wireless technologies and cellular communications such as the IETF, 3GPP[3] (4th generation cellular communications), ITU[4][5] (worldwide radio channel regulation) and 5GPPP[6][7] (5th generation cellular communications). The software is designed to provide a main operation mode which follows a strictly linear flow diagram. This flow diagram includes the following steps: a) parameter initialization for the simulation scenario, b) topology initialization for the user equipment devices and network components, c) simulation execution runtime (with rich graphical user interface that provides interaction capabilities and d) measurement of various important KPIs and generation of reports for the results. Multithreaded techniques are utilized to provide rich visualization of the operation of the network and the underlying playground. In addition, real time analytics in the form of time series, probability density function plots and rasterized colored heat-maps can be enabled to measure different aspects of the system.

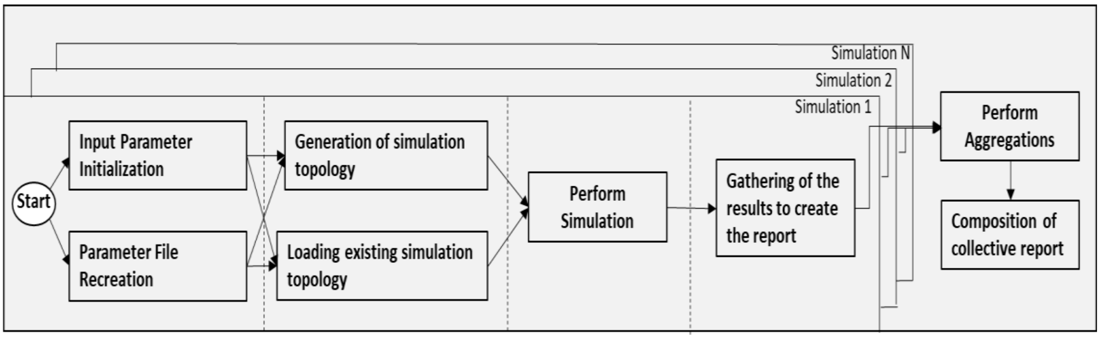

Simulation N
Simulation 2
Simulation 1
Start
Input Parameter Initialization
Parameter File Recreation
Generation of simulation topology
Loading existing simulation topology
Perform Simulation
Gathering of the results to create the report
Perform Aggregations
Composition of collective report

Figure 7 - Flow diagram of the simulator

2.3 Functional Modules of the Simulator

The HetNet system level simulator is a modular piece of software which is based on connected software components that exchange synchronous message containing the multi-layer information. The summary of all the modules combined form the realistic simulation of the cellular network that provides accurate calculations of the environment and entities. In this chapter we will analyze thoroughly all these modules and provide description for their functionalities.

2.3.1 Topology and Mobility Module

The placement and location of mobile terminals in space, with respect to the positions of the various network elements, is a key input parameter of a system level simulation[3][4]. Large scale network scenarios such as large, dense urban, multi kilometer coverage areas with a mixture of buildings and streets require the usage of element placement models that are based on measurements from various European capitals (e.g. Paris or Berlin). These models are sometimes mentioned as population density maps and they play a key role in the generation of traffic volume through space.

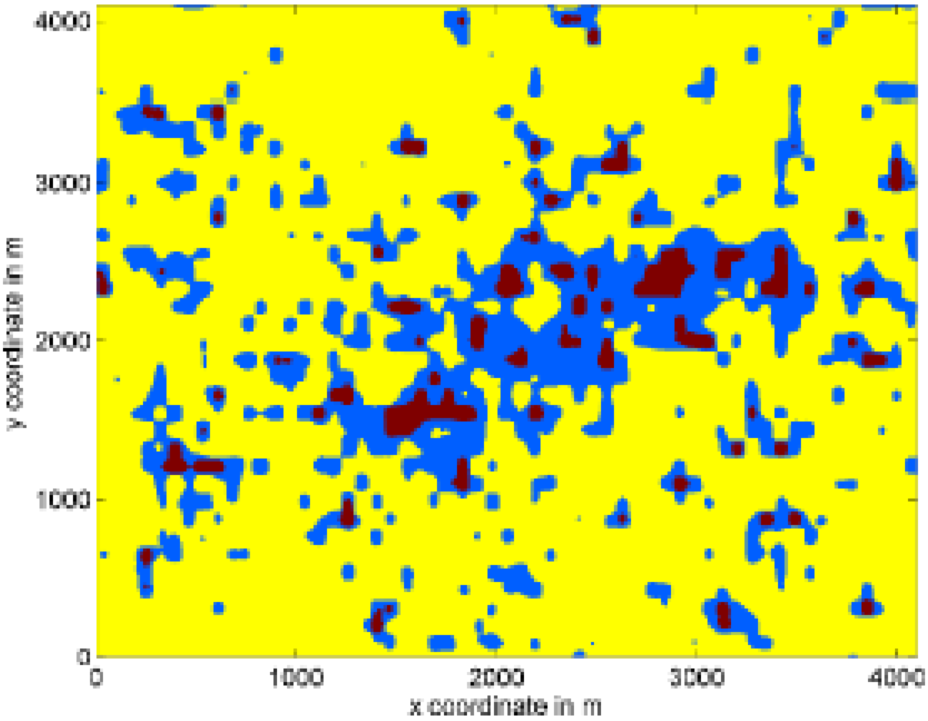


Figure 8 - Example of a "realistic" telecommunication traffic demand map based on 3GPP models[5]

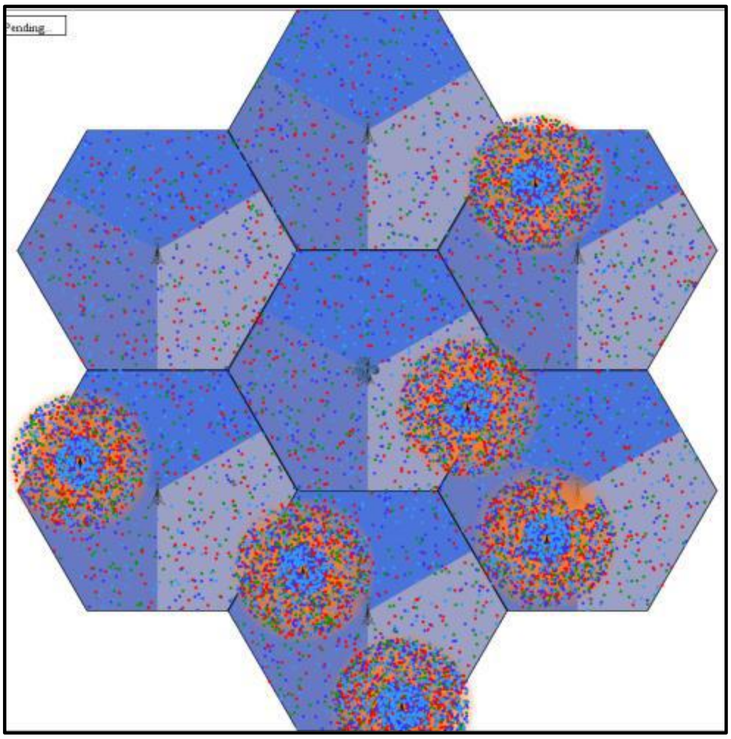


Figure 9 - Example of Topology generated by the simulator

In these examples (Figure 8 , Figure 9) We see how a realistic 2D city map can be mapped to a simulation scenario with mixed radio elements (blue are sector antenna and orange are Pico cell radiuses)

2.3.2 Physical Layer Module – Channel Modeling

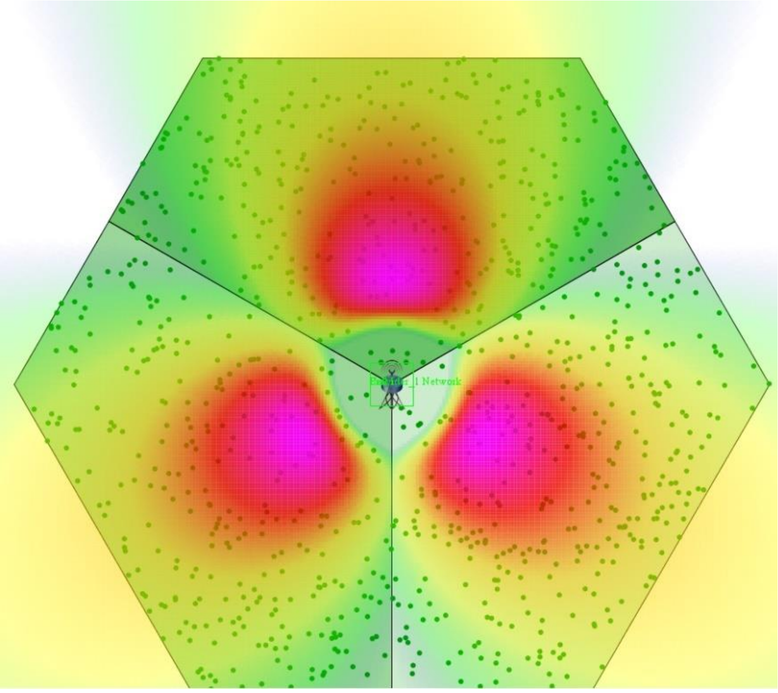


Figure 10 - Directional Gain of sector type antenna used for Macro eNodeB components

The radio channel transmission model performs accurate link budget calculations taking into consideration a fully capable 3-D antenna model. This includes a) Path-Loss model using multiple log-distance values for various environments based on 3GPP and other literature[5][6] (e.g. GreenTouch), b) Probabilistic Shadowing model to include the time-related occurrence of obstacles that remove the line of sight from the transmission, c) inclusion of additive / reductive interference from the surrounding radio transmission systems and multi-path reflections that result into the "fast-fading" phenomenon, d) models for transforming the radio quality into effective symbol and bit-rate at the data link layer, e) location-based absorption for various inddor environments (e.g. solid metal / concrete buildings that can lead to 20dB+ transmission losses for 2100Mhz transmission frequency)

Element Type	Pathloss Model
Macro eNodeB (Large range)	128.1 + 37.6log10(R), R km
Pico cell (Smaller range)	140.7 + 36.7 log10(R), R km

The simulator's graphical model can render for each different point of the playground (Figure 10, red to yellow colored interpolation of gain in dB) the result of the directional gain of the 3D antenna model.

2.3.3 Physical Layer Module – Probabilistic Shadowing

The shadowing module of the simulation engine[5][6] can greatly affect the outcome of the quality of service measurements for a simulation as it provides significant variability in the acquired throughput. A statistical model [6]is based on references that follows the log-normal distribution for shadowing coefficients that are combined with the geolocation of the user using the spatial correlation method. To achieve this the playground of the simulation is split into 50meter (adjustable) tiles that contain a random but unchanging shadowing value for any user equipment terminal inside. The terminals share the spatial shadowing value and therefore their radio environment becomes correlated (spatially). The lognormal distribution used for sampling the tile grid can take various parameter values based on the area type (i.e. DU, UR, SU, RU) and can range from 10 to 4 dB standard deviation and -5 to +5 dB mean value. The 2nd component of the shadowing model is linked to every single pair of UE – Network device and will be referred to as the "pair-wise" shadowing component. To produce higher shadowing diversity, each terminal has its own shadowing sampled value for each network element providing uniqueness in its radio environment and better formation of the statistical value of the empirical model. For the final shadowing value, a combination of the spatial and the pairwise component is used (divided by square root of 2). This value is then used for every new calculation of the radio quality and provides a solid basis for the shadowing component of the total link budget. The reduction of those dB's is a sum of the direct, refracted and reflected waves that can occur in this kind of radio environments. The pre-computed shadowing values are a very time-efficient technique that sacrifices very little accuracy in large scale system level simulations.

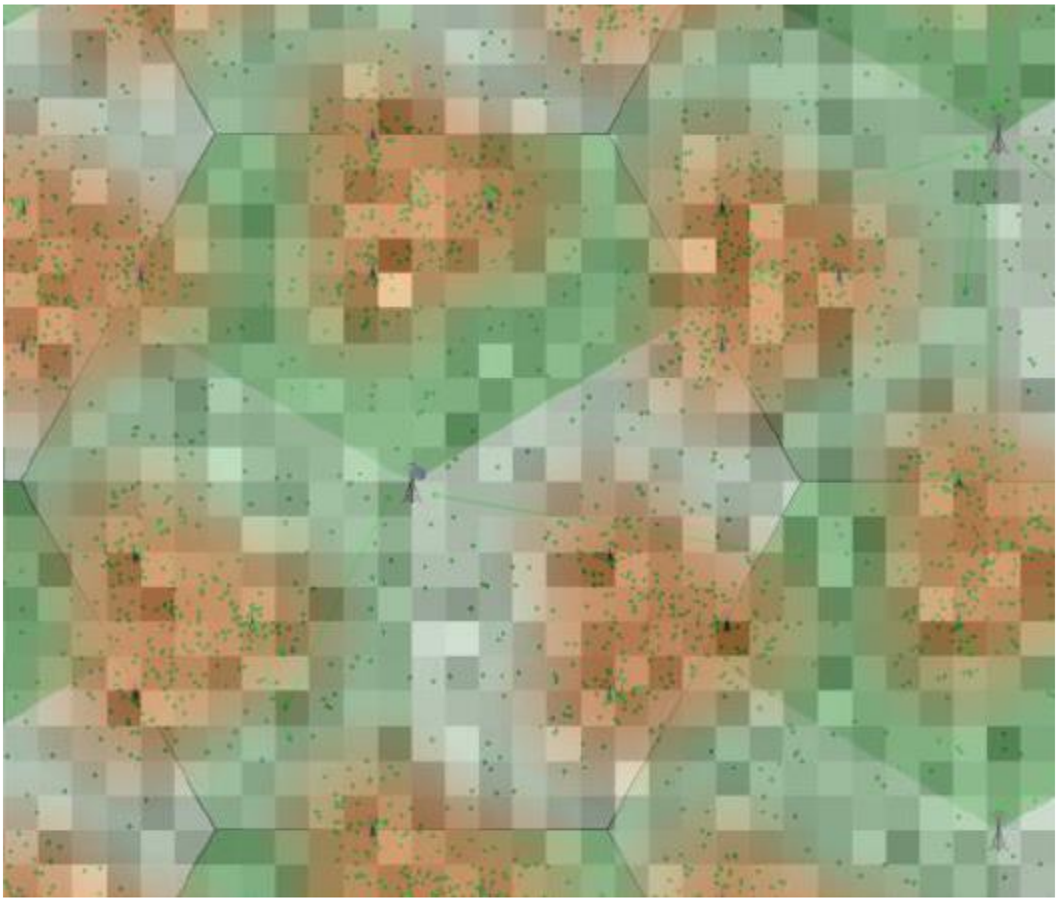


Figure 11 - Geospatial placement of the tile shadowing areas leads to accurate shadowing calculations that reproduces the real distributions

Simulator-generated visualization (Figure 11) can show in a greyscale projection the value of the shadowing that is computed for each tile of the playground.

2.3.4 Physical Layer Module – Channel Spectral Efficiency and Net Throughput

For the physical layer module that is responsible for converting the channel quality (i.e. RSSP, RSRP and INR/SIR/SINR ratios) we are utilizing a pre-computed mapping curve model[5][6]. This model included all the symbol level error and distortion effects as a function of the achieved SINR on the receiver and can also be parametrized by different MIMO or SISO configurations (including 1x1, 2x2, 4x2 and 8x2 stream configurations). The accurate interference and noise power calculations along with the link budget are being calculated into a total net throughput that includes coding losses, timing losses for signaling frames and losses based on fast fading. The spectral efficiency curve can also be visible in the following (Figure 12 – green for 2x2, blue for 4x2 and red for 8x2 mimo configurations)

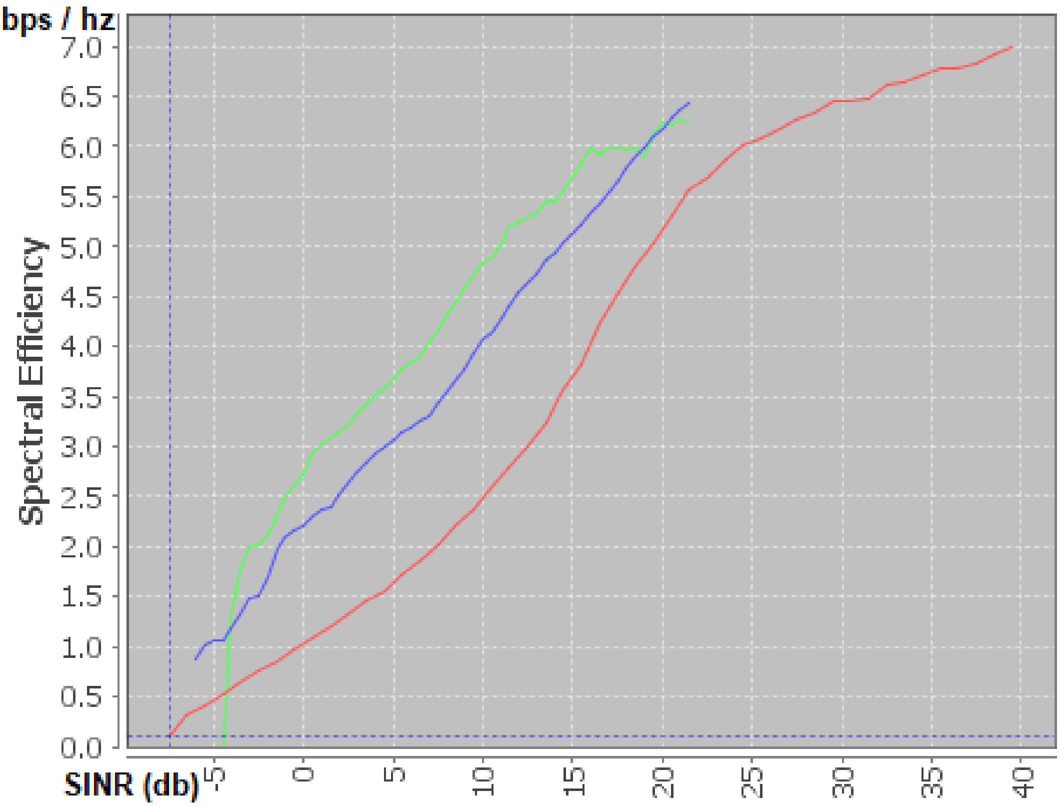


Figure 12 - Simulator's mapping curve spectral efficiency model showing the spectral efficiency as a function of SINR

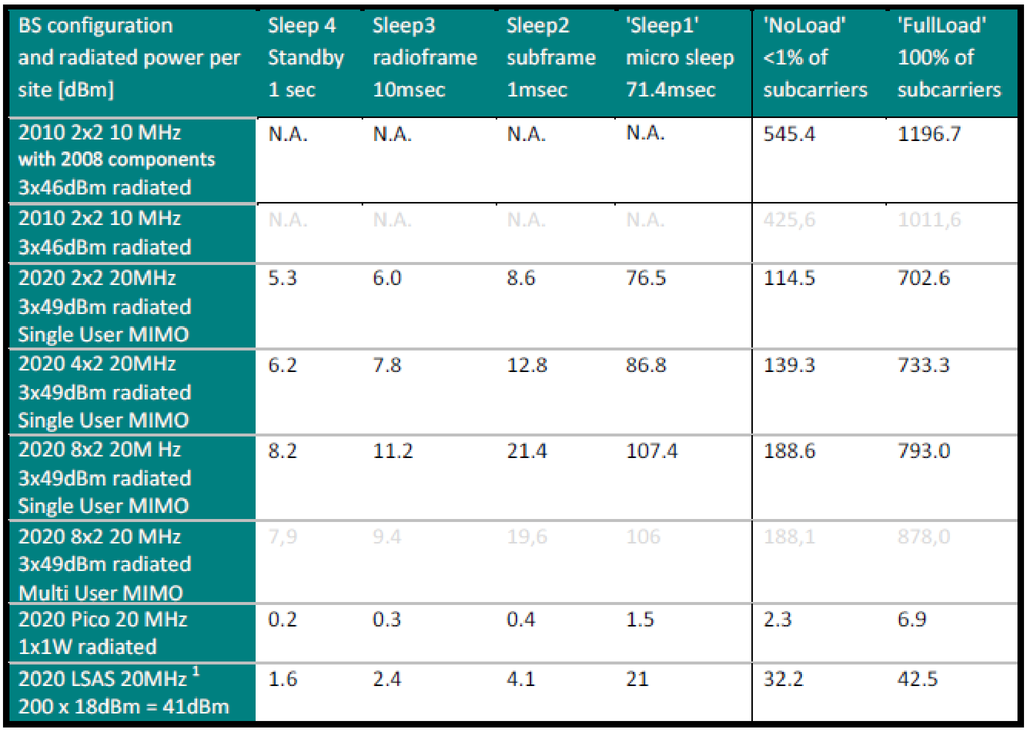
2.3.5 Physical Layer Module – Channel Interference Model

Channel interference [11]is one of the most important negative aspects of the cellular telecommunication channel. The continuous reuse of the same frequency (and time) resource results in different radio elements interfering with each other which contributes to the degradation of the link quality and can result in error, lower achieved throughput and /or outage. In the simulation software, we are approaching the interference phenomenon with two different models. First, we implemented an interference model that was based on a statistical estimation based on the neighboring cells load (usage). This resulted in accurate estimations on a macro scale (total amount of interference) but failed to capture the impact of interference in a link-to-link level of detail. Since this tool was designed to implement and test the impact of algorithms that can occur in the lowest granularity, such as radio resource allocation algorithms on the scheduler level of RRC, we redesigned the interference module in a different approach. The second implementation takes into consideration tree additional new parameters: time frames measured in air frames (0.5 – 1 ms), resource blocks measured in subcarriers per band and space (propagation of the interfering transmission). In this way we are continuously calculating with high accuracy (and complexity) the instantaneous interference that occurs in every user equipment terminal of the simulator that is actively operating. This captures a realistic value for the effective (achieved) SINR of the terminal that can be fed into the data link layer net throughput mapping curve mechanism described previously. Additional optimizations to the implementation were also included in the simulator in order to improve reduce the complexity of the interference calculation. The interference calculation frequency is adjustable, but the default value is set to be 100 ms (which is the interval of operation for every network device) and inside this interval we are assuming that the active load for each element is the average previous load. We are also using a bitwise transmit mask for every scheduled resource block (time and frequency entry) that can be orthogonal or overlapping.

2.3.6 Energy Module – Calculation of power consumption

The scope of this PhD dissertation is to measure important operational KPIs for various scenarios in order to optimize them. Such a KPI is the total power consumption of the substrate network which directly maps to the OPEX of the infrastructure and consequently to the EMF footprint of the hardware. The origins of the power consumption can be split into two different categories: a) power consumption that results from the cellular network's physical layer – data link layer (antennas and transmission components) and b) power consumptions from the back-end infrastructure including the processing servers and the supporting links (copper wire, microwave links or optical fiber). Studies performed[8][9][10] on these systems have shown that the back-end power consumption tends to remain stable and linearly linked to the number of elements included. On the other hand, the radio transmission related parts of power consumption can show very high variance based on multiple parameters of the air medium. Simulation models of the literature, real data acquired from existing infrastructure and theoretical energy models can lead to very accurate calculation models to be incorporated into the software platform. The power consumption model is a linear model with a fixed offset and limit. The dependent variable of the model is the element's load (i.e. the amount of air frames within a specific time window that were used for transmission). In 4th generation LTE, this load is often measured in Resource Blocks (temporal and frequency pairs). The final form of the calculation formula is Minimum Energy Consumption + 100% * Power Consumption Margin. Quantitively the minimum to maximum consumption ratio have a factor of between 2 and 5 in difference. This drastically impacts the active power consumption of LTE simulations based on the radio transmission state. Impacting the radio transmission circumstances (e.g. improving any aspect that un-loads the cell whilst maintaining the same effective bandwidth results in high benefits for energy consumption). Another aspect that influences the power consumption of simulations is the usage of multiple antennas (input and output) to utilize MIMO configurations. Finally, some power consumption models use reference hardware that have different shutdown policies when they are being inactive or move to a sleep state. All these can be seen in the considered power values below.

Table 2 - Linear Power Consumption Model Coefficients [7][8]



BS configuration and radiated power per site [dBm]	Sleep 4 Standby 1 sec	Sleep3 radioframe 10msec	Sleep2 subframe 1msec	'Sleep1' micro sleep 71.4msec	'NoLoad' <1% of subcarriers	'FullLoad' 100% of subcarriers
2010 2x2 10 MHz with 2008 components 3x46dBm radiated	N.A.	N.A.	N.A.	N.A.	545.4	1196.7
2010 2x2 10 MHz 3x46dBm radiated	N.A.	N.A.	N.A.	N.A.	425,6	1011,6
2020 2x2 20MHz 3x49dBm radiated Single User MIMO	5.3	6.0	8.6	76.5	114.5	702.6
2020 4x2 20MHz 3x49dBm radiated Single User MIMO	6.2	7.8	12.8	86.8	139.3	733.3
2020 8x2 20M Hz 3x49dBm radiated Single User MIMO	8.2	11.2	21.4	107.4	188.6	793.0
2020 8x2 20 MHz 3x49dBm radiated Multi User MIMO	7,9	9.4	19,6	106	188,1	878,0
2020 Pico 20 MHz 1x1W radiated	0.2	0.3	0.4	1.5	2.3	6.9
2020 LSAS 20MHz [1] 200 x 18dBm = 41dBm	1.6	2.4	4.1	21	32.2	42.5

2.3.7 RRC Layer Module – Handover Models

An important feature of cellular networks is the capability to provide wireless access in moving (mobile) users in the served area. This is achieved by the involvement of signaling procedures[11] that perform the service migration of one eNodeB to another, also named as the "handover" procedure. Because of the importance of this procedure and also because of the high impact it has to the HetNet systems, a handover-specific module has been developed that will be active in all simulations. The basic operational principle behind the handover algorithm is the provisioning of measurements from the LTE cell and the UE device in order to identify the best targets as handover candidates. This function is repeated for as long as the user is active (i.e. in the RRC Connected state). The active handover algorithm will trigger a handover operation that will change the user's active cell to the best candidate based on the received signal strength indicator (RSSI). This procedure is implemented in the system using two different approaches (as found in the literature[11]). The first handover algorithm implementation is called the "Threshold-based handover". In this implementation, a handover threshold value is used as a "trigger" for the best cell selection. If the active cell's RSSI is below the designated value, the user will then change cell to the best possible. This methodology can generally control the effective range in which each cell will operate and absorb traffic, however it requires different parametrization for different area types since other ISDs will result into different value ranges for the received signal strength (especially in the cases of SU and RU area types. For this purpose, we have also implemented the second case of handover algorithm, namely the "Hysteresis-based handover".

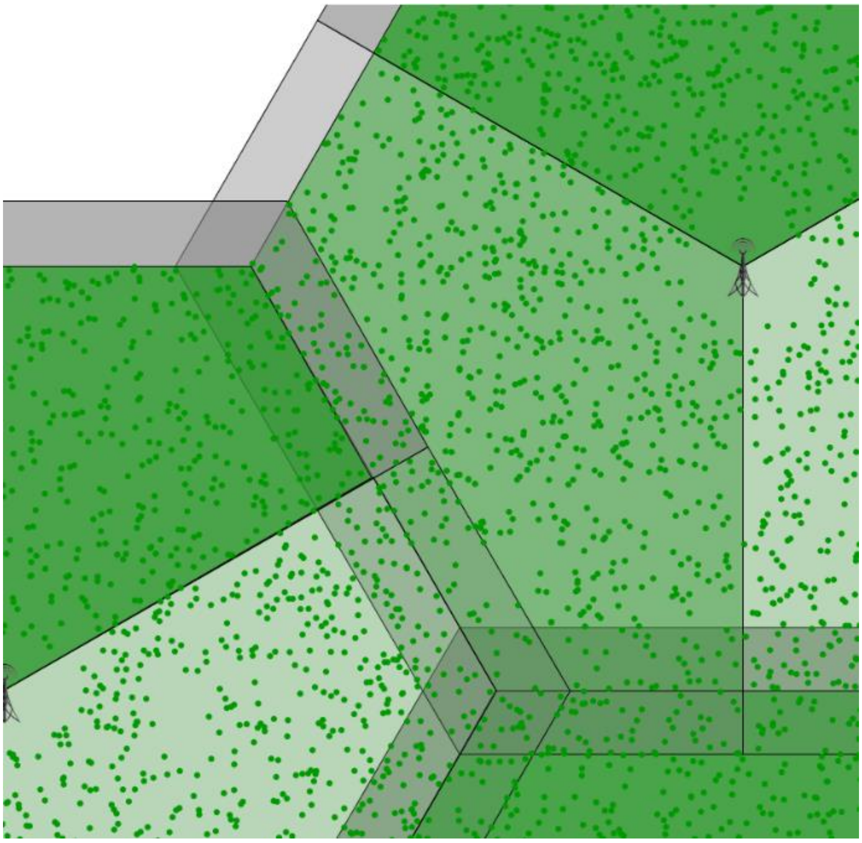


Figure 13 - Hysteresis additional RSSI margin

Hysteresis-based handover compares all the available handover target's RSSI with the current (serving) cells value. If their difference (hence "hysteresis") is greater than the hysteresis configuration value, then a handover event is being triggered. This approach is self-tuned since the difference of the values is being used. An additional part of the handover module is the inclusion of the cell bias or CIO configuration parameter. This scalar configuration is used as a virtual RSSI gain for a specific cell or cell-UE pair. Adding an imaginary value into the RSSI measurement feedback, the network's handover algorithm is being "manipulated" in order to achieve manually triggered handovers and manage the network in a more controlled way. The simulator has a special visualization module (Figure 13) to display the active hysteresis values for each cell.

2.3.8 RRC Layer Module – Radio Resource Allocation & Scheduler Models

LTE networks have a built-in notion of quality of service for each of their active (serving) users. The reason behind this is that this technology has been developed as a commercial technology that will be used in a paid access manner. For this reason, the scheduling and radio resource allocation layer needs to be a fully controlled and extensible environment in which different policies can be applied. In other networks, this layer can sometimes operate in either best effort or take decisions based solely on the physical layer’s restrictions (e.g. Wi-Fi). The radio resource allocation module in LTE networks is governed by the RRC layer which dictates how the lower level protocols (i.e. RLC) will handle the pending traffic. The quality of service is provided by classification of the generated traffic into several classes, example classes are “default”, “high-priority”, “low-latency” etc. based on the specific quality of service characteristics they have. The scheduler module is a submodule of this system which handles the placement of the requests to transmit into the time queue. Users can either receive their transmissions simultaneously (split into different resource blocks) or they can have a round-robin access into the downlink / uplink channel. For the various classes mentioned previously, different policies need to be followed in order to ensure successful transmission.

2.3.9 Application Layer Module – Traffic Demand Model

For the telecommunication traffic demand module, we have implemented a programmable downlink / uplink traffic generation scheme which we can create various different traffic profiles that correspond to different inter-packet arrival times, packet size variations (constant vs distribution based vs specific sequences for protocols such as TCP / RTSP) and also different expectation and/or timeout values for the application layer (which results into delivery failures). Reference scenarios[12][13][14] set an average per Km or per user equipment device rate and we can easily translate this specification to values of such a model. In the lifecycle of a simulation, every user equipment device generates a transmission job of the selected model by sampling a Poisson generation function. The requests are sent into the buffer of the system and translate to data being transmitted through the air interface. In the standardization literature, we see a simplified version of FTP transfer model being used thoroughly for various simulations. This traffic model is using two different types of packets generated with different probabilities:96% of the packets generated correspond to signaling of web site micro-transactions (ajax requests, REST , JSON, XML payloads of http messages) of an average size 10KB and the rest 4% of packets correspond to the initial load of the web site resulting in an average ~2MB size packets. This is based on statistical study on a very large number of websites and mobile applications used currently and accessed through mobile internet. We can refer to this traffic model as "2020 WWW-FTP model"[7][8] which will be used in the following chapters as a reference simulation traffic model. As a quality of service parameter, the expected delays for each of the two types of packets are 4000ms (for the initial, large packet) and 20 ms (for the smaller signaling packets) respectively. In the diagram below we can see a time diagram of packet arrivals and their transport delay as a function of time.

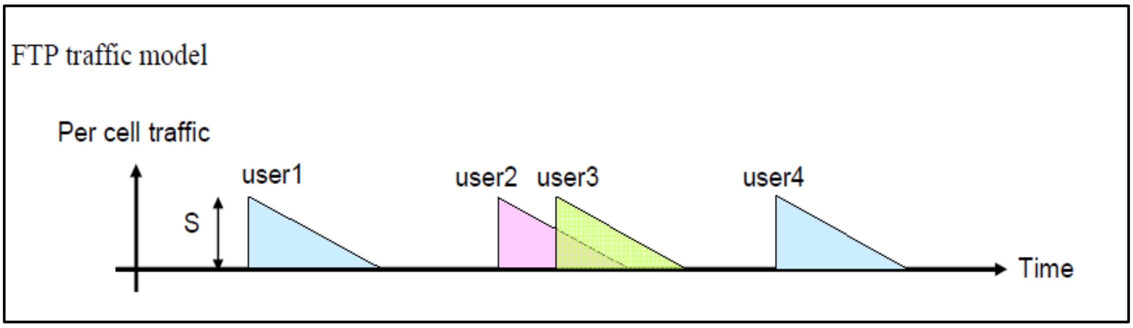
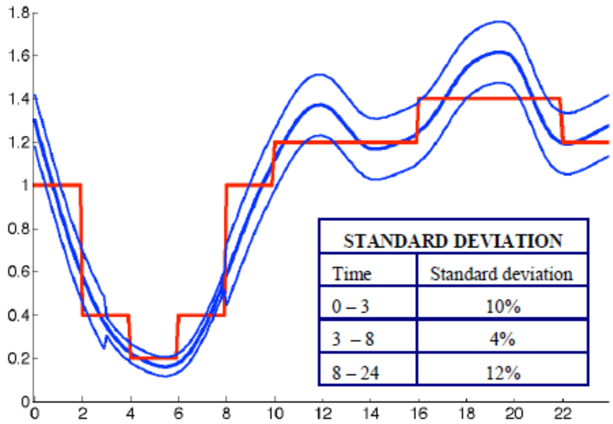


Figure 14 - Traffic Model Implementation in the Simulation Software[6]

2.3.10 Application Layer Module – Hourly Traffic Demand Profiles

The usage of a 4th generation mobile terminal device is influenced by various aspects of the user's everyday life cycle. This daily / weekly cycle changes the way they use their devices and the amount of traffic they generate. This can act as its own separate research topic, however for the scope of this study we will focus only on a set of aspects that can be easily incorporated and integrated into the simulation tool. A time-related study[6] of daily profiles for different area types (DU, UR, SU, RU) show (Figure 15 – one line per case that shows the multiplication weight per hour) that there is a coherent daily traffic profile that acts as a "weight" for the usage of mobile internet. We have included an extension in the traffic model of the simulation software that will translate the current (active) simulation time into such a weight based on the simulation area, therefore enforcing the demand curve of the external reference measurements.



STANDARD DEVIATION	
Time	Standard deviation
0 – 3	10%
3 – 8	4%
8 – 24	12%

TRAFFIC VALUES	
Time	Value
0 – 2	100%
2 – 4	40%
4 – 6	20%
6 – 8	40%
8 – 10	100%
10 – 16	120%
16 – 22	140%
22 – 24	120%

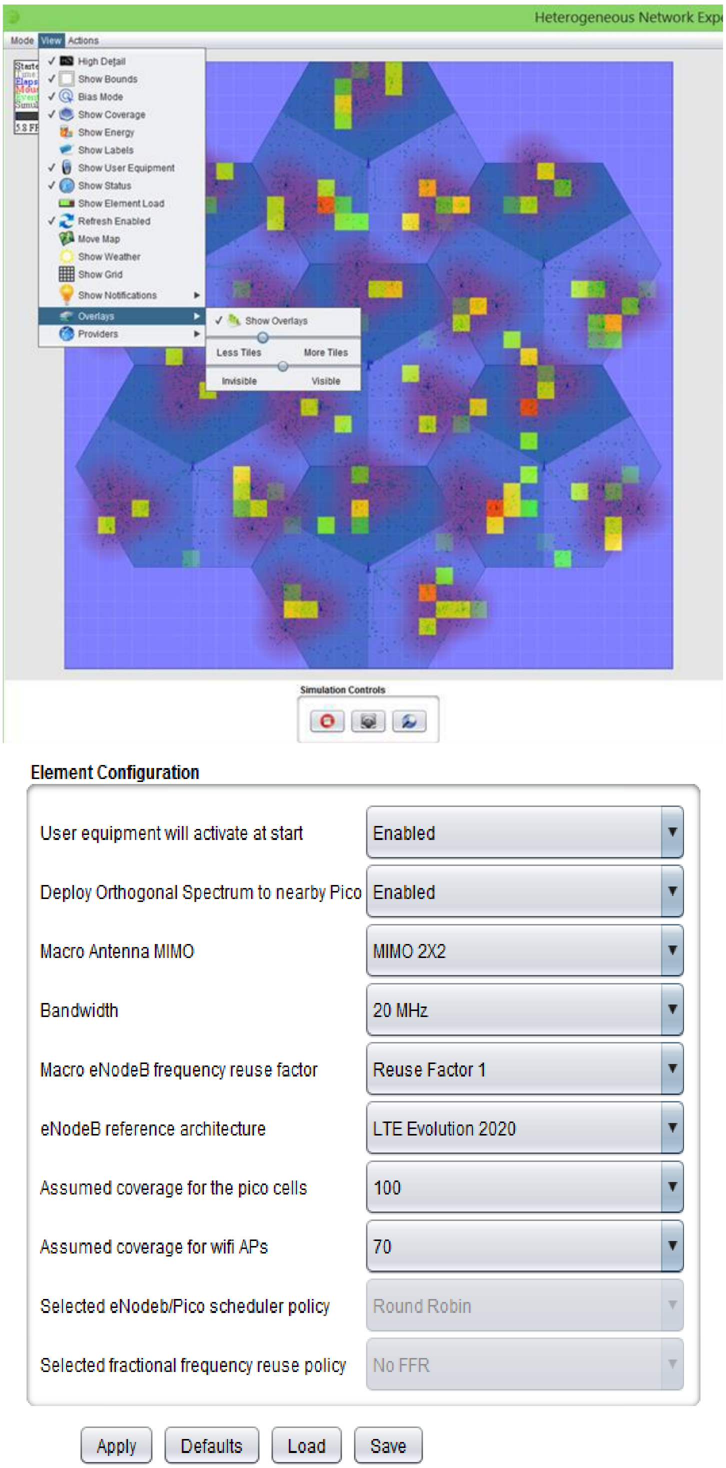
Figure 15 - Daily Traffic Demand Profile for a 3GPP-based cellular network[6]

2.3.11 Network Management Module – Application of optimization actions

Apart from the regular operation of the cellular network, various network management interfaces must be utilized in order to optimize / tune or deteriorate the simulation's performance KPIs. The developed libraries of the simulator allow for a rich parametrization the UE devices and all the types of the heterogeneous network elements with either a provided configuration file in .JSON format or by utilizing the GUI (developed in Java Swing) in order to edit settings of the various elements. In order to develop algorithms that will be activated in the runtime and act as management agents, appropriate programming styles have been used based on other simulation software (e.g. NS2 / NS3 / OpNet). In the specifications of a simulation scenario, a user can activate a management action or schedule it to occur after an event or after a specific point in time. In this way, the various proposed algorithms will produce usable and reproducible measurements that will not require the user's interaction from the GUI.

2.4 Graphical User Interface

The design of the System level simulator for this study includes a graphical user interface to give the user the capability to adjust a simulation and monitor its runtime (before performing the simulations in an exhaustive and automated manner). The programming language used for this GUI is the Java Programming Language (Simple Edition version 8). The design of the UI is maximizing the user's visible information and the controls that can affect the simulation.



Element Configuration

User equipment will activate at start	Enabled
Deploy Orthogonal Spectrum to nearby Pico	Enabled
Macro Antenna MIMO	MIMO 2X2
Bandwidth	20 MHz
Macro eNodeB frequency reuse factor	Reuse Factor 1
eNodeB reference architecture	LTE Evolution 2020
Assumed coverage for the pico cells	100
Assumed coverage for wifi APs	70
Selected eNodeb/Pico scheduler policy	Round Robin
Selected fractional frequency reuse policy	No FFR

Figure 16 - Indicative View of the tool's graphical user interface

2.4.1 Network Element Resource Allocation Visualization

A specialized visualization module was designed to allow the user to monitor each substrate element's active radio resource allocation. Different allocation schemes[17][18] are assigned a different background color to be used as a replacement color for the network element's radius. The function for the generation of the colors is a hashing algorithm applied on the available frequencies and the resulting hash value is then used to perform an HSB color conversion (Hue-Saturation-Brightness color model).

