Efficient Access Network Selection and Data Demand Prediction for 5G Systems

University of Cape Town

By: Udita Paul

Supervisor:

A/Prof Olabisi E. Falowo

Dissertation submitted to the Department of Electrical Engineering in partial fulfilment of the requirements for the Master of Science in Electrical Engineering degree at the University of Cape Town

April 2018

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

As the candidate's supervisor, I have approved this dissertation for submission.

Name: A/Professor Olabisi E. Falowo

Signature:

Date: 06/04/2018.

APPLICATION FORM

Please Note:
Any person planning to undertake research in the Faculty of Engineering and the Built Environment (EBE) at the University of Cape Town is required to complete this form **before** collecting or analysing data. The objective of submitting this application *prior* to embarking on research is to ensure that the highest ethical standards in research, conducted under the auspices of the EBE Faculty, are met. Please ensure that you have read, and understood the **EBE Ethics in Research Handbook** (available from the UCT EBE, Research Ethics website) prior to completing this application form: http://www.ebe.uct.ac.za/ebe/research/ethics1

APPLICANT'S DETAILS		
Name of principal researcher, student or external applicant		Udita Paul
Department		Electrical Engineering
Preferred email address of applicant:		Plxudi00@myuct.ac.za
If Student	Your Degree: e.g., MSc, PhD, etc.	MSc
	Credit Value of Research: e.g., 60/120/180/360 etc.	180
	Name of Supervisor (if supervised):	Olabisi Falowo
If this is a researchcontract, indicate the source of funding/sponsorship		Click here to enter text.
Project Title		Efficient Access Network Selection and Data Demand Prediction for 5G Systems

I hereby undertake to carry out my research in such a way that:

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

SIGNED BY	Full name	Signature	Date
Principal Researcher/ ~~Student/~~External applicant	Udita Paul	Signed by candidate	**22 Jan 2018**

APPLICATION APPROVED BY	Full name	Signature	Date
Supervisor (where applicable)	Olabisi Falowo		~~Click~~ here to ~~enter~~ a date. 22/01/2018
HOD (or delegated nominee) Final authority for all applicants who have answered NO to all questions in Section1; and for all Undergraduateresearch (Including Honours).	S. CHOWDHURY Click here to enter text.	Sunetra Chowdhury	23/1/18 Click here to enter a date.
Chair : Faculty EIR Committee For applicants other than undergraduate students who have			

Declaration

I know the meaning of plagiarism and declare that all the work in the document, save for that which is properly acknowledged, is my own. This thesis/dissertation has been submitted to the Turnitin module and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

I am now presenting the thesis for examination for the Degree of MSc in Electrical Engineering. I also grant the University free license to reproduce the above thesis in whole or in part, for the purpose of research.

Name: Udita Paul

Signature: Signed by candidate

Date: 13/03/2018

Acknowledgements

I would like to thank the almighty God for giving me the strength to carry out this research. I am grateful to my supervisor, A/Prof. Olabisi Falowo, for his tireless and priceless guidance throughout the entire research. I also convey my thanks to all the members of the Communications Research Group (CRG) for their constructive criticism.

I am especially grateful to my colleague, Jiamo Liu, for constantly being by my side during the course of my research. I would also like to thank Dr. Joyce Mwangama for the help she provided on different technical aspects related to my research.

Lastly, I would like to thank my parents, Amal and Sabita, my sister, Adrita, and my extremely dear friends, Niranjana and Sampath for all the prayers, support and love they have shown me over the years.

Abstract

The massive proliferation of sophisticated mobile terminals with advanced capabilities have led to an enormous surge in the demand for mobile broadband data. Also, the recent popularity of bandwidth intensive applications such as Netflix and YouTube has contributed to this demand for the wireless resources. In order to cope with this massive demand, fifth generation (5G) of wireless network is on the verge of deployment. This new generation of the wireless networks would pose different challenges for both subscribers and service providers, and the challenges need to be carefully addressed. Due to the diverse nature of the subscribers of mobile broadband, one network element is inadequate to meet the imposed requirements. Subscribers vary in terms of their usage of wireless resources as well as their preferred content. Deployment of the 5G systems promises the introduction of multiple tiers of heterogeneous networks within its architecture. This means radio access technologies (RATs) of various kinds (2G, 3G, 4G, 5G and Wi-Fi) would have to co-exist and aim to bridge the gap between the supply and demand for data. Subscribers, equipped with multi-mode or multi homing mobile terminals, can connect to one or more RATs to receive the required services. They also often run multiple applications simultaneously and as such, it must be ensured that the best access technology is assigned to a particular subscriber to maintain quality of experience and service. As such, an algorithm need to be devised that selects the best network to provide ubiquitous coverage to different types of users, running various kinds of applications, under dynamic network conditions.

The network and infrastructure providers, on the other hand, face the need to meet up with the demand for data that the subscribers in different coverage regions require. In the 5G system, traditional proprietary hardware performing dedicated network functions such as packet gateway and service gateway would be replaced by softwarized virtual network functions (VNFs). These VNFs would need to be hosted in the data centres and would require computational power to process the subscribers' traffic originating in an area. Therefore, data centres are set to play a key role in the provisioning of service in 5G systems. However, before establishing a data centre in a region, the traffic profile of that region need to be carefully studied to determine the optimal position and dimension of the facility. Furthermore, as cellular traffic differs depending on the time of the day, accurate prediction models are required to forecast future traffic demand to ensure dynamic and proper utilization of resources.

This thesis aims to propose solutions to address these problems that subscribers and infrastructure providers face. Firstly, an algorithm is proposed to select the best access network for a subscriber running single or group of applications. Deviating from the existing access selection schemes in the literature, which consider the RAT-selection problem in an environment where accurate information related to key metrics such as latency and throughout is always available, the proposed algorithm models the problem in a completely fuzzy environment. As wireless networks are highly dynamic systems that are not only very unpredictable but also susceptible to sudden changes (for example malfunction of a particular RAT rendering it unusable), fuzzy systems are most adept in representing them. In the proposed algorithm, a new branch of fuzzy logic, Intuitionistic Fuzzy (IF) logic, is used with a popular multi-criteria decision making (MCDM) algorithm -Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), to formulate a network selection problem. The IF-TOPSIS scheme is designed to accurately take in various parameters such as network conditions, different number of applications and user preferences to select the ideal network for different types of subscribers. The second part of this thesis aims to solve the problem associated with establishment of data centre and utilization of its resources. As the cellular traffic exhibits strong spatial and temporal dependencies, it becomes necessary to analyse the traffic before establishing an infrastructure like a data centre. Existing literature do not consider real world traffic while determining the best location and dimension of 5G data centres. In this thesis, a real world traffic data set is first analysed to understand the variations that are present in different regions within a city. Based on the traffic analysis, the ideal placement of the data centre is formulated as a facility location problem and solved using the Weiszfeld's algorithm. Additionally, based on the traffic analysis, the optimal dimensions of the data centre in different regions are heuristically obtained. Finally, machine learning algorithms are employed to obtain future traffic demand values to aid dynamic allocation of data centre resources. Simulation results are presented to show the effectiveness of the proposed schemes.

Table of contents

List of figures

List of tables

Nomenclature

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
AHP	Analytical Hierarchy Process
AMPS	Advanced Mobile Phone Service
ANN	Artificial Neural Network
AP	Access Point
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
AT & T	American Telephone and Telegraph
BS	Base Station
C	Cost per megabyte
CA	Carrier Aggregation
CDR	Call Detail Record
CNN	Convolutional Neural Network

CoMP	Coordinated Multiple Point
COPRAS	COmplex PRoportional ASsessment
COTS	Commercial off-the Shelf
CRT	Cooperative Relay Transmission
DiA	Distance to Ideal Alternative
EDGE	Enhanced Data Rate for GSM Evolution
ELECTRE	Elimination et Choix Traduisant la REalite
ETSI	European Telecommunication Standard Institute
FDMA	Frequency Division Multiple Access
FFNN	Feed Forward Neural Network
FIS	Fuzzy Inference System
Gb	Gigabit
Gbps	Gigabit per second
GPRS	General Packet Radio Service
GRA	Grey Relational Analysis
GRU	Gated Recurrent Unit
GSM	Global System for Mobile Communication
H	High
HSDPA	High-Speed Down Stream Packet Access
HSPA	High-Speed Packet Access
HWN	Heterogeneous Wireless Network
I	Important
IF	Intuitionistic Fuzzy
IM	Instant Messaging

IoT	internet of Things
IP	Internet Protocol
Kbps	Kilo bit per second
L	Low
LTE	Long Term Evolution
LTE-A	Long Term Evolution- Advanced
LSTM	Long Short Term Memory
M	Medium
MAPE	Mean Absolute Percentile Error
MB	Mega Byte
Mbps	Megabits per second
MCDM	Multi-Criteria Decision Making
MCGDM	Multi-Criteria Group Decision Making
MEW	Multiplicative Exponential Weighting
MH	Medium High
MIMO	multi-input Multi-output
ML	Medium Low
MMS	Multimedia Message Service
MNO	Mobile Network Operator
MT	Mobile Terminal
MTS	Mobile Telephone Service
MULTIMOORA	MULTIplicative-form with Multi-Objective Optimization Ratio Analysis
NFV	Network Function Virtualization

NHBBO	Non-Biogeography based Optimization
NRT	Non Real Time
PD	Delay
PFS	Parallel Fuzzy System
PL	Loss rate
PJ	Packet Jitter
PoA	Point of Attachment
QoS	Quality of Service
QoE	Quality of Experience
RAN	Radio Access Network
RAT	Radio Access Technology
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RSSI	Received Signal Strength Indicator
S	Security
SARIMA	Seasonal Autoregressive Integrated Moving Average
SAW	Simple Additive Weighting
SDN	Software Defined Networking
SINR	Signal to Interference and Noise Ratio
SMS	Short Message Service
SLA	Service Level Agreement
std	Standard Deviation
T	Throughput