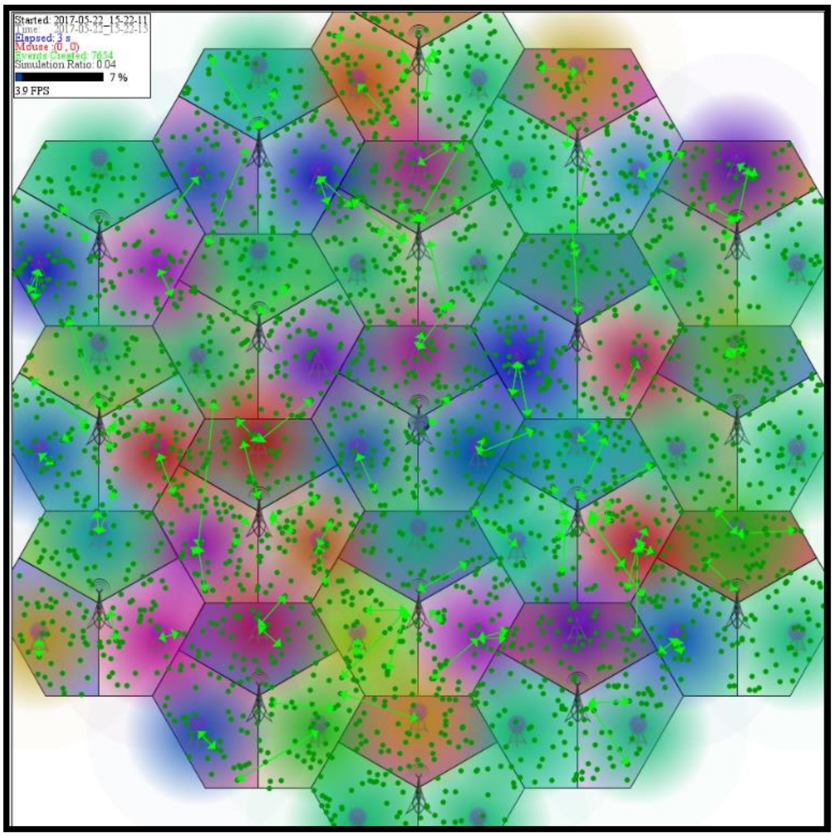
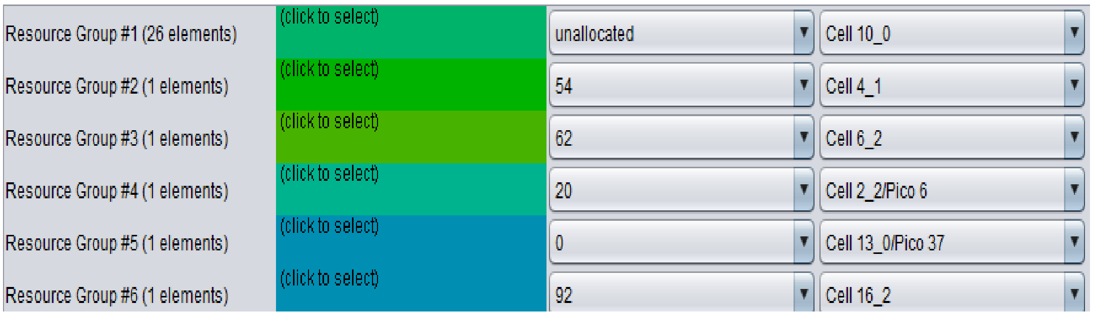


Figure 17 - Visualization of radio resource allocation using colored radius

During the simulation's runtime, the initial allocation is immediately switched to the designated active allocation according to the radio resource management implementation and continuous to change if necessary (e.g. if dynamic resource allocation SON is activated).

An additional visualization capability for the active allocation is a customized matrix view which shows the per element resource block allocation using the same color coding as the playground background colors. The horizontal axis of the visualization is the element id and the vertical axis is occupied if the resource block is allocated to the specific element at the present (simulated) time.



Resource Group #1 (26 elements)	(click to select)	unallocated ▼	Cell 10_0 ▼
Resource Group #2 (1 elements)	(click to select)	54 ▼	Cell 4_1 ▼
Resource Group #3 (1 elements)	(click to select)	62 ▼	Cell 6_2 ▼
Resource Group #4 (1 elements)	(click to select)	20 ▼	Cell 2_2/Pico 6 ▼
Resource Group #5 (1 elements)	(click to select)	0 ▼	Cell 13_0/Pico 37 ▼
Resource Group #6 (1 elements)	(click to select)	92 ▼	Cell 16_2 ▼

Figure 18 - GUI for resource allocation

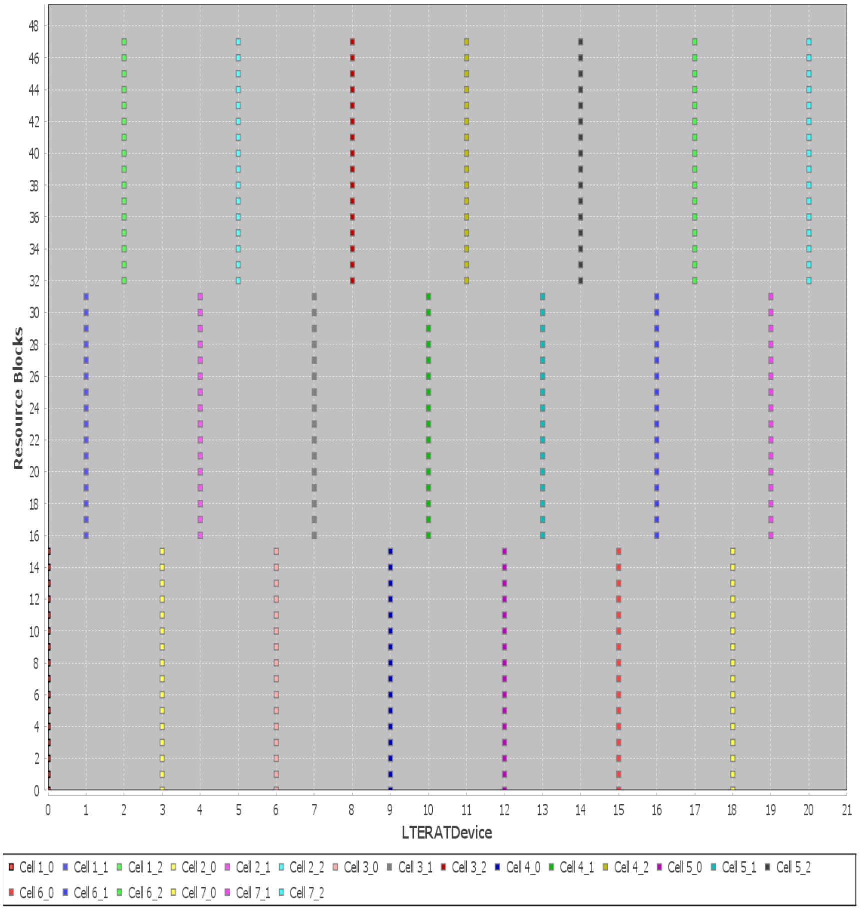


Figure 19 - Radio Resource Group visualization (reuse factor 3)

These capabilities are a key aspect of the simulation tool in a sense that they allow the safe development (with rich graphical environment that allows for easy debugging) of radio resource allocation algorithms. The best performing radio resource allocation algorithms can effectively extend the 4G standard's capabilities and improve this technology.

2.5 Validation of the Simulation Software

In this section we will provide the methodology that was followed to validate the accuracy of the simulation software along with a set of KPIs / measurements that will be used to compare with other referenced 3GPP-based simulators for the predefined reference scenarios of the various technical reports / specifications of IETF/3GPP

2.5.1 Validation Methodology

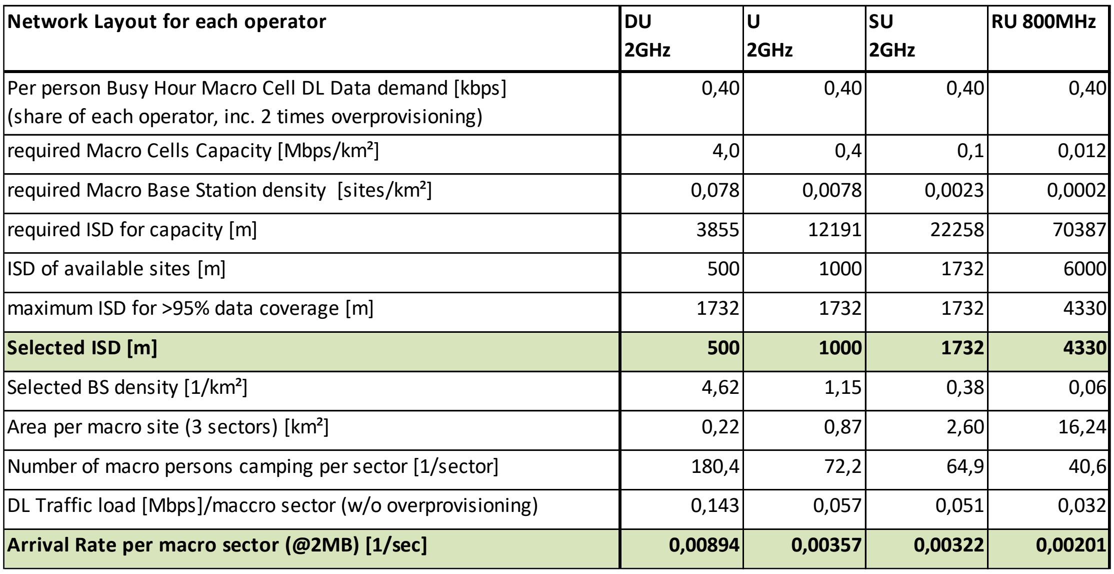
The most important part of developing a simulation software that will replicate real environment conditions is the simulation result validation. In order to ensure that the tool reproduces trustworthy results we first must isolate some key scenarios that will be used to perform the calibration measurements. Thankfully, 3GPP has defined a set of standardized reference scenarios to be used for both mature (reference to year 2010) and future (projections to year 2020) with enough geographical diversity (all different designated area types namely dense urban, urban, sub-urban, rural). The next step is to select the proper network KPIs that will be used to perform the value comparison. According to methodology defined[4][5][6] in the GreenTouch Consortium and other 3GPP partners, the metrics will be converted into probability density function and their integral – the cumulative distribution function. The different CDFs will then be passed through a statistical significance calculation function that will show the statistical likelihood between the results generated from this software and other existing simulation software from other organizations (e.g. ALUD, Orange, POLIMI).

2.5.2 Simulation Calibration KPIs

The KPIs that will be measured in the various simulations are basically three and are split into different categories (different underlying modules of the simulator). Firstly, the Coupling Loss CDF is the distribution of all the propagation losses for each UE device in the simulation. This targets the mobility and geometry module of the simulation as it is responsible for the 3D- distance calculations, the propagation loss, and the motion of the UE devices. It also checks the shadowing and absorption modules (mentioned in previous chapters) to ensure that the simulation tool has identified correctly the users with "good" and the users with "bad" radio environments. Second KPI is the SINR CDF, it is based on the coupling CDF but it also incorporates the noise calculation module (based on the reference technology of the receiver) and the interference calculation module (which in itself uses its own coupling measurements). Another key aspect of the SINR calculation is the addition of the generated traffic to the users. In order to achieve high accuracy in interference we need to replicate the same amount of traffic that will result into neighboring interference conditions. SINR is a key measurement for the physical layer of the software as it reflects the end quality of propagation that the link will have to translate into effective throughput. The last calibration KPI is the normalized achieved throughput CDF. This statistical measurement is the final QoS measurement of the total simulation scenario and it provides accuracy in the link level (layer 2) of the system. Based on the SINR-to-Throughput mapping curve and the implementation of the radio scheduler of the system, the achieved net throughput is used to serve the underlying telecommunication traffic and therefore provides a safe basis to build more calculations (of application layer and other higher-level protocols).

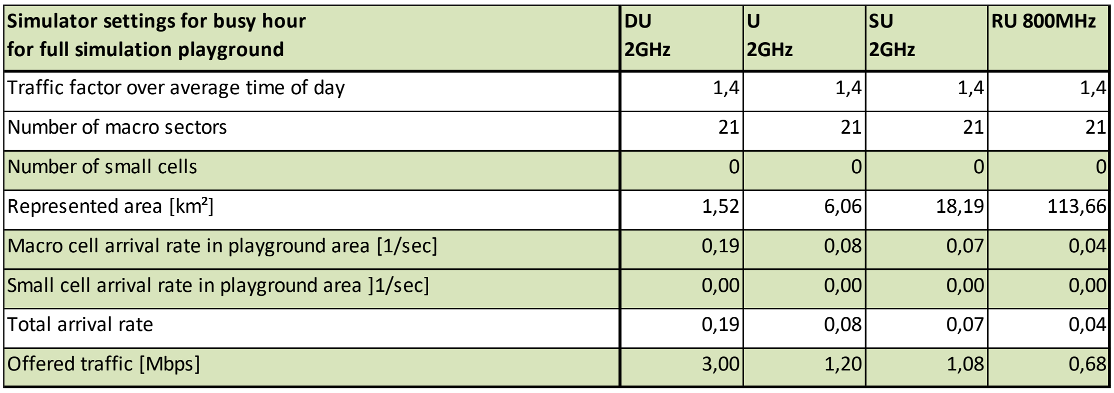
2.5.3 Simulation Scenario Parameters

The simulation parameters that are used for the calibration process are split into two categories: network layout and number of elements such as inter-site-distance an and traffic model parameters such as packet arrival rate, avg packet size and. These both contribute in different aspects to the generated results of the simulation as they change the operation of various functional modules from propagation to network operation.



Network Layout for each operator	**DU 2GHz**	**U 2GHz**	**SU 2GHz**	**RU 800MHz**
Per person Busy Hour Macro Cell DL Data demand [kbps] (share of each operator, inc. 2 times overprovisioning)	0,40	0,40	0,40	0,40
required Macro Cells Capacity [Mbps/km²]	4,0	0,4	0,1	0,012
required Macro Base Station density [sites/km²]	0,078	0,0078	0,0023	0,0002
required ISD for capacity [m]	3855	12191	22258	70387
ISD of available sites [m]	500	1000	1732	6000
maximum ISD for >95% data coverage [m]	1732	1732	1732	4330
Selected ISD [m]	**500**	**1000**	**1732**	**4330**
Selected BS density [1/km²]	4,62	1,15	0,38	0,06
Area per macro site (3 sectors) [km²]	0,22	0,87	2,60	16,24
Number of macro persons camping per sector [1/sector]	180,4	72,2	64,9	40,6
DL Traffic load [Mbps]/maccro sector (w/o overprovisioning)	0,143	0,057	0,051	0,032
Arrival Rate per macro sector (@2MB) [1/sec]	**0,00894**	**0,00357**	**0,00322**	**0,00201**

Figure 20 - Network Layout parameters for each different area type



Simulator settings for busy hour for full simulation playground	DU 2GHz	U 2GHz	SU 2GHz	RU 800MHz
Traffic factor over average time of day	1,4	1,4	1,4	1,4
Number of macro sectors	21	21	21	21
Number of small cells	0	0	0	0
Represented area [km²]	1,52	6,06	18,19	113,66
Macro cell arrival rate in playground area [1/sec]	0,19	0,08	0,07	0,04
Small cell arrival rate in playground area]1/sec]	0,00	0,00	0,00	0,00
Total arrival rate	0,19	0,08	0,07	0,04
Offered traffic [Mbps]	3,00	1,20	1,08	0,68

Figure 21 - Simulator settings for full simulation (busy hour) per area type

2.5.4 Dense Urban Scenario Results

For the evaluation of the Dense Urban scenario featuring hexacombs of 500m inter-site distance and all DU-related propagation parameters (absorption, shadowing, population density and traffic model) it is evident from the coupling loss that the differences between each tool are small close to +-0.5db. In the case of the distribution of SINR for the playground, the tool products a curve that has slight difference with the rest of the tools at the [-3,1]db range and also close to the [19,…] where all tools seem to deviate with different behaviors. The mean normalized downlink throughput deviates for at least 3% at the ranges 0-0.3 and 1-1.5.

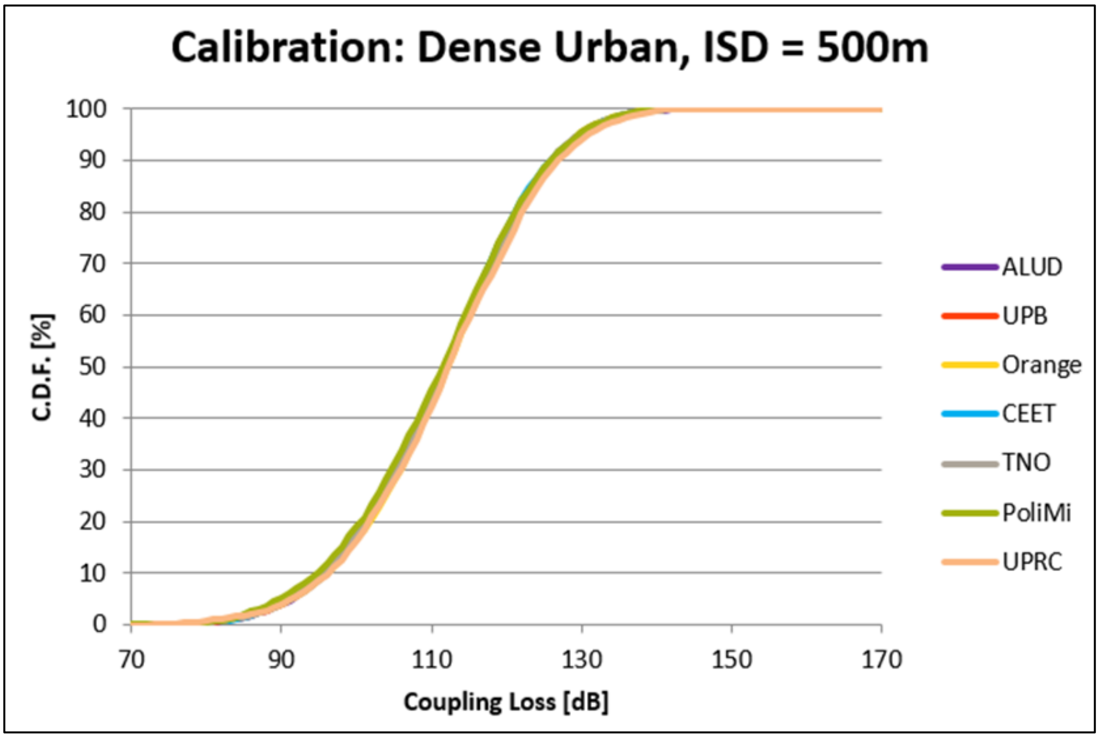


Figure 22 – DU Calibration Results of the simulation software for coupling loss

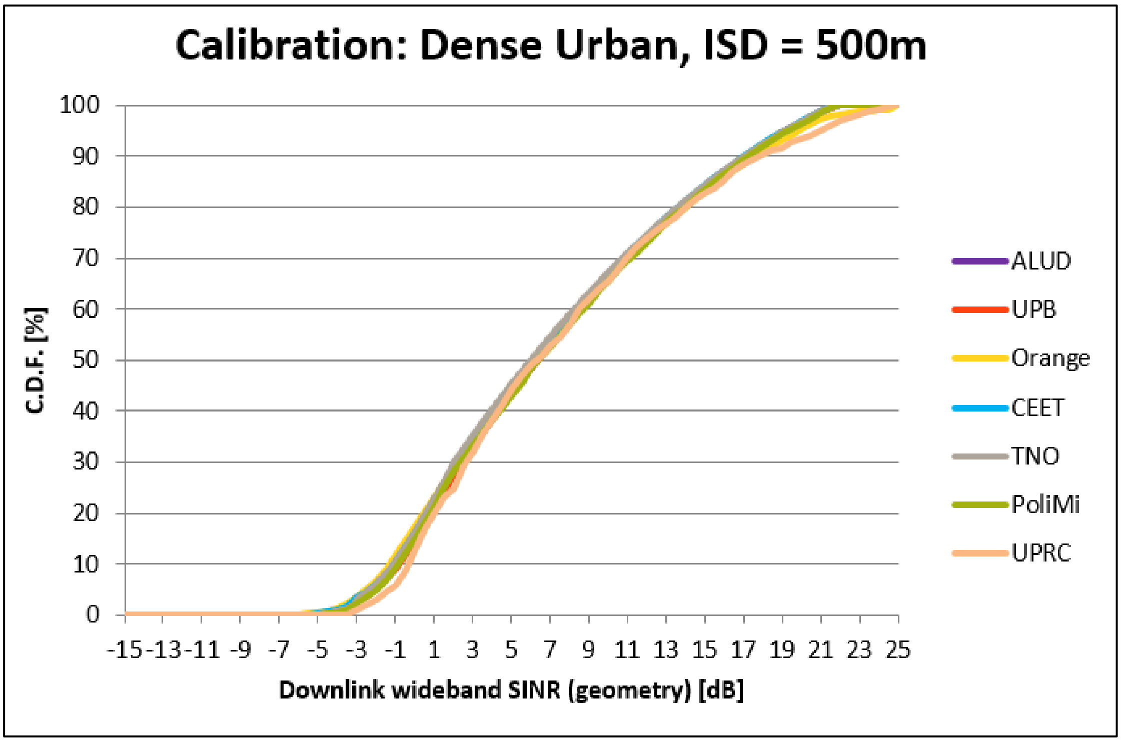


Figure 23 – DU Calibration results of the simulation software for SINR

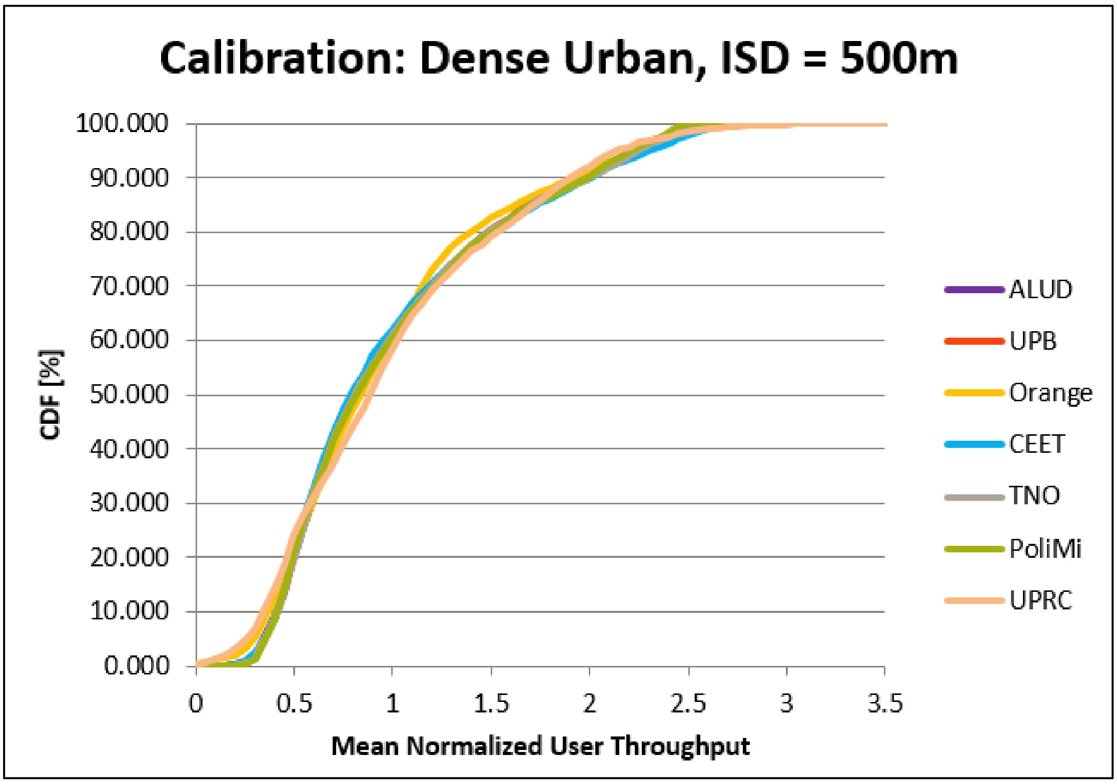


Figure 24 – DU Calibration results of the simulation software for average normalized throughput

2.5.5 Urban Scenario Results

In the urban case featuring hexacombs with 1000m inter-site distance we also see a very close distribution for all the simulation software at the coupling loss kpi. The distribution of noise and interference is showing an increased range of ~1.5 db with highest values found at the Orange simulation software and the lowest values for the PoliMI and UPRC simulator. Finally, for the Mean normalized user throughput we see for the same ranges as the DU case (0-0.3 and 1-1.5) an identical deviation of ~3%.

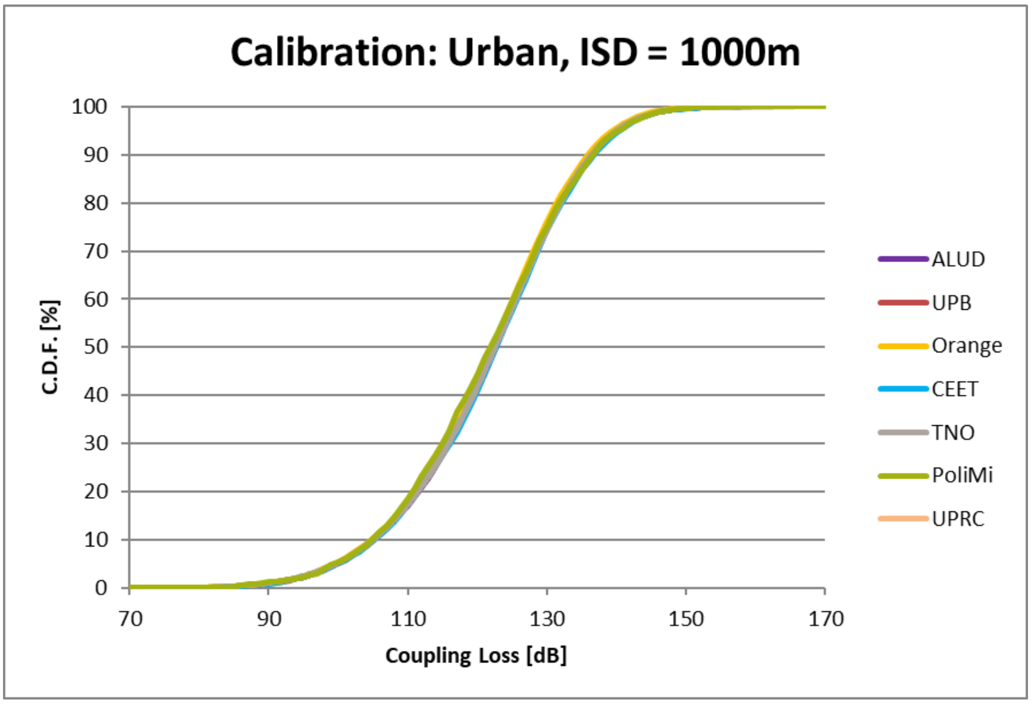


Figure 25 – UR Calibration results of the simulation software for coupling loss

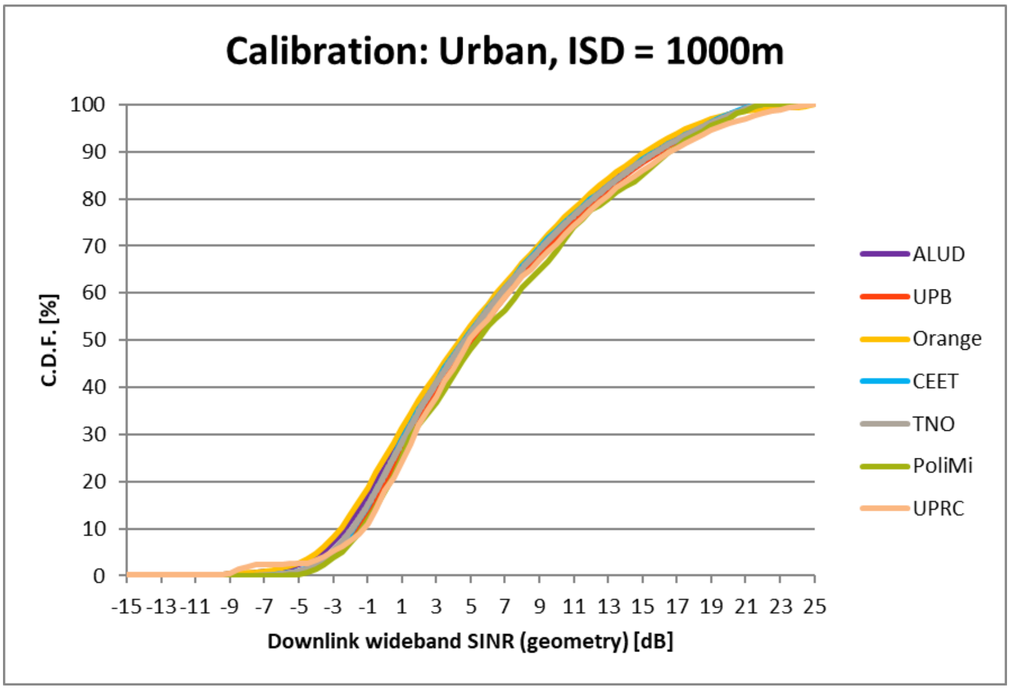


Figure 26 – UR Calibration results of the simulation software for SINR

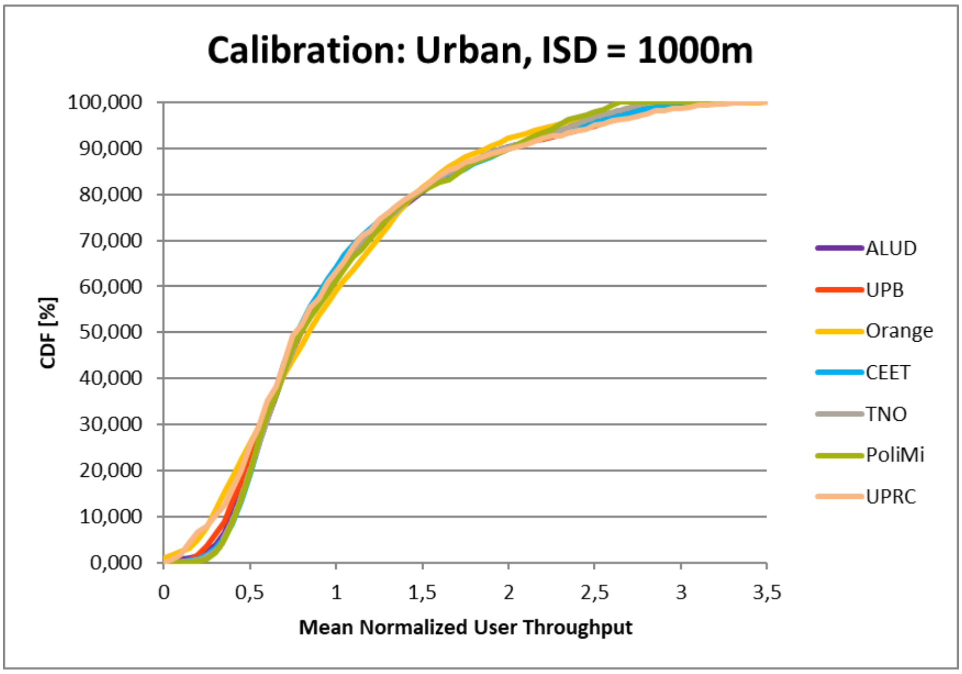


Figure 27 - UR Calibration results of the simulation software for average normalized throughput

2.5.6 Sub-Urban Scenario Results

In the sub-urban case featuring hexacombs with 1732m inter-site distance we see a slight deviation on the coupling loss curve at the 120db losses from the PoLIMI simulator. For the interference distribution measurements, we see that the margin of decline increased to ~2dB and we see distribution variations from the Orance, CEET and UPRC tools on various db sections. Finally, for the normalized throughput curve we see the same pattern that applies for the DU and UR case.

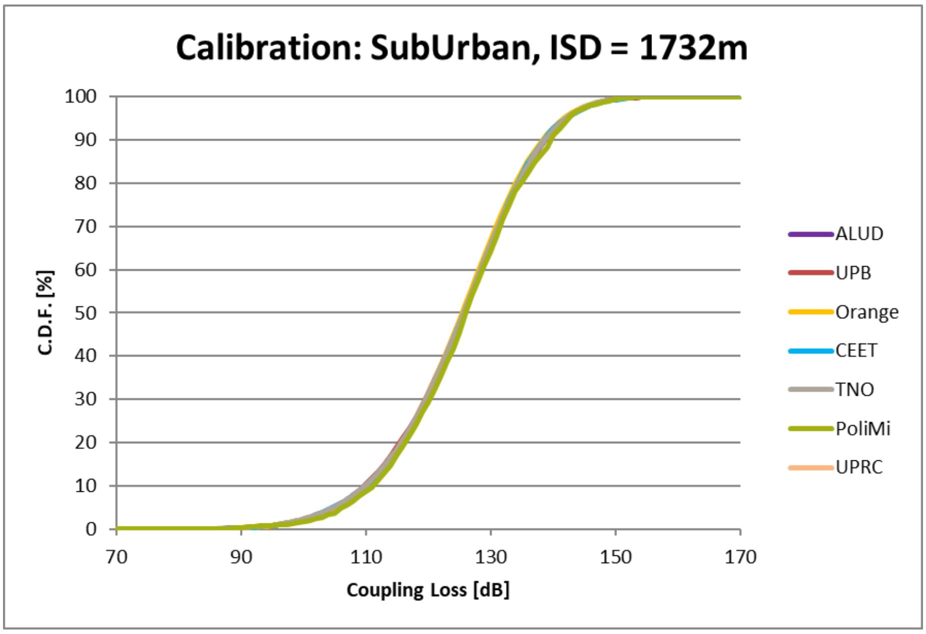


Figure 28 - SU Calibration results of the simulation software for coupling loss

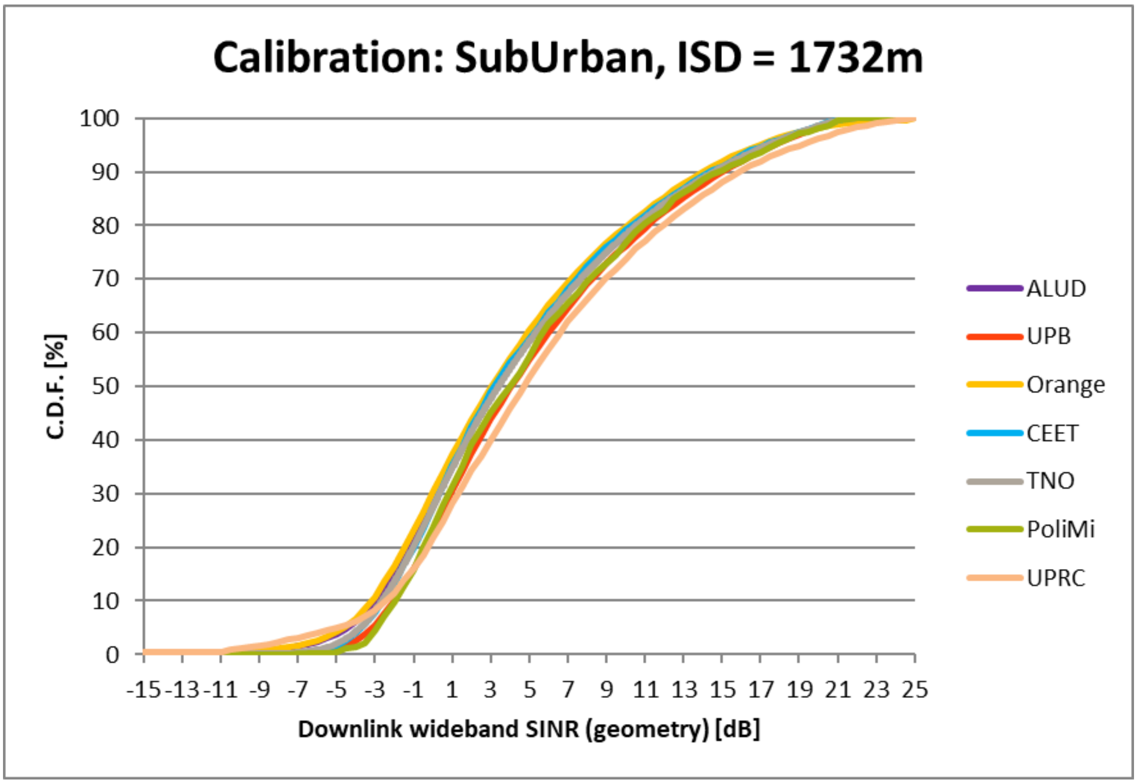


Figure 29 - SU Calibration results of the simulation software for SINR

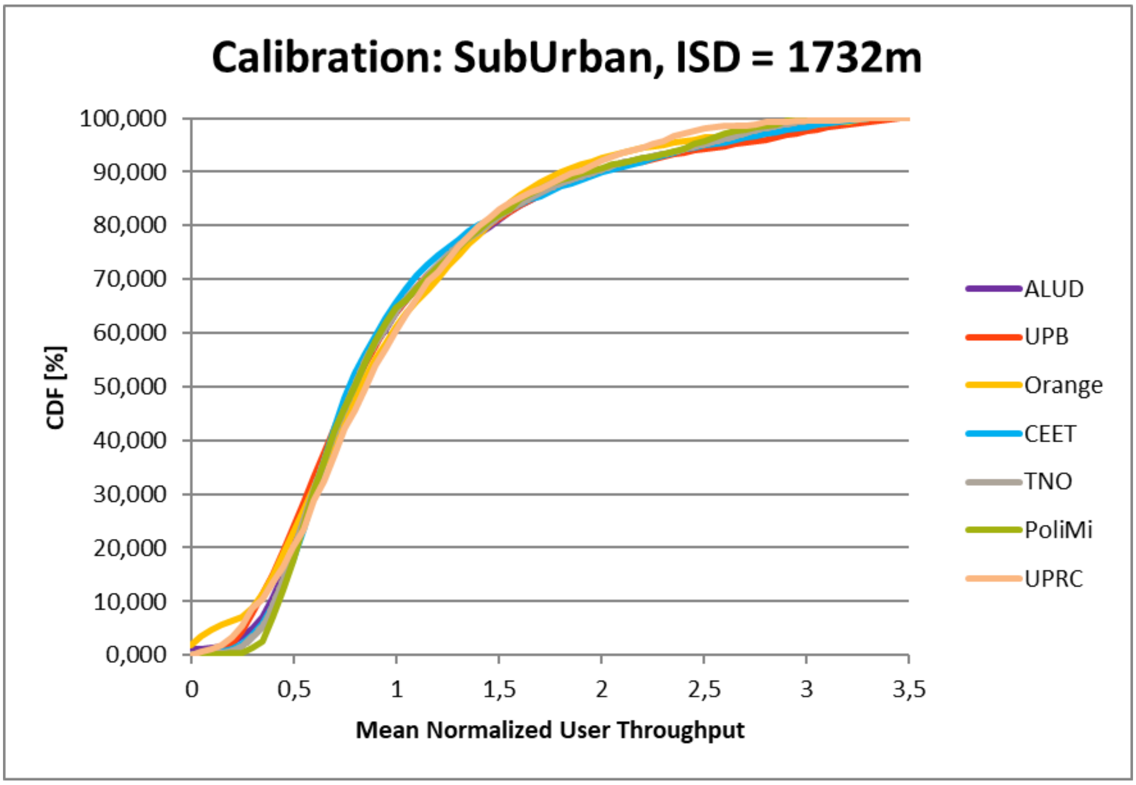


Figure 30 - SU Calibration results of the simulation software for average normalized throughput

2.5.7 Rural Scenario Results

The final area type that is used for the calibration is the largest when it comes to coverage area. Rural has an inter-site distance of 4.33 km covering a playground of over 15x15km hexacomb. Slight variations again in the coupling loss of low significance for all tools provides solid verification for the propagation model. In the zone with high interference, (low SINR) we see that the UPRC tool deviates for approximately 8% until the -3 db point. This can be due to the low number of samples generated (because of the low population density of the rural area type). In addition, the normalized throughput curve deviates for 2-3% in a number of different points without causing any concerns for simulation result faults.