TDMA	Time Division Multiple Access
TIM	Telecom Italia
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
U	Unimportant
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
VH	Very High
VI	Very Important
VHO	Vertical Handover
VIKOR	VlseKriterijumska Optimizacija I Kompromisno
VL	Very Low
VoIP	Voice over Internet Protocol
VNF	Virtual Network Functions
VVH	Very Very High
VVL	Very Very Low
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WTP	Willingness to Pay
WWAN	Wireless Wide Area Network

Chapter 1

Introduction

1.1 Introduction

The mobile wireless networks (MWNs) have been under a constant state of evolution and development to improve the applications' quality of service (QoS) and the subscribers' quality of experience (QoE). Recently, the MWN traffic has been increasing rapidly due to the increased availability of different types of mobile terminals (MT) with advanced capabilities. Subscribers now possess multiple of such MTs which allow them to access different bandwidth intensive applications, popularities of which are on the rise as well. In 2016, it was reported that the number of mobile device connections reached the figure of 8 billion [1]. This figure is projected to increase to 11.6 billion by 2021. The demand for data in 2021 is also set to reach 49 exabytes per month, from 10 exabytes per month in 2017 [1]. Therefore, it is no longer feasible for traditional homogeneous networks consisting of single macro base stations to cater for the data requirements of users under its coverage. To efficiently meet up with the subscribers' demands, heterogeneous wireless networks (HWNs) have emerged. HWNs essentially integrate different radio access technologies (RATs such as macro cells, small cells and Wi-Fi Access Points) that complement each other in terms of coverage area, mobility support, bandwidth and price to provide ubiquitous coverage while maintaining high QoS.

Currently, research is in progress for the development and deployment of the fifth-generation (5G) mobile networks. Among many of its features, 5G is envisioned to incorporate multi-tier of heterogeneous networks to efficiently meet requirements for the explosive growth of data traffic [2]. 5G will also be responsible for providing services not only to the traditional mobile broadband users, but also to many vertical industries such as automation, health, energy and other industries. To cater for the service demands of such a diverse subscriber group, 5G is proposed to have at least three logically separated portions of the network, with each portion called a 'network slice'. Network slicing would enable operators to separate a physical network into multiple virtual networks tailored to meet requirements of different user groups [3].

In addition, 5G is set to introduce a transition of the wireless communications industry from the traditional hardware based approach to a more software based environment. Key technologies such as Network Function Virtualization (NFV) [4] and Software Defined Networking (SDN) [5] are set to act as key enablers for this transition. These technologies would split the control plane from the data plane and would be able to deploy network functions in commercial off the shelf (COTS) hardware. Essentially, in a 5G architecture, a network function such as the service gateway would no longer be required to have a dedicated hardware of its own. Instead, these network functions would be hosted by various data centers as virtual network functions [6]. As with the introduction of any new generation of wireless networks, 5G would require additional modifications and enhancements in the available user equipment, application requirements, RATs and access schemes in conjunction with the introduction of new internet services by content providers [7]. Both users and service providers would be faced with unique challenges that need to be properly solved to achieve smooth end-to-end service provisioning. .

From a subscriber's point of view, selection of the ideal access network becomes an important issue. RATs differ from each other in a number of parameters such as the offered bandwidth, coverage area, cost per resource, operating frequencies and packet delay. In addition, the network conditions and pricing policies of these RATs also vary significantly during different times of the day. Subscribers also are of different types with some being more frequent users of heavy bandwidth demanding applications (such as HD real time video streaming applications) than others. In such a dynamic wireless environment, the selection of the best access network becomes a crucial decision making challenge. To ensure that the best access network is indeed selected for a particular subscriber, robust RAT selection schemes need to be implemented.

On the flip side, the demands imposed by the subscribers need to be adequately met by the service providers to ensure customer satisfaction and prevent churn. As the industry shifts towards a more softawarized environment, computational power is required to maintain proper operation of all the network functions. This computational power in 5G systems will be provided by the data centers. As such, these data centers need to be optimally placed and properly equipped to cater for the demand experienced within their coverage areas. To achieve that, traffic profiles in different zones need to be analysed to extract any pattern that might be present. In addition, with the help of past traffic data, prediction models need to be designed to better aid dynamic and proper utilization of resources to meet the future demand.

This thesis aims to address some of these important challenges that different players of the wireless communications industry are currently faced with. The main focus of this thesis can be enumerated as follows:

- Development of a well-rounded access network selection scheme that selects the optimal access network for different subscribers under dynamic network condition.
- Design the optimal location and dimension of a key 5G infrastructure: data center and subsequent prediction of future data demand to facilitate proper allocation of wireless resources.

In the rest of this chapter, the motivations behind this research, research questions, the key research objectives, research methodology, research scope & limitations, the contributions made, and a general outline of the thesis organization are presented.

1.2 Motivations

As different types of subscribers, with powerful mobile terminals, are inundated with many available access networks, the selection of the ideal network becomes a vital problem. In addition, as the demand for wireless resources continue to increase, it also becomes important to properly analyse cellular traffic traces of different areas to facilitate the establishment, design and operation of key network infrastructure. The following factors are the motivations behind the work conducted in this thesis:

- In the presence of many wireless access networks in the dynamic HWN where users also vary in their requirements, efficient access network selection schemes need to be implemented. In designing such an algorithm, the unpredictable nature of the HWN needs to be adequately modelled. Furthermore, user preferences along with network conditions also need to be incorporated to ensure proper QoS.
- The traffic demand experienced in different areas varies with space and time. While residential areas might experience higher traffic volume during night, office areas experience more during the day time. As such, traffic profiles of different areas need to be carefully evaluated to understand the level of demand experienced. With the traffic analysis in hand, facilities such as data centers can be properly designed to satisfy demands imposed by subscribers. As wireless resources are scarce, proper prediction of future data demand is needed to ensure utilization of resources.

1.3 Research Questions

In this thesis, the following important research questions are examined:

- In designing an efficient network selection algorithm, how can dynamic and unpredictable nature of the HWN be modeled?
- How can an access network selection scheme be implemented that takes into consideration multiple parameters such as the network condition, user preferences and requirements of single or multiple applications while selecting the best available RAT in a HWN?
- How does the traffic varies within different areas in a major city?
- How can optimal location and dimension of a 5G regional data center be determined?
- How can future predictions of traffic demand be made based on the previous data to aid dynamic utilization of resources?

1.4 Research Objectives

The key objectives of this work can be summarised as follows:

1. Develop a RAT selection scheme that categorizes subscribers into different classes based on their applications' requirements and selects the network best suited to meet the required applications' QoS. This RAT selection scheme will consider subscribers initiating single or multiple calls in a HWN by integrating user preference, application requirements, subscriber classes and dynamic network conditions (such as network cost per resource) in a completely fuzzy environment. Furthermore, the scheme will be robust and capable of overcoming ranking abnormality.
2. Analyse the traffic experienced in different areas within a city and based on the analysis determine the ideal location and best dimension for a key 5G infrastructure in form of data center.
3. Apply machine learning algorithms on previous days' demand data to obtain a forecast for the next day's traffic demand.

1.5 Research Methodology

In this work, two issues have been addressed. Firstly, an access network selection scheme is presented that focuses on the selection of the ideal network for different classes of users running multiple applications. This scheme takes into consideration the dynamic and fuzzy nature of the HWN and incorporates user preferences, application requirements and dynamic network condition to determine the best network. A novel mathematical concept in from of Intuitionistic Fuzzy (IF) Sets with multi-criteria decision making (MCDM) algorithm are used to map the dynamic wireless environment and the RAT selection process for different types of subscribers.

Secondly, the work thoroughly investigates a real world traffic dataset to examine the various traffic profiles that different zones within a city experiences. With the information obtained from these traffic profiles, data centers are optimally placed in different zones to provide computational power required to process the traffic demand. The location of data centers are obtained using Weisfeld's algorithm. The actual capacity of each data center is also determined heuristically depending on the traffic demands experienced in different areas. In addition, performances of several state-of-the-art machine learning algorithms are analysed to obtain future predictions of traffic demands to better aid the allocation of resources.

1.6 Scope and Limitations

First part of this work considers different types of applications to select the best access network among a group of alternatives. The developed scheme is suitable for use in the HWN and can be used in any wireless environment that consists of multiple alternatives. Five independent and distinct RATs with several applications of different classes are considered without loss in generality.

This scheme is also suitable for both single and multiple mobile network operator (MNO) environment. The users are assumed to be equipped with multi-mode MTs allowing them to seamlessly select a single network interface.

It is assumed that different the subscriber classes have different willingness to pay (WTP) which is agreed upon on each subscriber's service level agreements (SLAs). The subscriber classes and their attributes are assumed to be clearly stated in the SLAs.

The RAT selection scheme judges each of the considered RATs on five criteria namely: packet delay (PD) (expressed in milliseconds), packet jitter (PJ) (measured in milliseconds), packet loss rate (PL) (measured as a percentage), throughput (T) (measured in megabits per second) and security (S).

For the second part of the work, Telecom Italia's (TIM) open Big data set for Telecommunications has been used. This dataset contains call detail records (CDRs) of users within a 556 square kilometer area in the city of Milan and Province of Trento.

The CDRs are highly normalized and this work performs all the associated data processing to obtain the required information. Individual user information is not possible to be obtained from this data set.

1.7 Contributions to knowledge

The main contributions of this thesis are contained in the author's publications listed below:

1. U. Paul and O.E Falowo, "Efficient RAT Selection for Group Calls using Intuitionistic Fuzzy TOPSIS in Heterogeneous Wireless Networks", in *Proceedings of the* 13^{th} *IEEE AFRICON International Conference*, 18-20 September, 2017, Cape Town, South Africa.

2. Udita Paul, Jiamo Liu, Sebastian Troia, Olabisi Falowo and Guido Maier, " Traffic-Profile and Machine Learning Based Regional Data Center Design for 5G Network" to be submitted in *IEEE Transactions on Network and Management Services.*

1.8 Thesis Outline

This thesis is structured as follows:

- Chapter 2 gives a background and description of different generation of wireless networks. Furthermore, existing methods related to the two areas covered in this thesis namely: access network selection and data demand prediction, have been explained in details in this chapter.

- Chapter 3 presents a network selection algorithm for a subscriber initiating multiple calls. As the HWN is a dynamic environment where accurate information of parameters is impossible to obtain, this chapter presents a network selection algorithm for a user initiating group calls in a fully fuzzy environment.

- Chapter 4 utilizes a real world cellular traffic dataset to first analyse different traffic profiles that different geographical regions experience over the course of a day. Additionally, information extracted from the traffic analysis is used to optimally place and design a regional data center that would be responsible for meeting up with the traffic demand in that area. Finally, using state-of-the-art machine learning algorithms, future traffic demands are predicted to ensure dynamic utilization of resources.

- Each chapter addressees a unique research problem and finally, the thesis is concluded in chapter 5 with some recommendations for future works.

Chapter 2

Background

This chapter presents background material related to the work conducted in this thesis. It is organized as follows. Section 2.1 presents an overview of the evolution of wireless cellular network. Section 2.2 discusses handover management and access technology selection in HWNs. Section 2.3 presents concepts related to cellular traffic analysis and prediction. Section 2.4 gives a summary of this chapter.

2.1 Evolution of Wireless Cellular Networks

The evolution in wireless cellular networks has taken place in successive stages over the years leading to several generations of wireless networks. The commencement of the evolution can be traced to the first introduction of the Mobile Telephone Service (MTS) by the American Telephone and Telegraph (At & T) in 1946. This was followed by the introduction of the first commercial cellular phone concept by the Advanced Mobile Phone Service (AMPS) Inc. The introduction of AMPS signalled the beginning of the First Generation (1G) wireless networks. This generation of cellular network was designed only for voice communication. In the 1980's, second generation (2G) of cellular wireless communication was heralded by the efforts made by the European Telecommunication Standard Institute (ETSI) in developing the Global System for Mobile Communication (GSM). The 2G was dissimilar from the 1G as it used digital voice signal rather than analog signal. Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA) were employed on circuit switching voice and data transmission in 2G wireless networks.

The launch of the GSM attracted massive attention and led to the substantial increase in the global mobile wireless subscription. With the rising popularity, the GSM network service operators recognized an increase in demand and usage of the Short Message

Service (SMS) within their networks. To cater for this demand, an upgrade on the 2G network system was made to develop the 2.5G. This semi generation introduced the technology of General Packet Radio Service (GPRS) which essentially enabled packet switching on top of the existing GSM infrastructure. The GPRS technology provided a modest data rate of 14.4kbps and allowed usage of basic internet services such as email, web browsing and file transfer, in addition to maintaining voice services. However, the need to obtain better data rates remained and this led to the further modification of the 2G networks. Enhanced Data Rate for GSM Evolution (EDGE) was introduced and this provided Multimedia Message Service (MMS), in addition to the previously existing internet services. This upgrade provided significant boost in the data rate by supporting 115 Kilobits per second (Kbps) -156 Kbps rate of data transmission [8].

The introduction and subsequent rapid demand for internet services dictated higher data rate and Internet Protcol (IP) based network services. This served as an impetus and paved way for the 3^{rd} Generation (3G) of the wireless network. Universal Mobile Telecommunications System (UMTS) emerged as the standardized architecture of the 3G and it was introduced by the Third Generation Partnership Project (3GPP). However, the 3G standard simply did not render the previous editions of the wireless network useless; rather, it was made to be backward compatible with the GSM and the EDGE technologies. This effectively made it possible for this new generation of wireless network to co-exist heterogeneously with existing technologies. The 3G was based on packet switching and offered mobile subscribers multitudes of application services. Some of the landmark features of the applications involved simultaneous voice and data transmission, international roaming and advanced web services. Alongside the applications, 3G supported data rates which ranged from an impressive 2 Megabits per second (Mbps) while stationary and 384 kilobits per second (Kbps) while in motion[9].

Further modifications on the 3G networks were made by the 3GPP with subsequent introduction of the High-Speed Packet Access (HSPA) standard which supported the High Speed Down Stream Packet Access (HSDPA) technology. This feature allowed the upgraded 3G technology to support downlink speed of up to 14 Mbps. Parallel to the cellular networks, the Wireless Local Area Network (WLAN) standard also gained significant popularity in this generation of wireless networks. Several WLAN standards such as IEEE 802.11a/b/g/n allowed users to experience high data rate of 11-54Mbps. In addition, the significant ease of installation of these standards has led to their massive deployments in homes as well as highly populated public areas such as offices, airports, cafeteria and stadiums. The 3GPP also implemented standardized protocols that hs enabled offloading of cellular traffic to the WLAN standards which eased congestion on cellular networks, achieved load balancing and improved user experience [10],[11].

The fourth generation (4G) wireless networks were designed on Long Term Evolution (LTE) under the 3GPP Release 8 [12]. This generation of wireless network was crucial to cater for the increasing demand for subscribers' data, bring improvement in applications' QoS while improving subscribers' QoE and necessity to design low-complexity packet-optimized network. LTE allowed the use of some novel technological innovations such the as Multi-users Multi-input Multi-output (MIMO) Antenna, Coordinated Multiple Point (CoMP) transmission, Carrier Aggregation (CA) and Cooperative Relay Transmission (CRT) [13]. The use of such technologies gave the 4G powerful features such as multi service platform for broadband wireless access, improved user mobility, very low latency and high data peak rate of impressive 1 Gigabits per second (Gbps). 4G integrates all wireless communications technologies such as the GSM, GPRS, EDGE, HSPA and Wireless Fidelity (Wi-Fi). An upgrade of 4G was introduced in 3GPP release 10 and it was termed as the Long Term Evolution Advanced (LTE-A) Networks.

LTE-A is expected to pave way for the upcoming fifth generation (5G) wireless networks. A further release is expected from the concerned authorities on or before the year 2020. The 5G networks are predicted to feature new technologies such as Device-to-Device Communications, NFV, SDN, Machine-to-Machine communication and Internet of Things (IoT). It is also envisioned to support multitude of services such as smartphone applications, autonomous driving, health, logistics and massive IoT. As application requirements of each of these categories vary in terms of latency and throughput, 5G would incorporate a concept called 'network slicing' [14]. This would effectively create a virtual partitioning of wireless resources with each partition/slice dedicated to provide services to a specific type of user. Doing so would guarantee better QoS for different types of applications with dedicated improvement in customers' QoE while maximizing security.

Figure 2.1 depicts the several stages that wireless communications have progressed through in time. As mentioned earlier, in a HWN wireless environment, the dissimilar RATs from different generations of wireless network coexist and subscribers with multi-mode or multi-homing MT can connect to one or multiple of these RATs in order to maintain their session activities. In presence of so many alternatives, a MT often needs to be handed over to a RAT of the same technology (horizontal handoff) or different technology (vertical handoff). The term handoff or handover essentially refers to the transfer of a MT from one base station (BS) to another. As handoff is an integral part

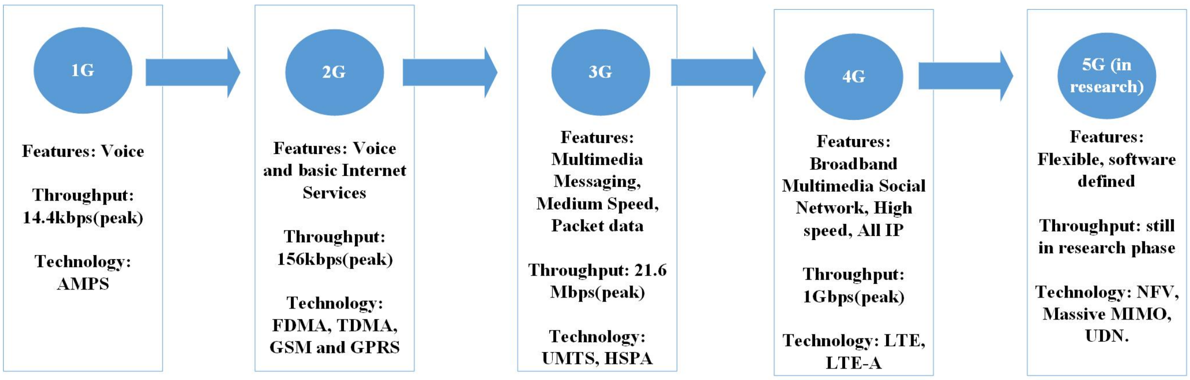

Fig. 2.1 Evolution of wireless communications through each generation

of this thesis, in the next section, handover management and access network selection in HWN is explained.