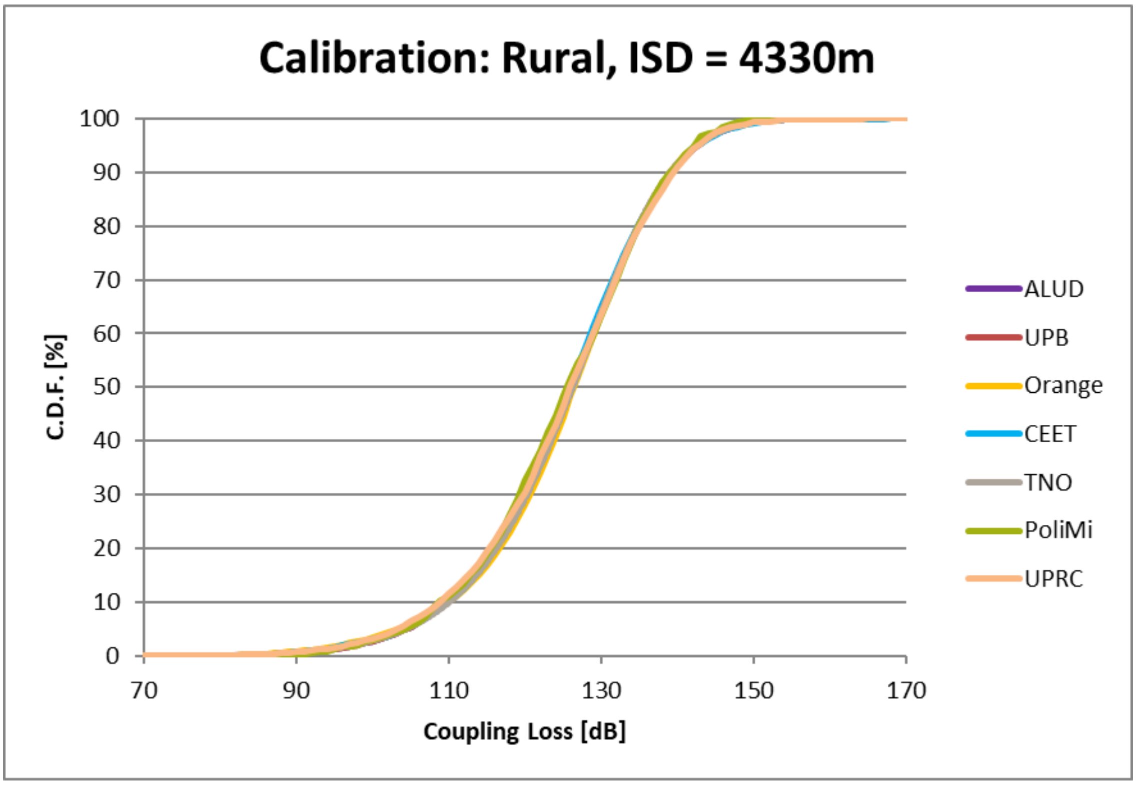


Figure 31 - RU Calibration results of the simulation software for coupling loss

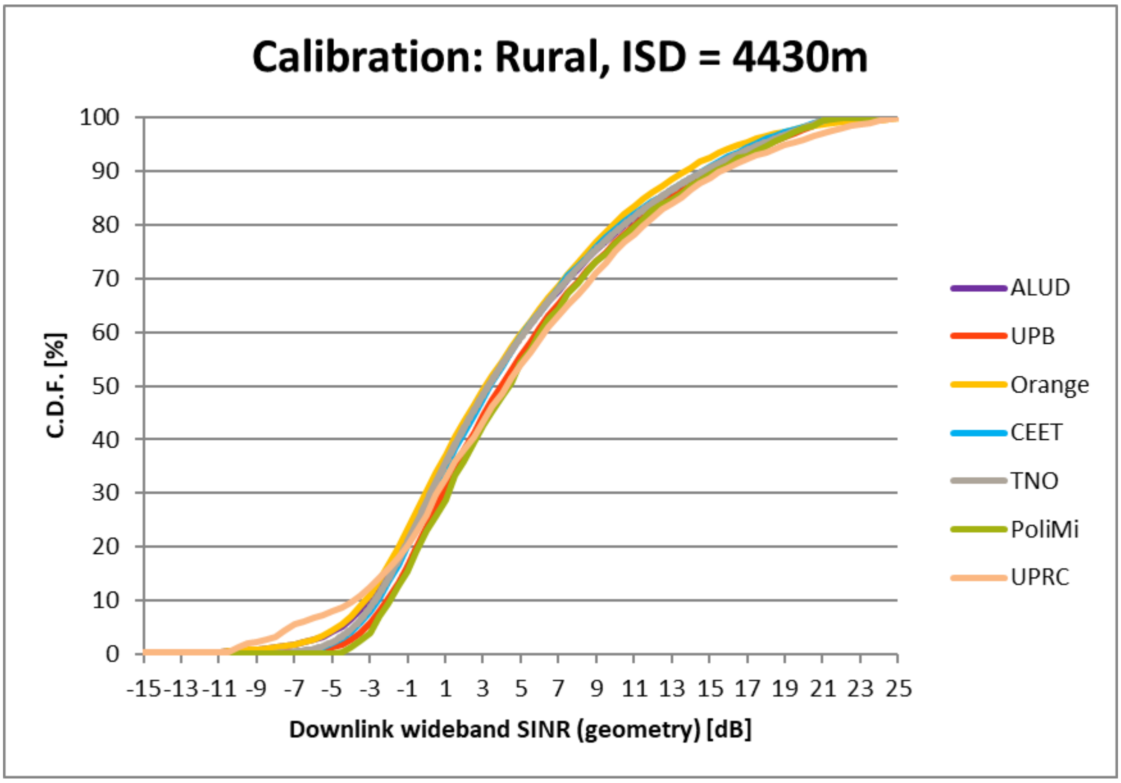
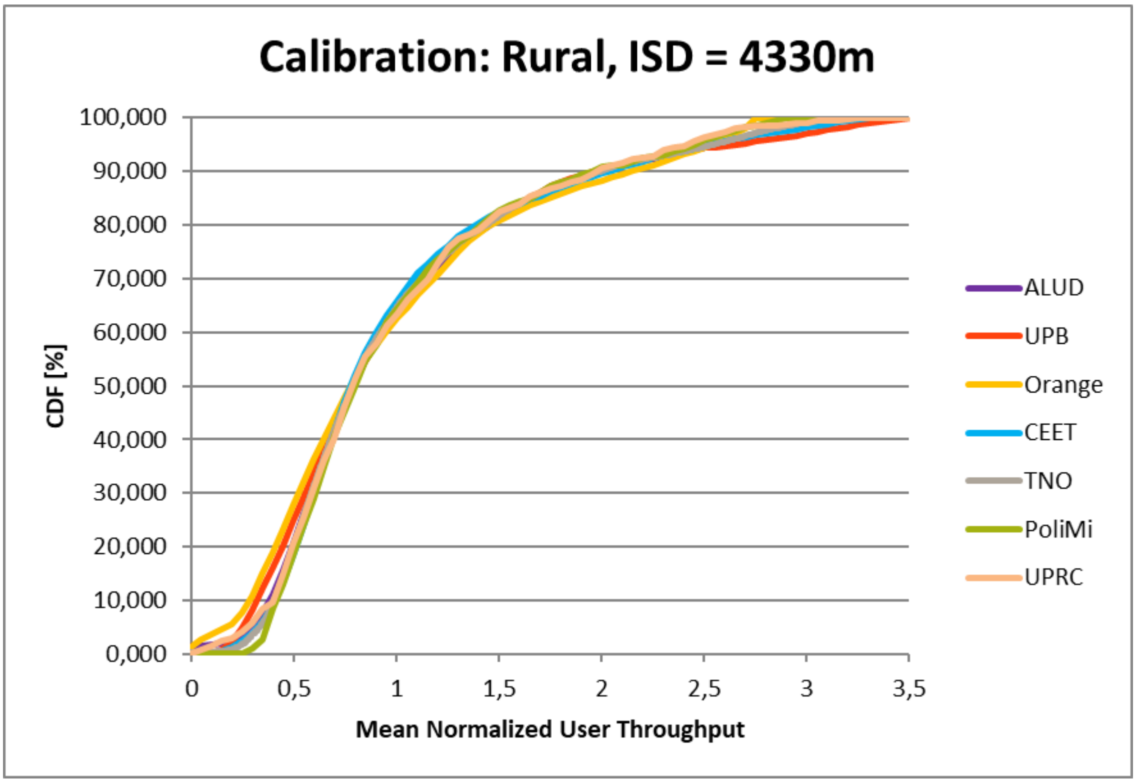


Figure 32 - RU Calibration results of the simulation software for SINR



2.6 Conclusion

In this chapter, it was shown that a valid simulation and knowledge-building software was necessary to move further with research on the optimization in cellular networks.

Such software was developed according to the specifications of other similar simulation environments with consideration for the upcoming chapter’s requirements in configuration and information building capabilities. After the completion of the software development, a thorough calibration / validation process was followed as to ensure that the software is a simulator of the real-world conditions. The successful calibration process is key to allow for new, beyond-state-of-the-art simulation scenarios that will allow us to prove that extensions in the technology are required to further advance the quality, efficiency and success of the 4G+ era. After performing many simulations in all the designated scenarios and comparing them with the respective results of the rest of the simulators, the simulation tool was found as a successful environment for 4G simulations. This means that new scenarios with different parametrization of the various aspects can be tested, and the results can be reliable enough to support new algorithmic schemes as optimization and improvement of existing technology.

2.7 Chapter References

[1] A. Margaris, Bachelor’s Thesis, “Simulation of 4G-LTE Cellular networks: Resource Allocation Algorithms”, University of Piraeus, 2012

[2] A. Margaris, Master’s Thesis, “A study on the Optimal Placement of the Decision-Making Entity in LTE Mobile Networks”, University of Piraeus ,2014

[3] GreenTouch, “End to End Network Architecture and Progress Measurement towards the Factor 1000 Improvement in Network Energy Efficiency” of the GreenTouch technical committee, Jan. 7, 2013.

[4] GreenTouch, Mobile Architecture Doc 1 - Models & Methodology, Version 3.0, Mobile Architecture/Metric Group

[5] GreenTouch, Mobile working Group, Mobile Communications WG, Architecture Doc2: Reference scenarios, May 8, 2013

[6] GreenTouch, Mobile working Group, Mobile Communications WG, Architecture Doc2A: Update on Modelling Parameter, December 11, 2014

[7] GreenTouch Mobile Communications WG, Architecture Doc3: GreenTouch Technical Solutions, 28 May 2014

[8] GreenTouch, Mobile Communications WG, Architecture Doc4: Energy Efficiency of GreenTouch Technical Solutions, version 0.2, 26 May 2015

[9] O. Arnold, F. Richter, G. Fettwis, O. Blume, “Power Consumption Modeling of Different Base Station Types in Heterogeneous Cellular Networks”, Future Network and Mobile Summit 2010 Conference

[10] K. Aho, T. Henttonen, J. Puttonen, L. Dalsgaard, T. Ristaniemi, “User Equipment Energy efficiency versus LTE Network Performance”, International Journal on Advances in Telecommunications, vol 3 no 3 & 4, year 2010

[11] Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects, 3GPP TR 36.814

[12] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2010-2015," white paper, June 1, 2011

[13] H.J. Kolbe, O.Kettig, E.Golic,” Monitoring the Impact of P2P Users on a Broadband Operator's Network”, 2009 IFIP/IEEE International Symposium on Integrated Network Management, 2009

[14] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, “Big Data: The next frontier for innovation, competition and productivity”, McKinsey Global institute, May 2011

[15] J.J.Arsanjani,M. Helblich, W. Kainz, A.D. Boloorani, "Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion", International Journal of Applied Earth Observation and Geoinformation, December 2011

[16] H.Li, J.Hajipour, A.Attar, V.C.M. Leung, "Broadband Access with fiber-connected massively distributed antennas architecture", June 2011, IEEE Wireless Communications Magazine

[17] K. F. Man, K. S. Tang, S. Kwong, "Genetic Algorithms: Concepts and Applications", IEEE Transactions on industrial electronics, Vol. 43, No. 5, October 1996

[18] A.Gosavi. Simulation-Based Optimization: Parametric Optimization Techniques and Re-inforcement Learning, Springer, New York, NY, Second edition, 2014.

Chapter 3 – Maximization of Energy Efficiency with static management and network redesign

3.1. Introduction

Energy efficiency as a management policy for HetNets involves the collection of metrics that are related (or contribute) to the power consumption footprint of the telecommunication infrastructure. All cellular communication systems require energy in order to provide the radio data transfer services they offer (either through the air or the backhaul interfaces). A large amount of energy is used by this network through devices such as controllers, servers and routers, various elementary network elements[1][2] that are necessary for the promised quality level of communication. The main metric that will be the source of analysis (and from each other metrics derive) is instantaneous and cumulative power consumption, measured in Watt (or KWatt) by all network components. The cumulative power consumption is the total energy required for the operation of the system and is measured in Joule. Other very important energy-related metrics are created by transformations (ratios and other formulas) namely the average power consumption, variance and standard deviation of power consumption, different time units for total power such as Kw/h (for 1-hour duration measurements). Energy costs alone are not enough to perform successful optimization actions. In order to make sure that our actions do not affect the quality of the network we need to include QoS KPIs in our evaluation. One very important KPI that is used in quality of service calculations is the achieved instantaneous throughput of each user equipment device. KPIs that use the energy consumption in combination with network link quality (in the form of instantaneous throughput) are two ratios: network intensity (measured in Kbit / J) , a measure that shows how efficiently every used Joule is converted and the inverted ratio , energy efficiency (measured in J/Kbit) translating into how much energy is required to achieve curtain levels of QoS service. Finally, one important KPI for measuring the impact of configuration and algorithms on quality of service is the packet drop rate (or call drop rate for GSM calls) and the average per packet delay. To summarize the metrics that must be monitored for energy efficiency optimization, we split them into three categories: a) Energy related that are expected to be reduced (i.e. Power consumption and total energy, b) QoS related that are expected to remain unchanged within expected ranges (i.e. Average throughput, packet or call drop rate, and c) Composite KPI that mix A and B type KPIs and they are expected to either increase(intensity) or decrease(efficiency). The total result of the analysis will be the overall effectiveness of the algorithms.

3.1.1 Energy Efficiency in HetNets

Energy efficiency in the complex environment of HetNets is one of the key KPIs of the reduction of cost and increase of element health. From a network management perspective, it is one of the three core categories of optimization problems for the 4th generation LTE. This type of telecommunication networks can have a very high variance in their power consumption profiles based on the demand in network access and the number of users that occupy the underlying area. These systems also have very strict specifications on the network transmission quality KPIs (mentioned in previous chapter[3][4]). Failure to comply with these specifications is not acceptable as it is immediately visible to the users of the network resulting into considerate loss of income for the telecommunication operators. To accommodate for this (even if done so in an inefficient way), operators perform actions of resource overprovisioning which results in high energy consumption, imbalanced usage of spectrum and added complexity in infrastructure management. Overprovisioning also results in low spectral efficiency, due to the fact that the extra spectrum is only utilized in extreme demand situations. The root cause of the high-power consumption in HetNets is overloaded network elements. Network load also leads to interference generation, which amplifies the problem. Since the air frame buffers need to forward the information to the user terminals, they need the best possible radio environment in order to achieve the highest transmission qualities. In the same time, constantly transmitting cells increase the overall interference which lowers the quality. This phenomenon is continuously deteriorating and leads to very high packet loss and consequently outage. Initial design of a cellular network tries to overcome congestion based on the current specifications. Unfortunately, since the installation of the infrastructure and the initial design process, the network demands evolve, especially after the introduction of smart devices (smartphones) with increased capabilities and embedded extensibility. In order to tackle this situation intelligently, we need to analyze the daily "load" KPI of reference areas for specific population densities and apply optimization according to these historical data. Empirical models and forecasting models can often be a part of the design process, providing with future projections of the incoming traffic demand but their accuracy diminishes as new parameters/aspects are being brought and technology evolves. For HetNet infrastructures, we can select from several intelligent configuration options that benefit from this knowledge-building. One large category of solutions revolves around the manipulation of the several element's load (load-balancing, load transfer, etc.). Since element load has such a large contribution to the power consumption KPI, we can apply load-balancing techniques like placing new types of network coverage elements that will absorb traffic partitions from the eNodeB elements. In order to ensure their co-existence in the HetNet environment, proper parametrization must be performed in the handover algorithm parameters to ensure the amount of offloading that they can absorb is adequate and worth the investment and extra complexity. In addition, the exact location of the new coverage elements must be selected strategically in order to align with the various traffic centers (hot-spots).

3.1.2 Simulation Scenario Topology

The simulation scenario (Figure 33) that will be used for this chapter is based on reference scenarios mentioned in 3GPP specifications[3][4][5][6][7]. One of the key characteristics of the input parameters is the inhomogeneous distribution of the population density. Dense Urban environments tend to present with such load geometries due to the different urban facilities (like transportation, markets etc.). This specific scenario has detailed instructions for the placement of the UE terminals and their generated traffic. In order to follow them, classes of UE devices must be declared with different position characteristics.

- Ambient Users that will correspond to the 40% of the total population. They will be uniformly placed in the playground with a density of 1 ue / m^2
- Hot-Zone Users that will correspond to the 40% of the total population. They will be focused in hot-zones spread throughout the playground with a density of 1 hot-zone / km^2. Each hot-zone will enforce a population density value of 2 ue / m^2
- Hot-Spot Users that will correspond to the 20% of the total population. Hot-Spots will be generated inside the hot zones with a density of 2 hot-spot / km2. Within a small radius from the center of the hot spots, ue devices will be generated with a density of 4 ue / m^2

The scenario also specifies the placement of the eNodeB 3-sector antenna elements for the 4G coverage. A 1500x1500 meter playground closely fits a full 6 eNodeB hexacomb of cells (18 elements in total) if we use the standardized inter-site-distance between the sites of 500m (Dense Urban). In order to maintain the uneven distribution of users, we will not apply any mobility model to the UE terminals as it would be very complex to design motion patterns that will obey the constraints.

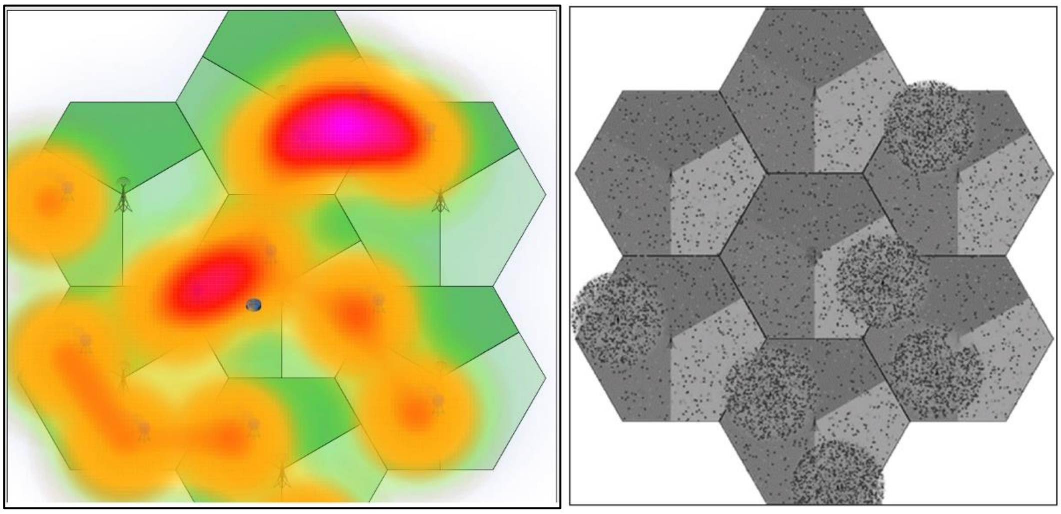


Figure 33 - Population density heat-map for the specific simulation scenario

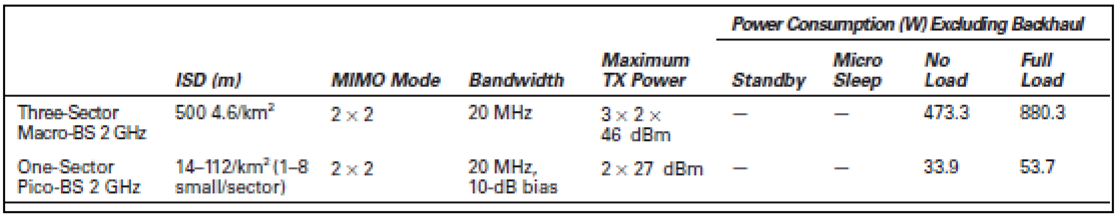
3.1.3 Simulation Traffic Model

For this simulation scenario, we will be using a traffic model that derives from the initial specifications of the simulation software. The selected traffic model in the WWW-FTP projection for 2020 traffic, which is composed from two different packet sizes, small packets (~10KB) with high frequency (96%) of arrival for signaling, updates and AJAJ response and large packets(~2MB) with low frequency (4%) and large payload for the initial mobile application / web page load. The average traffic demand (throughput / m^2) generated from this model will be 10Mbps/m^2. The arrival probability distribution for each UE device is a separate stochastic Poisson generation process. This will ultimately result in the desired inhomogeneous distribution of the traffic (according to the specified user population density) with lower traffic demand on the ambient user area (most area of the playground) increased traffic in the hot zones and most of the traffic in the hot spots. In addition, for the quality of service requirements of the user traffic, we will be enforcing the 2020 expected delay thresholds that are specified in the literature [6]which specify 100ms maximum delay for the signaling small packets and 500 ms for the large packets. This delay will be the network transport delay as propagation and processing delay is not included in the system level simulator software. Any packet serving delay higher than the expected results into application-level rejection (packet drop) and subsequently, RRC session failure.

3.1.4 Simulation Network Element Characteristics

The available network elements that will be used for this simulation (both for the initial scenario and for the optimization phase) will be reference hardware in the predefined simulator assets. The sector antenna[6] of the eNodeB will use a) maximum transmit power of 49dBm (for 2x2, 4x2 and 8x2 MIMO with diversity gain) b) 3D sector antenna model with max gain 14 dBi c)15-degree vertical tilt (Dense Urban inter-site distance) d) half-beam width 70 degree. For omni-directional antennas that will be used for Pico elements the simulator will be configured to maximum transmit power of 30 dBm and 5 dB omnidirectional gain (at the surface level). For the power consumption model of the network elements, the eNodeB will use the standard load-based model with minimum power consumption of 473.3 watt and maximum 880.3watt. The pico cell will use an order of magnitude less energy with parameters of 33.9 and 53.7 watt respectively. For the transmit bandwidth, we are assuming 20Mhz bandwidth slices for each of the Pico and eNodeB elements. For the initial handover parametrization, we are setting the Pico bias value to 10 dB (virtual gain)

Table 3 - Macro cell and Pico cell simulation characteristics



					Power Consumption (W) Excluding Backhaul			
	ISD (m)	MIMO Mode	Bandwidth	Maximum TX Power	Standby	Micro Sleep	No Load	Full Load
Three-Sector Macro-BS 2 GHz	500 4.6/km^2	2 × 2	20 MHz	3 × 2 × 46 dBm	—	—	473.3	880.3
One-Sector Pico-BS 2 GHz	14–112/km^2 (1–8 small/sector)	2 × 2	20 MHz, 10-dB bias	2 × 27 dBm	—	—	33.9	53.7

3.1.5 Simulation Optimization KPIs

The selected performance metrics that will be used to evaluate the effectiveness of the optimization schemes will be based on the analysis of the previous chapters. More specifically they are separated in contextual categories: a) Energy characteristics that are linked to the OPEX of the cellular network such as Power consumption, total energy of the equipment, b) Telecommunication link quality characteristics such as average per UE downlink throughput and average packet transmission delay. Another important quality of service parameter is the cell edge throughput, namely the achieved throughput from the lowest 5% (percentile) of the throughput's distribution. Finally, the packet download success ratio (%) is another important indicator of quality. c) Mixed characteristics that combine OPEX and quality characteristics (a+b) such as energy efficiency and energy intensity

3.2 Problem Solution

We are exploring various approaches to improve the energy efficiency of the setup in such a HetNet environment. These approaches can be a) Multi-Operator infrastructure sharing[9][10][11][12][13], b) Uniform Placement of Pico cell elements in the playground to offload the cells[14][15][16], c) Intelligent Placement of Pico cell elements to target the traffic in the high-density zones, d) other methods of QoS improvements such as increase in effective bandwidth of cell elements. For the next chapters we will apply these principles to reduce the EMF footprint and power consumption while maintaining the QoS of the simulation in adequate levels

3.2.1 Power Improvements - Operator Infrastructure Sharing

Modern HetNets operate simultaneously in most civilian zones using different slices of the 4G spectrum. These networks are operated by different network providers and are isolated orthogonal networks that, due to the stochastic nature of the traffic distribution, are in different load states. Usually 2 to 6 network operators (depending on the country) coexist in the various domestic zones. All these operators are overprovisioning their network equipment with radio resources (equal spectrum) and network elements (close or same in number) in order to be able to serve opportunistic traffic spikes. In addition, the geographical nature of the serving zone restricts the providers from reducing the number of network elements they use for coverage and signal penetration constraints (i.e. locations of the cellular coverage map must always have a minimum service capability). However, traffic analysis on historical data and projection models show us that the same substrate network area can be served by one network operator with the sacrifice of partial quality of service degradation. Deactivating or switching the infrastructure to sleep modes while 1 (out of 4) providers would serve the traffic could lead to important power consumption benefits (as seen by the min-max values of power consumption of the elements). Since the 4G standard requires a fixed quality of service level, we need to explore additional QoS improvements.

3.2.2 QoS Improvements - Pico cell placement and bandwidth increase

Infrastructure sharing imposes a radical change in the active available resources to be used for the radio transmission. This greatly affects the provided quality of service and rises the need for technological countermeasures. Pico cell elements have been shown in rich literature as a method of energy efficient high-quality coverage technology.

Based on the reference Pico cells technology (shown previously) a new network design that would include them in the coverage area would have minimum energy efficiency and power consumption impact whilst keeping the quality of service in the same (ever higher) levels. Various schemes can be followed as for the placement of the Pico cells in the playground depending on the knowledge that we assume we can utilize. In a simplistic case, Pico cells can be placed in an evenly distributed manner (much like the hexacombs of the cellular network). This however will not be the optimum placement as the pico cell elements work most effectively when placed in a location with high population density. The second and more advanced placement scheme is to utilize the known user distribution (as created by the simulation software) to target the overloaded zones (denoted as hot-zones or hot-spots) with Pico cells. That way the small coverage of the Pico cells wil be sufficient to offload the eNodeBs and therefore reduce the power consumption of the infrastructure. Another important aspect of this optimization is configuration on the handover algorithm of the network. Since pico cells wil be placed in a densely covered area by nearby sector antenna eNodeBs, we need to make sure that users will handover their service effectively. The Pico Bias parameter works as a virtual power gain, masquerading the closest Pico cell's transmit power as the strongest received without the need of manual handover.

3.2.3 QoS Improvements – Available Bandwidth increase

Another method to maintain high levels of quality of service after performing infrastructure sharing and element turn-off is bandwidth transfer. As analyzed in previous chapters, multiple operators own different slices of the available 4G bandwidth. By shutting down these elements (via power control or infrastructure sharing) this bandwidth becomes unused and therefore underutilized. This bandwidth can then be transferred dynamically in the other cells increasing the capability in which they serve the area. Although the increase in bandwidth will reduce the service time (and subsequently the load of the cells) It is important to note that it will also increase the power consumption of each cell. Large part of the power consumption comes from the linear amplifiers of the digital transmission and higher energy requirements will be needed for double or triple the bandwidth. The conclusion is that, although spectrum transfer is an option for extreme cases of QoS degradation, it needs to be performed selectively in isolated zones in order to not reverse the energy efficiency effects of the optimization.

3.3 Application of the Proposed Scheme – 2-Phase Optimization

Having analyzed the hypothetical outcome of the various solutions to the energy efficiency problem, we propose the following optimization methodology(Figure 34). Initial simulations will be run for the network's performance before optimization: this state will be now on referred to as "reference scenario". The next step will be to apply infrastructure sharing in the same simulation scenario. This means merging all the users from different network operators into a single operator and shutting down the operation of the elements from the rest. The removal of all the eNodeB elements will consequently lead to great reduction in power consumption of the network but the increased traffic transferred to the single operator will lead the infrastructure to very high load and likely high outage. This scenario will then be used to test the various offloading techniques that we mentioned in the previous scenarios in order to maintain the energy benefits and improve the quality of service into an adequate stage. Uniform placement of Pico cells will be used in addition with the increase of the effective bandwidth of the eNodeB elements. Finally, the intelligent placement of the Pico cells inside the hot zones is expected to provide the best solution to the problem. Variations of the Pico cell placement configurations will be tested and compared in order to find the instance in which the quality standards (predefined by reference scenario) will be met (or surpassed). In addition, increase on the available bandwidth will also be tested as a candidate to absorb the QoS deterioration caused by the infrastructure sharing.

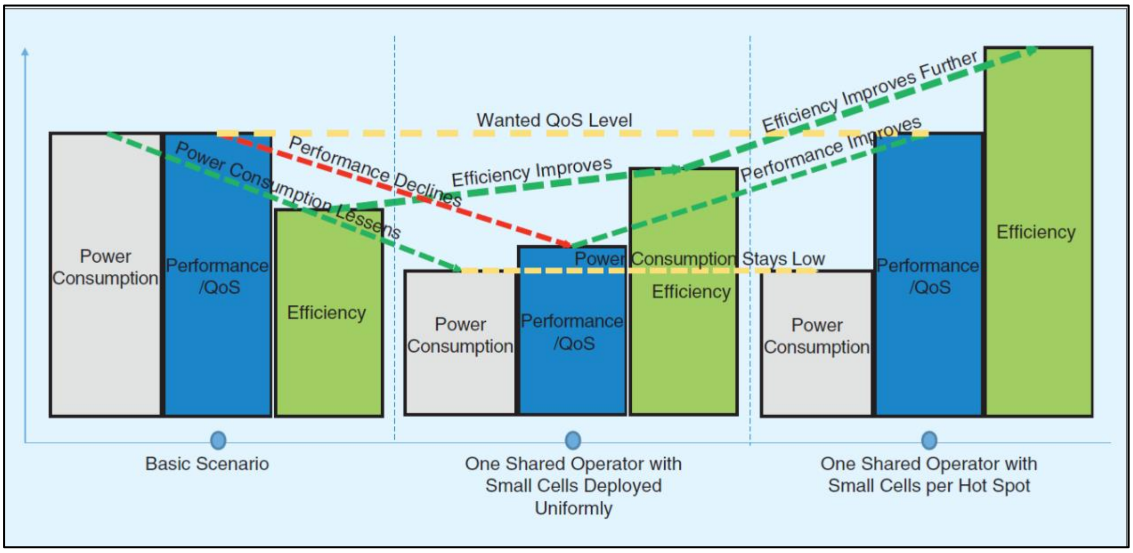


Figure 34 - The proposed scheme performance evaluation

3.4 Performance Evaluation

In this chapter we will analyze the results from the various executions of the simulator software in order to identify the problems generated in the reference scenario and validate that the proposed scheme is showing evidence of improvement in the operation of the cellular network.

3.4.1 Evaluation Methodology

The complexity of simulating such a large reference scenario (including multiple radio access network elements and user equipment devices) is forcing us to perform small simulation samples of various states in which the network can be set. According to methodologies from other simulations, we can split the daily traffic profile of a network's profile into small sub-simulations of different traffic demand amplitude (weight). After gathering all these measurements, we then complete the profiling of the total simulation by performing a weighted average summation of the results. These "load weights" take the values of 20%, 40%, 100%, 120%, and 140% (peak rate). In addition, simulations for various other environments (such as Urban, Sub-Urban or Rural) could also be included but will not be for the scope of this study. The selection of the focused KPIs will be based on the analysis performed in previous chapters. Energy efficiency, intensity, power consumption and total energy will be converted in a per km^2 density value for extrapolation purposes. For the quality of service measurements, average packet download time will be used as well as the packet failure rate (which will be a direct result of packets exceeding their delivery expectation rate). Finally, as an additional QoS KPI we will be using the cell edge throughput (equivalent to the fifth percentile of the throughput CDF). The gathering of the datasets is part of the implementation of the simulation and knowledge-gathering tool developed for the scopes of this dissertation. This includes the generation of the reports and graphs of the following chapters.

3.4.2 Result Analysis – Reference scenario

The first results that we will be analyzing are from the "reference scenario". This scenario included 4 simultaneous network operators serving a Dense Urban area of 1500 square meters of the specified user density and traffic model. In addition, each operator has 20 MHz of bandwidth at his disposal and serving with a separate 6-site (18 cell) hexacomb. All load levels were evaluated for all the selection of the KPIs and created the reference measurement for all the improvements. The results followed the expectations of the preliminary study: Quality of service is in the expected (adequate) levels whilst the resource usage and power consumption are in a high state. It is important to note that the 4 operator system benefits also from the different radio frequencies that result into zero to little inter-channel interference.

3.4.3 Result Analysis – Phase I - Infrastructure Sharing

Activating the "infrastructure sharing mode" from the simulation configuration, leads us to the results of phase 1. A total of 20Mhz allocated at operator 1 is used to serve the same number of users for the selected area. The disabled network elements of the simulation have dramatically decreased the power consumption(Figure 35), but the network is now at a state of increased load. Quality of service begins to radically deteriorate especially for the case of the peak hour (140% load), (Figure 37,Figure 38). However, at this stage we have the least amount of network elements and the smallest required bandwidth.

3.4.5 Result Analysis – Phase II – Intelligent Pico placement

Phase 2 is split into all the different configurations of location and number of Pico cell placement. The idea is to use the simulation software in order to identify the optimum number of Pico cell elements for the performance "repair". After reaching the boundary of 7 Pico cells per hot spot (with hot spot placement) we see that we have successfully recovered (and even improved in cases of edge throughput) the performance KPIs with a benefit of 55% in power consumption (Figure 35, Figure 36) of the network.

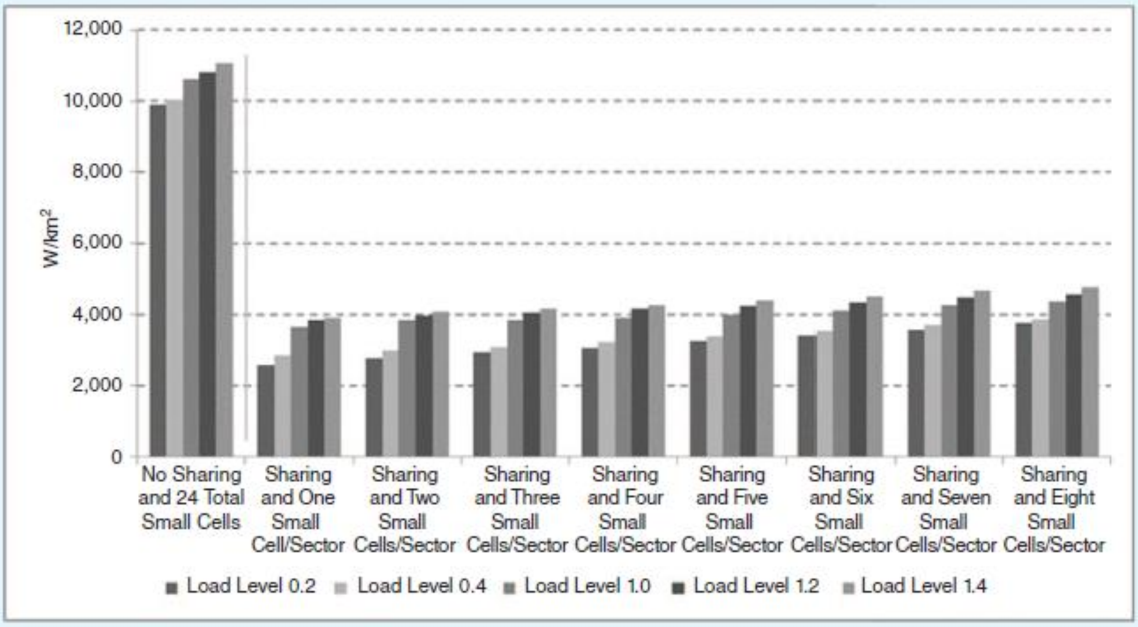


Figure 35 - Power consumption for each simulation scenario

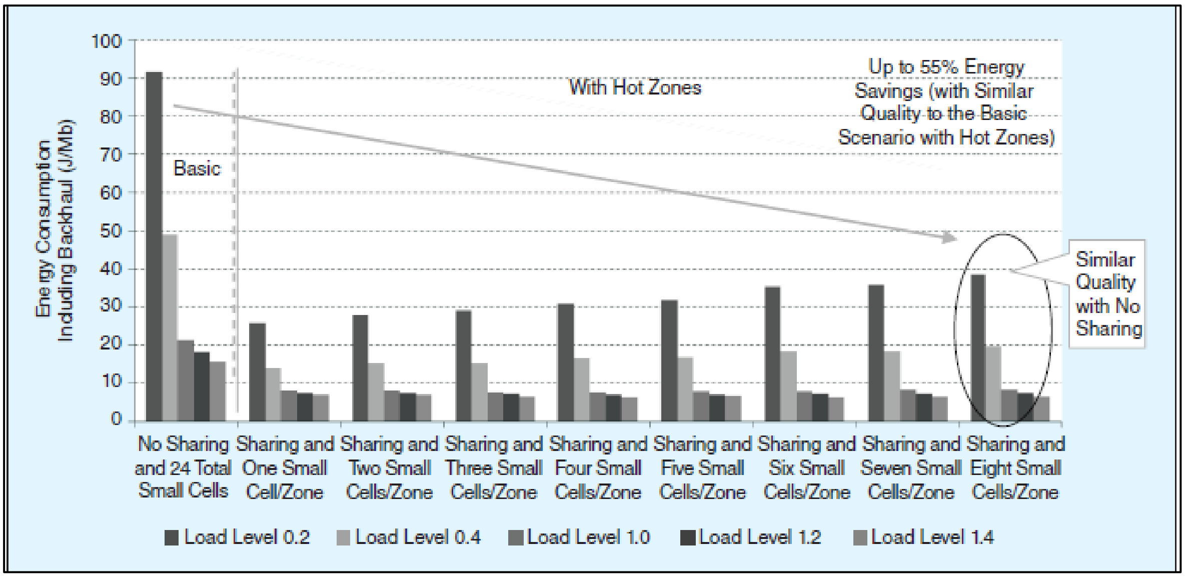


Figure 36 - Energy efficiency for each simulation scenario

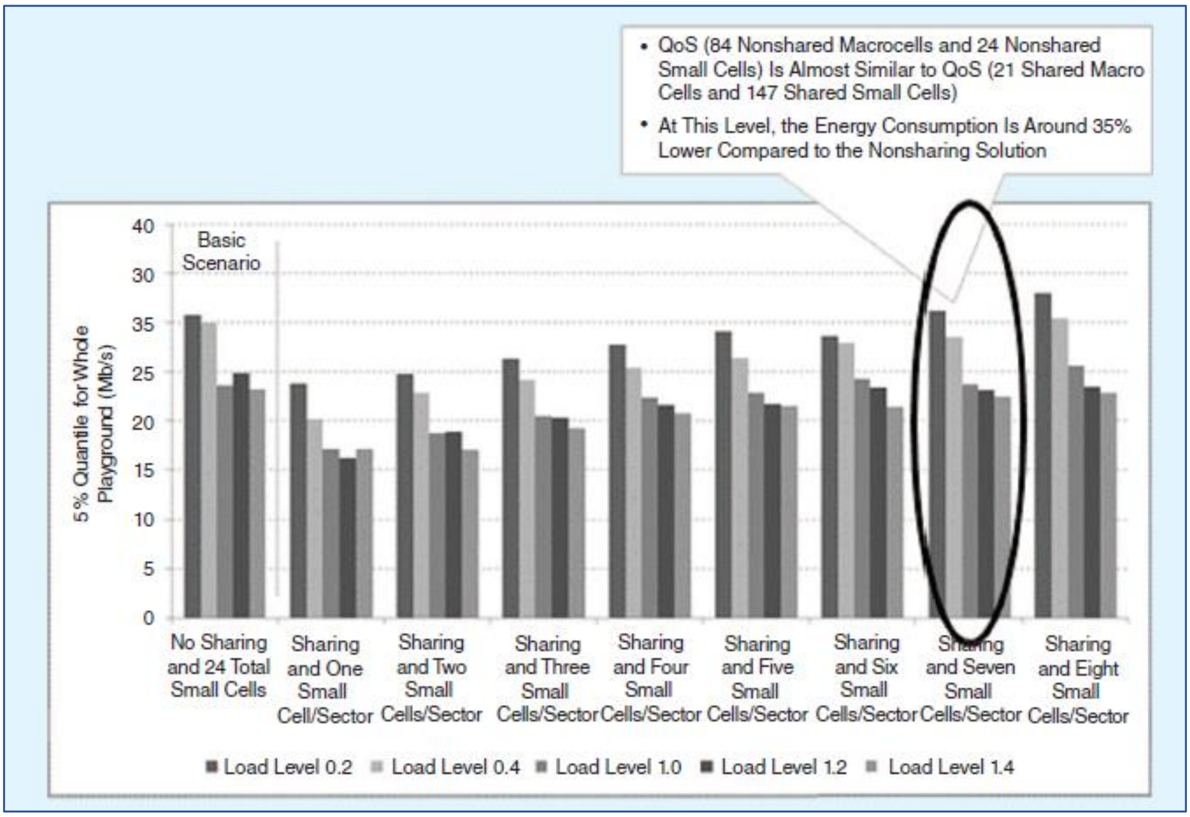


Figure 37 - Edge throughput (5%) for each simulation scenario

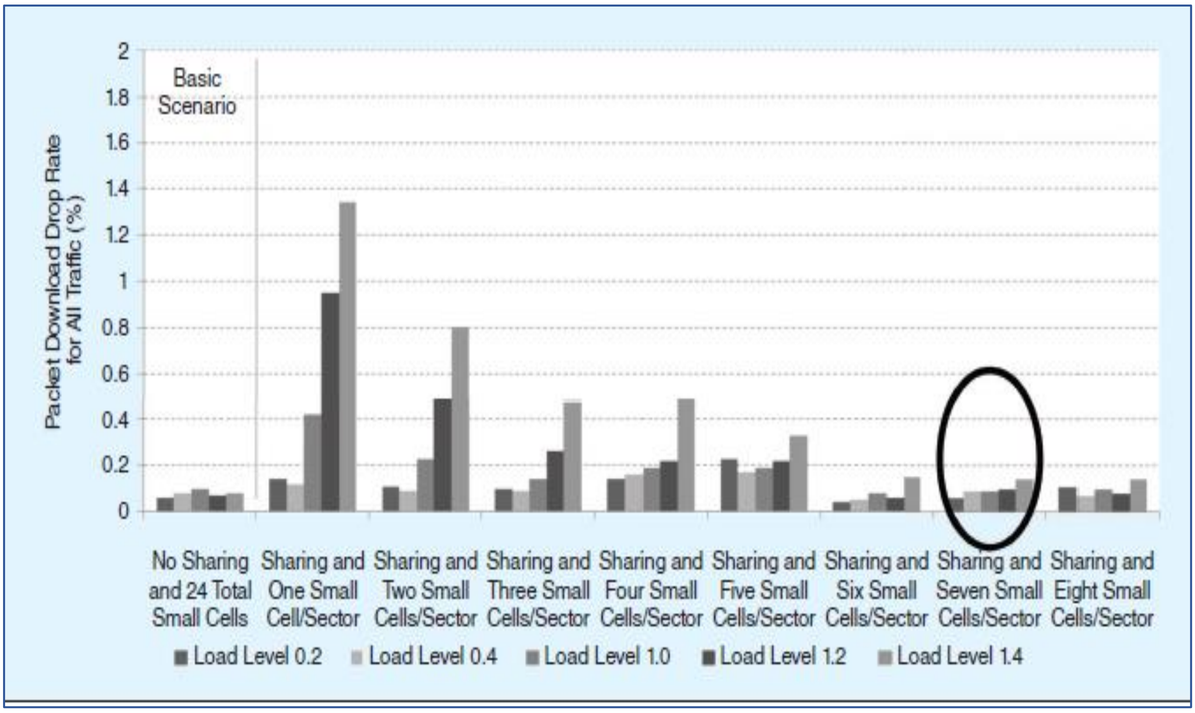


Figure 38 – Packet drop rate for each simulation scenario

3.3.3 Results from bandwidth increase

As expected from our initial analysis, the increase of the eNodeB's effective bandwidth resulted in an increased power consumption for all different load levels. This result however is still valuable because the network quality of service increased dramatically surpassing the "reference scenario" initial values. This means that scenarios like this could be a considerable input for future network configurations with increased available bandwidth and smaller number of operating network elements.