2.2 Handover Management and Radio Access Technology Selection in HWNs

2.2.1 Classification of Handoff

Handover is a vital process in ensuring the user's required QoS and appropriate usage of the scarce wireless resources. The process of handover becomes more important when a user equipment (UE) is in motion and changes its point of attachment (PoA) from one BS to another. Handover management as such, allows for the seamless migration of the MT from one cell to another. There are several benefits to handover with optimal radio resource utilization and proper load balancing being the major ones [15]. Figure 2.2 [16] illustrates the different types of handoff that exists. The major three classifications of handoff are explained below:

1. The process of handoff can be classified into horizontal and vertical depending on the type of nertwork technologies involved [17]. Horizontal or intra-system handoff takes place when the MT changes its PoA to a different cell of the same access network. An example of this type of handover is when a user moves between two geographically adjacent cells of a 4G network. For homogeneous networks, a horizontal handoff is performed when the link quality falls below a certain preset threshold. On the other hand, vertical or inter-system handoff occurs when the

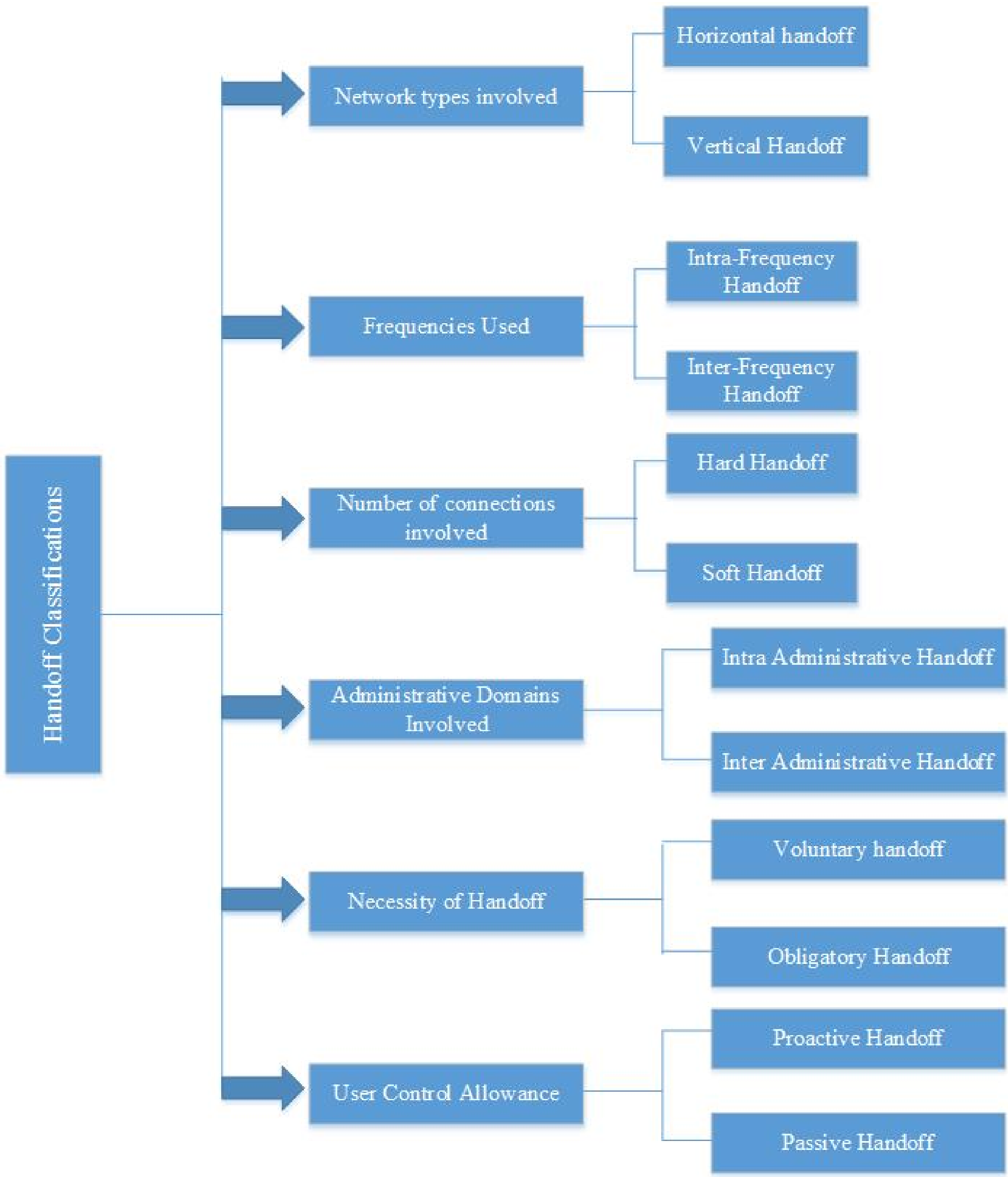

Fig. 2.2 Types of handoff [16]

MT seamlessly translates between the network interfaces of two different wireless access networks. An example would involve a MT changing its PoA from a 4G BS to a 3G BS or a Wi-Fi access point (AP). Unlike in homogeneous networks, handover in HWNs is initiated based on complex multiple criteria which involve network criteria and users' preferences.

2. Handover can further be classified into hard and soft handoff. Hard handoff employs the technique called "break before make" whereby the MT's attachment with the former PoA is first broken before a new attachment is established. Soft handoff utilizes "make before break" technique that allows A MT to maintain the old connection while making a new one. This form of handoff allows the MT to connect with multiple access network [18].

3. Handover can also be terminal controlled, terminal-assisted and network-controlled. These types of handoffs depend on the player that makes the handoff decisions [19]. Terminal assisted handoffs require the collaboration of the terminal and network in making the handoff decision in selection of the new PoA. Terminal controlled and network controlled handoff decisions are made by terminals and network respectively.

2.2.2 Characteristics of Seamless Handover

Some of the desirable characteristics of handover can be described as follows:

1. Speed : Once determined a handoff is to be performed, it needs to be executed as fast as possible to prevent degradation or interruption of service.

2. Reliability: The occurrence of handoff must not fall short in meeting the required QoS of the MT.

3. Number of handoffs: Handoffs must be performed in as limited number as possible. This is due to the fact that excessive handoff incurs overhead operational cost for the operators in addition to providing poor QoS to the MT.

4. Multiple criteria: Handoffs should not be performed based on a single criterion and should involve a group of criteria to ensure intelligent selection of the best access network for a particular MT.

2.2.3 Radio Access Technology Selection in Heterogeneous Wireless Networks

Several criteria are evaluated to make the selection of the best network for a MT in the HWN. These criteria can be related to the network, terminal, user preference and application. Each of these type of criteria are explained briefly as follows:

1. Network-related Criteria: These criteria are related to the status of the access network being considered as a possible PoA for the MT. Some of the network related criteria include signal-to-interference and noise ratio (SINR) and Received Signal Strength Indicator (RSSI). These metrics give an indication of the considered network's availability in terms of signal strength. Some of the other parameters related to network involve network load (a dynamic criteria which gives a measure of a network's present traffic scenario), network cost (a dynamic criteria which indicates the cost of receiving service from the considered access network) and network security. Another crucial metric used to judge an access network is the degree of the QoS parameters that it is capable of supporting. For example, the throughput offered by a LTE connection would be far superior to that offered by a 2G connection. Some of the QoS metrics include packet delay (PD), packet jitter (PJ), throughput (T) and packet loss (PL).

2. Terminal-related Criteria: As the MT is one of the players involved in the access network selection process, its status is also very crucial in selection of the best RAT. The most important feature of the terminal is its energy consumption while connected to a wireless interface. Some RATs might require higher battery power than others. Terminal speed, processing power and capacity could also be considered under this category.

3. User preference-related criteria: As the purpose of RAT selection is to enhance the user satisfaction of the wireless service, it is important to integrate user preferences while making the selection decision. Users may often have preference of the network they want to use to run their session activities. Also, users may have different priorities on different applications while running them in group. Such preferences of the subscribers are termed as user preference-related criteria.

4. Application related criteria: User applications generate the demand for mobile data. As applications of diverse nature such as streaming, interactive gaming and instant messaging exist, the requirements of each of these applications to run smoothly

vary significantly. While some imposes stringent demand in terms of throughput (e.g real time video streaming), others require very small delay (Voice over IP applications). Therefore, before a network is selected for a MT, the criteria related to the applications run on the MT need to be considered.

Many approaches are used in the selection of the best RAT among a pool of alternatives. As network selection involves integration of multiple criteria, MCDM algorithms have proven to be very useful in solving the network selection decision problem. Popular MCDM algorithms include simple additive weighting (SAW) [20], multiplicative exponential weighting (MEW) [21], Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [22], grey relational analysis (GRA) [23], VlseKriterijumska Optimizacija I Kompromisno (VIKOR) [24], Elimination et Choix Traduisant la REalite (ELECTRE) [25] and analytical hierarchy process (AHP) in its various forms [26–28].

Methods other than MCDM techniques have also been utilized to infer the best RAT for different subscribers. Among others, techniques such as game theory [28], utility theory [29], and fuzzy logic [30] have proved to be effective in this field . These algorithm however falls short in some aspects when compared to MCDM algorithms while inferring the best RAT.

Utility can be defined as the perceived value a subscriber obtains from using a good or service. It can be modeled in several ways and the network selection algorithms that use utility theory assign different utility functions to different users [31] or different criteria [32]. These functions are meant to model a subscriber's level of satisfaction upon receiving the services of a particular RAT. A common problem associated with employing utility theory is the increasing complexity of the algorithm with the increase in users or criteria. Even though the assumptions made in the utility theory render the algorithms simple, they also end up making the algorithms insufficient to meet up with the requirements [33].

Game theory is a branch of economics that can be employed for a scenario involving multiple participants or players (users and operators in case of HWN). The participants in game theory are modelled to compete either unselfishly (co-operative game) or selfishly (non co-operative game) for a common pool of resources. The solution or resolution of the game is achieved when the strategy for every player yields an equilibrium, also known as the Nash equilibrium. This theory has been used to model the network selection problem in some work in the literature. However, some shortcomings exist among these methods. Firstly non cooperative games assume that players are rational and would not act against the benefit of others [34], [35]. This notion however does not always hold as many at times subscribers focus on obtaining the best possible service at least cost while operators seek to maximize their profits. Also to reach the Nash equilibrium, games tend to last for many iterations leading to computationally complex algorithms that become too difficult for practical implementation. Also sometime, the games lead to a sub-optimum solution termed as the Pareto-optimum which is not always desirable.

Fuzzy logic and fuzzy inference systems have also been used tools to perform the selection of the best RAT. Fuzzy systems are able to model the uncertainty and presence of incomplete information in the HWN. A number of literatures have been able to combine fuzzy logic with MCDM techniques. However, some shortcomings exist in literatures that use fuzzy logic or fuzzy inference system (FIS) as a lone method to select the best network. The rules associated with the FIS prevents consideration of a large number of alternatives as the complexity of the system increases with each alternative [30], [36]. Due to similar reasons, only a number of criteria are also considered. In a HWN, many alternative RATs exist which need to be ranked against a significant number of alternatives. The use of FIS thus falls short in achieving that.

The MCDM algorithms have proven to be mathematically simple but efficient in determining the ranking of available RATs while judging each one of them against a set of criteria. They have also been used in a lot of literatures to aid the selection of the best RAT among a number of alternatives. Due to their efficiency in the area of access network selection, this thesis employs one of the popular MCDM technique known as TOPSIS. In the next section, the frameworks of some common MCDM algorithms are explained.

Frameworks of MCDM Algorithms

In addition to their application in wireless communications, MCDM algorithms have been widely adopted in the decision making problems in the fields of medicine, economics and other engineering problems.. All MCDM algorithms follow the basic steps of specification, normalization, ranking and selection [37] as depicted in figure 2.3. Each of these steps are described below.

Specification step: In the first step of MCDM algorithms' mathematical framework, a decision matrix is formulated. The alternatives are rated against the criteria to represent the performances of the considered alternatives for each of the criteria. An example could be seen in case of a HWN consisting of four alternative RATs of different technologies (UMTS, LTE, Wi-FI and 5G) which are to be judged against 5 criteria (PD, PJ,T,PL and security). In the specification stage, the values of each of these RATs for the each of the considered criteria would be specified and represented in form of a matrix given in

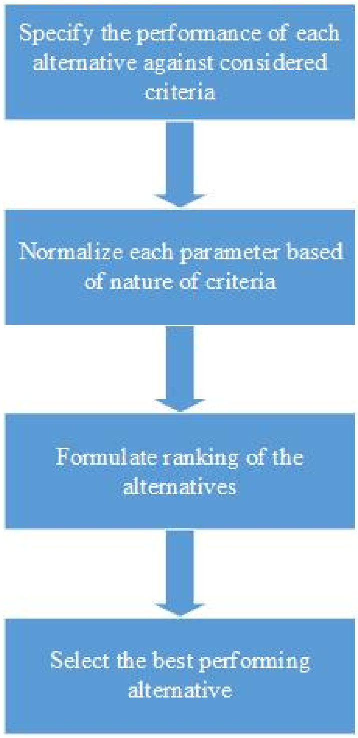

Fig. 2.3 Steps in MCDM algorithms

equation 2.1.

$$\begin{array}{c c} & \begin{array}{cccc} C_1 & C_2 & \dots & C_M \end{array} \\ \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_N \end{array} & \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NM} \end{pmatrix} \end{array} \tag{2.1}$$

where $A_1, A_2, A_3, ..., A_N$ represents the set of N alternatives available for selection and $C_1, C_2, C_3, ..., C_M$ refers to the set of M criteria which the alternatives are evaluated against. Each of the elements in the decision matrix, a_{ij} represents the performance of the i^{th} alternative against the j^{th} criterion. In case of the HWN example, if C_1 is PD and A_1 is LTE, the element a_{11} would be the performance of the LTE network against the criterion of PD (which could be between 50-100 milliseconds).

Each of these criteria however have different levels of importance in the context of a decision problem. For example, the packet delay criterion is more important for a VoIP call than the throughput criterion. Therefore, the relative importance of each of the criteria needs to be specified. This results in a set of weight values assigned to the considered criteria. These weights could be subjective or could be derived mathematically depending on the nature of the problem being solved. The set of weights, W, can be denoted as :

$$W = \{w_1, w_2, w_3, ..., w_M\} \tag{2.2}$$

where w_M is the priority given to the M^{th} criterion.

Normalization step: As MCDM algorithms need to deal with different alternatives with multitude of parameters, each of which is measured in its own unit/dimension (e.g PD is measured in milliseconds while T is measured in megabits per second), they cannot be evaluated in the same fashion. Judging them on their native units would inevitably lead to an inaccurate selection of a candidate as the best alternative. To achieve a common ground for the purpose of evaluation of all the alternatives, a popular method termed as 'normalization' is employed. Normalization transforms each of the criteria values within a common range e.g $[0, 1]$. Several techniques to achieve normalization are used in MCDM algorithms. Some of the common ones are listed below:

1. Euclidian Normalization: This method achieves normalization by dividing the specified rating of an alternative against a criterion by its norm. This can be achieved by using the expression given below:

$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{N} a_{ij}^2}} \tag{2.3}$$

where r_{ij} is the normalized value of the element a_{ij} in the decision matrix. This form of normalization depends on the values of other alternatives for the same criterion to achieve normalization.

2. Linear Max-Min Normalization: This form of normalization categorises criteria into two: benefit and cost. The benefit criteria are such that their higher values yield desirable outcome while the opposite holds true for the cost parameters. The normalized values for the benefit criteria can be represented by the equation below.

$$r_{ij} = \frac{a_{ij} - {a_j}^{min}}{{a_j}^{max} - {a_j}^{min}} \tag{2.4}$$

For the cost criteria, the normalization is achieved using the expression in equation 2.5.

$$r_{ij} = \frac{{a_j}^{max} - a_{ij}}{{a_j}^{max} - a_j min} \tag{2.5}$$

where ${a_j}^{min}$ is the least value possessed by an alternative for the j^{th} criterion and ${a_j}^{max}$ is the maximum value for the same criterion.

3. Linear Sum Based Normalization: In this method, the rating of each of the alternative is divided by the sum of the rating of all the alternatives for the considered criterion. This normalization is achieved by using the equation below.

$$r_{ij} = \frac{a_{ij}}{\sum_{i=1}^{N} a_{ij}^2} \tag{2.6}$$

4. Gaussian Normalization: In this normalization technique, two factors are considered namely: i) Average rating shift and ii) different rating scales. These two factors are integrated to achieve rating variance of each of the elements of the decision matrix. The gaussian normalization is carried out using the formula below:

$$r_{ij} = \frac{a_{ij} - \bar{a}_i}{\sqrt{\sum_{j=1}^{N}(a_{ij} - \bar{a}_i)^2}} \tag{2.7}$$

where $\bar{a}_i$ stands for the average rating of the i^{th} alternative.

Ranking and selection step: After achieving normalization of each of the element of the decision matrix, different MCDM algorithms utilize different mathematical framework to achieve ranking and subsequent selection of the best alternative. The next section provides the overview of some of the popular MCDM techniques used in RAT selection.

RAT selection using MCDM methods

MCDM methods have been a popular methodology in the area of RAT selection due to their computational simplicity coupled with ability to assess wide range of alternatives and criteria. In this section, some of the commonly used MCDM techniques used in the literature are presented.

Simple Additive Weighting: Simple additive weighting (SAW) is also termed as the weighted linear combination or scoring methods. This technique and its modified forms are simple and popularly used to solve MCDM problems [38]. This method revolves around the concept of weighted average. The ranking of an alternative (i.e one of the available networks) in SAW is obtained by a weighted value of the normalized criteria of that alternative. For example, the weighted average of the i^{th} alternative among N alternatives can be obtained by taking the product of each of its rating for the considered criteria (a_{ij}) and the assigned weight, w_j of each of the criterion. The summation of each alternative's weighted product for all the criteria is then carried out. The alternative that has the largest weighted product score emerges as the best network. The mathematical framework of SAW can be illustrated in the following steps:

1. A decision matrix is created similar to the one presented in equation 2.1. With a finite set of N alternatives and M criteria, normalization of each of the elements of the decision matrix is carried out to ensure all the criteria are made dimensionless. There are several techniques that can achieve normalization as described in the previous section. Using any of those techniques the values a_{ij} can be obtained representing the normalized value of each of the alternative's rating for the considered criteria.

2. The weight values for each of the considered criteria, w_j is obtained.

3. The SAW rank score, S^i_{SAW} of each of the alternatives is evaluated using the equation below:
$$S^i_{SAW} = \sum_{j=1}^{M} w_j \times a_{ij} \tag{2.8}$$

4. The final step of SAW involves selecting the alternative with the maximum rank score. The selected alternative would correpond to the best RAT among the available alternatives.

$$S^i_{SAW_{best}} = \arg\max_{i \in N} \sum_{j=1}^{M} w_j \times a_{ij} \tag{2.9}$$

SAW has been used to perform the task of access network selection in some work. Among others, authors in [39], have used SAW to develop a distributed vertical handoff algorithm to select network for users present in the same area of cellular network and WLANs coverage. The main objective of this scheme was to reduce the processing overhead of the MTs by handing over the computational duties from the MTs to the visiting networks (WLANs). If a MT had required higher bandwidths, throughput, lower cost of data or suffered from low terminal capacity, it would broadcast the information related to these metrics to the nearby WLANs. The WLANs would carry out the necessary calculations and convey the handoff decision to the MT which would in turn relay that to the cellular network. The cellular network upon receiving the information, if needed, performed the VHO of the MT to the optimum WLAN. SAW was used here to facilitate the handoff decision process.

Despite its simplicity, SAW possesses some inherent drawbacks which can be summarised as follows:

- SAW only deals with an environment where accurate information related to the alternatives and the criteria are available.
- It is susceptible to ranking abnormality that causes ranking order to incorrectly change if the best or worst performing alternative is dropped from the candidate list. This phenomenon is explained in details later on.

Technique for order preference by similarity to ideal solutions: Another popular MCDM technique is Technique for order preference by similarity to ideal solutions (TOPSIS) which was introduced by Yoon and Hwang [40]. The traditional TOPSIS algorithm begins with a decision matrix, M in form of $(Y \times Z)$, where set $A \in [A_1, A_2, ..., A_Y]$ denotes the available alternatives and $C_1, C_2, ..., C_Z$ denote the number of considered criteria in the decision making process. Each element of the decision matrix, a_{ij} represents how well the i^{th} alternative fares against the j^{th} criterion in the decision matrix. TOPSIS algorithm follows Euclidean normalization and breaks down the problem in a set of ideal solution and worst solution. The optimum choice becomes the one with the largest distance from the worst solution (and shortest distance from the ideal solution). The steps used in TOPSIS algorithm are described as follows:

1. The decision matrix is first defined and appropriate weight of each of the considered criterion is obtained using any suitable method. The weight vector can be represented as:
$$W = \{w_1, w_2, ..., w_Z\} \tag{2.10}$$
where
$$\sum_{j=1}^{Z} w_j = 1 \tag{2.11}$$
2. Each element of the decision matrix, a_{ij} is normalized using the Euclidean normalization technique given as:
$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{Y} a_{ij}^2}} \tag{2.12}$$
where r_{ij} is the normalized form of each of the elements of the decision matrix.
3. The weighted normalized element, s_{ij}, is then computed to integrate each original element of the decision matrix with the respective weights of the considered criteria.
$$s_{ij} = w_j \times r_{ij} \tag{2.13}$$

4. In the next step, the Positive Ideal Solution (PIS), I^+, and Negative ideal Solution. I^- are evaluated as:

$$I^+ = \{(\max_i s_{ij} | j \in J_1), (\min_i s_{ij} | j \in J_2)\} = \{s_1^+, s_2^+, ..., I_j^+, ..., s_Y^+\} \tag{2.14}$$

$$I^- = \{(\min_i s_{ij} | j \in J_1), (\max_i s_{ij} | j \in J_2)\} = \{s_1^-, s_2^-, ..., s_j^-, ..., s_Y^-\} \tag{2.15}$$

   where J_1 and J_2 are benefit and cost criteria respectively.