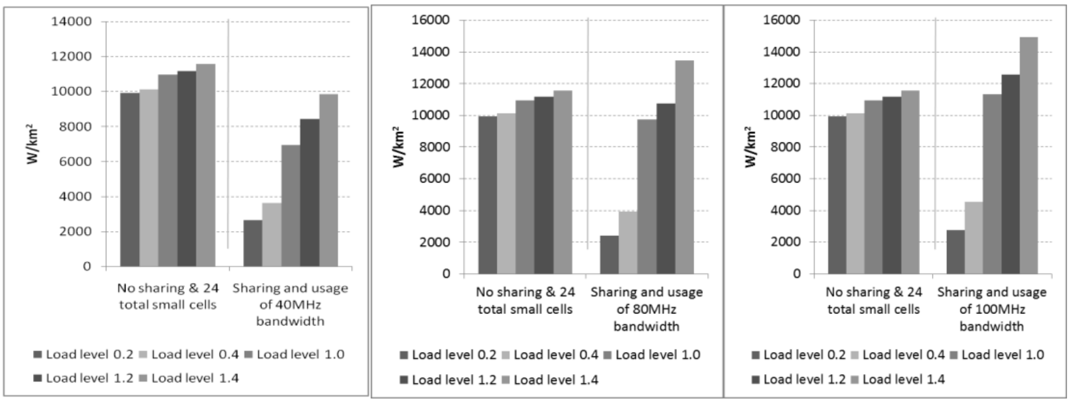


Figure 39 - Power consumption for various bandwidth allocations

3.4 Conclusion

Energy efficiency is a very important goal in the cellular network design and operation lifecycle. In this chapter we have shown that initial network design can lead to important losses in power and energy efficiency which would normally be handled in an inefficient way. Intelligent redesign of the cellular network, knowledge building from historical data of traffic usage and advanced radio coverage elements such as Pico cells can be a combined methodology to dramatically reduce the power consumption of this system without sacrificing quality of service or efficiency. Simulations in reference scenarios confirm that these solutions are effective even in the most densely populated areas, producing large amounts of stochastic traffic. These findings greatly solidify the importance of Knowledge-building in the design and configuration of cellular networks.

3.5 Chapter References

[1] F. Richter, G. Fettweis, M. Gruber, and O. Blume, "Micro base stations in load constrained cellular mobile radio networks," in Proc. IEEE 21st Int. Symp. Personal, Indoor Mobile Radio Communications Workshops (PIMRC Workshops), Istanbul, Turkey, Sept. 26–30, 2010.pp. 357–362.

[2] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling different base station types in heterogeneous cellular networks," in Proc. Future Network Mobile Summit, Florence, Italy,June 2010, pp. 1–8.

[3] GreenTouch, "End to End Network Architecture and Progress Measurement towards the Factor 1000 Improvement in Network Energy Efficiency" of the GreenTouch technical committee, Jan. 7, 2013.

[4] GreenTouch, Mobile Architecture Doc 1 - Models & Methodology, Version 3.0, Mobile Architecture/Metric Group

[5] GreenTouch, Mobile working Group, Mobile Communications WG, Architecture Doc2: Reference scenarios, May 8, 2013

[6] GreenTouch, Mobile working Group, Mobile Communications WG, Architecture Doc2A: Update on Modelling Parameter, December 11, 2014

[7] GreenTouch Mobile Communications WG, Architecture Doc3: GreenTouch Technical Solutions, 28 May 2014

[8] GreenTouch, Mobile Communications WG, Architecture Doc4: Energy Efficiency of GreenTouch Technical Solutions, version 0.2, 26 May 2015

[9] T. Frisanco, P. Tafertshofer, P. Lurin, and R. Ang, "Infrastructure sharing and shared operations for mobile network operators from a deployment and operations view," in Proc. IEEE Network Operations Management Symp., Apr. 7–11, 2008, pp. 129–136.

[10] J. P. Pereira and P. Ferreira, "Infrastructure sharing as an opportunity to promote competition in local access networks," J. Comput. Networks Commun., vol. 2012, Article ID 409817, 11 pages, Feb. 2012.

[11] D. Meddour, T. Rasheed, and Y. Gourhant, "On the role of infrastructure sharing for mobile network operators in emerging markets", Comput. Networks, vol. 55, no. 7, pp. 1576–1591, 2011.

[12] F. Berkers, G. Hendrix, I. Chatzicharistou, T. de Haas, and D. Hamera,"To share or not to share?," 14th Int. Conf. Intelligence Next Generation Networks, Oct. 11–14, 2010, pp. 1–9.

[13] Evolved Universal Terrestrial Radio Access (E-UTRA), RAN Sharing enhancements, 3GPP TR 22.852

[14] Evolved Universal Terrestrial Radio Access (E-UTRA), Network sharing; Architecture and functional description, 3GPP TR 23.251

[15] Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects, 3GPP TR 36.814

[16] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2010-2015," white paper, June 1, 2011

Chapter 4 – Quality of Service Optimization utilizing dynamic radio resource allocation mechanisms

4.1 Introduction

Optimum radio resource allocation in 3GPP cellular networks (especially HetNets)[1][2][3] is a multi-factor problem that is widely studied in the available academic literature. It is also an important study item of the standardization technical teams of 3GPP for specifications that will impact the performance of the technology's future releases. From a digital communications perspective, 4G networks use a physical layer of OFDM multiplexing (OFDMA for multi-user environment) which is essentially an evolution of the GSM (2^{nd} gen.) FDMA / TDMA 2-dimensional implementation. In principle, time and frequency resources are split into two dimensions that can be split in a discrete manner. Time is split into the minimum transmission window which generally results in higher frames of 0.5 or 1 milliseconds. The frequency dimension is split into the minimum possible bandwidth (180-200 Khz subcarriers) that can contain the information symbols. The difference between 2G and 4G is in the wave form. 4G is using the OFDM pulse also known as subcarrier that is an improvement from various perspectives. The 2D pair entries group together in what is referred to as "resource-block" which is the minimum allocation that an active user terminal can acquire from the system for transmission. Algorithms that select which resource block will be allocated to which user (and for which application) are generally called radio resource allocation policies and are implemented in the scheduler of the LTE layer 2 system.

4.1.1 Dynamic Resource Allocation in HetNets

In HetNets, multiple high-level objectives can determine the drive for dynamic radio resource allocation policies[4][5]. As a general rule, the default policy in which networks provide access to the shared is through the means of equal resource allocation, sometimes referred to as "even" or "round-robin" allocation. This allocation scheme however does not include any intelligence since it will provide equal quality of service to users with the same radio environment conditions. Round robin allocation can either be a full buffer per user allocation, in which every time frame of transmission a single user is utilizing the channel or it can be a sequential radio resource distribution for simultaneous users. If we include additional information to the radio resource allocation algorithms, we then have a number of "intelligent" radio resource allocation policies. Firstly, we have the spectral efficiency objective which translates to achieving the highest possible throughput / service per km^2 / per Hz sometimes referred to as "greedy". This translates to a good conversion rate of the raw resource (purchased frequency band) into network throughput. However, the reality of this goal is that it will benefit only user equipment devices that experience the best radio environment. The spatial diversity of cellular networks in cities and all geographical area types does not allow for direct line of sight telecommunication links. The cellular technology is designed to penetrate various zones such as buildings, alleys, industrial zones and transportation facilities at the cost of signal strength and network errors. Distributing resources to the small part of the coverage distribution that receives the best radio quality will lead to outage for many users and will cause the technology to fail in one of its most important goals: coverage rate. In the opposite hand of this approach we have the opposite policy of overprovisioning users with bad radio quality. This is sometimes referred to as the "fairness" policy. By doing so we are reducing the quality of service diversity between users at the cost of radio resources. Users with good radio quality receive low resources and these resources are then transferred to the users with bad radio conditions. Ultimately the users can then experience equal services. One key element of efficient radio resource allocation is the physical layer monitoring function. Through the means of measurements and LTE signaling, the network must be aware of the transmission link quality of each user and its corresponding access device. Utilizing this information for decisions that involve the scheduling of packets and radio resource allocation is sometimes referred to as cross layer optimization or cross layer information exchange. For the radio resource allocation problem, simple localized measurements performed between the UE and the eNodeB do not suffice. The most important factor of quality deterioration if LTE networks is inter-channel interference as it results from the densification of the element deployment. Since the information of every single transmission and occupied resource block is not included in the backbone signaling, it is believed that a centralized approach to efficient radio resource allocation is better. Centralized SON functions can provide a general overview of the network area's instantaneous (and future) interference profile based on the knowledge of the network element's load and current resource allocation. The signaling of the UE terminal with the eNodeB also provides geospatial information for total knowledge of the geography of the problem. That can lead to estimations of the INR, SNR and SINR ratios for each UE device accurate enough and fast enough that can be applied in real time dynamic radio resource allocation. For the next chapters, we have selected a specific simulation scenario in order to test and contradict the various dynamic resource allocation algorithms of the literature. We have also a proposed algorithm that we believe will greatly improve the quality of service and the fairness of the HetNet infrastructure and positively contribute to the technology's evolution.

4.1.2 Simulation Scenario

The selected network simulation (Figure 40)is a Dense Urban area type large scale simulation playground including 19 eNodeB sites (3 sectors each). We are using a HetNet infrastructure, with homogeneous placement of 9 Pico cells inside every sector eNodeB resulting in a total of 228 radio access elements (out of which 171 are small cells). The radio configuration of these elements is 4 full LTE bands resulting in 20 MHz (or 100 resource blocks) for each cell. The selected network topology is generated by the hexacomb network layout generator which resulted in a central 6-cell hexacomb with 13 additional eNodeB sites on the circumference. The operational band for the Pico cells is the LTE 3.5GHz band (different resource slice than the eNodeB which operated in the 1800 MHz LTE band) so that they can greatly penetrate the users in spite of the dense placement in the center of the antenna sector of the macro base station.

Table 4 - Network Element characteristics of the simulation

BS	MIMO mode	Bandwidth
Macro	2x2	20Mhz
Small	Omni-antenna	20Mhz

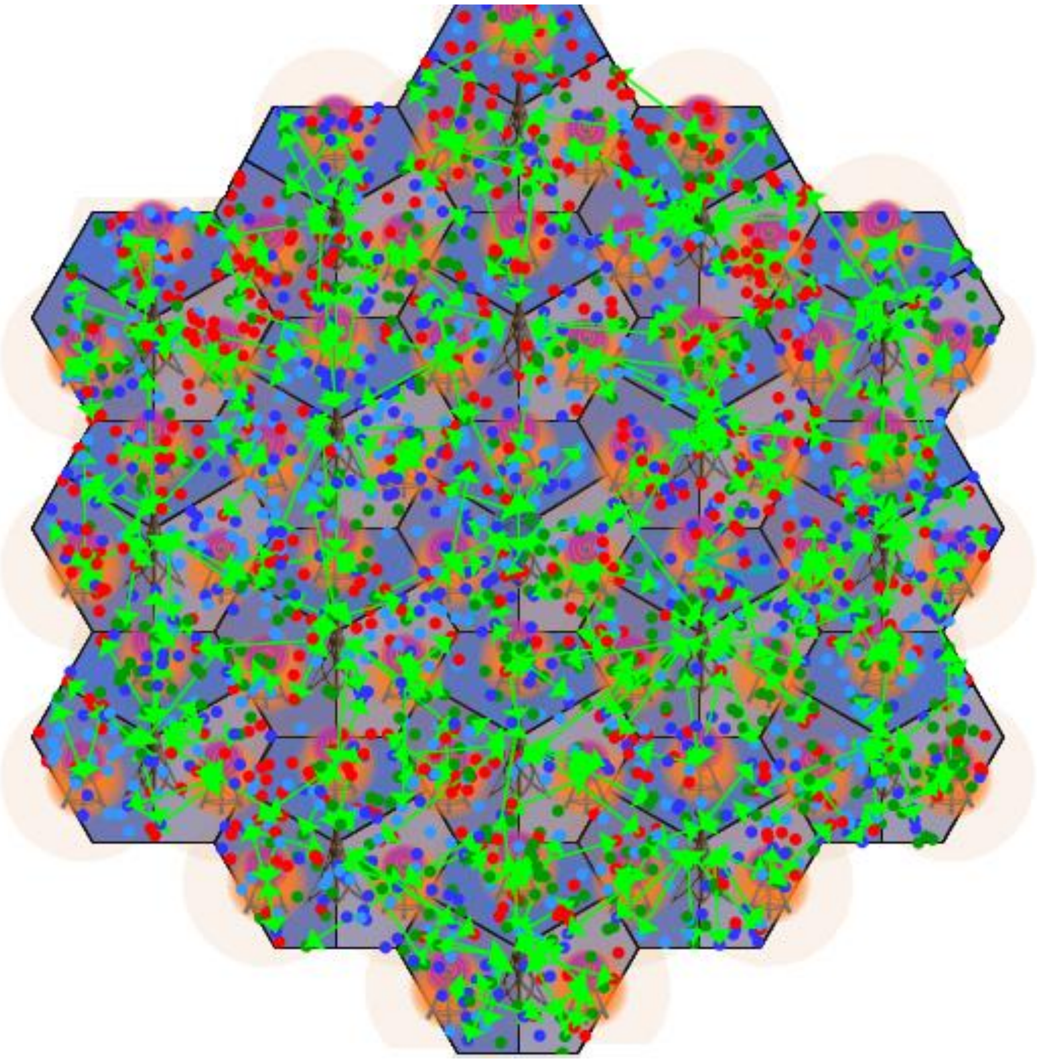


Figure 40 - The simulation scenario topology

4.1.3 Simulation Traffic Model Variations

For the next experiments that will determine the optimum dynamic radio resource allocation scheme, a set of different test cases have been designed in order to thoroughly cover all possible load levels. By parametrization of the "average per user daily requests" of the simulators FTP Application layer emulation, we end up with 5 different scenario variations to be evaluated varying from 2280 to 14400 packets per day per user. The number of UE devices used for this simulation is 5000 uniformly distributed in the playground. For the packet size requested for transmission, we will be using a fixed 2MB (large packet) in order to achieve the wanted 4 to 20 MB / minute per user load level.

Table 5 - Simulation test cases

Test cases	Users	Sessions/Day/User	Packet size
1	5000	14400	2Mbytes
2	5000	11520	2Mbytes
3	5000	8640	2Mbytes
4	5000	5670	2Mbytes
5	5000	2280	2MBytes

4.1.4 Simulation Service Classes

In addition to the simulation traffic model, we will be splitting the packet transmission entries into different categories of packet priority in order to include this also into our benchmarking. In a sense we want to be able to parametrize the radio resource allocation to also include a quality of service class in the level of service. Quality of service classes can be found in various literature references such as the US three tiered authorization framework[6][7]. An example of different QoS classes can be a) general access users (GAA) which is the default access to the medium and will have no throughput guarantee (as well as no delay thresholds). The GAA class will be the most likely class to activate and will result in the majority of the packets generated in the system. The second class (b) can be referred to as the priority access layer users (PAL) which have higher priority than the GAA users. Finally, we have the (c) class Incumbent Access or IA users which have mission critical transmissions and demand the highest priority. These classes will be adopted in the much simpler descriptions of low, medium and high priority. For the simulation we will randomly distribute various generated packets as to emulate real world conditions where users request premium and/or non-premium services simultaneously by using their phones.

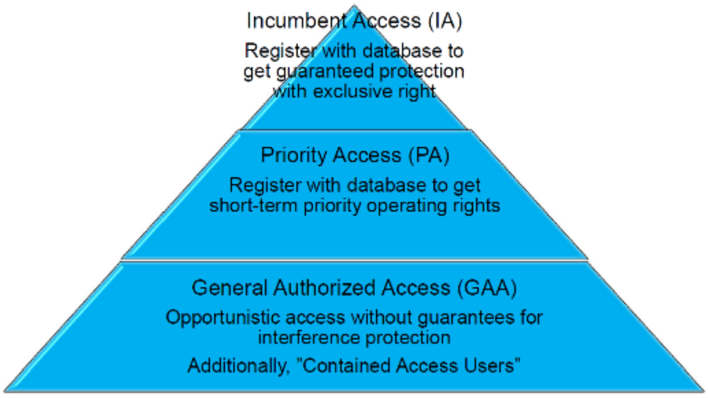

Incumbent Access (IA)
Register with database to get guaranteed protection with exclusive right
Priority Access (PA)
Register with database to get short-term priority operating rights
General Authorized Access (GAA)
Opportunistic access without guarantees for interference protection
Additionally, "Contained Access Users"

Figure 41 - Service class example hierarchy

4.2 Problem Solution

For the solution of the optimum radio resource allocation problem from an average quality of service perspective, we will enumerate the literature's most referenced [8][9][10][11][12]implementations and analyze their operation separately. We will then analyze more advanced techniques and our suggested cross-layer optimization approach. Afterwards, we will select 2 cases from the SOTA and perform simulations in order to compare its performance with our suggested solution for the various priority levels and load levels defined previously.

4.2.1 Radio Resource Allocation Algorithms

As mentioned in previous chapters, there are several approaches [13][14][15][16][17][18][19][20]to the radio resource allocation problem depending on what is your ultimate optimization goal.

4.2.2 Full buffer radio resource allocation

The most simplistic radio resource allocation approach is the full buffer or full allocation approach. Allow all elements to use all resources simultaneously. This simplistic approach is only effective if inter-site interference is very low and the traffic demand is also within normal ranges. Maximum load will result in disastrous interference levels and this allocation can only benefit users with the best radio quality conditions (greedy).

4.2.3. Orthogonal frequency reuse with reuse factor

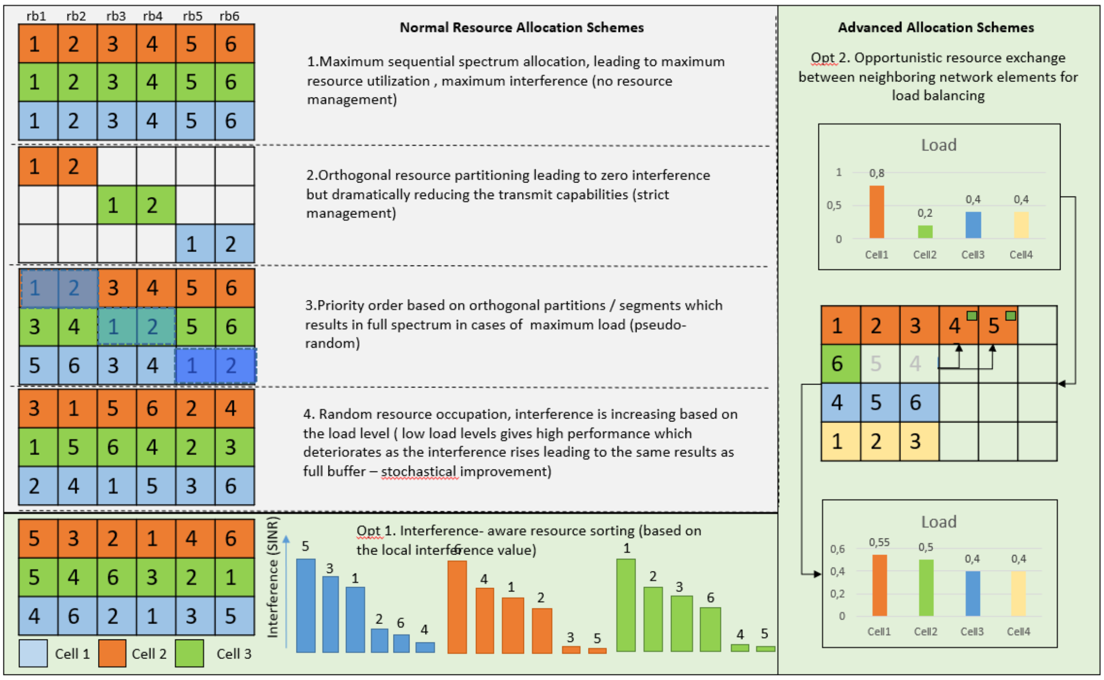
Orthogonal radio resource allocation with reuse factor (e.g. 3) is a traditional method of radio resource allocation to reduce inter-channel interference. Neighboring cells are allocated with different slices of the total bandwidth, resulting in no overlapping frequencies and no interference (for the neighboring cells). This greatly benefits the users with the worst radio conditions which will receive service levels of the best possible quality (assuming already bad signal strength from serving cell. The combined solution of these two algorithms is the prioritized orthogonal full allocation scheme. Each cell has all the total resources available, however it prioritizes an orthogonal set over the total amount. In low load situations, this will result in fair and high-quality communications. While the conditions change and demand increases, the quality of transmission will deteriorate and reach the 100% load limit like first case.

4.2.4 Random resource allocation

Random resource allocation is a simple methodology that requires very little to low planning and knowledge of the existing system. Regardless of that, it can perform as good as the orthogonal reuse algorithm of the literature due to the combination of randomness and the stochastic nature of cellular traffic. The conflict probability, however, increases according to the average load of the surrounding cells (like cases 1 and 3) and its theoretical limits are the same. From the SOTA algorithms, the random algorithm will be implemented as the best possible solution for both high performance and fairness.

4.2.4 Advanced algorithms for radio resource allocation

Other dynamic channel allocation schemes are found in the literature; however they are more intrusive for the HetNet systems in a way that much customization is required for their success.



Normal Resource Allocation Schemes

1.Maximum sequential spectrum allocation, leading to maximum resource utilization , maximum interference (no resource management)

2.Orthogonal resource partitioning leading to zero interference but dramatically reducing the transmit capabilities (strict management)

3.Priority order based on orthogonal partitions / segments which results in full spectrum in cases of maximum load (pseudo-random)

4. Random resource occupation, interference is increasing based on the load level (low load levels gives high performance which deteriorates as the interference rises leading to the same results as full buffer – stochastical improvement)

Opt 1. Interference- aware resource sorting (based on the local interference value)

Advanced Allocation Schemes

Opt 2. Opportunistic resource exchange between neighboring network elements for load balancing

Figure 42 - An overview of LTE radio resource allocation algorithms

The first algorithm is the load balancing resource exchange algorithm. It starts at an initial stage like algorithm the orthogonal reuse partitioning – even partitioning and starts a control loop that aims at balancing the load of all the network elements. The key operation is the resource block exchange between the various cell elements. If an eNodeB's load increases, the algorithm will require resources from a neighboring cell in order to balance them out. The extra resource will be used for as long as the traffic demands it and then it will be traded back to maintain equilibrium. This algorithm is very effective in principle; however, it has curtain drawbacks that are tightly coupled with the cellular technologies in general. This algorithm falls under the category of distributed SON functions. Each eNodeB will act as a local agent to solve a local problem. To do so it needs a communication protocol for interfacing with its neighboring cells. Via this signaling route, the cell will then communicate the demand for additional resources and standardized methods must be able to implement such an exchange. This requires a lot of extensions for the standard cell-to-cell communication cell (X2 interface). Another problem with this algorithm is the extreme case. No amount of resource exchange can solve the traffic demand overload problem. It will however even the problem out as good as possible and can be used in conjunction with handover optimization and traffic steering schemes.

4.2.5 The proposed resource allocation scheme

The suggested algorithm that will be used for the basic comparison with all the other literature is the cross-layer-interference-aware DCA algorithm. By the means of measurement collection in a centralized SON instance, the network is able to deduce the SINR (INR or IR) of each different eNodeB – UE pair in the playground. It will then rank the available resources with a weight that will mark either "clean" or "dirty" frequency block. If we exclude the computational complexity from the equation, we are expecting this algorithm to outperform the random allocation algorithm of the literature. The reason is that whatever the load case is (low or high) we will always have a notion of the best and worst resource to provide. This will be as accurate as the interference calculation models can be based on the radio feedback of the UE terminal and propagation models. It will also be based on the current transmit load of the neighboring cells therefore it will collect multi-source information from the network. In general, these algorithms have a "core" flow diagram that they follow that can be seen in the following flow diagrams. We see that extensions can be performed in the channel assignment step of the diagrams.

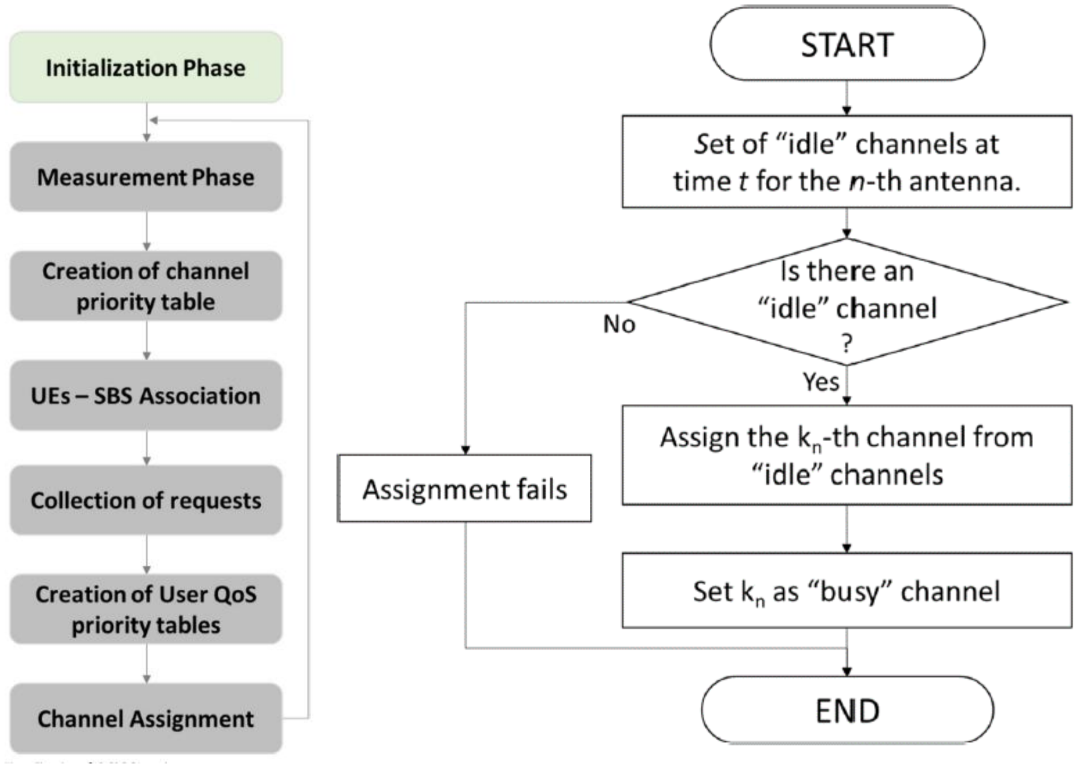

Initialization Phase
Measurement Phase
Creation of channel priority table
UEs – SBS Association
Collection of requests
Creation of User QoS priority tables
Channel Assignment
START
Set of "idle" channels at time t for the n-th antenna.
Is there an "idle" channel ?
No
Yes
Assignment fails
Assign the k_n-th channel from "idle" channels
Set k_n as "busy" channel
END

Figure 43 - General Radio Resource Allocation Flow Diagram[12]

4.2.2 SOTA algorithm: Random Resource Allocation

The random resource allocation algorithm is a simple but effective DCA scheme. It provides all eNodeB elements with full radio resource capabilities and tries to achieve lower SINR values by using a random resource allocation. Randomness is an effective and cheap way to allocate resource in an orthogonal manner, especially for low load values of the network elements. As the load of each element increases however, the radio transmit conflicts are rising leading into the same situation as the full buffer serial allocation. The key to this algorithm is which scenario it will be applied and what actual level of element load will be achieved. Below we can see a flow diagram of the algorithms implementation as it is included in the various radio resource algorithm implementations of the simulation software.

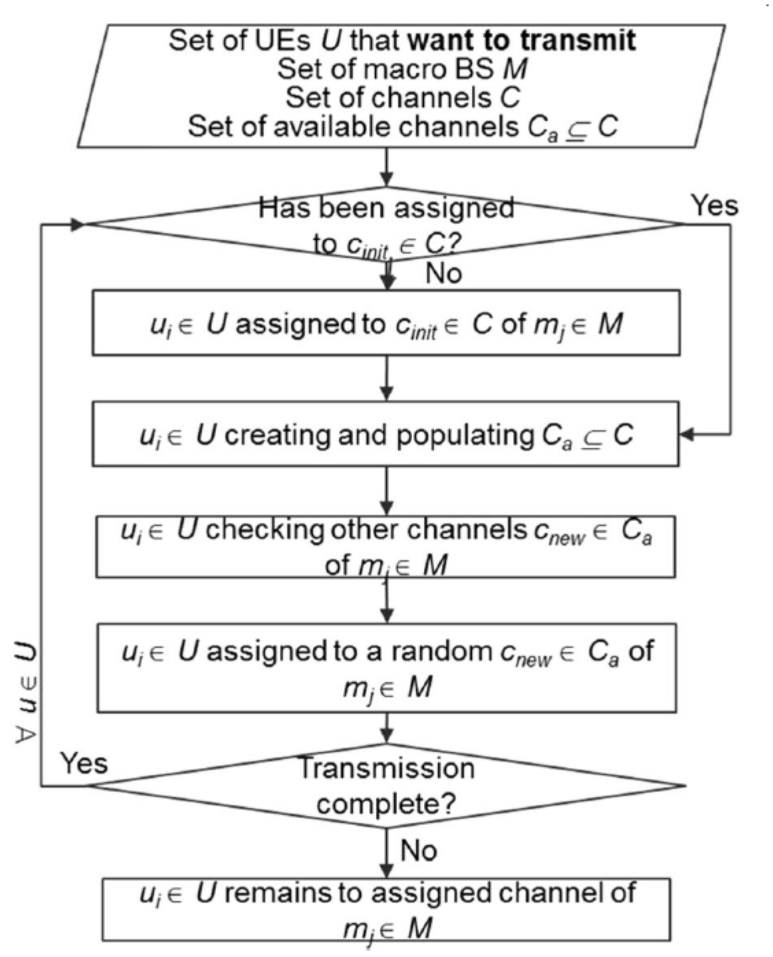
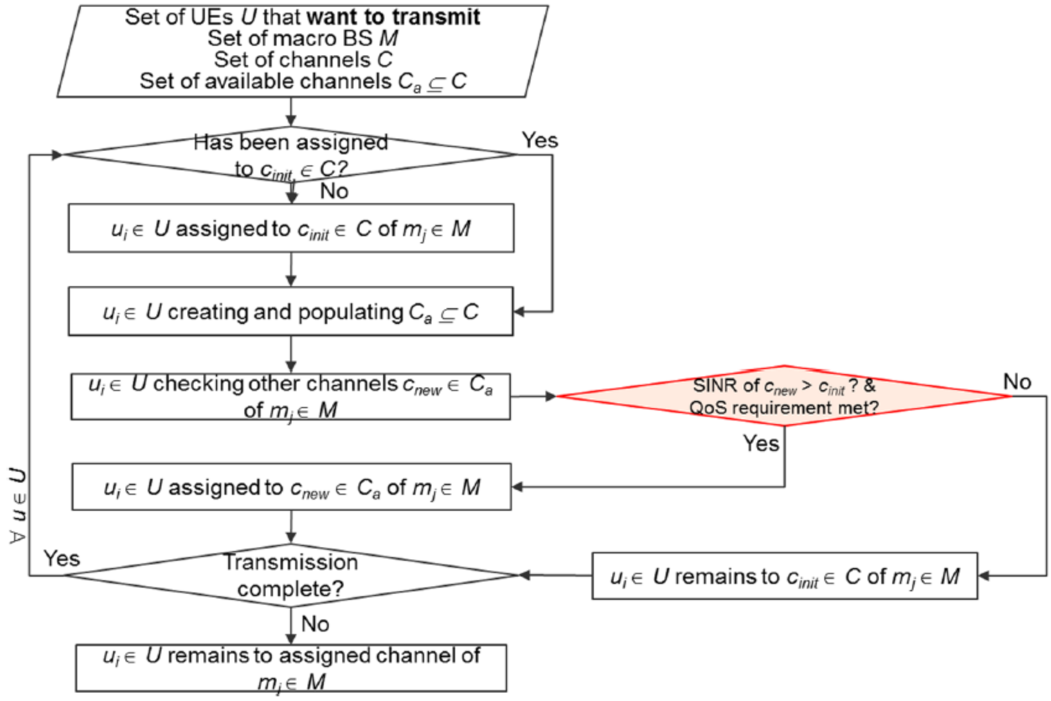

Set of UEs U that **want to transmit**
Set of macro BS M
Set of channels C
Set of available channels $C_a \subseteq C$
Has been assigned to $c_{init} \in C$?
Yes
No
$u_i \in U$ assigned to $c_{init} \in C$ of $m_j \in M$
$u_i \in U$ creating and populating $C_a \subseteq C$
$u_i \in U$ checking other channels $c_{new} \in C_a$ of $m_j \in M$
$u_i \in U$ assigned to a random $c_{new} \in C_a$ of $m_j \in M$
$\forall u \in U$
Yes
Transmission complete?
No
$u_i \in U$ remains to assigned channel of $m_j \in M$

Figure 44 - Flow diagram for the random DCA algorithm[12]

4.2.3 Proposed Algorithm: Interference-Aware Resource allocation

Interference-aware resource allocation is our suggested DCA scheme for the solution of the fair allocation problem. It is an algorithm for evaluating locally every different resource block according to an estimation of the SINR that the transmission will produce. This is done by information exchange from a centralized SON function on the network controller of the HetNet. If the complexity of such an algorithm is not restrictive for real time usage, it can prove to be the best in terms of fairness and relative quality of service per user. It will also allow for quality of service class enforcement for the various use cases we have analyzed in the previous chapters. The flow diagram of this algorithm is an extension of the simple resource allocation diagram in which all the resources are rated with a different SINR coefficient. In cases of extreme load, all the resources will be used and then the full buffer mode will arise. However, the notion of resource ranking and the "best" or "worse" resources might produce better results in the high vs low priority service requirements.



Set of UEs U that **want to transmit**
Set of macro BS M
Set of channels C
Set of available channels $C_a \subseteq C$
Has been assigned to $c_{init} \in C$?
Yes
No
$u_i \in U$ assigned to $c_{init} \in C$ of $m_j \in M$
$u_i \in U$ creating and populating $C_a \subseteq C$
$u_i \in U$ checking other channels $c_{new} \in C_a$ of $m_j \in M$
SINR of $c_{new} > c_{init}$? & QoS requirement met?
No
Yes
$u_i \in U$ assigned to $c_{new} \in C_a$ of $m_j \in M$
$\forall u \in U$
Yes
Transmission complete?
$u_i \in U$ remains to $c_{init} \in C$ of $m_j \in M$
No
$u_i \in U$ remains to assigned channel of $m_j \in M$

Figure 45 - Flow diagram form the interference-aware DCA algorithm[12]

4.3 Performance Evaluation

In this section we will analyze the performance KPIs of each of the following three DCA algorithms: The basic algorithm (hereby named SOTA or A for short) which will include full spectrum allocation for reference purposes.

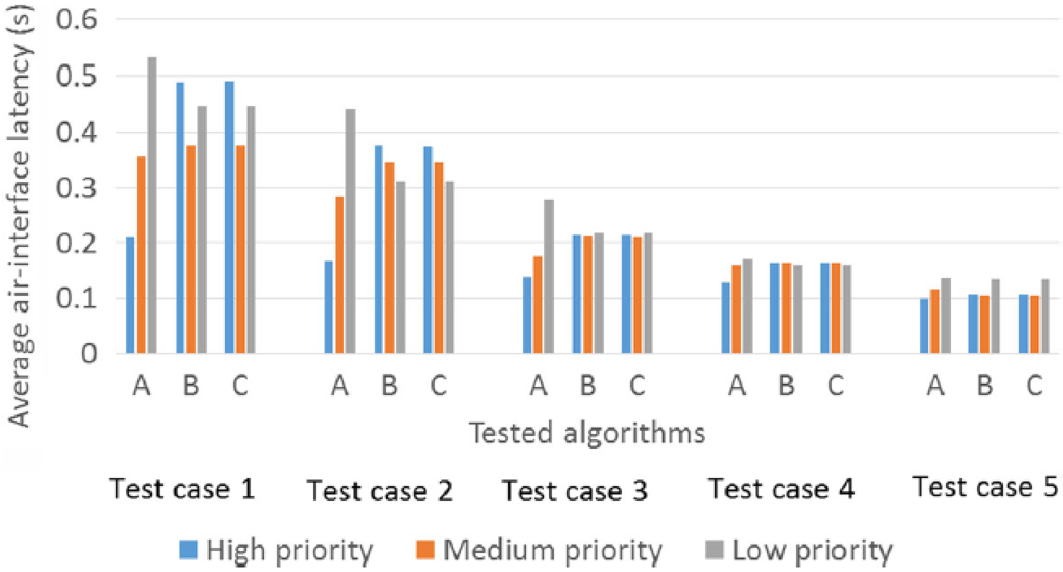


Figure 46 - Average air-interface latency for each test case and algorithms A, B, C

The SOTA random algorithm denoted as B which will be compared to our proposed algorithm and the algorithm with QoS priority named the "interference-aware algorithm" or C. The selected KPIs that will be compared are average (per user) measurements of QoS in order to achieve both high performance and fairness for all the UE terminals. The KPIs that will be displayed are the average packet latency (for the specific simulation traffic model) and the normalized per user throughput. These results will also be split into different histograms for different packet classes (as described in the simulation test cases chapter).