5. The Euclidean separations of each of the alternatives are then carried out. The distance of the i^{th} alternative from the PIS is given by:

$$I_i^+ = \sqrt{\sum_{j=1}^{Z} (s_j^+ - s_{ij})^2}, \quad i = 1, 2, 3, ..., Y. \tag{2.16}$$

   and

$$I_i^- = \sqrt{\sum_{j=1}^{Z} (s_j^- - s_{ij})^2}, \quad i = 1, 2, 3, ..., Y. \tag{2.17}$$

6. The last step of TOPSIS algorithm involves determination of closeness coefficient, C_i, of the i^th alternative. The alternative with highest value of C_i is deemed to be most suitable alternative for the problem. The value of C_i can be obtained by:

$$C_i = \frac{I_i^+}{(I_i^+ + I_i^-)} \tag{2.18}$$

TOPSIS algorithm has been extensively utilized in the literature to select the best available access network. TOPSIS, however, also suffers from the phenomenon known as the ranking abnormality. This occurs when the best or worst ranked alternative suddenly disappears from the selection pool. This absence of alternative(s) leads to a change in the ranking order and subsequent selection of a wrong alternative as the best. Some of the work in the literature have attempted to overcome/minimize the effect of ranking abnormality in TOPSIS. In [41], Lahby *et al* proposed use of diff-Analytical Hierarchy Process (AHP) with TOPSIS in order to select the access network suited to meet the QoS requirement of different subscribers. Six different criteria namely: cost per byte, available bandwidth, packet loss, security, packet delay and packet jitter and three RATs in form of UMTS, WiFi and Worldwide Interoperability for Microwave Access (WiMAX) were considered for the network selection problem. The weights of each criterion was obtained using a modified version of the AHP and the TOPSIS algorithm was employed to obtain the ultimate ranking of the alternatives. Their simulation results showed that the proposed algorithm managed to reduce the ranking abnormality than the conventional AHP and TOPSIS algorithm.

In [42], the authors combine fuzzy logic with TOPSIS algorithm to select the most energy efficient RAT in HWN. They modeled QoS requirements of different types of applications by using parametrized utility functions. The utility functions also aided in elimination of ranking abnormality from TOPSIS algorithm. User preferences for different applications were input into the scheme using linguistic variables. The overall ranking and the selection of the best network was performed using fuzzy-set representation of the TOPSIS algorithm.

The authors in [43] modified conventional TOPSIS algorithm to formulate a network selection framework involving seven criteria (cost per byte, total bandwidth, allowed bandwidth, utilization, packet delay, packet jitter and packet loss) and five available alternatives (UMTS, IEEE 802.11b, IEEE 802.11a and IEEE 802.11n and 4G). They identified traditional method of normalization as a chief cause of occurrence of ranking abnormality in TOPSIS. They proposed an iterative approach whereby the worst performing network is eliminated after each iteration. After that, if all the remaining alternative networks' closeness coefficients were too close to each other for one to be selected as the best, the network with the lowest cost was selected as the preferred option. Through simulations, they demonstrated the effectiveness of their algorithm in eliminating ranking abnormality.

Senouci *et al* in [44], replaced the traditional normalization technique in TOPSIS with utility functions to normalize the values of each of the considered criterion. The use of utility functions ensured that the normalization of each element of the decision matrix was carried out irrespective of the values of other elements. Their simulation results proved the effectivity of their proposed mechanism in eliminating ranking abnormality in TOPSIS algorithm.

VlseKriterijumska Optimizacija I Kompromisno: VlseKriterijumska Optimizacija I Kompromisno (VIKOR) is another effective MCDM tool first introduced in [45]. Two parameters namely: i) Maximum group utility and ii) Minimum individual regret, are crucial to every VIKOR algorithm. The ranking in VIKOR is obtained by taking these two factors into consideration. The steps involved in VIKOR algorithm can be summarized as follows:

1. The decision matrix is specified with each element , y_{ij} of the matrix representing the performance of the i^{th} alternative, among M alternatives for the j^{th} criterion. The weight of the considered criteria are also obtained. The weight decision vector

can be represented as:

$$W = \{w_j, j = 1, 2, ..., N\} \tag{2.19}$$

where N is the total number of attributes.

2. The elements in the decision matrix is normalized using the Euclidean normalization tactic given by:

$$n_{ij} = \frac{y_{ij}}{\sqrt{\sum_{i=1}^{M}(y_{ij})^2}} \quad i = 1, 2, ..., M. \tag{2.20}$$

where n_{ij} is the normalized form of the element and M is the total number of alternatives.

3. An updated matrix with the normalized values is then created.

4. In the next step, the best (n_j^+) and worst (n_j^-) values of all the criteria are computed. The criteria are divided into benefit criteria (J_1) and cost criteria (J_2). The best and worst values are obtained as follows:

$$n_j^+ = \max_i\{n_{ij}, J_1 = 1, ..., n\}; \quad n_j^- = \min_i\{n_{ij}, J_1 = 1, ..., n\} \tag{2.21}$$

$$n_j^+ = \min_i\{n_{ij}, J_2 = 1, ..., n\}; \quad n_j^- = \max_i\{n_{ij}, J_2 = 1, ..., n\} \tag{2.22}$$

5. The computation of utility, S_i and the regret measure R_i for each of the i^{th} alternative is carried out next. They are calculated using the expressions below:

$$S_i = \sum_{j=1}^{N}\left[\frac{(n_j^+ - n_{ij}) \times w_j}{n_j^+ - n_j^-}\right] \tag{2.23}$$

$$R_i = \max_j\left[\frac{(n_j^+ - n_{ij}) \times w_j}{n_j^+ - n_j^-}\right] \tag{2.24}$$

6. VIKOR algorithm's ranking index value, Q_i is then calculated using the following formula:

$$Q_i = \lambda\frac{(S_i - S^-)}{S^+ - S^-} + (1 - \lambda)\frac{(R_i - R^-)}{R^+ - R^-} \tag{2.25}$$

where $S^+ = \max_i S_i$, $S^- = \min_i S_i$, $R^+ = \max_i R_i$ and $R^- = \min_i R_i$. λ represents the weight of the strategy of the maximum group utility (usually taken to be 0.5).

7. Upon obtaining the Q, S and R values of all the alternatives, the candidates are arranged in descending order of each of these parameters. The optimum alternative

becomes the one with minimum Q score and is marked as $Q^{[1]}$. However, two very important conditions must hold true for this alternative to be selected:

- Acceptable advantage: $Q^{[2]} - Q^{[1]} \geq \frac{1}{N-1}$, where $Q^{[2]}$ is the second best alternative and N is the total number of criteria. If this condition is not met, a set of compromised solutions , $Q^{[1]}, Q^{[2]}, ..., Q^{[K]}$ are proposed. $Q^{[K]}$ is determined by: $Q^{[K]} - Q^{[1]} < \frac{1}{m-1}$.
- Acceptable stability in decision making: The alternative $Q^{[1]}$ needs to be best ranked by S_i and R_i as well. Else, alternatives $Q^{[1]}$ and $Q^{[2]}$ are proposed as alternative solutions.

Some work in the literature have used VIKOR algorithm to formulate the network selection problem. Like TOPSIS, VIKOR also suffers from the ranking abnormality problem. Baghla *et al* in [46] used several normalization and weighting techniques to judge three alternative networks (UMTS, WLAN and WiMAX) against six criteria (cost per byte, delay, jitter, available bandwidth, security and packet loss). The network selection was performed using the VIKOR algorithm for three sets of traffics (background, conversational, streaming and interactive). Their simulation results showed that Euclidean normalization technique provided the best performance with reduced ranking abnormality and number of handovers for the considered traffic classes.

A common feature noticed among all these MCDM techniques is their dependencies on accurate and exact information related to the criteria and alternative while selecting an ideal alternative from a group. However, as the HWN is a highly unpredictable and dynamic environment where accurate information are hard to come by, algorithms are required that can effectively deal with the ambiguity. Developing such an algorithm is one of the goals of this thesis.

2.3 Traffic Analysis and Prediction in Wired and Wireless Networks

Another objective of this thesis is to analyse and predict the cellular traffic volume that different coverage zones within a major city experiences over a duration of time. Given the widespread usage of the mobile broadband, cellular data trace potentially provides more information related to network behaviours than the traditional voice traffic. Understanding traffic pattern can facilitate network planning while establishing key components such as data centers. In addition, with the help of forecasting algorithms, previous traffic pattern can be used to obtain future prediction values to better aid utilization of resources.

Many work has been conducted in the areas of traffic analysis and prediction in the literature. Popular algorithms used in the field of network traffic forecast can be broadly classified as linear and non linear models. The most commonly used linear models are the Autoregressive Moving Average (ARMA)/ Autoregressive Integrated Moving Average (ARIMA) and HoltWinters algorithm. On the other hand, machine learning algorithms have proved to be very effective among non linear models. The framework of these algorithms are briefly explained below.

ARMA/ARIMA: ARMA is a time domain approach that is used to forecast future values based on the current and/or past observation of parameters[47]. This models assumes the errors to be Gaussian White Noise, therefore there exists no autocorrelation between the errors [48]. As ARMA is a combination of the auto-regressive and moving average models, each of them can be used independently. The models are discussed below.

AR Model: This auto-regressive model states that the current output, y_t, is a linear combination of its past values. p^{th} order Auto-regressive model can be represented as:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \epsilon_t \tag{2.26}$$

The current output, y_t can therefore be denoted as:

$$y_t = \epsilon_t + \sum_{x=1}^{p} a_x y_{t-x} \tag{2.27}$$

ϵ_t here represents the Gaussian White Noise.

MA Model: The moving average model makes use of the q^{th} order and its relation with the current output is represented by the equation below:

$$y_t = \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + ... + b_q \epsilon_{t-q} \tag{2.28}$$

The current output for this model becomes:

$$y_t = \epsilon_t + \sum_{x=1}^{q} b_x \epsilon_{t-x} \tag{2.29}$$

The ARMA model combines the previous two models to for yield the present output of:

$$y_t = \epsilon_t + \sum_{x=1}^{p} a_x y_{t-x} + \sum_{m=1}^{q} b_m \epsilon_{t-m} \tag{2.30}$$

Using the above mathematical model, ARMA's prediction framework can be used to predict a future value for the $t+1$ time step using the following equations:

$$y_{t+1} = \epsilon_{t+1} + \sum_{x=1}^{p} a_x y_{t-x+1} + \sum_{m=1}^{q} b_m \epsilon_{t-m+1} \tag{2.31}$$

In general, ARMA's prediction for the h^{th} time step in future can be denoted by:

$$y_{t+h} = \epsilon_{t+h} + \sum_{x=1}^{p} a_x y_{t-x+h} + \sum_{m=1}^{q} b_m \epsilon_{t-m+h} \tag{2.32}$$

ARIMA model is similar to the ARMA model but uses an additional parameter in the form of d which is termed as the level of differencing parameter. These models have been used to make cellular traffic prediction in the literature. The authors in [49] investigated the performance of a modified version of the ARIMA known as the Seasonal ARIMA (SARIMA) on a real world 3G data set. Data samples were aggregated in both spatial and temporal domains and the forecasting accuracies of SARIMA was compared with other algorithms. SARIMA performed reasonably well in both spatial and temporal domains. However, in addition to being a linear model, ARIMA is often criticized for being computationally complex.

Holt-Winters Method: Holt Winters is a short term forecasting method and it is a combination of the Holt Method and the Winter method. This method is deployed to predict data that exhibits trend and possess seasonal component[50]. This method relies on three smoothing equations, individually called the level, the trends, and the seasonal smoothing equations. Holt-Winters additive prediction model can be additive and multiplicative in nature. Both these forms are explained below:

Holt-Winters Additive Model : This model is utilized when the size of the data remains independent of the seasonality. The forecast value for a one time step, t is obtained using:

$$F_t = L_{t-1} + B_{t-1} + S_{t-s} \tag{2.33}$$

To forecast m time steps ahead, the following equation is used:

$$F_{t+m} = L_t + mB_{tm} + S_{t-s+m} \tag{2.34}$$

where l_t, B_t and S_t denote the overall, trend and seasonal components of the smoothing equations respectively. These smoothing components at time t are given as:

$$L_t = \alpha(X_t - S_{t-s}) + (1-\alpha)(L_{t-1} + B_{t-1}) \tag{2.35}$$

$$B_t = \beta(L_t - L_{t-1}) + (1-\beta)B_{t-1} \tag{2.36}$$

$$S_t = \gamma(X_t - L_t) + (1-\gamma)S_{t-s} \tag{2.37}$$

where α, β and γ correspond to the level, trend and seasonal smoothing coefficients. The smoothing coefficients belong in the set $[0, 1]$. s represents the seasonal duration and X_t represents the actual value at time step t.

Holt-Winters Multiplicative Model: This model is employed when the size of the data depends on the seasonal pattern. The multiplicative forecast at time step m can be denoted by:

$$F_{t+m} = (L_t + mB_t)S_{t+m} \tag{2.38}$$

where L, B and S are given by:

$$L_t = \alpha\frac{X_t}{S_s} + (1-\alpha)(L_{t-1} + B_{t-1}) \tag{2.39}$$

$$B_t = \beta(L_t - L_{t-1}) + (1-\beta)B_{t-1} \tag{2.40}$$

$$S_t = \gamma(\frac{X_t}{(L_{t-1} + B_{t-1})}) + (1-\gamma)S_{t-s} \tag{2.41}$$

In the work conducted by the authors in [51] employed the Holt-Winters' exponential smoothing model to predict future traffic demand in the circuit switched services. They performed statistical analysis on a traffic data set and using historical traffic data of GSM/GPRS networks made prediction of future demands.

Machine Learning Algorithms: In recent years, machine learning algorithms have emerged to perform prediction tasks due to their ability to model highly complex non-linear data. In particular, artificial neural networks (ANN), since their origin [52], have gained popularity due to the advantages presented by their various models. ANNs have been employed to solve problems in diverse fields such as robotics [53], real time sign language translation [54], speech, handwriting and object recognition [55, 56]. ANN and its model has been particularly successful in dealing with complex data that are registered over a period of time. When these time dependent data are collected together, the measurements form what is known as a *Time-Series*. Part of the work conducted in this thesis uses hourly cellular traffic demand records which also belongs to the domain of Time-Series. It is therefore prudent to discuss some of the common ANN models that are used for time-series analysis and prediction.

Feed Forward Neural Network (FFNN) : Every neural network consists of units termed as the neurons which process information that are fed in as inputs to produce an output. Between the input layer and output layer, there could be various layers of neurons, with each layers containing different number of neurons. The layer(s) between the input and output of a neural network is called the hidden layer. Feed forward neural network is a simple form of ANN where the connections between different units or neurons do not form a cycle. They are called feed forward due to their ability to transfer information in only forward direction in the network.Figure 2.4 shows a feed forward neural network model with an input layer, an output layer and two hidden layers.

Feed forward neural networks take in an input set of X that contains n number of inputs $[x_1, x_2, ..., x_n]$. This set is also termed as the training set and their outputs are known. These inputs are then multiplied with various weight matrices, w, as they combine with various neurons in the hidden layer. In case of Figure 2.4, there are two (s_1 and s_2) and three (s_3,s_4 and s_5) neurons in the first and second hidden layer respectively. Depending on the task being carried out, the input of previous layer combines with the next layer with the aid of activation functions. Several popular activation functions include: hyperbolic tangent, rectified linear unit (ReLU) and sigmoid. The inputs, after their various interactions with the neurons produce the output set of Y. The weights are updated after each iteration as the network attempts to 'learn' over time and reduce the error that exists between the network's output and the actual output. A popular algorithm with which most neural networks manage to learn over time is the backpropagation algorithm [57]. Given its simpler forward architecture, feed forward neural network is not fully capable of learning the trend present on larger time series data set. This restricts their usage on time series analysis.

Convolutional Neural Network (CNN): CNN [58] has proven to be very effective in areas such as image processing. CNN is similar to feed forward neural network in that it also consists of neurons and learn-able weights. Unlike feed forward neural network, the CNN architecture arranges its neurons in three dimensions: width, height and depth. Instead of all the neurons in a layer being fully connected to the previous layer's neurons, in CNN, the neurons in a layer only connects to a smaller region of the layer before it. The three main types of layers present in the CNN architecture are called the convolutional layer, pooling layer and fully connected layer respectively. Each of these layers accepts

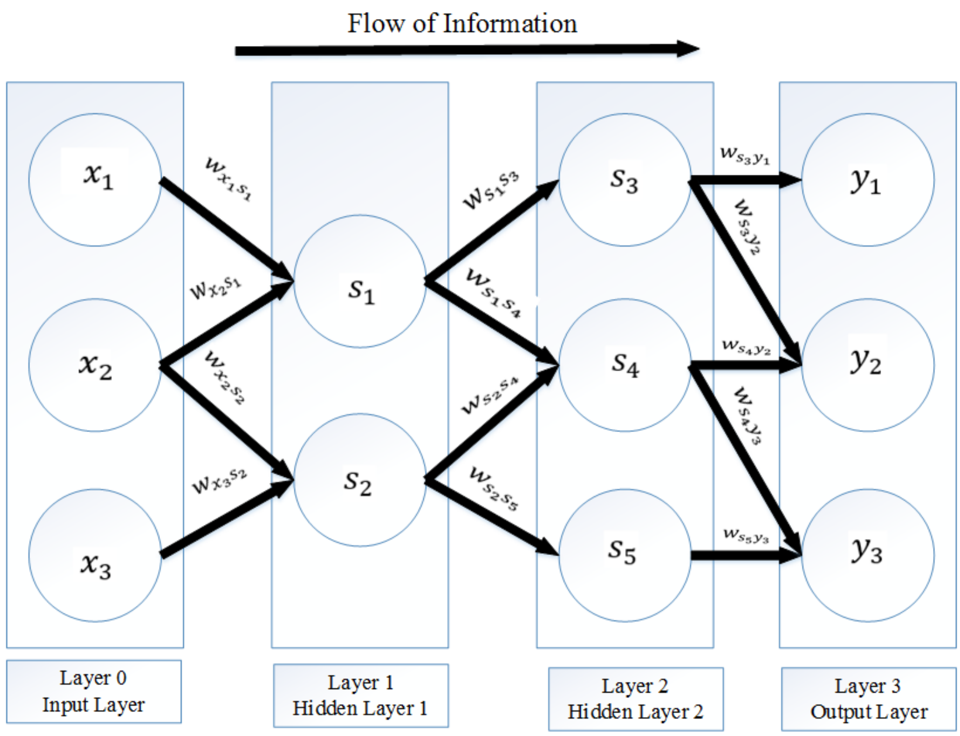

Fig. 2.4 Feed Forward Neural Network

an input in 3D and transforms it to an output that is also 3D in nature. Figure 2.5 show how an image is processed by the CNN in its various layers [59].