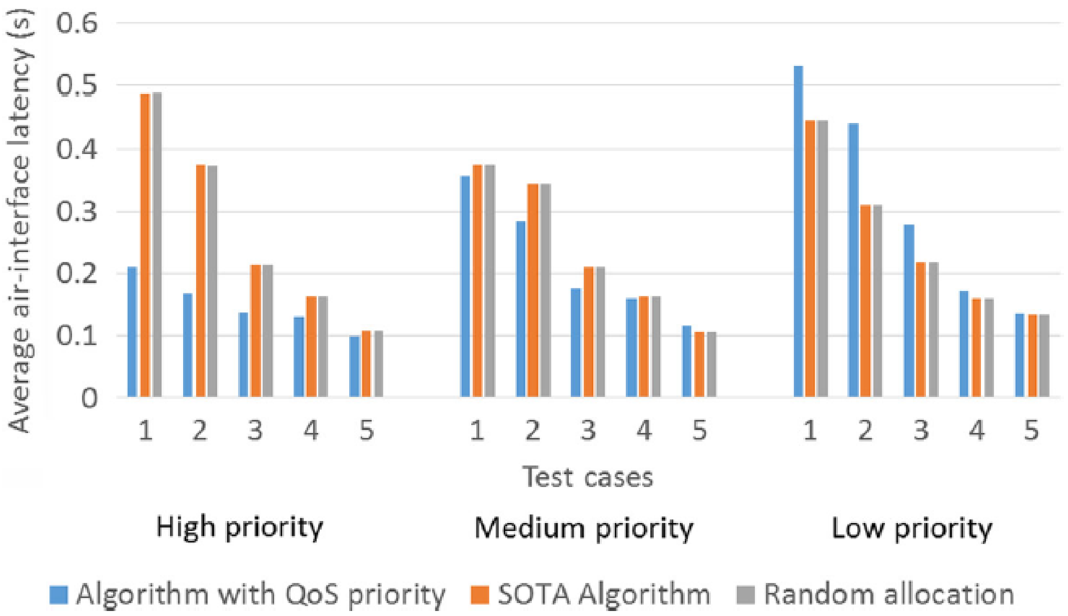


Figure 47 - Average air-interface latency per priority level

The first set of results to analyze is the average air interface latency. On average, it is shown that our proposed algorithm outperforms the other two algorithms (up to 50%) especially in high and medium priority services by giving them a performance boost. On the contrary, low priority services seem that they do not benefit as much as the other two. In the next figure, the results are sorted by priority levels, and the large benefit of our algorithm is more visible for high priority which is not the case for low priority services.

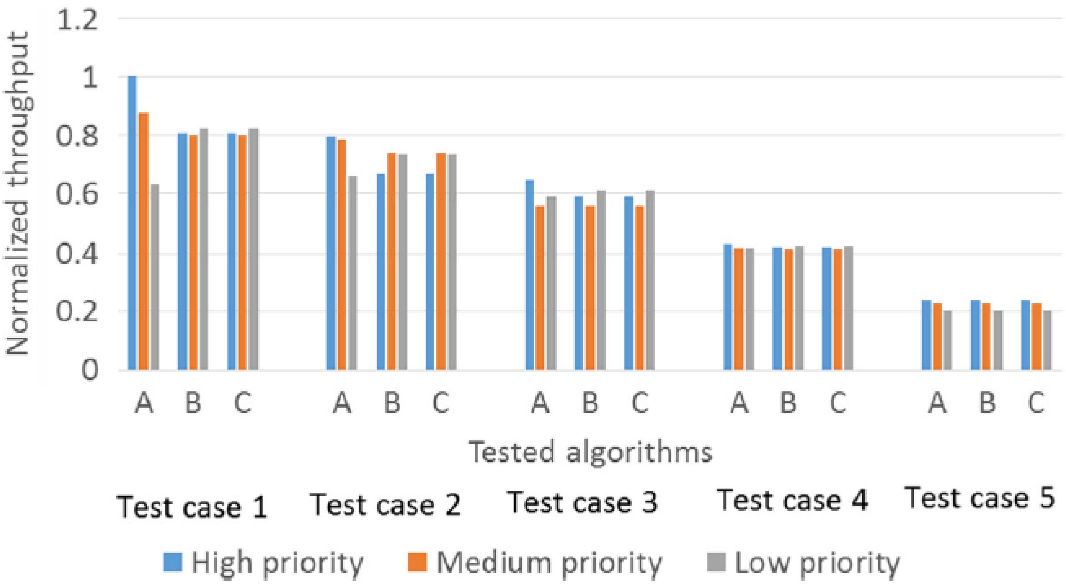


Figure 48 - Normalized Throughput for each test case and algorithms A, B, C

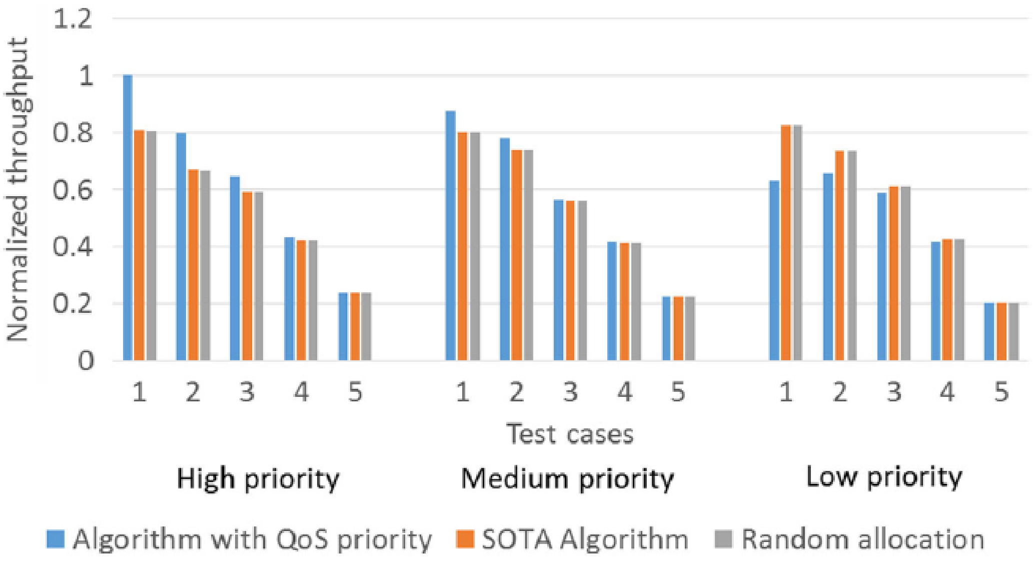


Figure 49 - Normalized Throughput for each priority level

The next result set illustrates the normalized throughput for each of the test cases and compared among each algorithm. It is evident that our algorithm performs better in almost every test case and especially in cases with higher loads (compared to less-loaded simulations). Switching the analysis perspective, we see that the next figure illustrates the normalized throughput as of service priority levels and here (as shown in latency charts), our solution seems to perform better especially in higher and medium priority services compared to low priority services.

The test cases used for this study where designed in order to investigate the performance difference between the state-of-the-art and our proposed solution. Our proposed solution was able to dynamically choose the optimum channel based on interference of the current position and thus allow each user to connect with higher speed and receive the file faster with less air interface latency. On the contrary, the algorithms that used for comparison on average were making the less optimal selection of the channels (without giving priority based on QoS requirements),hence the users were not able to download at full speed and with higher loss packet ratio, creating a continuously loop of poor selection of channels without being able to overcome this situation. Furthermore, there are some differences between random allocation of channels and SOTA algorithm when increased load is provided in the system. The random allocation has worst performance in high and medium priority services. The state-of-the-art algorithm performs better, and our proposed algorithm has the best performance especially in high and medium priority services.

4.4 Conclusion

Dynamic Channel allocation and radio resource management is part of the complexity that the large configuration space of HetNets introduces to network operators. Analysis of the knowledge provided by real world data, analytical models and simulations can be used as important tools to understand which DCA scheme is most effective under various circumstances. In this chapter we have seen that basic radio resource algorithms fail to handle situation of overloaded areas in large scale HetNets. Random DCA helps with the problem but fails to handle situations of critical load. Random DCA also has no notion of embedded quality of service classes, something that can prove useful if the network wants to prioritize different classes of service instead of treating all traffic as equal. A proposed, multi-context DCA algorithm that utilizes both radio quality and network load aspects to provide projections for the quality of each resource block is shown to outperform the SOTA and the random algorithm in both selected QoS and fairness indices. By including the interference into the computation for the resource allocation, we reduce the required resources, and this ultimately results into more energy efficient networks. These findings may push the design of the 4th and 5th Generation standards to include this additional information in their message exchange protocols and allow for further advances and achievements in terms of technological features.

4.5 Chapter References

[1] Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects, 3GPP TR 36.814

[2] Evolved Universal Terrestrial Radio Access (E-UTRA); Study on multiple radio access technology (Multi-RAT) joint coordination 3GPP TR 37.870 V13.0.0, June 2015.

[3] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2010-2015," white paper, June 1, 2011

[4] K. Aho, T. Henttonen, J. Puttonen, L. Dalsgaard, T. Ristaniemi, "User Equipment Energy efficiency versus LTE Network Performance", International Journal on Advances in Telecommunications, vol 3 no 3 & 4, year 2010

[5] E.G.Villegas, J.P.Aspas (advisor), "Self-optimization of Radio Resources on IEEE 802.11 Networks", PhD Thesis , Universitat Politecnica de Catalunya (UPC), July 2009

[6] M. Matinmikko, "US three-tiered authorization framework model overview", VTT Tecnhical Research Centre of Finland – Tekes Trial meeting, May 2014

[7] FCC, "Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550-3650 MHz Band", April 2015

[8] Y.L. Lee, T. C. Chuah, J. Loo, A.Vinel, "Recent Advances in Radio Resource Management for Heterogeneous LTE/LTE-A Networks" IEEE Communication surveys & tutorials, VOL. 16, NO. 4, Q4 2014

[9] P. Mogensen, W.Na, I.Z. Kovacs, F. Frederiksen, A. Pokhariyal, K.I. Pedersen, T. Kolding, K. Hugl, M.Kuusela , "LTE capacity compared to the Shannon Bound", Vehicular Technology Conference, April 2007

[10] H.E. Elmutasim, O.M.Elfadil, M.A.I Ali, M. Abas , "Fractional Frequency Reuse in LTE Networks" ,IEEE WSWAN, March 2015

[11] Mehbodniya, K. Temma, R.Sugai, W. Saad, I. Guvenc, F. Adachi, "Energy-Efficient Dynamic Spectrum Access in Wireless Heterogeneous Networks", IEE ICC 2015

[12] Y. Matsumuira "Interference-Aware Channel Segregation Based Dynamic Channel Assignment Using SNR-Based Transmit Power Control", IEICE Transactions vol e98 – 5, May 2015

[13] R. Matsukawa, T. Obara, F. Adachi, "A Dynamic Channel Assignment Scheme for Distributed Antenna Networks", IEEE, Vehicular Technology Conferenece, May 2012

[14] K.I. Pedersen, T. E.Kolding, F. Frederiksen, I. Z. Kovács, D. Laselva, P. E. Mogensen, "An Overview of Downlink Radio Resource Management for UTRAN Long-Term Evolution", IEEE, Communications Magazine, July 2009

[15] H.Kumar, D. Kumar, R. Srilakshmi, "Short term Spectrum trading in future LTE based cognitive radio systems", KSII Transactions on Internet and Information Systems Vol. 9, Jan. 2015

[16] P. Ameigeiras, J.N. Ortiz, P.A. Maldonado, J. M. Lopez-Soler,J.Lorca,Q.P. Tarrero, R.G. Perez, "3GPP QoS-based scheduling framework for LTE" , EURASIP journal, 2016

[17] T. Kemptner, T. Tsvetkov, "LTE SON-Function Coordination Concept", Seminar Innovative Internettechnologien und Mobilkommunikation SS2013, August 2013

[18] S.A. AlQahtani, M. Alhassany, "Comparing Different LTE Scheduling Schemes" , Wireless Communications and Mobile Computing Conference (IWCMC), July 2013

[19] S. Vassaki, A. Georgakopoulos, F. Miatton, K. Tsagkaris, P. Demestichas, Interference and QoS aware channel segregation for heterogeneous networks: a preliminary study (2016 European Conference on Networks and Communications (EuCNC), Athens, 2016), pp. 195–199

[20] A. Georgakopoulos, A. Margaris, K. Tsagkaris, P. Demestichas, Resource sharing in 5G contexts: achieving sustainability with energy and resource efficiency. IEEE Vehicular. Technol. Mag. 11(1), 40–49 (2016) 6. S.-J. Kim, I. Cho, Y.-K. Kim, C.-H. Cho, A two-stage dynamic

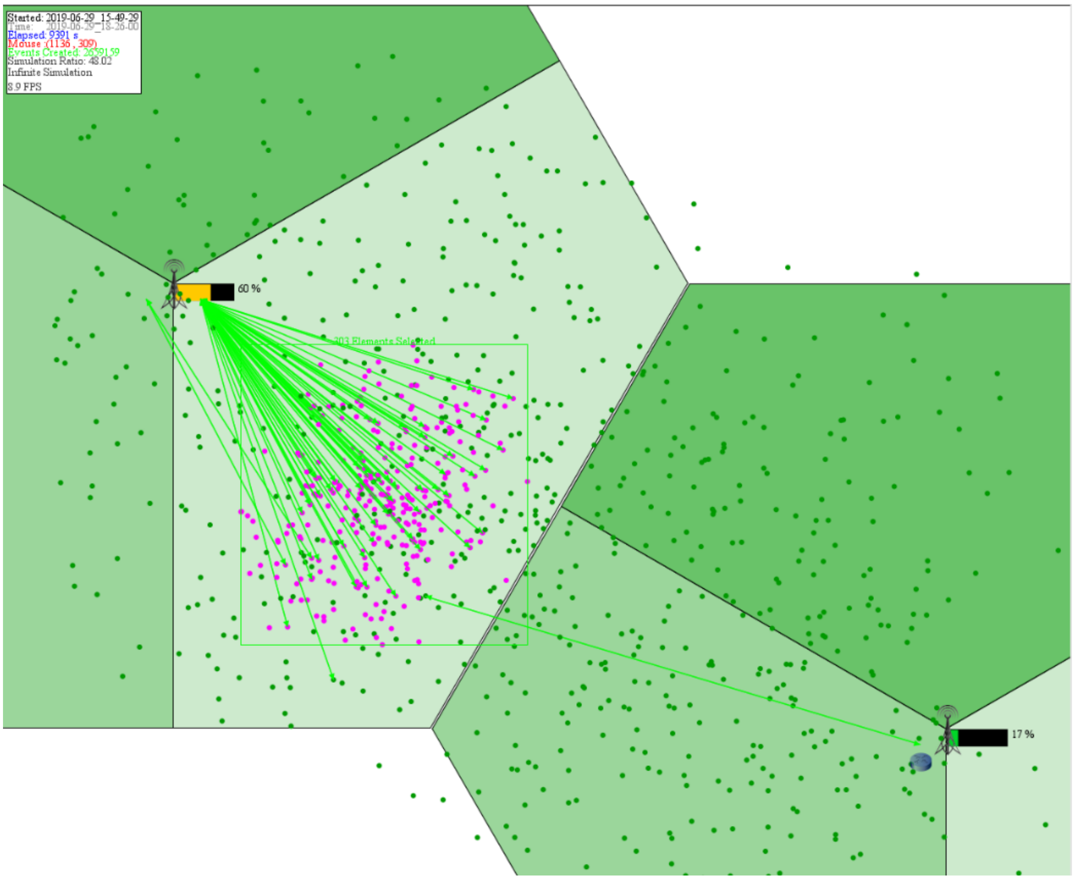
Chapter 5 – Congestion Prediction and Prevention using Machine Learning models

5.1. Introduction

Cellular network congestion is one of the most important problems in the infrastructure's lifecycle. Network congestion is the high usage rate of the network's resources which will result in a high interference low efficiency state. Planning techniques (used during the design of these networks) use estimates on the network's usage to provide resources and they fail to overcome the opportunistic, random and unexpected nature of the user's traffic. In addition, reactive SON-based approaches rely too much on control loops that have a very slow convergence rate (e.g. LB-SON). Key to the prevention of congestion is the existence of adequate management options that will be able to eliminate the congestion (by either resource increase or intelligent offloading) and the ability to have a constant probability of congestion or the means to predict incoming congestion events beforehand, therefore avoiding the problem from ever happening.

5.1.1 Simulation Scenario

For this chapter, we have designed a simulation scenario of a small isolated, 2-cell area that has constant user equipment motion. This can sometimes be part of a dense-urban city centre or an area near a city square / event area. Mobility and high user density in dense-urban environments is one of the key factors of congestion in many simulations.



```
Started: 2019-06-29_15-49-29
Time:     2019-06-29_18-26-00
Elapsed: 9391 s
Mouse :(1136 , 309)
Events Created: 2659159
Simulation Ratio: 48.02
Infinite Simulation
8.9 FPS
```

Figure 50 - Congestion Simulation Scenario

The users of the simulation are split into two categories: The ambient users, moving in a random manner around the playground of the simulation (marked as green circles in the graphics) and the concentrated users (marked as pink dots in the graphics) which move as groups would move in various social and other events (e.g. users coming out of transportation vehicles). For the traffic model of each of the two groups, we have selected the reference 2020 FTP traffic model used in the previous simulations. An addition is that the concentrated users traffic model is parametrized to have 100x increased small packet arrival rate. This is since these users have a lot more cell phone usage. Multiple scenario executions will be conducted in order to account for all different global traffic levels (20%, 40%, 100%, 120%, 140% etc.)

The outcome of this configuration is that the location in which the group of users selects to move results in congestion for the cells that are serving it as part of its coverage area. We can see this clearly in the measurement module of the simulator and the load indication on the screen. The congestion moves from one cell sector to another as a result of the high demand concentration.

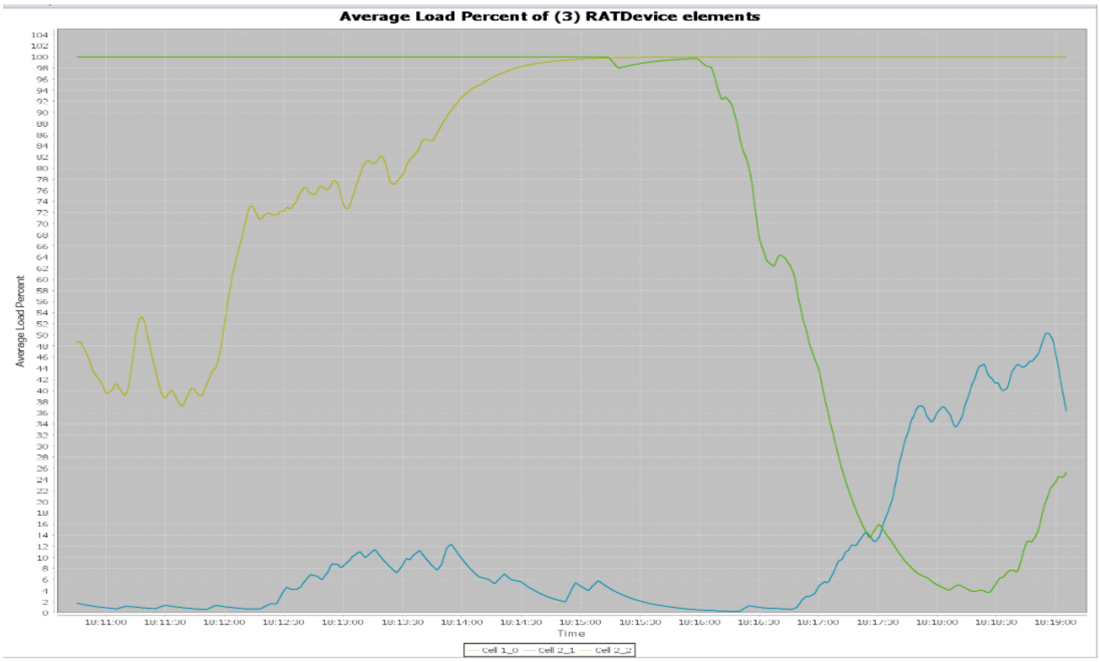


Figure 51 - Cell load congestion on the simulation scenario

For the other simulation scenario parameters, we are using the reference scenario for Dense Urban 2020 standard simulations[1][2] (including power models, antenna parameterization, spectral efficiency, MIMO mode, available LTE resources and Pico Cell technology).

5.1.2. Congestion Control mechanisms in HetNets

The mechanisms that can be included by management policies[3] in order to avoid network congestion are offloading mechanisms that are based on the mobility management subsystem. The idea is that a traffic spike can be handled by surrounding cell elements if proper parametrization of CIO and element bias values are changed (in real-time). These elements can either be eNodeB cells or Pico cells depending on the simulation scenario and the assumed area type. Parametrization of this sort will lead to active users being moved to less loaded cells and resolve the overload issue. This procedure can occur either in the inactive user terminal (using the relocation procedure) or in the RRC connected state (using the RRC reconfiguration procedure) also known as a handover. These mechanisms are traditionally activated by either manual configuration, in order to fix a specific mobility issue such as high traffic highways, or by activating a specific network KPI rule. These rules are simplistic threshold-based approaches that use either the current value, the change in the value or combination in order to trigger a change in the handover parameters. Whichever the case is, the problem is that the users experience a short but significant amount of negative radio quality because the algorithm is simply “reacting” to the problem. In order to prevent this, we need to apply a sense of predictive KPI modelling that will be used as a trigger for the same congestion control countermeasures.

5.2 Problem Solution

We have formulated an ensemble of machine learning models and network configuration as a means to intelligently prevent large amount of the congestion caused by the user equipment concentration. For this solution, we have studied thoroughly the existing SOTA literature for congestion predictive models (supervised and unsupervised), we have selected the best algorithm (namely the SOM , semi-supervised model) which is then trained in a collection of data from the selected simulation during both congestion and normal operation, and we have used the predictive model as an input trigger for the activation of congestion prevention counter-measures as described in previous chapter.

5.2.1 Unsupervised Machine Learning Models

Unsupervised machine learning models[5][6][7][8] is a sub-category of machine learning models that specialize in the discovery of hidden features and information in seemingly uncorrelated raw datasets. It usually involves a training process in which the model tunes its hyper-parameters based on a training dataset and an underlying optimization problem. Known unsupervised algorithms can be split into additional sub-categories such as distance-based clustering methods (K-means, X-means etc.), density-based clustering (DBSCAN, Optics etc.) and vector quantization algorithms such as Self-Organizing Maps and Growing Neural Gas. In the subsequent chapters we will focus more on the application of the vector quantization algorithms SOM and growing neural gas (GNG) which will allow us to build a robust, intelligent predictive engine for cellular simulated networks.

5.2.2 Growing Neural Gas (GNG)

In the literature of unsupervised learning techniques, vector quantization and clustering can also be achieved by using the growing neural gas algorithm[9]. Gas molecules tend to move to areas of lower pressure from zones of higher, and they tend to form links in cases of low distance between them. On the contrary high temperatures cause these links to break and their speed to increase. An artificial gas simulation also known as growing neural gas simulation can use this analog to move gas molecules into the most crucial spots of the shape of a data point cloud. The algorithm is transforming a number of random initial molecule points into a connected graph of vertices and edges that is capturing the "shape" of the underlying data (like it would capture the shape of the low-pressure container in the case of gas). This technique is very robust especially in higher dimensions that regular clustering techniques mostly due to the connections between molecules playing a role of gateways to different areas. The connections are also exhibiting gravitational forces and therefore achieving median values and even distribution of molecules. Various code implementation of the growing neural gas algorithm can be used to show that this algorithm can effectively reduce the size of a dataset into multiple orders of magnitude lower without losing information for decision making on the quantized data. In order to use this algorithm effectively as a clustering method, we need to identify the topological objects of our data. Various shapes that are distinguishable from one another or groups of objects that are linked can form cluster labels. Therefore, we can also extract information about what family of data points is the most fitting for the measurement of a new network element.

We have conducted experiments for the growing neural gas to see its strengths and weaknesses against various pathogenic n-dimensional datasets. For the data point generation, we are using N-dimensional shape distributions of particles and also the orthogonal interlocking rings dataset.

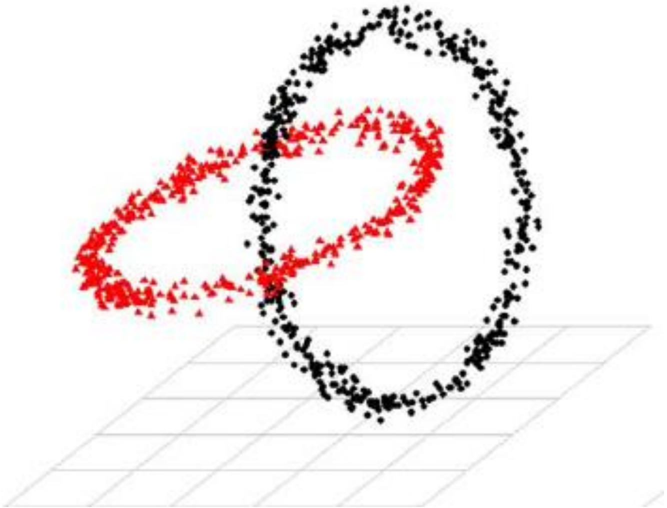


Figure 52 - 3D representation of the interlocking rings dataset[9]

This dataset consists of two groups of data points (2 rings or donut-shaped objects) that are linked, and they intersect from every direction of analysis. Distance-based clustering algorithms fail in this dataset because they do not perform a local search or agglomerative approach. The growing neural gas algorithm overcomes this by forming the rings itself using the gas molecule edges. These two rings are not in any contact in the 3-dimensional space and they are perfectly isolated in their own embedded surface.

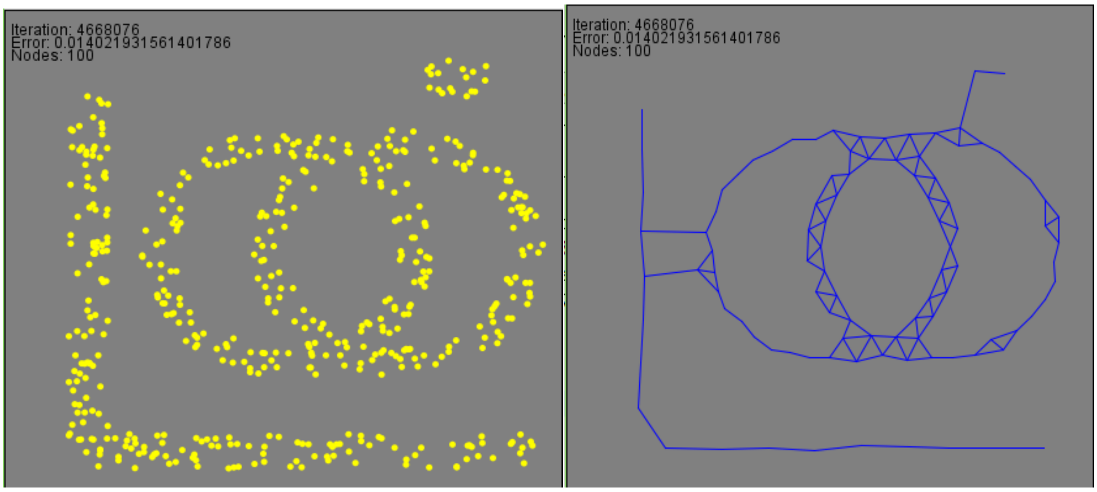


Figure 53 - a) Initial Dataset before algorithm execution b) GNG links identified

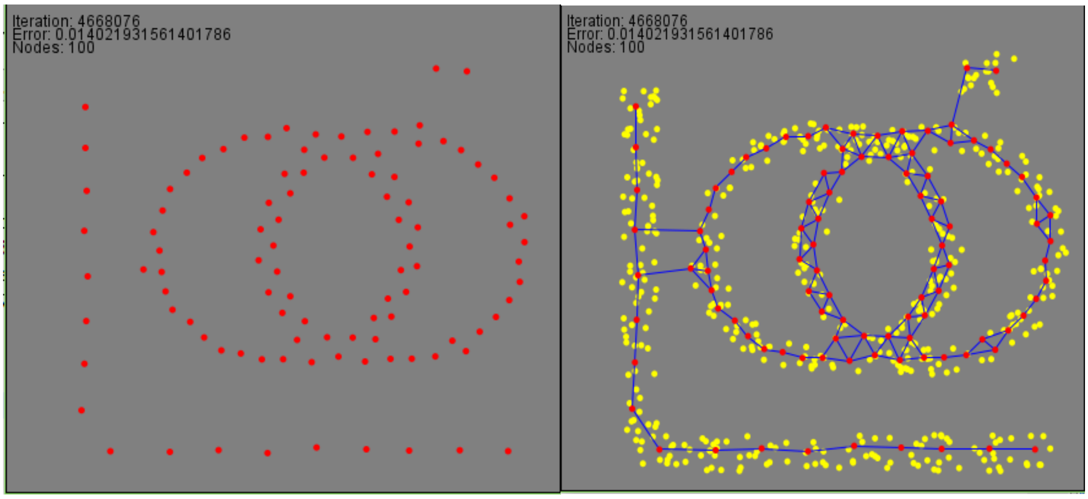


Figure 54 - a) GNG quantized data points b) GNG during execution

Measurements on actual HetNets may contain a lot of hidden information by the shape of their point clouds and the GNG algorithm can assist on identifying them effectively and using the result as an input for various management actions that will optimize the operation of the infrastructure. For a real-case application, the algorithm has a good efficiency and it is designed to work on real-time data with streams of new measurements constantly updating the topology of the underlying structure. This adds to its robustness and can accommodate for changes in the architecture and evolution of the HetNet which is a serious drawback for other unsupervised learning algorithms.

5.2.3 Self-Organized Maps

Self-Organized Maps is an unsupervised machine learning model[10][11][12][13][14][15][16][17][18] that specializes in the revelation of hidden structure behind collected data that can lead in the identification of meaningful data groups of hidden variables. This information can be crucial to understanding the underlying cause of a problem and its characteristics or it can even allow us to optimize curtain situations by using it as a predictive indication. SOMs are essentially the projection of a dataset in a 2-dimensional discrete grid of fixed dimension, with each different grid containing a vector value that is learned to be characteristic of the dataset. In order to find these vectors, a "hidden" simple neural network is optimizing the Euclidean distance KPI which results into the discovery of these key vectors. SOM projections also allow for easy optical detection of clustered data. A preprocessing pipeline is necessary on the data in order to acquire the maximum potential knowledge from the SOM algorithm, this preprocessing includes feature scaling and dimensionality control. The execution of the SOM algorithm begins with the random initialization of the discrete matrix. The collected data points from the real dataset are then "thrown" into the grid in a position that is closest to their Euclidean distance. This results in a "spreading" effect altering the values of the neighboring vectors. The locality of the SOM algorithm is creating a geography of the data while acting also as a noise-removing filter that focuses only on the essential part of the collected data. After enough iterations of the SOM algorithm, the structure of the data begins to take form and it can even be visual by 2D projection and coloring functions.

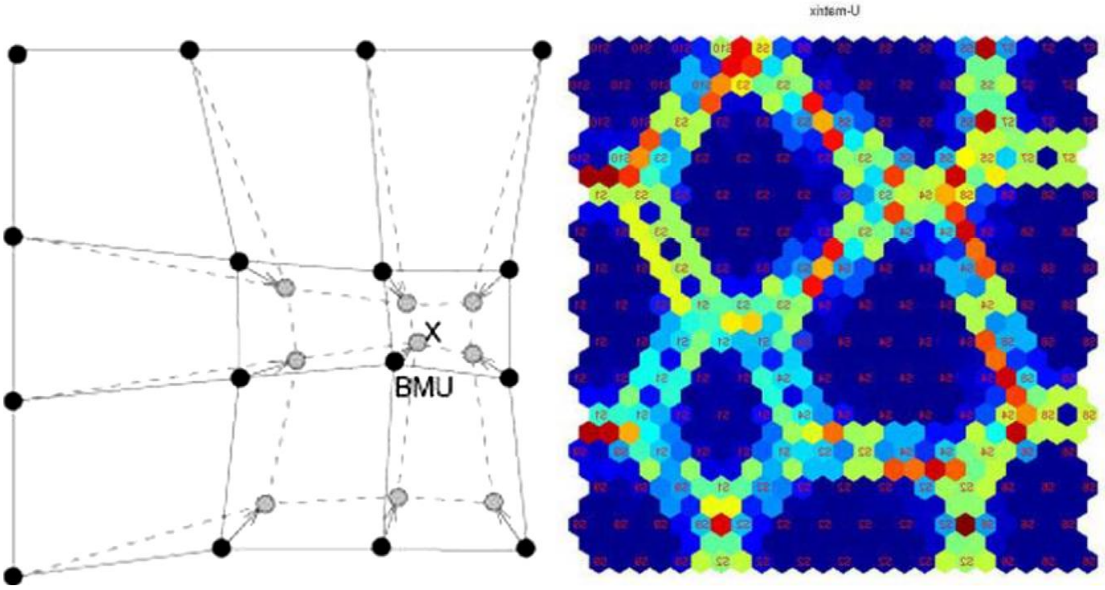


Figure 55 - a) SOM hidden neural network b) Color SOM map (example)[11]

5.2.4 Semi-Supervised Classification

In the existing literature[19][20][21][22], we have found that classification problems can be approached with a large variety of algorithms, each with its own benefits and drawbacks. Most of these algorithms is shown to have high sensitivity on the input dataset and particularly to the diversity of the input KPIs. Fixed feature length and sample count is causing a lot of models to require large customization in order to be used in this study. Dimensionality reduction techniques are shown to be selected as preprocessing steps, as much as vector quantization methods which reduce the input dataset to a smaller, more focused subset. Dimensionality reduction can vary from various mathematical operations (such as summation, averaging, various statistical properties) and also output of clustering techniques such as centroids, medoids and gas molecules. Semi-supervised models utilize benefits from both categories of ML models. It uses the existing structure and knowledge acquired from the unsupervised techniques to establish a geometry of the problem (i.e. a geometric expression of various hidden states of the system, some of which could be the problematic states) and then it uses a small samples of supervised data to label the sections into the prediction results. New data points are then placed inside the embedded geometry of the models, and their distance from the various supervised data points is used to determine the output of the classification algorithm

5.2.5 Cell Congestion Prediction using Semi-Supervised SOMs

In order to utilize the predictive capability of the semi-supervised SOM model, we need to plot a total workflow / lifecycle of the solution. As we see in the diagram (Figure 56), historical simulation data is being processed and fed into the predictive engine, there it is being filtered according to its correlations. The training of the SOM model occurs and then the Semi-Supervised model is formed based on the congestion samples acquired from the simulation. After that the model is ready for real time classification of values acquired from the live simulation.

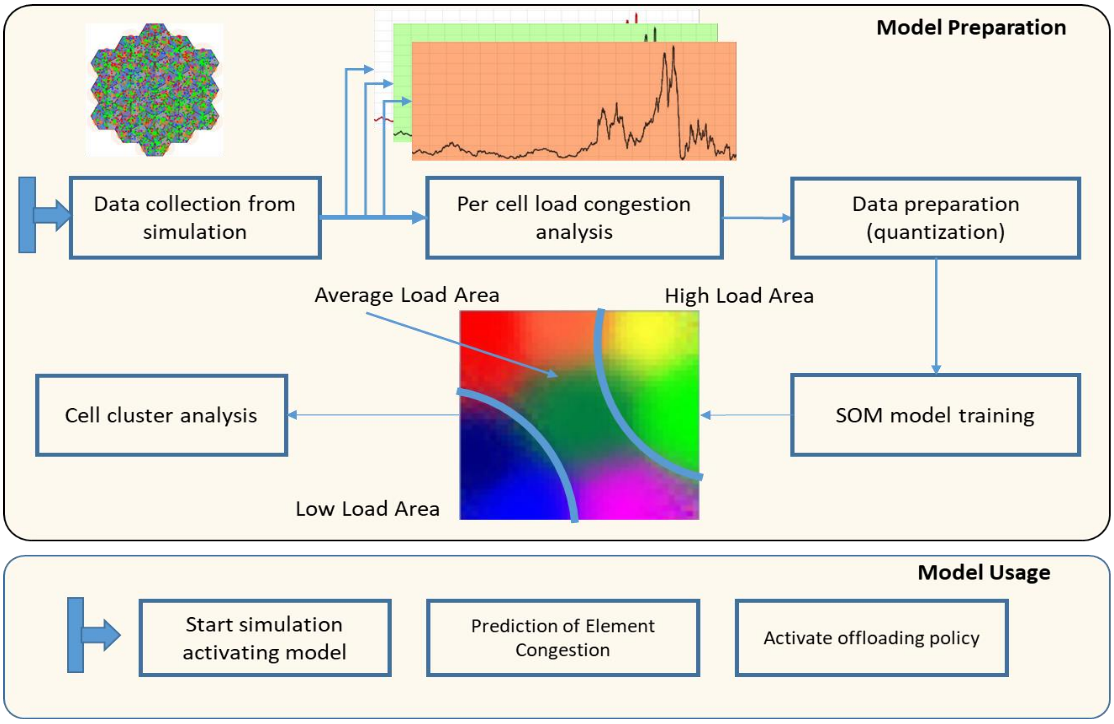

Model Preparation
Data collection from simulation
Per cell load congestion analysis
Data preparation (quantization)
Average Load Area
High Load Area
Cell cluster analysis
SOM model training
Low Load Area
Model Usage
Start simulation activating model
Prediction of Element Congestion
Activate offloading policy

Figure 56 - SOM usage methodology, from training to usage

5.3 Performance Evaluation

Results are split into three categories: A) SOM quantization and clustering plots during the training and after the finished training results. These will show us insights between the various simulated network KPIs and the congestion of the network. B) Predictive modelling results that is an isolated study for the accuracy of the semi-supervised classification subproblem of congestion prediction. C) Network KPI (simulation) results after applying the prediction as input for the trigger. There we will see the improvement or decline of the congestion rate KPI as a function of the reference and proposed algorithm.

5.3.1 SOM Predictive model metric correlation

After running an adequate amount of simulations in order to generate data, we then feed them to the SOM engine in order to generate the SOM maps. We see (Figure 57) that many implementations allow for probing on the training process to debug its effectiveness. After the first initial data points, small clustering of data occurs that quickly changes into more complex and intricate "valleys" and "mountains".

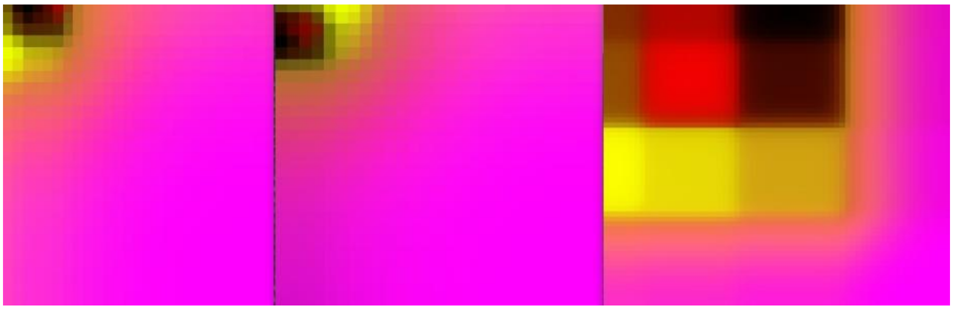


Figure 57 - Training Progress of the SOM map model[11]

We can swap the colouring (Figure 58) in order to show the shape that different metrics take.

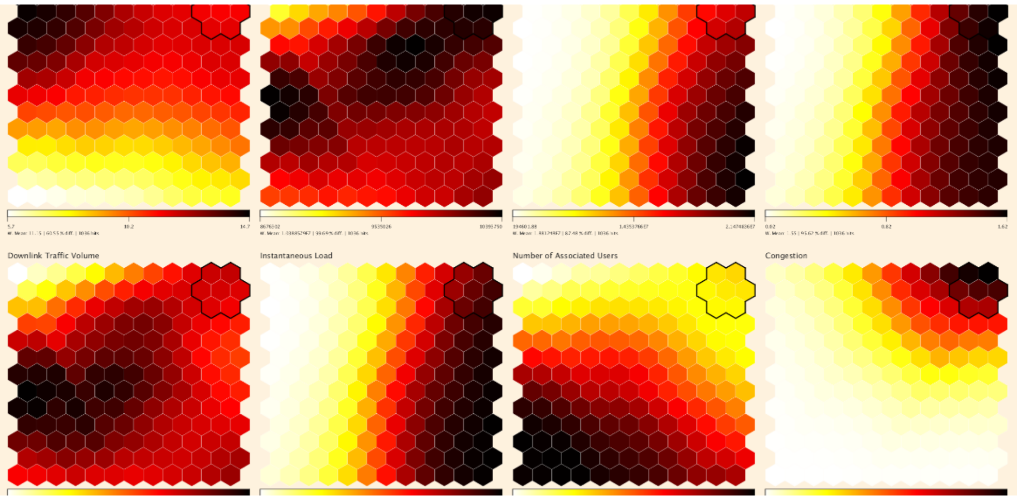


Figure 58 - Metric Correlation of congestion with other KPIs[11]

In order to understand the predictive model's insights, we need to visualize the U matrix of the target KPI (namely the "congestion") and highlight the areas with the highest value in order to see what corresponding values the other KPIs have. We see that congestion triggers with high edge user throughput, moderate average throughput, high number of users, high downlink traffic volume, very high instantaneous load and moderate number of associated users.

5.3.2 SOM Predictive model prediction accuracy

In order to evaluate the robustness of the SOM predictive model we need to use it to generate predictions for various traffic levels of the reference simulation scenario. In each case we count the time that the congestion label is predicted and then after a short period of time, congestion occurs. As we see from the overall collection of the evaluation results (Figure 59), it is evident that the predictive model accuracy diminishes on the validation set as the traffic increases. This can happen because of many reasons but the most apparent is that the high volatility and instability that the high traffic demand introduces to the system changes the statistical mechanics of the congestion.

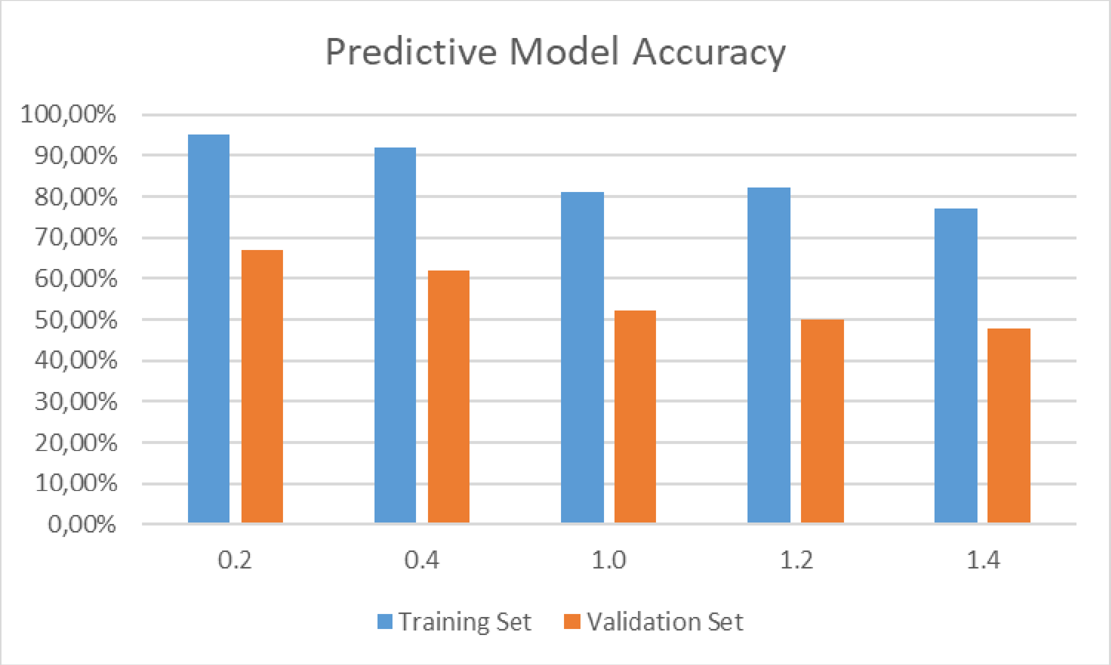


Figure 59 - SOM predictive model accuracy per load level

5.3.3 Network KPI results on simulation

The final set of results is the effectiveness of the end-to-end solution (Figure 60). Predicted congestion labels trigger activation of the offloading mechanism which in turn reduces the congestion rate of the simulation. This is being compared to the reference, threshold-based algorithm in order to see if there are any benefits in the proposed solution. We can see that in every case of load level, the predictive congestion avoidance solution outperforms the reference algorithm by a factor that varies from 5 to 2 (from low to high load levels respectively). This is a significant improvement of the system’s stability as it means that congestion is avoided even in this extreme case of sudden traffic spikes.

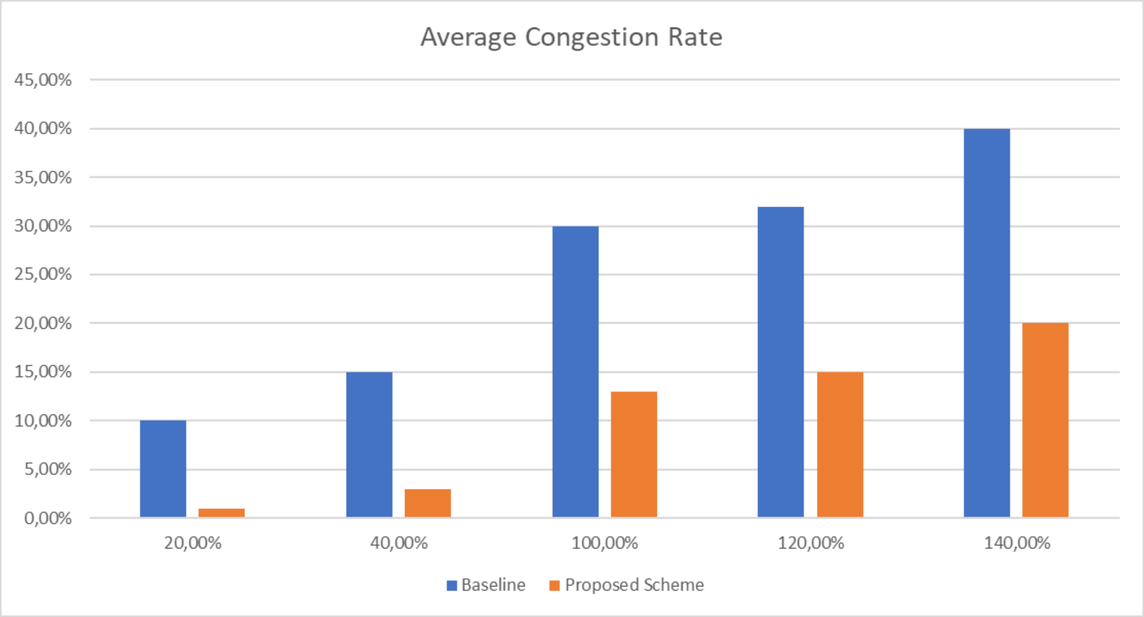


Figure 60 - Network congestion KPI measured for different load levels

5.4 Conclusion

In this chapter we have studied the advanced network optimization methodology of applied predictive modelling for network congestion prevention. This involved the detailed process of composing, implementing and training such predictive models in real measurements of a simulated 3GPP-based cellular network. We have also formulated a mechanism to incorporate the predictive output of such a model into a SON control loop that utilizes the future indication of a congestion as a trigger for network load reduction actions. We prove that the early indication of incoming congestion is a key aspect in improving the network stability aspect of the system which reduces the overall need for resource overprovisioning and therefore makes network elements cheaper and energy efficient. We also have shown that faulty assumptions and incomplete training is a pitfall that can lead into false prediction results and low effectiveness of the process. Having evaluated thoroughly the overall outcome of the predictive modelling we conclude that they can be an important addition in the future cellular communication technologies management plane.

5.5 Chapter References

[1] Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects, 3GPP TR 36.814

[2] Evolved Universal Terrestrial Radio Access (E-UTRA); Study on multiple radio access technology (Multi-RAT) joint coordination 3GPP TR 37.870 V13.0.0, June 2015.

[3] Bantouna, V. Stavroulaki, Y. Kritikou, K. Tsagkaris, P. Demestichas, K. Moessner, "An overview of learning mechanisms for cognitive systems", Journal on Wireless Communications and Networking ,2012

[4] A.Bantouna, G. Poulios, K. Tsagkaris, P. Demestichas, "Network load predictions based on big data and the utilization of self-organizing maps", submitted for publication to the Journal of Network and Systems Management, 2013, Journal Papers

[5] T. Kanungo, N.S.Netanyahu, C.D.Piatko, R. Silverman, A.Y.Wu , "An efficient k-Means clustering algorithm analysis and implementation", IEEE transactions on pattern and machine intelligence, vol. 24, NO. 7, July 2002

[6] L.Yingqiu, L. Wei, L. Yunchun, "Network Traffic classification using K-means Clustering", Computer and computational sciences conference, 2007

[7] P. Sasikumar, S. Khara, "k-MEANS CLustering in wireless sensor networks", Fourth International Conference on Computational Intelligence and Communication Networks, 2012

[8] S. Datta, C.Giannella, H. Kargupta, "K-means clustering over a large, dynamic network", Proceedings of the 2006 SIAM International Conference on Data Mining, 2006

[9] J. Holmström, "Growing Neural Gas – Experiments with GNG, GNG with Utility and Supervised GNG", Master Thesis, Uppsala University, Department of Information Technology Computer Systems, August 2002

[10] T. Sarazin, H. Azzag, M. Lebbah, "SOM clustering using Spark-MapReduce", IEEE 28th International Parallel & Distributed Processing Symposium Workshops, 2014

[11] P. Stefanovic, O. Kurasova, "Visual analysis of self-organizing maps", Nonlinear Analysis: Modelling and Control, Vol. 16, No. 4, 488–504, 2011

[12] M.L. Westerlund, "Classification with Kohonen Self-Organizing Maps", Soft computing, Haskoli Islands, April 2005

[13] K. Tsagkaris, A. Bantouna, P. Demestichas,"Self-Organizing Maps for advanced learning in cognitive radio systems" , Computers and Electrical Engineering 38(862–881), 2012

[14] P. Sukkhawatchani, W. Usaha, "Performance Evaluation of Anomaly Detection in Cellular Core Network using Self-Organizing Map", Proceedings of ECTI-CON, 2008

[15] K. Raivio, O. Simula, J. Laiho, P. Lehtimaki, "Analysis of mobile radio access network using the self-organizing map", Integrated Network Management. IFIP/IEEE Eighth International Symposium, 2003

[16] E. Bodt, M. Cottrell,P. Letremy, M. Verleysen, "On the use of self-organizing maps to accelerate vector quantization", Neurocomputing volume 56, January 2004

[17] D.Pratiwi, "The use of self-organizing map method and feature selection in image database classification system", International Journal of Computer Science Issues (IJCSI), May 2012, Vol. 9 Issue 3

[18] M.I. Chacon, P.Rivas-Perea, " Performance Analysis of the Feedforward and SOM Neural Networks in the Face Recognition Problem", Proceedings of the IEEE symposium on computational intelligence in image and signal processing (CIISP), 2007

[19] Z. Ghahramani, "An introduction to hidden markov models and Bayesian networks", Journal of Pattern recognition and Artificial Intelligence, 15(1):9-42, 2001

[20] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian Network Classifiers", Machine Learning 29, 131–163, 1997

[21] D. Grossman, P. Domingos, "Learning Bayesian Network Classifiers by maximizing conditional likelihood", ICMLProceedings of the twenty-first international conference on Machine learning, 2004

[22] K. Tsagkaris, A. Katidiotis, P. Demestichas, "Neural network-based learning schemes for cognitive radio systems", Computer Communications Volume 31, Issue 14, May 2008

Chapter 6 – Identifying groups of Network Elements and User Types by the means of clustering techniques

6.1 Introduction

Effective management of HetNet infrastructure can sometimes means successful grouping and simultaneous configuration of connected or similar network entities. As the size of a network rises, and multiple devices are being placed in various location of the serving area, it becomes harder and harder to precompute which network elements will behave in the same manner and exhibit the same type of performance incidents. In order to improve this, we need to introduce a methodology that will allow for a data-driven (measurement-driven) grouping of network elements that will correspond to same area types and same requirements for resources and/or management. These techniques (namely unsupervised learning or simply 'clustering') will be applied into two different perspectives of a telco infrastructure dataset. In the first case it will be used to identify groups of serving network elements (like mentioned before), and in the second case it will be used to identify different groups of users that have different requirements for data access.

6.1.1 Simulation Scenario 1 – Cell groups

For the first case we have developed a simulation scenario of a number of cell coverage elements that are serving the same amount of user equipment devices – generating a uniform amount of downlink traffic. This can be the case for random locations of the urban / dense urban area as different population densities can be identified in various zones. This is not always possible to include in the planning phase of these networks due to the constant evolution of the substrate area.

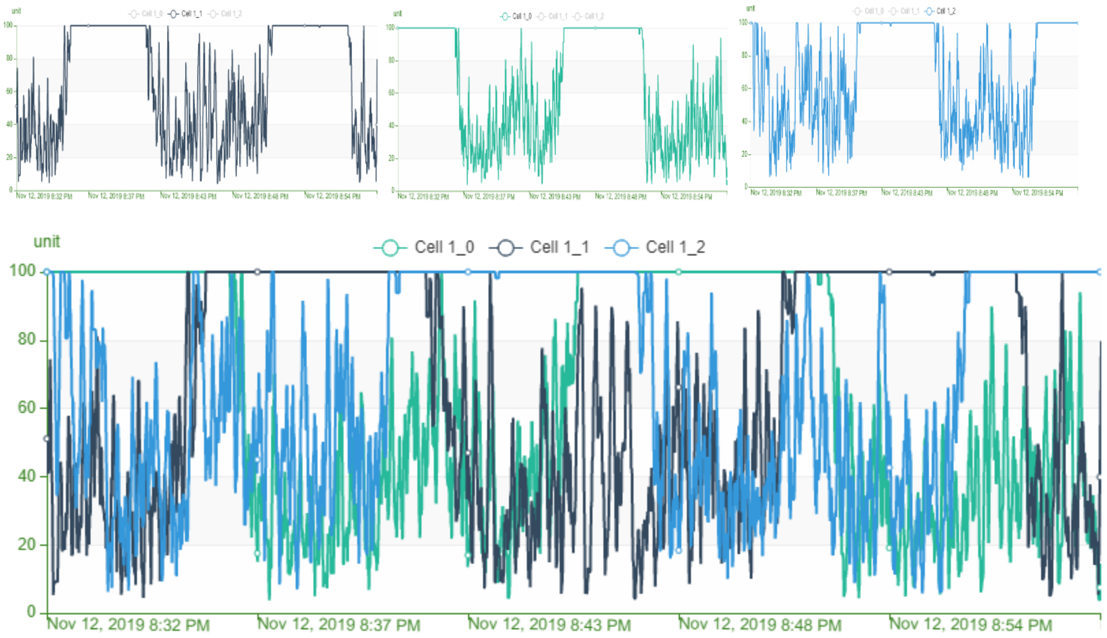


Figure 61 - Cells with similar network load KPI

6.1.2 Simulation Scenario 2 – User equipment device groups

For the second simulation scenario, we are focusing on the automated identification of users with different traffic demand profiles. Leveraging on the capabilities of the variable traffic model of the network simulator, we are separating the users into different categories of broadband access. Global management of the hole will not be as effective as a customized strategy on the ways to handle the resource allocation on different classes of users. Here we will be applying the clustering algorithm in order to identify

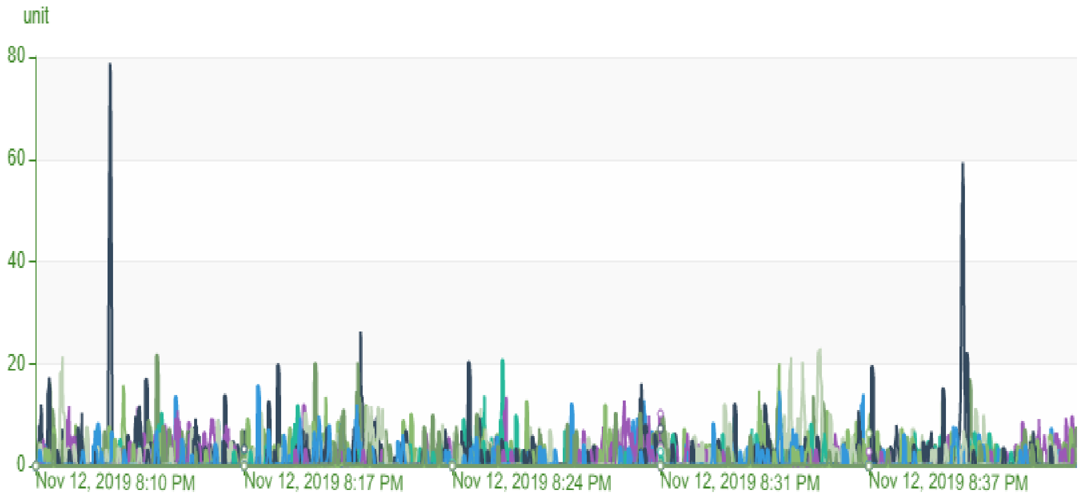


Figure 62 - Different groups of user equipment devices

6.2 Problem Solution

Rich literature exists[1][2][3] in the vast field of data clustering and unsupervised learning as seen in multiple publications and applied implementations. The generated dataset from the proposed simulation scenarios follows a generic structure that makes it possible to apply different approaches from different fields of the machine learning community. The first category of algorithms that we will discuss are Euclidean-distance based clustering techniques (e.g. K-means and X-means) that will characterize a set of measurements from a network element based on their geometric Euclidean distance

6.2.1 Data clustering using distance-based algorithms (K-means, X-means)

The baseline clustering algorithms of the literature are the K-means and its variation X-means distance-based algorithm[4][5][6][7][8][9]. It is an algorithm for identifying geometric centres of clustered data points for n-dimensional datasets that will then be used to characterize them. These centres, namely centroids, are the main characteristics of all the members of the cluster. The k-means algorithms require a prior selection of the number of clusters that it will generate. This is described to be a weakness as it is not always apparent from the use case that the data will have a known number of point clusters. Extensions of the k-means algorithms such as the X-means algorithm is using an internal optimization process to automatically determine the optimum number of clusters in a dataset.

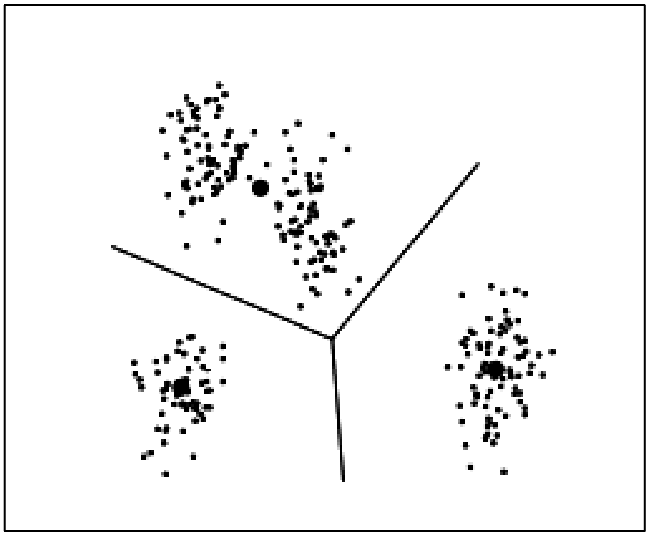


Figure 63 - distance based clustering[4]

In general – distance-based clustering algorithms suffer from the same types of drawbacks all revolving around non-linearly separable groups. Since the centroid is a point that is used to group the particles together, the separation hyperplanes are required to be part of the linear sub-space. This means that non-linear groups cannot be separated in an effective manner. Specific datasets[8][9] exist in the literature and illustrate such drawbacks.

6.2.2 Data clustering using density-based algorithms (Optics, DBScan and Gaussian Mixture Models)

In this category of clustering algorithms, the authors[10][11] are trying to identify the clusters of data points by a degree of reachability between various location of the dataset. This is solving the linear hyperplane problem introduced by the distance-based clustering family and can help with identifying more complex clusters. Sensitivity hyperparameters will dictate whether two nearby data points are connected and therefore belong to the same cluster. Another important aspect of these clustering algorithms is that they have an additional feature - the outlier cluster. Data points that get isolated and are found outside of the range of other clusters are marked as noise and being designated to the noise cluster.

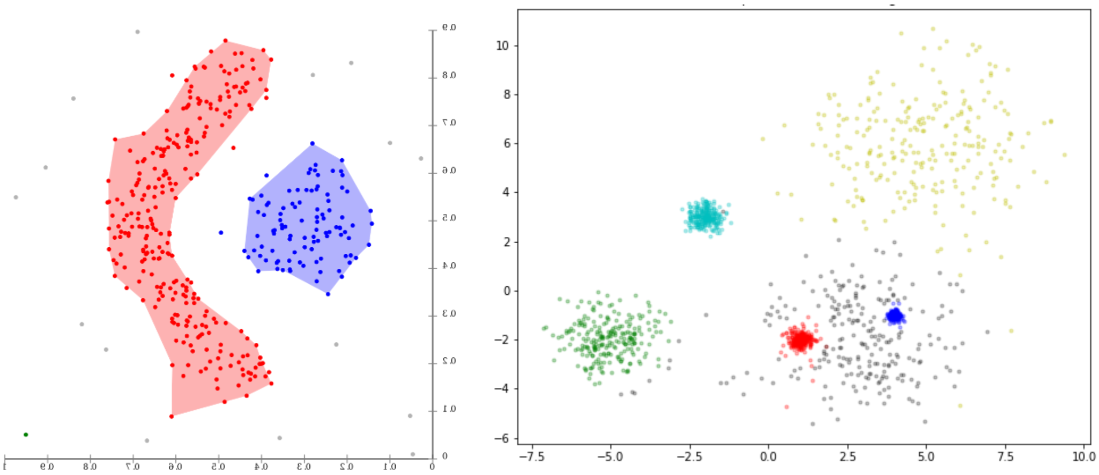


Figure 64 - Clustering using DBScan and OPTICS[10]

Gaussian mixture model clustering is based on gaussian distribution identification in the underlying data. It is implying that each subgroup of data points belong to different distributions (with other median and standard deviation) and therefore it can be separated and distinguished.

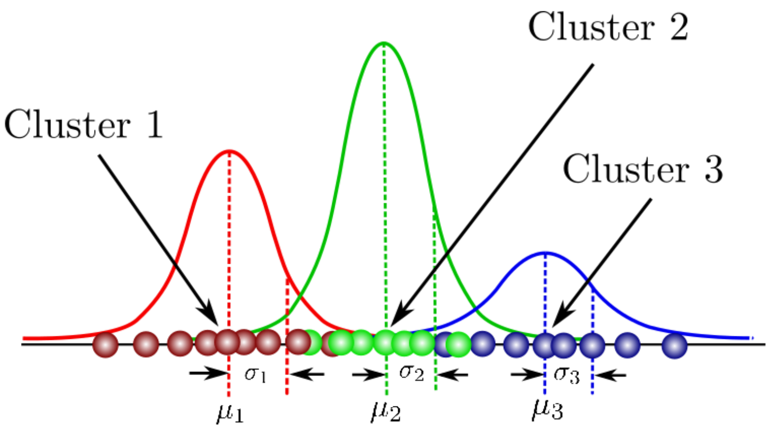


Figure 65 - Gaussian Mixture Model clustering[11]

6.2.3 Clustering and dimensionality reduction techniques

Clustering of data points is a process that is very sensitive to the high dimensionality of the dataset. In the cases of cellular network datasets, the large amount of measurements performed in various layers of the network such as traffic volume, packets, types of packets, alerts and alarms, procedure counters, indicators and other KPIs leads to data points of high dimension. Dimensionality reduction techniques can be used in two different ways with this type of dataset. The first case is at the input dataset pre-processing stage, were by applying the dimensionality reduction operation in order to acquire a more robust and accurate clustering result mostly due to the cleansing effect towards noise and correlated dimensions. In the second case, we are using the dimensionality reduction as a tool to debug or to make sense of the output of the clustering algorithm

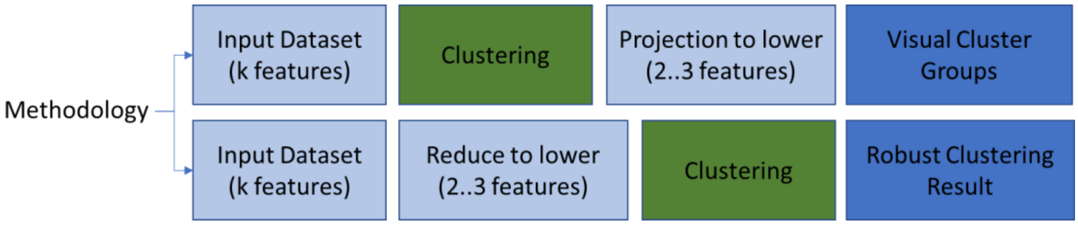

Methodology
Input Dataset (k features)
Clustering
Projection to lower (2..3 features)
Visual Cluster Groups
Input Dataset (k features)
Reduce to lower (2..3 features)
Clustering
Robust Clustering Result

Figure 66 - Methodologies of dimensionality reduction in clustering

6.2.3.1. PCA / ICA projection

Principal component analysis [12][13]is a method for projecting multi-dimensional datasets into smaller dimension (usually 2 or 3) capturing the maximum possible information in the form of colour distinction and separability. For this purpose, both PCA and ICA are trying to produce composite dimensions that satisfy different statistical measures of non-likelihood between the multi-dimensional samples. In PCA, various implementation revolving the co-variance matrix of the dataset such as SVD (singular value decomposition) result in dimensions that are separated based on the maximum variability of the vector values. In ICA, the notion of independence (meaning statistical independence) is used in the form of mutual information gain and entropy loss. The process aims at finding the appropriate transformation to provide the most accurate projection on an independent space or plane.

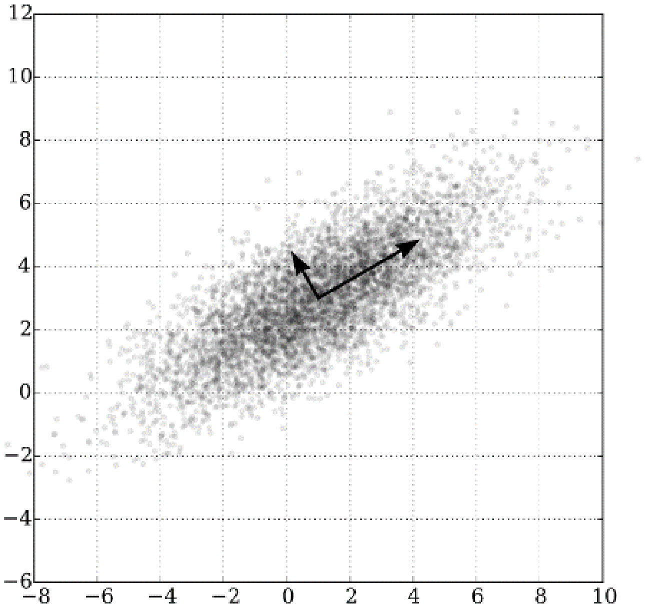


Figure 67 - Projection of multi-dimensional data into a 2-dimensional plane using PCA[13]

6.2.3.2 t-SNE projection

More advanced techniques of dimensionality reduction[14] for identification of data clusters have been developed that take advantage of statistical properties of that dataset in embedding space. Namely the Student's T distribution stochastic neighbourhood embeddings algorithm is such an implementation that is transforming a multi-dimensional dataset into a low dimension projection (the output dimension can sometimes be 2 or 3) by minimizing the Kullback-Leibler divergence distance between conditional probability distribution of the data points.

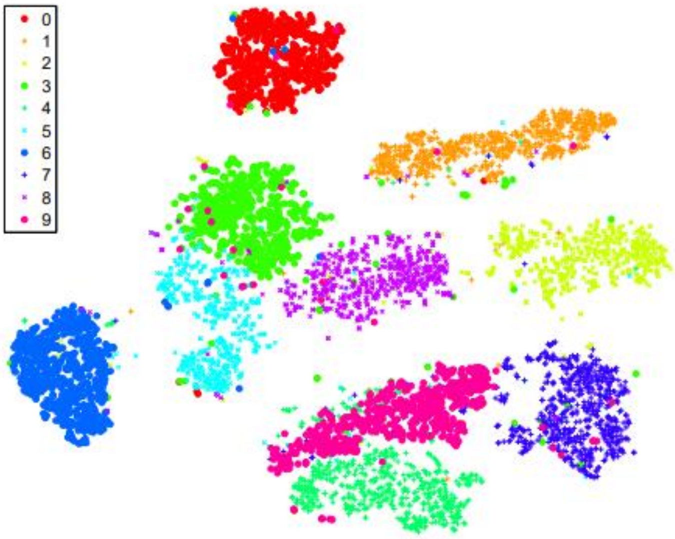


Figure 68 - Illustration of the 2D projection capabilities of t-SNE [14]

Applications of the tSNE algorithm have shown linearly separable clusters of data points in famous, complex dataset such as the MNIST dataset (consisting of a large number of hand-written digits).

6.2.4 Time Series Pattern clustering

Time series pattern clustering is a special family of clustering techniques that rely on different pre-processing for the execution of the clustering operation. In particular the idea is that data points should not group together based on their value but on their time evolution and shape of their evolution. In HetNets this is especially useful because of the various heterogeneous elements that they are being used. Small cells and eNodeBs operate in different network metric scales due to their different capabilities. However, in various use cases, they all exhibit temporal and periodic phenomena that are wanted to be identified and grouped together. Pattern clustering is using entity-wise scaling for the data points (using the min and max of the observed values for each element) and therefore it brings the data point vectors into a comparable scale. It can then work with any of the aforementioned clustering algorithms to produce different results than the original, total max min scaling that is being performed.

6.3 Performance Evaluation

The selected algorithms will be compared in their respective accuracy on grouping together the network elements and user equipment devices that have originated from the same group in the simulation environment. In case 1 – network element clustering – we are expecting the clustering algorithms to provide three different clusters for each different user equipment density zones generated in the simulation environment. In case 2 – user equipment devices clustering – we are expecting the clustering algorithms to provide us with user equipment devices clusters that correspond to four different classes of application usages. For the experiment we will be evaluating the number of cells assigned to their correct cluster. This will form the MAPE KPI (mean absolute percent error). We will average this KPI over all classes in order to create a weighted total MAPE

6.3.1 Network Element Clustering Performance Results

For the case of network element clustering, the simulation consists of 42 network elements split into three different clusters of different areas (low, medium, and high network usage).

Algorithm	*Low Cluster*	*Medium Cluster*	*High Cluster*	*Total*	*Noise*	*Low MAPE*	*Mid MAPE*	*High MAPE*	*Total MAPE*
Simulation	24	12	6	42	0				
K-means	15	24	3	42	0	-37,50	100,00	-50,00	154,17
X-means	21	18	3	42	0	-12,50	50,00	-50,00	79,17
Dbscan	25	11	5	41	1	4,17	-8,33	-16,67	18,06
Optics	24	10	6	40	2	0,00	-16,67	0,00	16,67
GMM	11	25	2	38	4	-54,17	108,33	-66,67	184,72

Table 6 - Network Element Clustering Benchmark

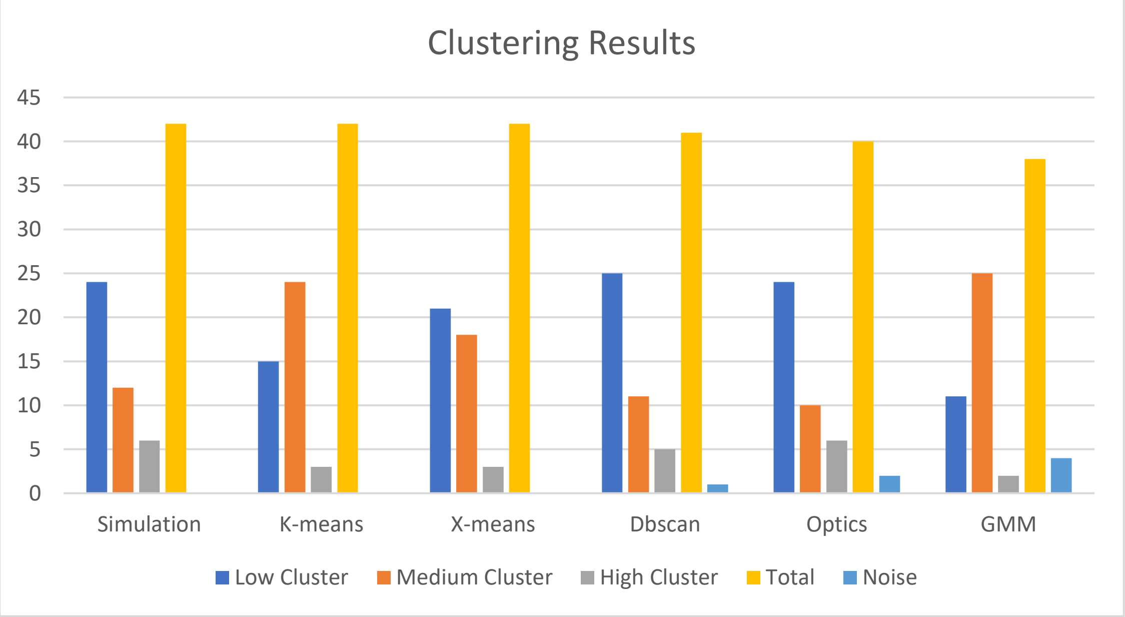


Figure 69 - Evaluation of clustering per algorithm used

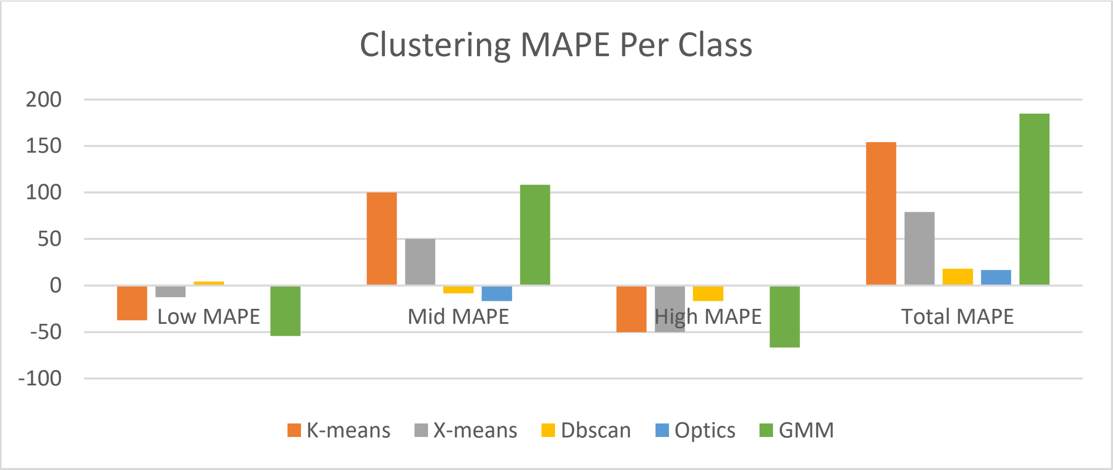


Figure 70 - MAPE per class per clustering method

For the case of network element clustering, the results of the evaluation indicate that the density-based methodologies, such as DBSCAN and OPTICS provide the best results for identifying hidden groups or subgroups of network elements. This is indicated in both the clustering per class and also the average MAPE of each class.

6.3.2 User Equipment Clustering Performance Result

For the user equipment device clustering approach, we are analyzing the result of identifying 4 different classes / levels of network usage.

Algorithm	*Low Cluster*	*Medium Cluster*	*High Cluster*	*Very High*	*Total*	*Noise*	*Low MAPE*	*Mid MAPE*	*High MAPE*	*Very High MAPE*	*Average MAPE*
Simulation	250	150	50	20	470						
K-means	215	161	59	35	470	0	14	-7,33	-18	-75	28,58
X-means	225	171	53	21	470	0	10	-14	-6,00	-5,00	8,75
Dbscan	195	180	45	19	439	31	22	-20	10,00	5,00	14,25
Optics	194	133	52	19	398	72	22,4	11,33	-4,00	5,00	10,68
GMM	201	174	71	14	460	10	19,6	-16	-36	30,00	25,53

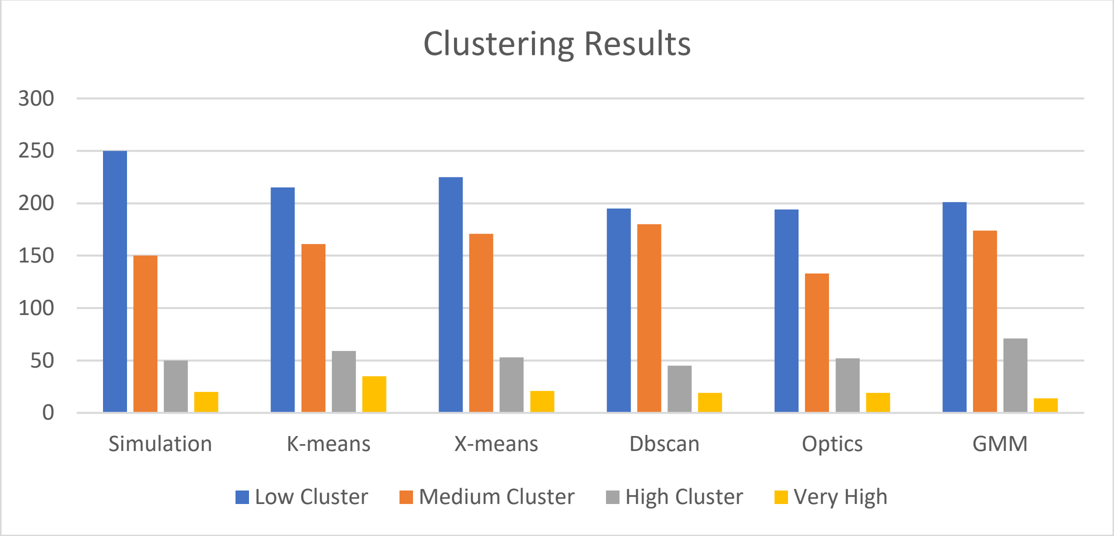


Figure 71 - Clustering results for UE Devices per algorithm

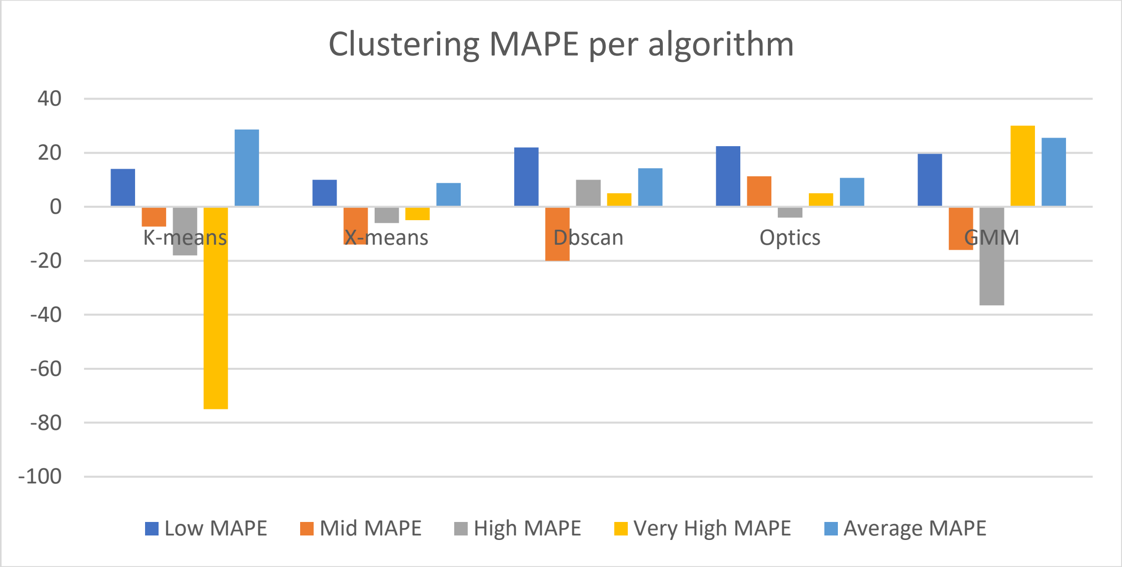


Figure 72 - Clustering MAPE per algorithm

In this case, we see that the overall performance of most algorithms is good, even for the cases of the Euclidean-distance-based algorithms (k-means, x-means). However, the best algorithm for identifying the user equipment device groups is shown to be the X-means algorithm yielding results very close to the ones of the optics.

6.4 Conclusion

Automated identification of network elements with the same behavior with reliable accuracy is a crucial asset to extend the capabilities of cellular networks. In this study we have selected two different cases of network elements clustering: serving element clustering (namely cells and Pico cells) and user devices clustering (split into different broadband access demand categories). For each case, we have exhausted the hyperparameter tuning and multiple implementations and found that both density-based and hyperplane-based approaches work for different problems. The solution for applying these methodologies with accurate results is to take into consideration all algorithms when coming up with a decision to group together elements and apply management actions.

6.5 Chapter References

[1] Mullner D. "Modern hierarchical, agglomerative clustering algorithms", CoRR, abs/1109.2378, 2011

[2] Rani U., Sahu S., "Comparison of Clustering Techniques for Measuring Similarity in Articles", IEEE CICT, 2017

[3]M. Hajjar, G. Aldabbagh "Using Clustering Techniques to Improve Capacity of LTE Networks", Proceedings of APCC2015 ,2015 IEICE

[4] S.P.Singh , Yadav A., "Study of K-Means and Enhanced K-Means Clustering Algorithm", International Journal Of Advanced Research In Computer Science, 4 (10), Sept–Oct, 2013, 103-107

[5] T. Kanungo, N.S.Netanyahu, C.D.Piatko, R. Silverman, A.Y.Wu , "An efficient k-Means clustering algorithm analysis and implementation", IEEE transactions on pattern and machine intelligence, vol. 24, NO. 7, July 2002

[6] L.Yingqiu, L. Wei, L. Yunchun, "Network Traffic classification using K-means Clustering", Computer and computational sciences conference, 2007

[7] P. Sasikumar, S. Khara, "k-MEANS Clustering in wireless sensor networks", Fourth International Conference on Computational Intelligence and Communication Networks, 2012

[8] S. Datta, C.Giannella, H. Kargupta, "K-means clustering over a large, dynamic network", Proceedings of the 2006 SIAM International Conference on Data Mining, 2006

[9] Oyelade, O.J, Oladipupo, O.O, Obagbuwa, I.C., "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", IJCSIS, vol7, No 1, 2010

[10] Yifeng Z., Kai L., Xueting X, Huayu Y., Lianfen H. "Distributed Dynamic Cluster-Head Selection and Clustering for Massive IoT Access in 5G Networks ", Applied Sciences, Vol 9. Issue 1, 2019

[11] Zhiqiang W., Cunha C., Mathieu R, Benoit F. "Comparison of K-means and GMM methods for contextual clustering in HSM", Procedia Manufacturing, Elsevier 2019, International Conference on Changeable Agile, Reconfigurable and Virtual Production (CARV)

[12] M. Üzümcü, A.F. Frangi, M. Sonka, J.H.C. Reiber, B.P.F. Lelieveldt "ICA vs. PCA Active Appearance Models: Application to Cardiac MR Segmentation ", Springer MICCAI 2003, LNCS 2878, pp. 451–458, 2003.\

[13] K. Baek, B. A. Draper, J. R. Beveridge, K. She, "PCA vs. ICA: A comparison on the FERET data set", Proceedings of the Joint Conference on Information Sciences, Vol 6., 2004

[14] L.Maaten, G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, vol 9, 2008

Chapter 7 – Forecasting of HetNet network KPIs

7.1 Introduction

In various aspects of design, operation and management for HetNets, an accurate forecast of the key network metrics is crucial for many operational decisions. Network forecasts can be used for fault prediction, congestion avoidance, future planning or adaptation / deprecation of new /old technologies in the complex stack of HetNet and Ultra-Dense deployments. Forecasting algorithms of different families and specifications exist in the literature of engineering, financial analysis and networking and each has its own benefits and also limitations as to its usage. It is the purpose of this chapter to study how these algorithms can fit into cellular hourly and daily KPI evolution prediction in order to further strengthen optimization algorithms that take such KPIs as input. This will result in "predictive" flavour of Planning and SON functions that will provide better results than reactive approaches.

7.1.1 Time granularity

HetNet KPI measurements are generally conducted in a continuous automated manner therefore resulting into very large amounts of data stored in the system. These measurements are then aggregated in higher granularities in order to maintain a meaningful history. The aggregation of these KPIs is conducted by gathering measurements of a specific time window (which can sometimes be 5,15,30 minute or 1,3,6,12,24 hour) depending on the system. It is in the nature of HetNets and cellular networks in general that different phenomena and patterns will be visible in different time granularities. Before conducting any forecasting operation, we need to make sure that our input data has the wanted "resolution" in order to identify re-occurrences of curtain patterns and outcomes. In general, cellular network traffic exhibits a number of expected behavioural patterns in relation to the time. Daily profiles are daily patterns that show us how the network load changes in the different "zones" of the day. It usually has one pattern for working days (Monday to Friday) and a different pattern for the weekend (Saturday-Sunday). Weekly profiles change only during holidays seasons (e.g. Christmas and Easter). Monthly and seasonal profiles are also highly correlated with the holidays season and especially in areas with high summer tourism (e.g. Greece, Italy) where a very large number of visitors are gathered in various zones such as recreational areas, parks, beaches etc. All these different perspectives of looking at network data provide us with different network KPI waveforms due to the aggregation (either averaging or summing, depending on the KPI).

7.1.2 Entity granularity

Another important aspect of input data selection for forecasting is the entity or element-wise granularity of measurements. In HetNets and general in large scale network infrastructures, we have a rich hierarchical ladder of interconnected components. A measurement of the various network KPIs is generated in the lowest of granularities but due to the large frequency of measurements and the need to keep large history, it is then being aggregated to a point higher in the network hierarchy. For 4G/LTE components this hierarchy follows this flow of information: Radio unit (namely the antenna that is serving an area), LTE Cell (consisting of various radio units , especially in MIMO or CoMP deployments) , LTE Site (consisting of a number of LTE cells that can either correspond 1-1 with a radio unit or include more), LTE location Area Code consisting of a number of LTE sites, LTE prefecture which consists of a number of area codes and finally total area (e.g. Athens, Chania, Argolida) consisting of a number of area codes and their respective sub-components. Other entity aggregation capabilities consist of grouping cell elements together with their vendor, their technology attributes such as M|IMO configuration, or their respective area types such as Dense Urban, SU, RU.

7.1.3 Aggregated Prediction

In order to predict KPIs for the aggregation of a set of elements, two approaches can be made each with their respective pros and cons. The agglomerative approach means that we are performing a forecast for the lowest of granularities for each element, and then we are performing aggregations in either time or entity in order to scale the prediction up. Performing forecasting in the lowest granularity requires more intricate and sensitive models in order to capture the high variability and changes in the values of the actual measurements. After the aggregation however, some of this microphenomena get smoothed out leaving only the bigger picture of the group of elements (or total network) that we wish to analyse. In the second approach, we are performing the aggregation on the wanted network KPIs first, and then the final aggregated dataset is fed into more generic and simple prediction models in order to provide a better sense of the "trend" and potential evolution of the smoothed out KPI.

7.1.4 Network Data Measurement Issues

The generation of cellular network data measurements is being performed by the control plane network of the LTE backhaul. This is being performed by the means of period measurement requests that get executed in a distributed manner and then aggregated in higher elements such as the EMS (entity management system) and then aggregated and forwarded to the top level element NMS (network management system). This procedure cannot be always reliable as multiple changes occur in the duration of the network uptime. This results into datasets with a lot of different missing sections of KPIs, entities, and various discrepancies that can damage the outcome of a predictive model in detrimental ways. Each forecasting model has certain sensitivity to noise and irregularities and special handling of these cases in a per-model basic must be performed. The summary of these methodologies are found in the literature as imputation methods, and they vary from simple data generation techniques to intricate models that perform learning on the original data in order to complete the missing sections.

7.2 Problem Solution

In order to perform effective forecasting of network KPIs in either the lowest or the highest granularities, we need to enumerate the processing capabilities that the literature and also the open source libraries and projects can provide[1][2][3][4][5]. In general, the forecasting problem begins with an exploration process of the different time series samples that are provided in the dataset. The series can be investigated by using visualization tools (e.g. MS Excel, MATLAB) or using integrated data science environments such as Jupyter notebooks (python), Tableau software platform etc. Additional statistical tests can be performed in the time series that will analyse the series as a random variable.

Important indices such as autocorrelation and heteroscedasticity can either generate the optimum values for predictive model hyperparameters or hint as to which model is the most suitable for the specific dataset.

7.2.1 Forecast using Linear, Polynomial and Harmonic decomposition

Time series of network KPIS have different shapes and time evolution patterns for different technologies and cases of system operation[1][2]. These shapes are generally complex but can be summarized into simpler higher-level functions if prediction detail is not as important as a general understanding of the "Trend" they enclose. The computation of a series trend is usually the output of a univariable linear regression procedure that leads into different curve shapes being fitted. Most commonly functions used for trendlines are linear, Nth order polynomials (Figure 61)and sinusoid (Figure 74)functions. These can be computed and then used for any time value wanted to generate a prediction. The accuracy of this prediction is as accurate as the likeness of the original series to the decomposed series.

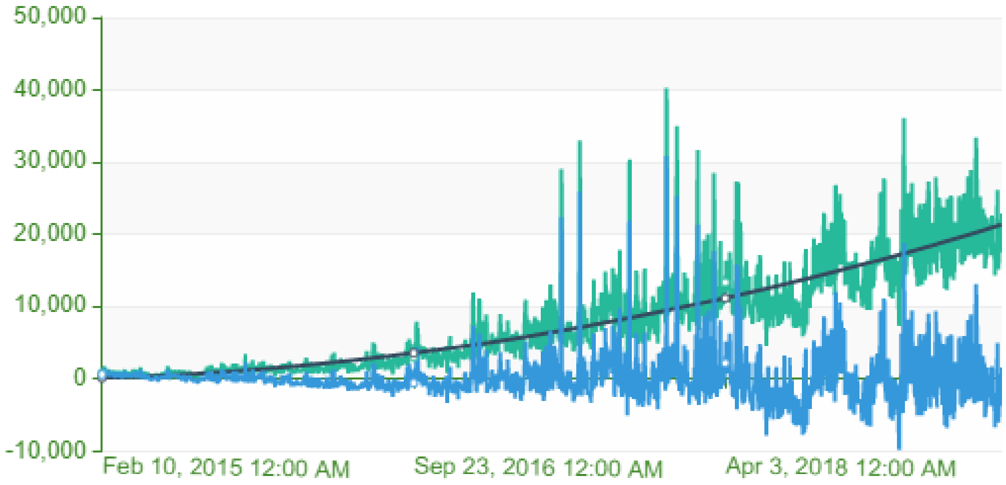


Figure 73 - Polynomial detrending, original (green), trend (black), detrended (blue)

Basic function decomposition is most commonly used in the pre-processing phase of a forecasting ensemble technique. Calculation of the trend is followed by divisive or subtractive detrending, resulting into a new (residual) time series that no longer contains a trend component. This allows for predictive models that specialize into other series to perform better, focusing on the evolution of the detrended data points. In the case of harmonic decomposition, repetitive application of harmonic fits and detrending can degenerate into a simplistic FFT decomposition due to the orthogonality of each different sinusoid function being subtracted from the original series.

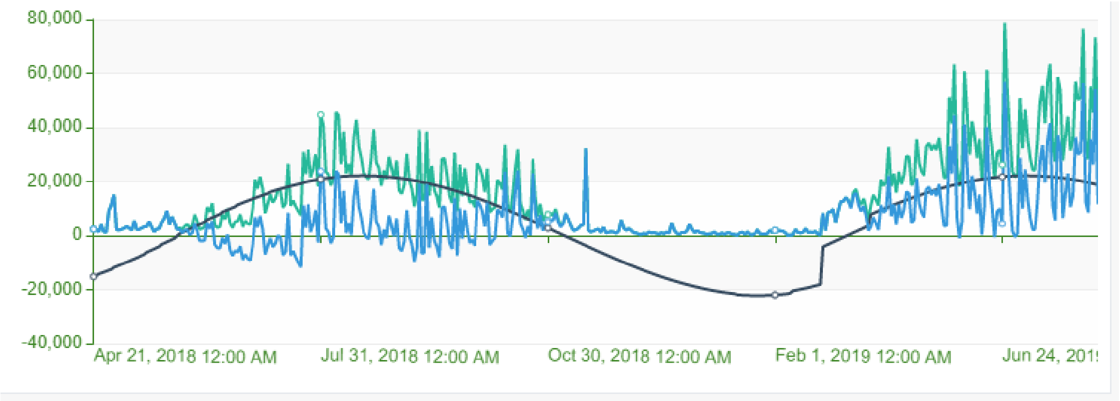
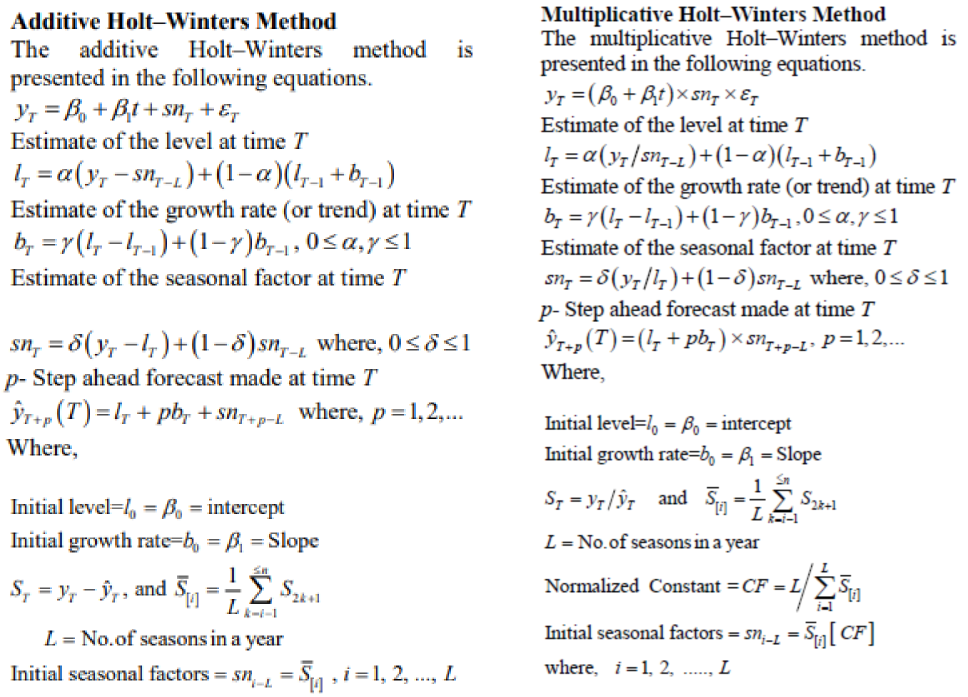


Figure 74 - Harmonic detrending, original (green), trend (black), detrended (blue)

7.2.2 Forecast using Holt-Winters Method

The holt winters method[4] (Figure 75)is an autoregression prediction model that is widely used in time series forecasting of financial and retail data. It is based on the basic principle of digital signal decomposition. Analysing the history results into different components that act independently and they are being predicted independently. In the end the final predictions result from their combination. The algorithm has two core implementations, the additive and multiplicative which differentiates in the methodology that is being used for the composition ensemble.



Additive Holt–Winters Method

The additive Holt–Winters method is presented in the following equations.

$y_T = \beta_0 + \beta_1 t + sn_T + \varepsilon_T$

Estimate of the level at time T

$l_T = \alpha(y_T - sn_{T-L}) + (1-\alpha)(l_{T-1} + b_{T-1})$

Estimate of the growth rate (or trend) at time T

$b_T = \gamma(l_T - l_{T-1}) + (1-\gamma)b_{T-1}, 0 \le \alpha, \gamma \le 1$

Estimate of the seasonal factor at time T

$sn_T = \delta(y_T - l_T) + (1-\delta)sn_{T-L}$ where, $0 \le \delta \le 1$

p- Step ahead forecast made at time T

$\hat{y}_{T+p}(T) = l_T + pb_T + sn_{T+p-L}$ where, $p = 1, 2, \ldots$

Where,

Initial level=$l_0 = \beta_0 =$ intercept

Initial growth rate=$b_0 = \beta_1 =$ Slope

$S_T = y_T - \hat{y}_T$, and $\bar{S}_{[i]} = \frac{1}{L} \sum_{k=i-1}^{\le n} S_{2k+1}$

$L =$ No.of seasons in a year

Initial seasonal factors $= sn_{i-L} = \bar{S}_{[i]}$, $i = 1, 2, \ldots, L$

Multiplicative Holt–Winters Method

The multiplicative Holt–Winters method is presented in the following equations.

$y_T = (\beta_0 + \beta_1 t) \times sn_T \times \varepsilon_T$

Estimate of the level at time T

$l_T = \alpha(y_T / sn_{T-L}) + (1-\alpha)(l_{T-1} + b_{T-1})$

Estimate of the growth rate (or trend) at time T

$b_T = \gamma(l_T - l_{T-1}) + (1-\gamma)b_{T-1}, 0 \le \alpha, \gamma \le 1$

Estimate of the seasonal factor at time T

$sn_T = \delta(y_T / l_T) + (1-\delta)sn_{T-L}$ where, $0 \le \delta \le 1$

p- Step ahead forecast made at time T

$\hat{y}_{T+p}(T) = (l_T + pb_T) \times sn_{T+p-L}, \; p = 1, 2, \ldots$

Where,

Initial level=$l_0 = \beta_0 =$ intercept

Initial growth rate=$b_0 = \beta_1 =$ Slope

$S_T = y_T / \hat{y}_T$ and $\bar{S}_{[i]} = \frac{1}{L} \sum_{k=i-1}^{\le n} S_{2k+1}$

$L =$ No.of seasons in a year

Normalized Constant $= CF = L \Big/ \sum_{i=1}^{L} \bar{S}_{[i]}$

Initial seasonal factors $= sn_{i-L} = \bar{S}_{[i]}[CF]$

where, $i = 1, 2, \ldots\ldots, L$

Figure 75 - The additive and multiplicative holt winters implementation pseudo-code

Holt winters method also has three (3) hyperparameters, one for each individual smoothing phase, that are mostly found using a heuristic hyperparameter search scheme such as grid search. Because they are continuous real parameters from [0,1], a sampling interpolation methodology must be followed in order to avoid long execution times for the algorithm. The best fit (Figure 76) of the model on an adequate validation period will result in the optimum hyperparameters for the total model.

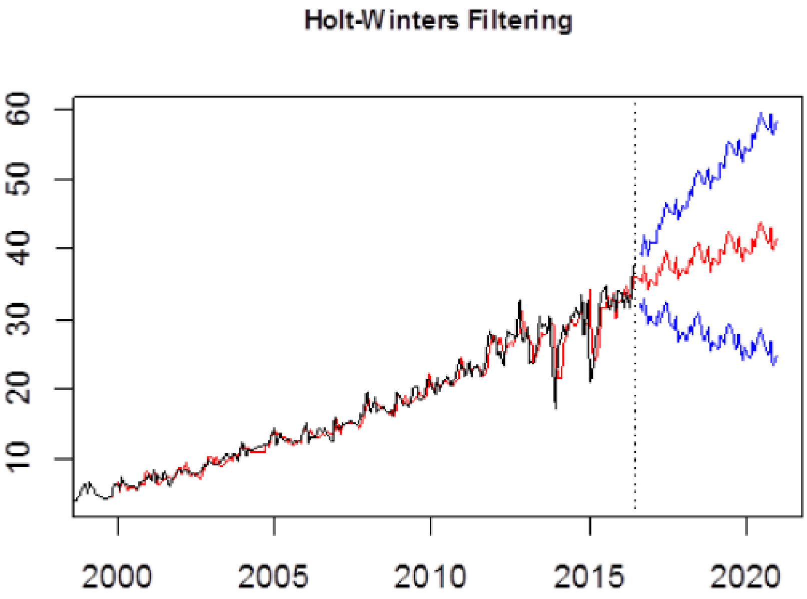


Figure 76 - Holt-Winters filtering prediction example

The output of the prediction result from the holt-winters model is generated by combining (additive or multiplicative) of the various subcomponents it generates(Figure 77). The components generated are a) seasonal, b) trend, c) level and d) xhat as seen in the example chart.

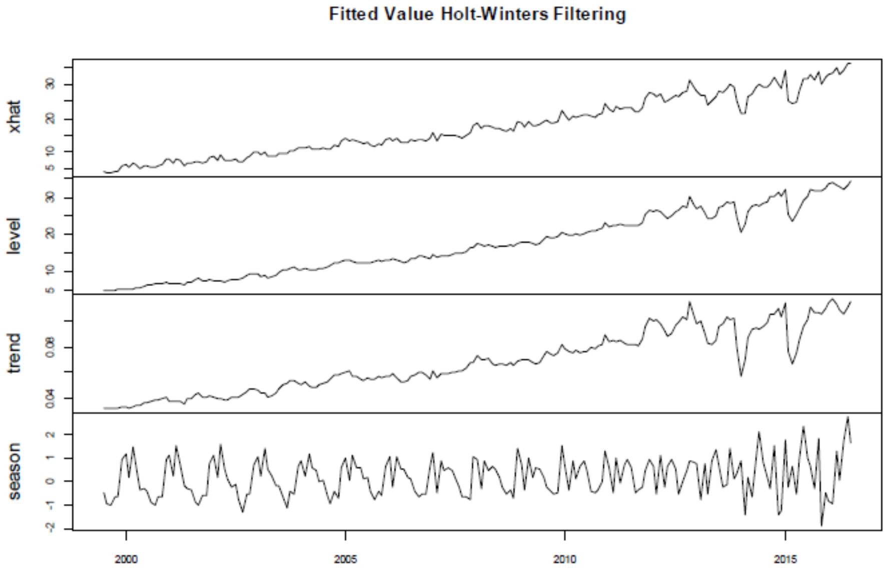
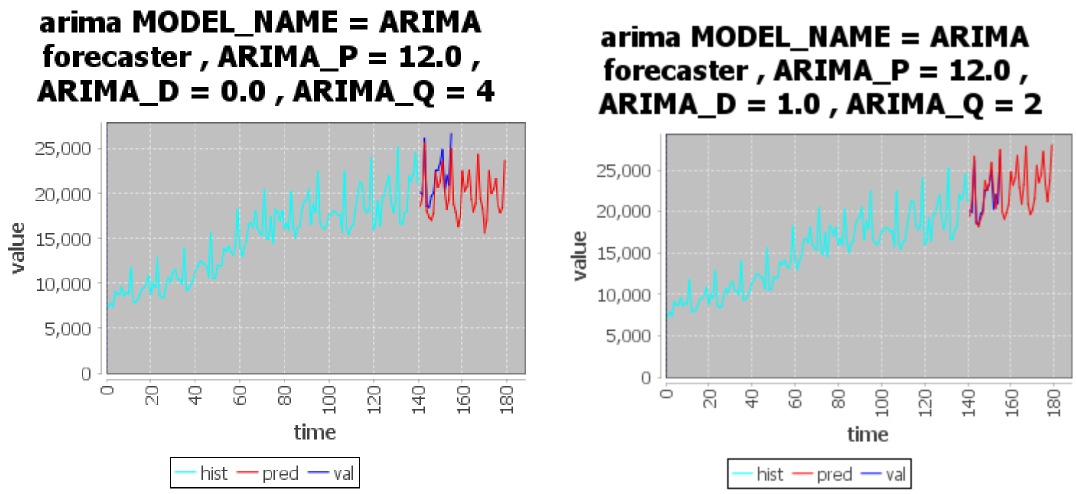


Figure 77 - decomposed time series components of the holt-winters model

7.2.3 Forecast using ARIMA / SARIMA / SARIMAX models

The ARIMA model family [5][6]is another autoregressive method for generating predictions based on computations performed in the historical data. It stands for Auto-regression (the first term that contributes to the computation), Integration (the second term of the computation) and Moving-Average (the 3rd term). Variations of the ARIMA model are the SARIMA which introduces also a seasonal component in order to repeat the feed-forward process and the SARIMAX which adds the 'X' for Exogenic. Exogenic series extended the algorithm from autoregression to MISO regression (multiple-input-single-output). This means that for the calculation of the wanted KPI, we will be using an additional time series along with its history. In some cases of network forecasting, a hexogenic mask of network events (which are greatly correlated with the selected KPI) can lead to a dramatic increase in the accuracy of the SARIMAX model. Hexogenic series will also be studied further in the time regression models that will be investigated in this study. ARIMA is using 3 hyperparameters to tune its calculation layer, namely P,D,Q which are positive integer parameters. They are low in complexity, so they are usually included in a grid search hyper parameter heuristics scheme (Figure 78).


arima MODEL_NAME = ARIMA forecaster , ARIMA_P = 12.0 , ARIMA_D = 0.0 , ARIMA_Q = 4

arima MODEL_NAME = ARIMA forecaster , ARIMA_P = 12.0 , ARIMA_D = 1.0 , ARIMA_Q = 2

Figure 78 - Example of forecasting using the ARIMA model

ARIMA is a widely used forecasting model for the capabilities that its variations provide. It has the potential to match many different series with different evolution profiles and it also can capture linear and approximations of polynomial trendlines.

7.2.4 Forecast using Empirical distributions

Empirical Distribution prediction is a statistical forecasting model that specializes in periodic or pattern-line time series forecasting. It splits the data into different sub-series based on a generated time characteristic such as day of week, week of year, month etc. These different series are then treated as statistical distributions. In order to perform a forward forecast, we generate a new timestamp and are selecting the distribution (or set of distributions) that are associated with the specific time instance. We then perform random sampling on the distribution (Figure 79), generating a sample from the history of the series and combining it with the other samples. In the end we have created the forecast from existing data points of the past of the entity. The simplicity and effectiveness of this algorithm lies in various statistical properties that are being maintained during this procedure. However, this algorithm fails to capture other function forms such as linear or polynomial trendlines. It is also sensitive to random noise which can greatly reduce the predictions accuracy.

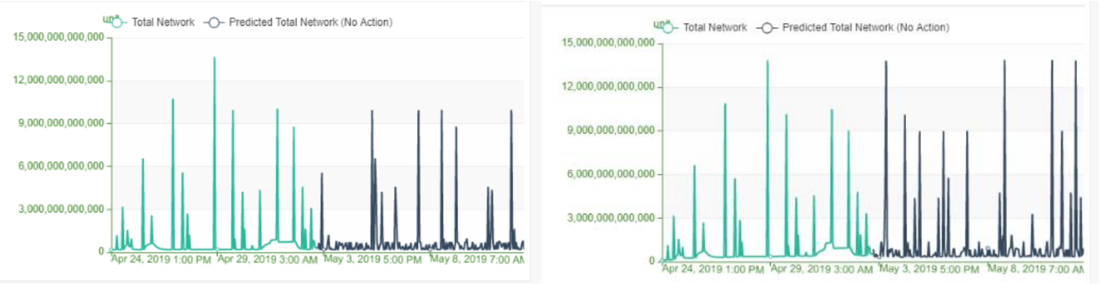


Figure 79 - Forecasting using empirical distribution sampling

7.2.5 Forecast using Regression Trees (Random Forests/ Gradient-Boosted Trees)

Tree models[7][8][9] is a big family of machine learning models that are using the tree data structure as a method of taking a decision in either classification or regression problems. The tree is being fitted into the input data with different optimization goals (e.g. stability, balance, lowest number of children nodes) and each leaf represents one of the possible decisions of the prediction problem (Figure 80). The effectiveness of tree models has led to their evolution by adding more and more individual trees into an ensemble array. Random forests and gradient boosted trees (with XGB as one of their most well documented implemented) are tree ensemble methods that use a large number of pre-fit trees to come into a prediction conclusion. They are one of the best non-neural network predictors and can be used in correlation with time and other exogenic factors in order to generate forecasting outputs.

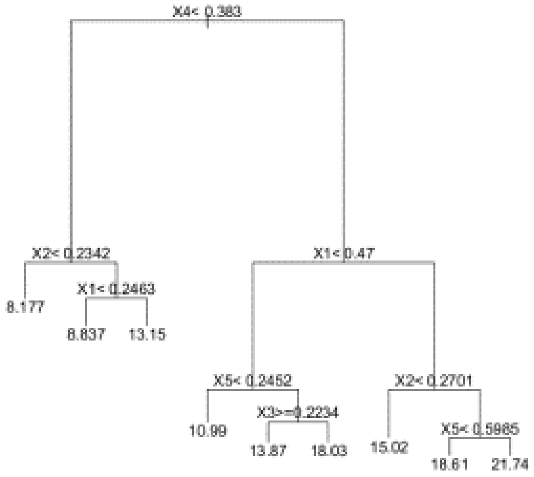


Figure 80 - example of a tree regression model[8]

One of the drawbacks of tree regression models is their ineffective modeling of the various trendlines (e.g. polynomial and linear). This becomes due to the fact that trendlines generate data points outside of the value domain of the input dataset whereas in tree models the predicted variable comes strictly from inside the input dataset. In a sense tree models are an organized redistribution of the initial data points (with their interpolated values). In conjunction with trend learning and detrending, tree models can be one of the most effective models to achieve fine detail forecasts.

7.2.6 Forecast using Neural Network Regression

Neural networks[10][11][12][13][14][15] are universal function approximator models. They consist of multiple computation network nodes that are connected via a feed forward mechanism. Their architecture (Figure 81) is inspired from the biological neurons that exist in every neural biological system. While connections in a neural network depict summation of values, circles represent transfer functions that transform the input data into different continuous functions. Neural networks have a notion of architecture which consists of several hyperparameters: a) number of neurons, b) number of layers, c) type of layers, d) type of activation functions. Different instantiations of these groups of hyperparameters can lead to different functional approximations of regression or classification problems. Neural networks also have a very large number of hyperparameters that are named weights. Weights are being multiplied at the output of each node to the generated value and they change depending on the input data via the learning process. Neural networks come with their own custom-built hyperparameter tuning methodology that is inspired from the neural networks itself. It uses gradient descent, a computation method to use the prediction error as a correction factor for the weight hyperparameters. The error is being transmitted backwards from the end nodes of the feed-forward graph towards the start, where the input features are placed. The procedure of correcting the weights of a neural network by back-wards traversal is referred to the literature as the back-propagation algorithm and is the cornerstone of the success of neural networks as a model. The complexity that neural networks capture is one of the highest that exists in the present literature and a lot of research on their evolution, deep neural networks, is used to model highly accurate complex phenomena

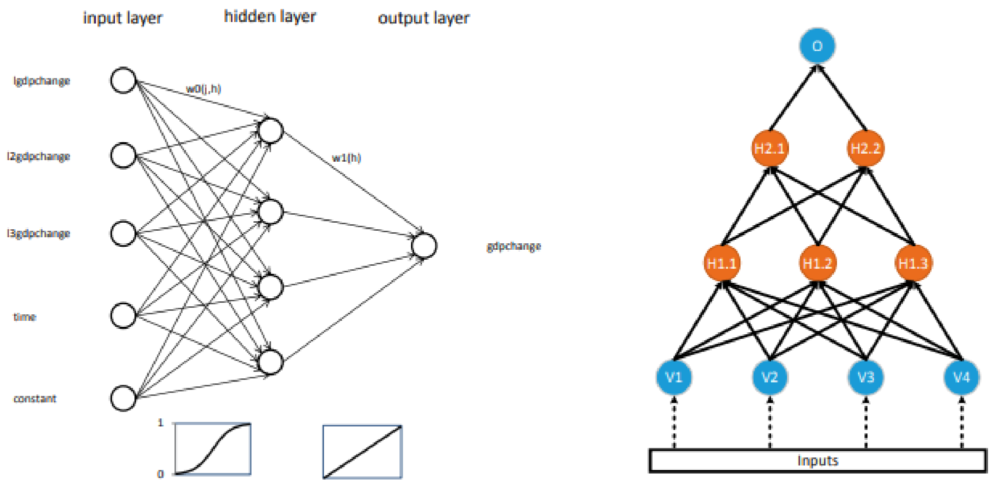


Figure 81 - Neural Network example architectures[13]

7.2.7 Fitting and hyperparameter tuning of the models

Machine Learning models mentioned in the previous chapters each possess a number of hyperparameters that can change the behaviour or output of the model. These hyperparameters can be categorical (e.g. additive or multiplicative detrending), numerical categorical (e.g. Arima P values from the set 1,5,12) or continuous integers and doubles which can be described as values ranges (e.g. Holt Winters A parameter with offset 0, limit 0 and step 0.05).

Table 7 - Forecasting models hyper-parameter space

Model	Parameter	Type	Values
ARIMA	P	Integer, Range	1,2,3, …,15
	D	Integer, Range	1,2,3, …,15
	Q	Integer, Range	1,2,3, …,15

Trend Fit	Order	Integer, Range	1,2,3, … 15
	Mode	Categorical	Harmonic, Polynomial
	Edit Mode	Categorical	Additive, Multiplicative
Holt-Winters	A	Double, Range	0,1 step 0.01
	B	Double, Range	0,1 step 0.01
	C	Double, Range	0,1 step 0.01
	Seasonality	Integer, Categorical	12 (monthly), 4 (seasonally), 24 (hourly), 365 (yearly)
	Mode	Categorical	Additive, Multiplicative
Distribution Fit	Mode	Categorical	Daily, Weekly, Monthly, Seasonally, Yearly, Total Distribution
	Aggregation Mode	Categorical	Sample, Min, Max, Mean
Regression Trees (RF / GB)	Number of Trees	Integer, Range	1,5, …, 1000
	Fit criterion	Categorical	Gini, Entropy
	Max tree length	Integer, Range	5,10, …,100
Artificial Neural Networks	Architecture Style	Categorical	Triangular, Symmetrical
	Activation Function	Categorical	Relu, Identity, Sigmoid, tanh
	Number of Layers	Integer, range	0,5, …,50
	Regularization / overfitting control	Categorical	L1, L2 norm, dropout, drop connect
	Solver	Categorical	Batch gradient descent,

			Stochastic gradient descent, minibatch
	Minibatch Size	Integer, range (optional)	5, 50, …, 500
	Error metric	Categorical	MSE, XENT, Negative Log Likelihood

The product of all the possible models, hyperparameters and their values can sometimes be found in literature as the configuration or scenario space (Table 7) of the optimization process. In order to select the best model instance from the available, we need first to select the wanted prediction error KPI, one that will most accurately depict the wanted output from the prediction model. The application of the error function can only be applied in a section of the existing time series (i.e. part of the history). This subset of the original time series history is commonly referred to as the validation set, or validation period (for time series). The process (Figure 82) of iterating through the available model instances and searching for the optimum configuration is commonly implemented by the means of a grid-search algorithm. In grid search each possible configuration is being assessed and therefore all available instances are being trained. In cases where the scenario space is very large or consisting of too many numerical range variables, other approaches can be followed in order to reduce the number of models evaluated. Random search, random walks, Greedy search, simulated annealing, genetic algorithms, taboo search, ant colony, bee colony, particle beam search, gradient descent, gaussian optimization are suboptimal optimizers that can help tackle the large scenario space of models. In addition, the search for the optimum configuration is a fully parallelizable problem

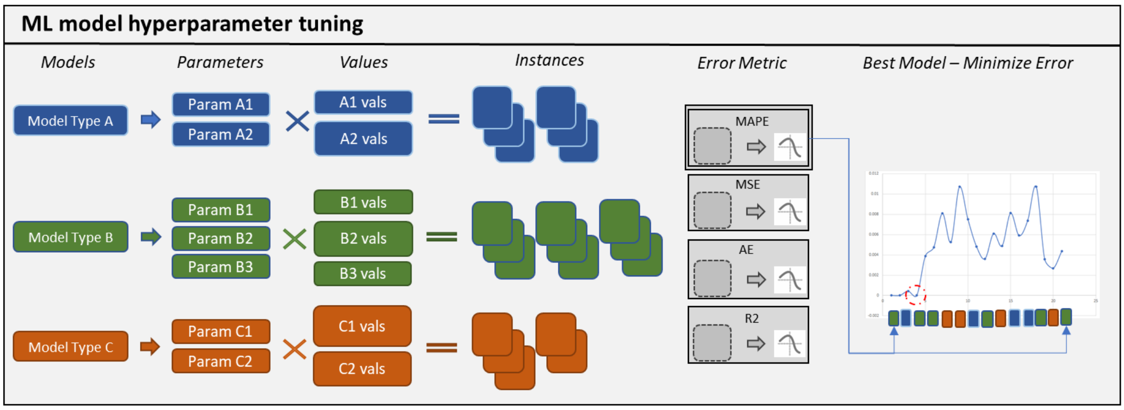


Figure 82 - Selecting the optimum model instance from various ml implementations

which can utilize a number of big data technologies methodologies and libraries such as Apache Spark, Apache Hadoop MapReduce and multi-processor parallelization on GPUs. During the hyperparameter search we can closely monitor the progress of each different current best model's result. By visualizing the progression and improvements (Figure 83) in the accuracy of the model, we can then decide to forcefully interrupt the hyperparameter tuning background procedure manually.

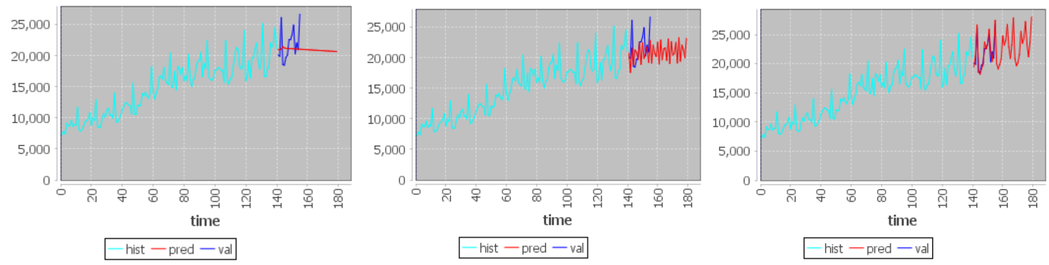


Figure 83 - Progress of the hyperparameter tuning function

7.2.8 Data Imputation schemes

As we mentioned in the problem statement, the dataset that is used in the input of a forecasting model can sometimes have missing values in various isolated or consecutive timestamps. This is generally referred to as the imputation problem and many methods exist in the literature[1] for solving this issue.

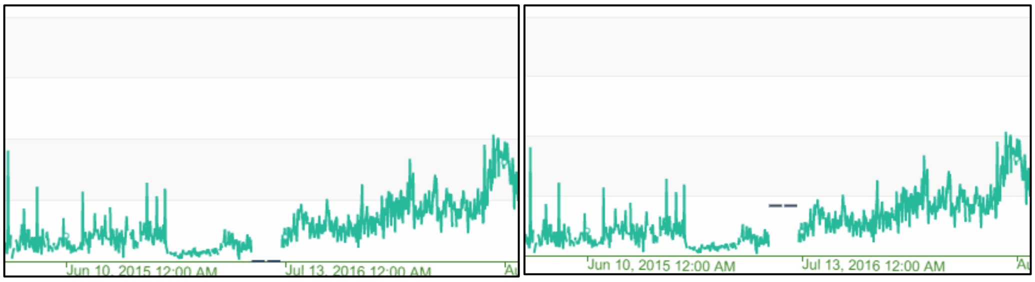


Figure 84 - Imputation with single value replacement, a) zero, b) average (black)

However, since all these new data points are not part of the original dataset, there can be no guarantees that they will not negatively impact the results of the machine learning models. Data imputation schemes are split into three core categories, statistical, smoothing and deterministic. In the statistical data imputation schemes, we have, average value imputation (replacing missing values with the average - Figure 84), distribution learning and sampling, distribution learning and average-sub-distribution (Figure 85). In the cases of imputation by smoothing, the exponential moving average (EMA) and the sliding window moving average are one of the most common methodologies (Figure 86).

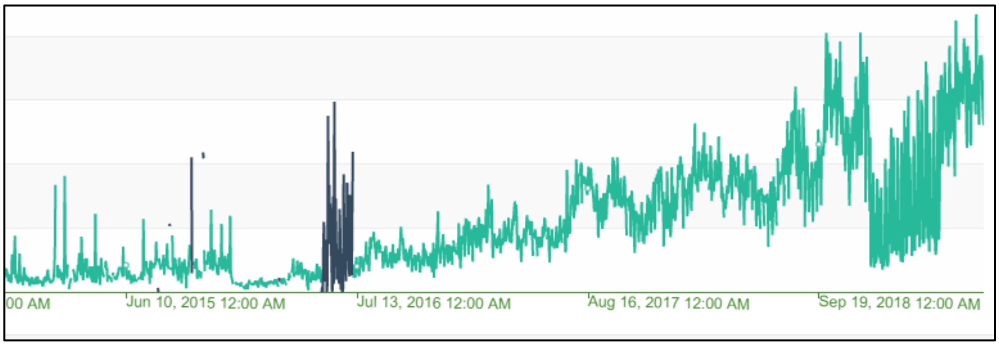


Figure 85 – Imputation using distribution sampling (black)

The generated values are the result of the smoothing operation applied multiple times until it reaches the number of missing points. In the deterministic imputation methods, we have single value replacement (most common value, zero value, max, min value), linear imputation (replacing all missing data points with linear interpolation of the latest two values), Bezier curve and other polynomial interpolation which are commonly used in various plotting / charting libraries and open source implementations as well. Imputation also needs to determine the time frame in which it will perform the imputation.

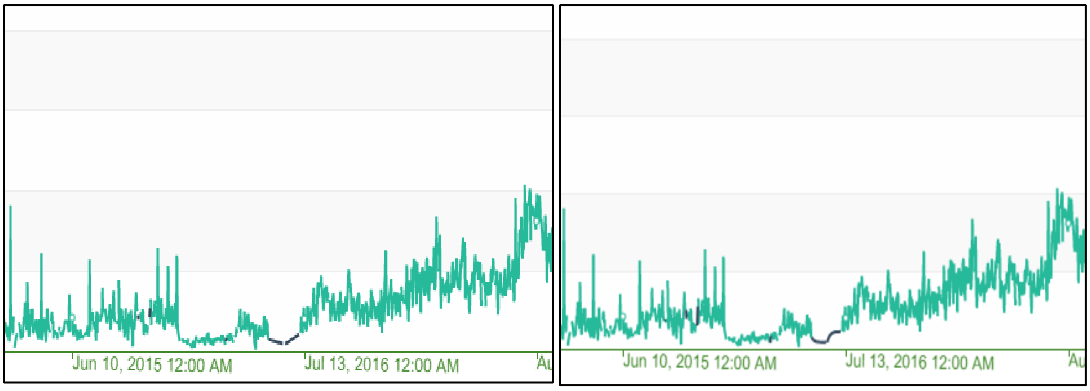


Figure 86 - Imputation using linear and EMA model (black)

In a dataset consisting of multiple elements, the start, end and interval of each element can be determined either individually or globally. In the global case, the minimum and maximum timestamp of all elements is used for each individual element’s imputation range. In the local, we are simply using the interval of the dataset to fill the missing data from internally of the series. The latest method generates the least problems in the machine learning models that are sensitive to noise but is worse in models that require time alignment between the various elements.

7.2.9 Proposed Forecasting pipeline

For the evaluation of each algorithm, we are proposing a framework in which multiple computational sections will be performed sequentially until the final prediction is being generated. These computational sections consist of the previously mentioned techniques, namely the filtering, aggregation, data imputation and forecasting categories and their respective children algorithms. Each operation will either add, subtract or edit one of the series of the inference instance. Operations may also append meta-data in the inference instance (e.g. the linear coefficient of the trendline) in a way that it can algorithmically and programmatically be accessed by future operations. This pipeline will take in considerations all mentioned model weakness and strengths and try to maximize their effectiveness while having the ultimate goal of achieving the highest accuracy on the forecast for each network element KPI.

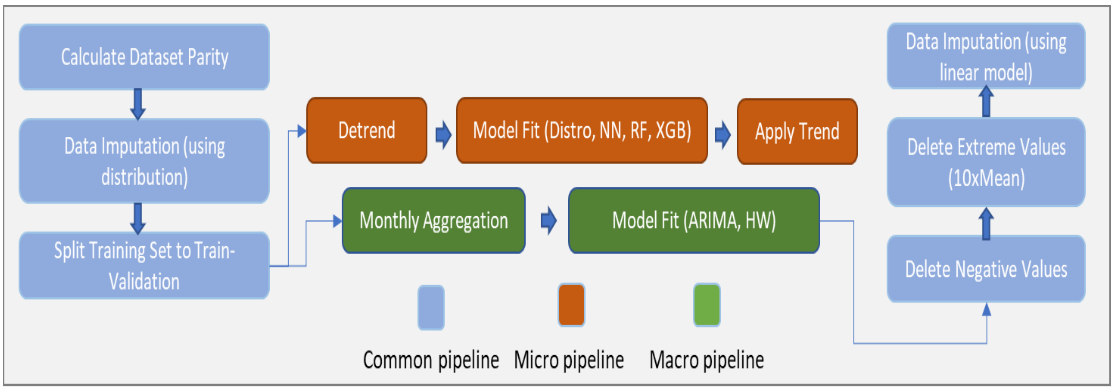

Calculate Dataset Parity
Data Imputation (using distribution)
Split Training Set to Train-Validation
Detrend
Model Fit (Distro, NN, RF, XGB)
Apply Trend
Monthly Aggregation
Model Fit (ARIMA, HW)
Data Imputation (using linear model)
Delete Extreme Values (10xMean)
Delete Negative Values
Common pipeline
Micro pipeline
Macro pipeline

Figure 87 - The proposed Forecasting pipeline

For the micro-forecasting case, where high detail of daily and weekly phenomena is required, we will be using a preprocessing pipeline which will include two different machine learning models used simultaneously. Trend analysis will be performed in the input data using polynomial or harmonic detrend (order 0, 1, 2, ...). Then the trend will be removed and passed on to the pattern fit models (distribution, neural network, random forests or gradient boosting trees). After the pattern models are trained, we perform the prediction phase on the new timestamps (forecasting range). Then we apply the trend (calculated in the previous phase) and shift the prediction on the trendline. Afterwards there is a cleaning stage in which we can automatically control some extreme values of the dataset. These cleaning rules will then be followed by another imputation method, replacing all the removed values by their linear interpolation. This is found to provide the best forecasting result as it can handle multiple anomalies that exist in such network KPI datasets.

7.3 Performance Evaluation

For the performance evaluation of the listed algorithms, we will select two different cases of forecasting problems. Forecasting on the microscale of network phenomena in the lowest time and entity granularity and forecasting on long-term aggregated network data for strategic planning and future predictions of aggregate values. In the cases of cellular networks, HetNet and other network infrastructures, both cases are needed in different scopes. Micro-prediction with high accuracy for short-term data can be used as input for real time optimization loops, fault prediction and anomaly detection whilst macro-prediction on network aggregate KPIs can be used for strategic planning, spectrum purchase and network rollout decisions.

7.3.1. Regression performance KPIs

For the evaluation of the accuracy of each predicting model, the literature [8][10][12]consists of multiple error indices (Figure 88) between the predicted values and the validation values. It is usual for forecasting problems to select the latest ‘n’ values of a dataset as its validation set. For this study we will be focusing on the MAPE error KPI which gives us an estimation of how much the prediction error is in relation to the actual value of the KPI. This allows us to estimate more qualitatively the performance of our model and how easily it can be used, regardless of its error, to assist in decision-making processes.

$$\text{Mean Absolute Deviation} \quad : \quad MAD = \frac{\sum |e_t|}{n}$$

$$\text{Sum Squared Error} \quad : \quad SSE = \sum e_t^2$$

$$\text{Mean Squared Error} \quad : \quad MSE = \frac{\sum e_t^2}{n}$$

$$\text{Root Mean Squared} \quad : \quad RMS = \sqrt{\frac{\sum e_t^2}{n}}$$

Mean Absolute Scaled Error:

$$MASE = \frac{1}{n}\sum_{t=1}^{n} \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n} |y_t - y_{t-1}|}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\text{Percentage Error} \quad : PE_t = \frac{(y_t - \hat{y}_t)}{y_t} \times 100$$

$$\textit{Adjusted } R^2 : \bar{R}^2 = 1 - \frac{(n-1)RSS}{(n-k)TSS} = \frac{(n-1)}{(n-k)} R^2$$

$$\text{Mean Percentage Error} : \; MPE = \frac{\sum PE_t}{n}$$

$$\text{Mean Absolute Percentage Error} : \; MAPE = \frac{\sum |PE_t|}{n}$$

Figure 88 - Calculation of various error KPIs for the forecasting models[1]

7.3.2 Performance Evaluation results for micro-scale forecasting

Micro-scale forecasting is performed mostly on bounded KPIs (with known maximum and minimum values) with a lot of time-related correlation, periodicity, random noise and other phenomena. In HetNets these KPIs can be, Instantaneous Uplink / Downlink throughput, Cell Load, Cell power consumption (watts). The most sensitive models of the literature are used in order to capture as much complexity as possible from the train data and replay it in the forecast. The autoregressive models lack the computational depth to emulate and learn the different phenomena, that’s why they sometimes can provide very bad results in this case.

Table 8 Average MAPE for the Micro-scale forecasting per model (Downlink Tput)

Model	**Avg MAPE (%)**
Neural Network	5
Gradient Boosting Tree	11
Random Forest	12
Distribution	42
Holt-Winters	69
ARIMA	75
Trend Fit	80

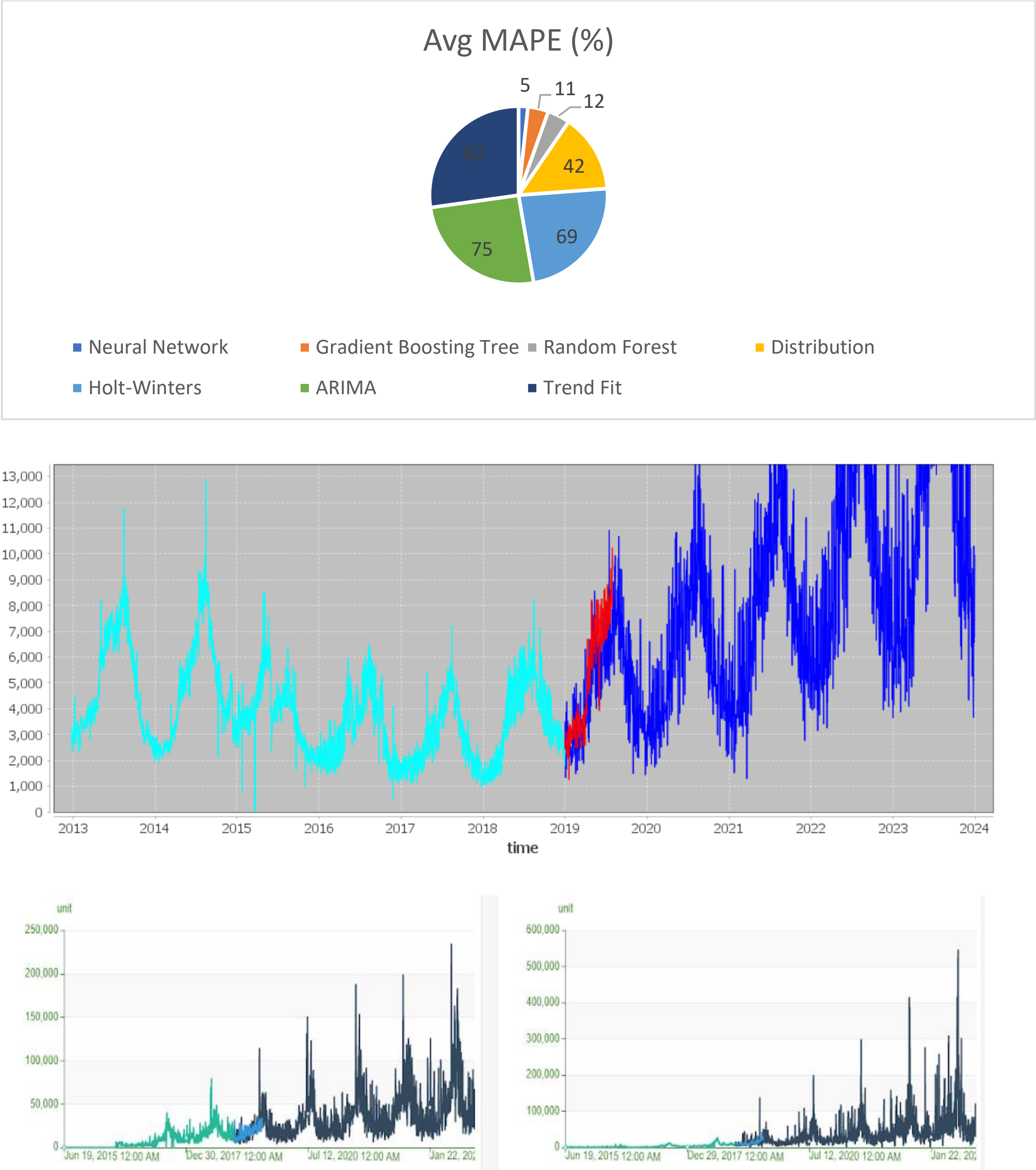


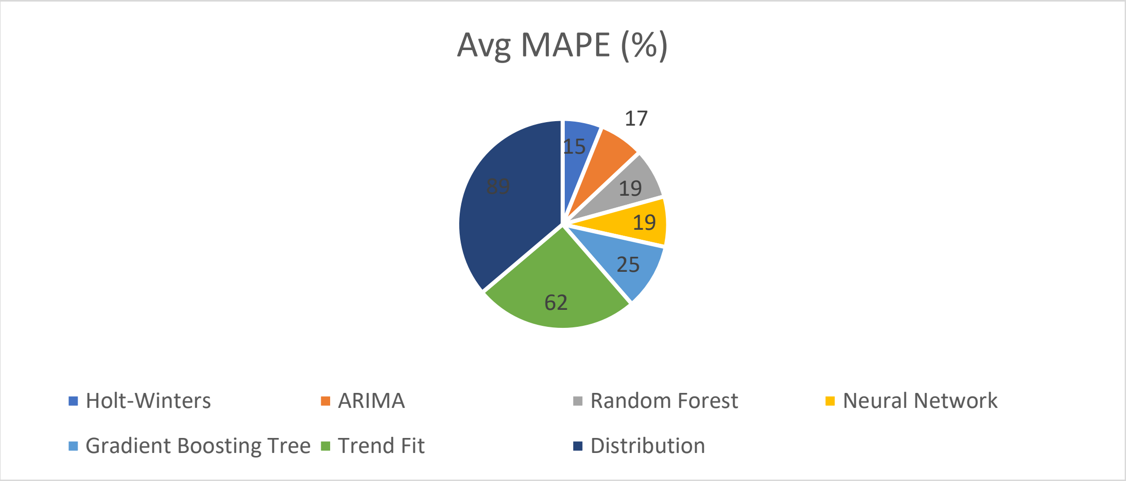
Figure 89 - Micro-forecasting indicative results

7.3.3 Performance Evaluation results for macro-scale forecasting

Macro scale forecasting is performed in network KPIs that are aggregate network gauges. These can be Downlink / Uplink total bytes (or TB), Uplink / Downlink packets, or total energy consumption. The entity aggregation for these KPIs are usually Site, Prefecture, cluster or region. The time granularity aggregation also is using sum and monthly time signatures. By applying this on the KPIs, the variations and fluctuations of the KPIs diminish. Also, the fact that these KPIs are increasing summations means that the trend component will play very huge part on the predictive accuracy. Auto-regressive methods have embedded trend-computation features. This is the reason why ARIMA and Holt-Winters provide the best predictions and therefore are the most effective as it can be seen in the average results.

Table 9 - Average MAPE for the Macro-scale forecasting per model (TB usage)

Model	Avg MAPE (%)
Holt-Winters	15
ARIMA	17
Random Forest	19
Neural Network	19
Gradient Boosting Tree	25
Trend Fit	62
Distribution	89



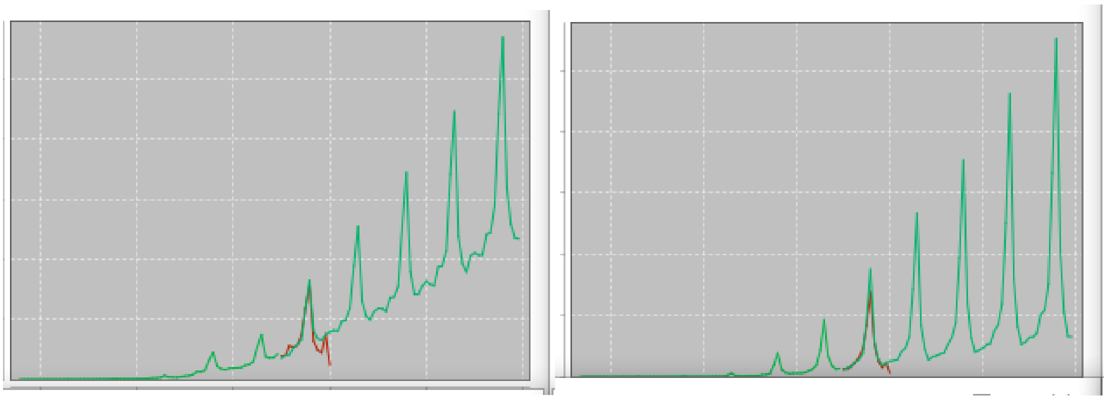


Figure 90 - Indicative Model Results for Macro-scale forecasting (TB-Usage)

7.3.4 Evaluation of training and execution times

In real case scenarios, it is required that all these aforementioned machine learning models are being constantly fed with new data and perform accurate predictions. Friction with implementation and usage for the scope of this chapter has led us into the conclusion that an additional study must be shown, one that depicts the complexity and therefore the time delay for each model on their training and prediction phase. For every model type, we will be calculating the average train (fit) time, and also the order of complexity for hyperparameter tuning and their multiplied KPI namely, the grid search delay order (i.e. the average delay that it would take for a grid search to find the optimum model over all possible hyperparameter scenario)

Table 10 - Performance Evaluation of Forecasting algorithms

Model	Delay (ms)	Hyper Parameter Order	Grid Search Delay Order
Distribution	50	10	500
Trend Fit	100	10	1000
Holt-Winters	183	1000	183000
ARIMA	210	1000	210000
Random Forest	1211	10000	12110000
Gradient Boosting Tree	2251	10000	22510000
Neural Network	5041	100000	504100000

From the general results, we see that the complexity of each model has a direct impact on the average delay(Figure 91). As we mentioned in the previous chapter, the distribution and trend fit are the lightest methods, followed by the ARIMA/Holt-Winters pair of autoregressive analysis. The tree models and the neural network model consist of complex internal optimization loops and this causes additional delay. Their delays also have a lot of variability that can occur from either the input dataset (train set) or the selection of hypermeters.

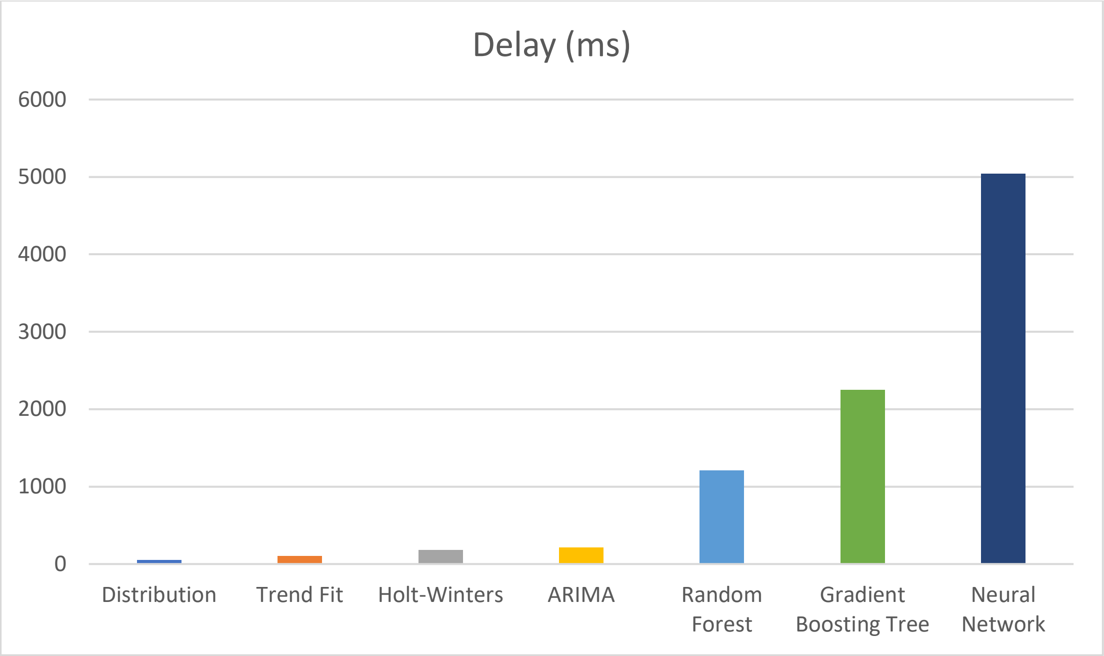


Figure 91 - Average delay per forecasting model

For a general dimensioning study, we need to see the order of hyperparameters of each model (Figure 92) and multiply it with the average delay per training (Figure 93). This makes the neural network a dominator model overall. However, the complexity of neural network architecture is not necessary a target for hyperparameter tuning. Reference architectures and duplicate architectures can sometimes reduce the scenario space into simpler, easier to traverse subspaces.

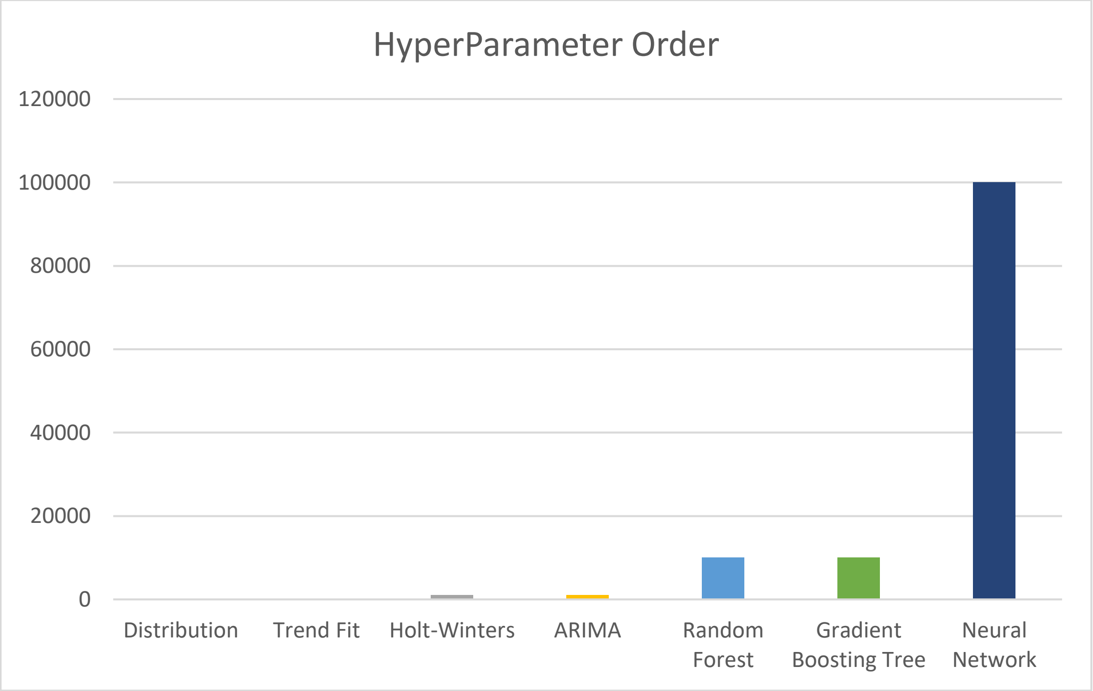


Figure 92 - Hyperparameter space order per model

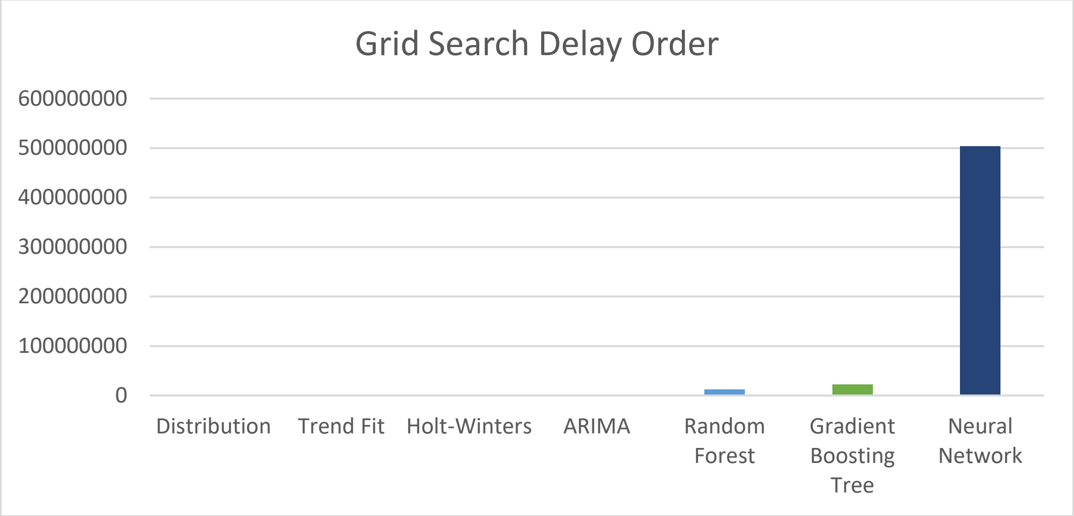


Figure 93 - Order of scenario complexity and average grid search delay

7.4 Conclusion

For this chapter, we have thoroughly evaluated a large literature of forecasting models and methodologies in order to provide insights as to which forecasting process is the most optimum for forecasting of micro-scale network KPIs and also for macro-scale aggregate KPIs. Each case has different characteristics and therefore requires different handling by predictive modelling. In our benchmarks, neural networks have been shown to be the most accurate predictors for micro-scale forecasting. This accompanied with polynomial detrending is performing in the best possible way for live network KPIs such as Uplink/ Downlink Throughput, Uplink / Downlink packet rate and Power consumption. For macro-scale forecasting, consisting of ever-increasing network gauges such as total Uplink /Downlink bytes, Uplink/Downlink packets and Energy consumption (total joules) we have found that simple models such as the holt-winters exponential smoothing model and ARIMA perform better or equal with these intricate models. This happens because the aggregation nature of the preprocessing pipeline is eradicating the random noise and micro-phenomena that better showcase the complex models. Another important conclusion is the time restriction that is being imposed in the case of the most complex models used. Neural networks and the various tree ensembles that we have tested have a relatively increased training time. This combined with their very vast number of hyperparameters, tells us that the extra points of MAPE increase require a lot of resource and time, something that will be available depending on the use case of the forecasting. If the forecasting is used in a real time cycle and requires constant retraining, then the accuracy can be sacrificed in order to provide with timely results to an existing optimization algorithm.

7.5 Chapter References

[1] M. Ruiter, "Using Exponential Smoothing Methods for Modelling and Forecasting Short-Term Electricity Demand", Bsc. Thesis, Erasmus University, Rotterdam, 2017

[2] D. Viberti, E. S. Borello, F. Verga "Pressure Detrending in Harmonic Pulse Test Interpretation: When, Why and How", Energies, Volume 11, 2018

[3] Prajakta S. Kalekar, "Time series Forecasting using Holt-Winters Exponential Smoothing", Computer Science, 2004

[4] Moyazzem H., Habibur R., Umma S., Tareq F. K., "Revenue Forecasting using Holt–Winters Exponential Smoothing", Research & Reviews: Journal of Statistics, Volume 5, Issue 3, 2016

[5] Kumar M., Anand M., "An application of time series ARIMA forecasting model for predicting sugarcane production in India", Studies in Business and Economics, Vol 9, 2014

[6] UK Centre for the Measurement of Government Activity "From Holt-Winters to ARIMA Modelling: Measuring the Impact on Forecasting Errors for Components of Quarterly Estimates of Public Service Output", Office for national Statistics, 2008

[7] Biau G., "Analysis of a Random Forests Model", Journal of Machine Learning Research 13 (2012) 1063-1095

[8] Dadez M, "Application of ensemble gradient boosting decision trees to forecast stock price on WSE", CEJSH, Issue 9, 2019

[9] Qingwen J., Xiangtao F., Jian L., Zhuxin X., Hongdeng J., "Using eXtreme Gradient BOOSTing to Predict Changes in Tropical Cyclone Intensity over the Western North Pacific", Atmosphere 2019, Volume 10

[10] Malte J., "Artificial neural network regression models: Predicting GDP growth", HWWI Research Paper, 2018

[11] Peng L., Peijun Z., Ziyu C., "Deep Learning with Stacked Denoising Auto-Encoder for Short-Term Electric Load Forecasting", Energies 2019, 12

[12] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, F. Martínez-Álvarez "Multi-step forecasting for big data time series based on ensemble learning", Knowledge-Based Systems, Elsevier 2018

[13] Thielbar M., D.A. Dickey, "Neural Networks for Time Series Forecasting: Practical Implications of Theoretical Results", North Carolina State University, Book, 2011

[14] G. D. Merkel, R. J. Povinelli, R. H. Brown "Short-Term Load Forecasting of Natural Gas with Deep Neural Network Regression", Energies 2018, 11

[15] J. C. Duarte, F. Rivas-Echeverria, "Time Series Forecasting using ARIMA, Neural Networks and Neo Fuzzy Neurons", University of the Andes, 2002

8 - Thesis Conclusions

8.1. Overview

The way forward for the next generation of cellular communications are complex large-scale infrastructures, HetNets, UltraDense networks and high-performance services. These systems are being analysed in chapter 1 of this thesis and after thorough analysis of the various technological aspects, we came into conclusion that this study requires an accurate simulation engine to design the use cases and measure the results. In this study we have also isolated several issues that rise in these reference architectures and researched the modern literature for solutions to problems and other improvements.

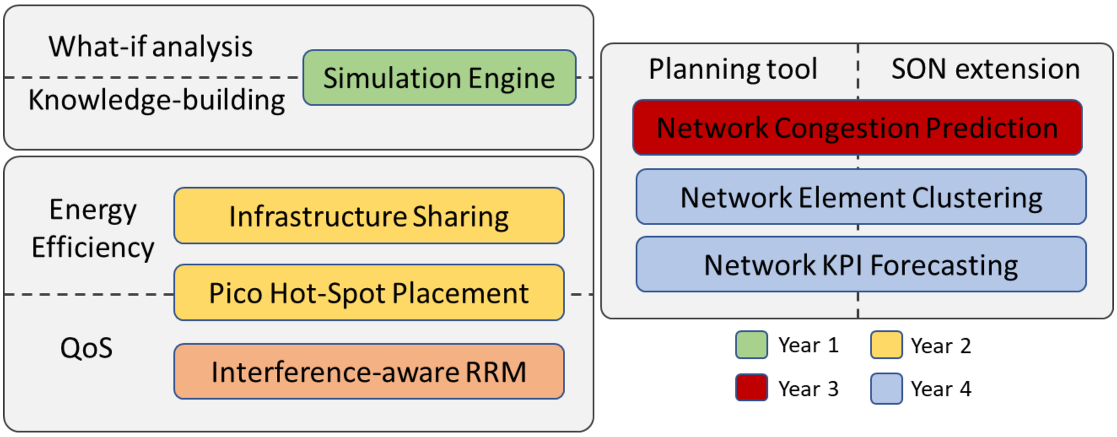


Figure 94 - Overview of this doctorate thesis outputs per year

Due to the complexity of this technology, it was identified that all different management and optimization approaches has benefits and should be used simultaneously in order to achieve the maximum performance and effectiveness. Firstly, cellular network simulation platform was shown to be critical to the execution of each different study that was performed. The accuracy of the data generation system allowed for the designed algorithmic solutions to perform according to the estimations and also revealed hidden issues that could not be foreseen by the theoretical analysis. Network design and infrastructure sharing between network operators played a huge part in the energy efficiency of future cellular operation scenarios as proven in chapter 3. Dynamic micromanagement of resource allocations by implementing an interference-aware SON proved to assist the LTE cell in providing the best possible quality of service to user equipment devices of various traffic demands. By extending the LTE network with predictive capabilities, we opened the possibilities for even further advancements in various management schemes.

Forecasting algorithms can enhance the operation of SON functions and planning operations for future evolution of various KPIs. Unsupervised grouping of network elements finds hidden behavioral patterns and structure in cells and user terminal devices allowing for targeted management. Finally, congestion prediction by the means of semi-supervised classifiers showed important increase in the network robustness of the HetNet scenarios executed.

8.2 Conclusions regarding Simulation engines for HetNets

In Chapter 2 of this thesis, we analyse the simulation engine that was designed specifically for this study. Based on the technological inputs from chapter 1, we tailored a custom Java-based large scale HetNet simulation tool that greatly exceeded the functionalities of existing simulation engines. The key aspects of the new software designed are: Scenario building flexibility including various technologies such as LTE Cells of different Antenna patterns, smaller cells of Pico cell technology, Multi-provider support, Wifi Access point emulation, full radio environment simulation including EIRP / SINR calculation and user equipment device simulation application usage using stochastic processes. This combined with multiple network KPI measurements and reporting capabilities (including charting, visual graphics on map playground and excel reports) gave us a research toolbox adequate to perform this study.

8.3 Conclusions regarding HetNet planning and Energy Efficiency

High energy consumption is a direct consequence of the expected traffic demand from the cellular networks. In chapter 3 we have analysed the possible improvements that can be derived from a series of redesign operations in a reference HetNet scenario. We have shown that cross-network-provider infrastructure sharing can be used to greatly reduce the total energy consumption of the network in the small cost of coordination between operators and also some performance losses. In the second stage of re-design we have analysed the underlying demand topology of the dense urban area and strategically placed Pico cells in the vicinities of various urban hot spots. Then we measured the same performance KPIs and saw that we have recovered and in some cased improved the network throughput and cell edge throughput (which had greatly deteriorated from the infrastructure sharing operation). We have also shown that an alternate solution (namely the increase of the cell spectrum) will not have the wanted results due to the increase in the demand for power in the cells high power amplifier unit.

8.4 Conclusions regarding SON functions and Quality of Service improvements

Het-Net quality of service is one of the most important aspects of their effectiveness as network infrastructures. Achieved throughput for user terminal devices in all the possible variations of radio conditions in conjunction with the diverse smartphone application environment means that LTE networks must be able to optimize the radio resource allocation and scheduling schemes in the fastest possible way. Existing radio resource allocation found in literature and the 3GPPP standards show that there is room for an improved RRA scheme that will take into consideration the SINR ratio in order to generate decisions for the per-user equipment real time resource allocation. Implementation of the state of the art algorithms and the proposed scheme in key simulation scenarios of the 4G network show that the proposed algorithm greatly increases the achieved throughput of the wireless network due to its better understanding of the overall network's degradation caused by the interference in each user terminal device. Considerations were also made for incorporating an a-priori user class tag that will further enhance this algorithm to greater user throughput gains.

8.5 Conclusions regarding Load balancing using Congestion prediction

In chapter 5 of this doctorate, we have analysed the literature for robust and efficient predictive methodologies in order to achieve reduction or elimination of network load congestion in specific 4G/5G network transmission scenarios. We have shown that unsupervised vector quantization algorithms such as the self-organizing map used in a semi-supervised prediction model can provide accurate congestion prediction indication. This indication can be used in a control loop that constantly affects the load of network elements by performing traffic steering via exploitation of the LTE handover mechanism. Integration of the predictive model in the SON algorithm has shown a dramatic decrease in the network load and the highest achievable values (congestion). Machine learning-augmented real time optimizations functions consist of various moving parts that require constant monitoring and finetuning. Also, the complexity of the learning process may impose hard limitations in the hardware that is performing these tasks in order to provide timely results. However, we have shown that the benefits from adapting this technology as part of a standard optimization procedure far outweigh the drawbacks.

8.6 Conclusions regarding Network Element / Device Clustering

In chapter 6 of this doctorate thesis, we have focused on the problem of identifying element groups with the same behaviour that can be clustered together to improve management and operations on a HetNet system. We have studied the literature of various approaches and focused mostly on unsupervised machine learning techniques because of their robustness and capability to understand hidden structures and patterns on various (telco and non-telco) datasets. We split the problem into two sub-problems and studied the algorithms under two hypotheses. The problem of identifying groups of LTE cells based on the density of their underlying user equipment devices and their network performance KPIs and the case of grouping together different user equipment devices that belong to different classes of broadband access usage. The results showed us that for the first case (per cell clustering), density-based clustering algorithms show the most promising results and are more accurate to their predictions. In the second case however, the case of the user equipment devices, we see that using Euclidean distance algorithms like X-means and K-means we can correctly identify the groups of users that belong to their corresponding network traffic demand model. This shows us that there is no single global algorithm that perform better for clustering of network KPIs and all different families should be checked and benchmarked to acquire the optimum grouping results.

8.7 Conclusions regarding Network KPI Forecasting

In chapter 7 of this doctorate thesis, we have identified the need for a predictive layer for various operational network KPIs in order to assist on real time management or large-scale network planning operations. We have studied the literature for the state of the art in forecasting models for time series data of different industries and forms. The problem of network forecasting was split into two subproblems, a) the forecasting for micro-management with sensitivity in every day fluctuations in the data and b) the forecasting for macro-management which relies on trendlines and long-term time evolution and can assist in planning tools that estimate traffic demand and load for large element aggregations. In order to better utilize the selected machine learning models, a data pipeline was devised for each of the two cases mentioned. The data pipeline consists of various time series processing components such as data imputation schemes, input filtering, output filtering, smoothing functions, evaluation functions for error metrics, and model hyperparameter tuning. In the end two different pipelines were isolated as the best, one for each scenario. Long-term predictions are found to be better approximated by using trend-sensitive models such as polynomial fit, ARIMA and Holt-Winters. In the cases of micro-prediction models, we have found that the pattern matching capabilities of distribution learning, tree models and neural networks is far more superior in terms of complexity than the autoregressive models. Finally, a performance benchmarking analysis has been performed in order to identify the fastest and slowest performing model. Neural networks sacrifice a lot of speed in order to obtain their complexity and accuracy whereas polynomial trend fit was found to be the fastest model to fit for all datasets. In general, speed of a machine learning will only be an impacting factor if these prediction models are being trained in real time during the operation of the network with constant retraining. Such rare cases can occur in embedding predictions as an input for a SON function or other optimization algorithm.

8.8 Consolidation and way forward

This doctorate thesis is approaching various different technological aspects of the current and future heterogeneous cellular network deployments. The complexity that these infrastructures impose result into various conflicting optimization goals and require advanced methodologies in order to provide robust and important improvements. For this study, we began analysing the current solutions on some of the key issues that occupy the literature, namely efficient 3GPP networks and intelligent radio resource allocation / scheduling. However, the way forward has led us into the new territory of Artificial Intelligence and machine learning. These methodologies were then used to either support or solve issues from various other aspects of the network such as congestion avoidance by predictive algorithms, identification of network element clusters and user equipment devices clusters and also forecasting of network KPIs on different granularities. Machine learning is proving to be a stable and robust tool to assist in the solution of various ICT technologies, expanding from systems and networks to financial systems, engineering, medical and commercial applications. Machine learning and AI improves sustainability of cellular infrastructure by reducing their resource usage, improving their operation and therefore resulting in less energy consumption and EMF reduction. In the future, it is expected to be incorporated in various forms as a component for industrial solutions and products.

APPENDIX A – ACRONYMS

Acronym	Explanation
A	
AI	Artificial Intelligence
AJAJ	Asynchronous JavaScript and JSON
AJAX	Asynchronous JavaScript and XML
APE	Absolute Percent Error
ARIMA	Auto-Regression, Integration, Moving-Average
B	
Bps	Bits per Second (throughput)
C	
CAPEX	Capital Expenses
CDF	Cumulative Distribution Function
CIO	Cell Individual Offset
COMP	Coordinated Multi-Point (LTE-A)
D	
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DCA	Dynamic Channel Allocation
DU	Dense Urban (Area Type)
E	
EMS	Entity Management System
EMF	Electro-Magnetic Force emissions
F	
FDMA	Frequency-Division Multiple Access
FTP	File Transfer Protocol
G	
GAA	General Access Application (Class)
GBT	Gradient Boosting Trees
GSM	Global System for Mobile Communications
H	

HetNet	Heterogeneous Cellular Network
HLO	High Level Objective
HSDPA	High Speed Downlink Packet Access
HTTP	Hyper-Text Transfer Protocol
HTTPS	Secure Hyper-Text Transfer Protocol
I	
IA	Incumbent Access (Class)
ICA	Independent Component Analysis
INR	Interference-to-Noise-Ratio
ISD	Inter-Site Distance
J	
JSON	JavaScript Object Notation
K	
KPI	Key Performance Indicator
L	
LTE	Long-Term Evolution of the 3GPP standard
LTE-A	LTE-Advanced
M	
MAPE	Mean Absolute Percent Error
ML	Machine-Learning
MIMO	Multiple Input Multiple Output (antenna)
MSE	Mean Square Error
N	
NME	Network Management System
O	
OFDM (A)	Orthogonal Frequency Division Multiplexing (Multiple Access)
OPEX	Operational Expenses
P	
PAL	Priority Access Layer (User Class)
PCA	Principal Component Analysis
PDF	Probability Density Function

Q	
QoE	Quality of Experience
QoS	Quality of Service
R	
RB	Resource Block
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RSSI	Received Signal Strength Indicator
RTSP	Real Time Streaming Protocol
RU	Rural (Area Type)
S	
SNR	Signal-to-Noise-Ratio
SINR	Signal-to-Interference-and-Noise-Ratio
SON	Self-Organized Network (Functions)
SOM	Self-Organized Map (Model)
SOTA	State of the Art
SU	Sub-Urban (Area Type)
T	
TCP	Transmission Control Protocol
TDMA	Time-Division Multiple Access
t-SNE	T-distributed stochastic neighborhood embedding
U	
UMTS	Universal Mobile Telecommunications System
UR	Urban (Area Type)
W	
Wi-Fi	Wireless Fidelity Alliance
WWW	World-Wide Web
X	
XGB	(e)Xtreme Gradient Boosted Trees

APPENDIX B – LIST OF PUBLICATIONS (FEB 2021)

Journal Publications	
1.	A. Georgakopoulos, A. Margaris, K. Tsagkaris, P. Demestichas, "Resource sharing in 5G contexts: Achieving sustainability with energy and resource sharing", accepted at IEEE vehicular technology magazine, January 2016, doi:10.1109/MVT.2015.2508319
2.	I.Belikaidis, S. Vassaki, A. Georgakopoulos, A. Margaris, F.Miatton, U. Herzog, K. Tsagkaris, P. Demestichas, "Context-aware Radio Resource Management Below 6 GHz for Enabling Dynamic Channel Assignment in the 5G era", accepted at EURASIP Journal, May 2017, doi: 10.1186/s13638-017-0946-8

Conference Publications	
1.	V. Foteinos, K. Tsagkaris, M. Michaloliakos, G. Poulios, T. Petropoulou, A. Margaris, K. Petsas, P. Demestichas, "Experimental Validation of Autonomic Traffic Engineering", in Proc. European conference on Network and Communications (EuCNC), 2015. doi:10.1109/EuCNC.2015.7194132

APPENDIX C – ABOUT THE AUTHOR

Short CV
Aristotelis Margaris was born in Athens, Greece in 1991. He has received the Diploma and Master’s degree in digital systems, from the University of Piraeus in 2012 and 2014, respectively. From May 2014 to 2015 he was research engineer at the University of Piraeus, Laboratory of Telecommunication Networks and Services, in the area of 4G network simulation and optimization. He has been involved in different research EU-funded FP7/ICT Projects, including “GreenTouch”, “iCore” and” XiFi”. He has also worked in Wings-ICT Solutions as developer and designer of handover optimization algorithms and opportunistic Wi-Fi Mesh networks. Since 2015 he is working as lead developer in machine learning and predictive analytics company “Incelligent”. His mains interests are software frameworks to enable AI transformation using the state-of-the-art AI algorithms alongside Big-Data platforms and multi-vendor compute clouds.