CNN, even with its success in areas of pattern, image and object recognition, remain sparsely used in the area of time-series analysis. This could be attributed to its slightly complicated architecture that, although makes it suitable for recognition tasks, does not aid in scenarios where trends need to be learned.

Recurrent Neural Network (RNN): RNN's main idea is to capture and store relevant amount of information from the input in a memory to use it while making a future

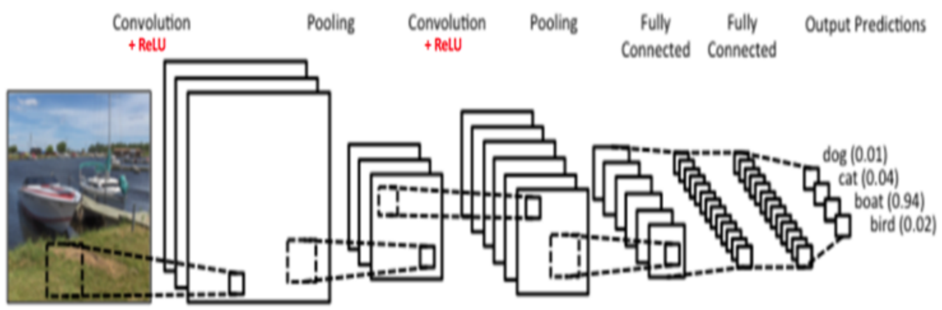

Fig. 2.5 Convolutional Neural Network

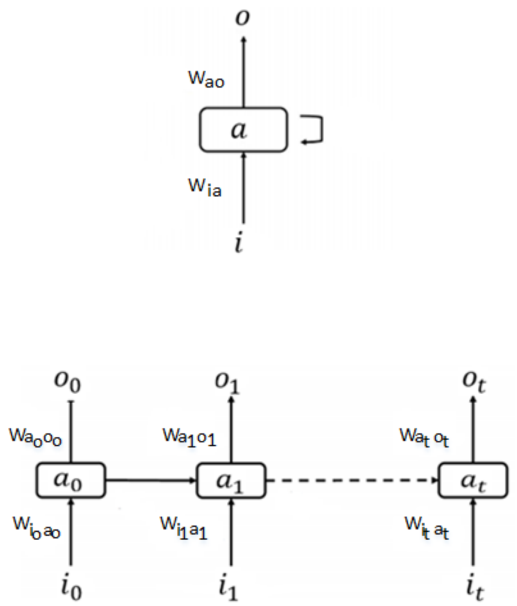

Fig. 2.6 Recurrent Neural Network

prediction for the output[60]. This is a fundamental difference between RNN and traditional feed forward neural networks that simply make use of only the present input to produce an output. RNNs are termed as recurrent as they perform the same operation on every element of a sequence whereby the output of the present step is heavily impacted by that of previous steps. Figure 2.6 illustrates a typical RNN model.

RNN takes in the input i, captures the hidden state a and produces an output of o at every time step t. The information from one step to the following is carried on by a loop. The W's stand for various weight matrices during the time steps. These matrices are changed during the training phase as the network is 'unrolled' for a certain number of time steps. As shown in Figure 2.6, this unrolling of network in time steps allow the RNN to learn information present in sequential data. The computation that takes place in every time step can be summarised as follows:

1. i_t serves as the input in time step t.

2. The hidden state a_t at time step t is calculated based on the previous hidden step and the present input. These two pieces of information are combined through the use of activation functions such as ReLU and tanh: $a_t = relu(W_{ia}i_t + W_{aa}h_{t-1} + b_a)$ or $tanh(W_{ia}i_t + W_{aa}h_{t-1} + b_a)$.

3. The output step at time step t is termed as o_t.

With different inputs i_t in different time steps same computations are performed with unrolled parameters W_{ia}, W_{aa} and W_{ao}. This attribute of the RNNs makes them extremely useful for smaller data set by avoiding over fitting. Two common RNN models in use now are the Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). Ae brief explanation of the working principles of these models along with the activation functions are provided below.

LSTM: The hidden state in traditional RNN does not provide enough control over how much of the past information should be kept and this leads to problems such as vanishing and exploding gradients [61]. To overcome such problems, LSTM models were designed to have two additional gates termed as the input and forget gates. The gating mechanism allows LSTMs to adequately model long-term dependencies present in complex non linear data. LSTM essentially learns the optimal parameters for its gates during the training phase, thereby determining the behavior of its memory. Interested readers are requested to read [62] for more details on LSTM.

GRU: Due to the presence of both input and forget gates, the LSTM model often becomes computationally expensive. GRU, a more recent edition of the RNN models, presents a simpler architecture where the input and forget gates are combined into a update gate. The basic idea of capturing and learning long term dependencies on time series data is however maintained in GRU as well. Detailed explanation regarding the GRU model can be found in [63].

RNN and its models have been extensively used in analysing and predicting time-series data. A detailed literature of works that employ these models in forecasting future values is presented in Chapter 4.

2.4 Chapter Summary

This chapter presented an overview of the progression of the wireless communication over the course of time starting from first generation (1G) networks to upcoming fifth generation (5G) networks. The technologies developed in each of the generations were explained and their specifications have also been provided.

In a HWN with so many types of RATs available, the selection of a access network and subsequent handover from the present network to the new network play a crucial role in ensuring maximum subscriber satisfaction. The different types and stages of the handover process have been explained. Also the desirable features of a seamless handover, which results in enhancing application's QoS and subscriber's QoE, have been stated.

As the selection of the best access network among a pool of alternatives is a complex task, a comprehensive study of different methods used in achieving this has also been presented. Some work carried out in the field of access network selection using various MCDM and other techniques have also been explained.

To address the challenges posed while making predictions of future traffic demand that MNOs face, several techniques are employed. The technique could be either linear or non-linear. Some of the popular algorithms have been discussed and some existing work done in the field have been highlighted.

Chapter 3

Efficient RAT-Selection Scheme for Group Calls using Intuitionistic Fuzzy TOPSIS in Heterogeneous Wireless Networks

3.1 Introduction

With massive proliferation of mobile devices having advanced capabilities coupled with increasing popularization of bandwidth intensive applications such as YouTube and Netflix, the wireless networks are experiencing a huge surge in data demand. As reported in [1], there will be 5.5 billion global mobile users, with each generating 5.7 GB/month of data. The majority of data (78%) is envisioned to be arising from video related applications. The 5G network is being developed with multiple features to cater for this astronomical demand for data while meeting stringent requirements of diverse user's applications. The introduction of 5G would signal the need for integration and interworking of 5G RATs with legacy RATs to provide ubiquitous coverage in future wireless network environment.

The MTs in HWNs require to migrate from one access technology to another, if need be, in order to maximize the benefits of being under the coverage of several alternatives. The RATs in HWN also tend to vary in terms of some parameters such network profile configuration, support for various applications, cost of service and offered security. Selection of the appropriate network by different users therefore becomes a crucial challenge. The MTs are also capable of simultaneously running multiple classes of applications (group calls/applications) comprising of independent calls such as interactive gaming, software download, video streaming and voice calls. Each of these calls can be placed simultaneously and are required to be sustained while the MT is connected to an access network within a HWN.

The HWNs consists of different types of access networks such as Long Term Evolution (LTE), Universal Mobile Telecommunications Systems (UMTS), Worldwide Interoperability for Microwave Access (WiMAX) and Wireless Local Area Networks (WLANs). Each of these access technologies vary in their performances in criteria such as packet delay, packet jitter, throughput, packet loss and offered security. As such, when a user initiates and places multiple applications simultaneously, each of these access networks perform differently in attempt to support the calls. The access network is required to provide the needed QoS for all the applications being run while enhancing user's QoE. In case the connected PoA fails to fulfil the required demand of the group of applications, it is important for an efficient network selection algorithm to sort and select the access network most suited to handle the applications. As such seamless vertical handover (VHO) schemes are needed to be designed which allows fast, smooth and seamless switching between different access networks with minimal delay.

Traditional single criterion access selection scheme falls short in combining multitudes of parameters that play an important role in the determination of the most suitable access network for different users. These different criteria can be adequately modelled as MCDM problems which is an optimization research technique developed to resolve multi-criteria decision problems. When group calls/applications are considered, the MCDM problem turns to MCGDM (Multi-Criteria Group-Decision Making). Many MCDM frameworks such as Simple Additive Weighting (SAW), Multiplicative Exponential Weighting (MEW), Grey Relational Analysis (GRA), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and VlseKriterijumska Optimizacija I Kompromisno (VIKOR) have been utilized to solve the network selection problem in HWN.

This chapter proposes a new and robust methodology using an advanced type of fuzzy set theory called Intuitionistic Fuzzy (IF) set to select the best access network for different types of users. The IF set provides an unique advantage over traditional MCDM algorithms by being able to model the unpredictability and unavailability of precise information that are inherent characteristic of HWNs. The main contributions of this chapter are summarised as follows:

- A comprehensive review of the existing network selection algorithms for group calls is presented.

- A branch of the fuzzy set: Intuitionistic Fuzzy set, with a popular MCDM algorithm - TOPSIS is employed to formulate a network selection scheme for different types of users in a completely fuzzified environment. The idea behind using this branch of fuzzy set is to mirror the ambiguous and unpredictable nature of the HWN.

- The ability of the proposed scheme to cater for the need of different classes of users initiating single and/or multiple calls is demonstrated.

- The performance of the proposed scheme in elimination of ranking abnormality that affect traditional MCDM schemes is evaluated.

This is the first work to utilize Intuitionistic Fuzzy logic with TOPSIS to formulate a network selection algorithm for a user simultaneously initiating multiple calls. This work provides a new angle to the area of access network selection by properly modelling the dynamic nature of the HWN.

The rest of the chapter is organised as follows: Section 3.2 presents the work related to the access network selection field. Section 3.3 introduces the IF-TOPSIS algorithm. Section 3.4 provides information related to the application of the IF-TOPSIS algorithm for users initiating single and multiple calls. Simulation results highlighting the performance of the proposed algorithm are presented and discussed in section 3.5. Finally, the chapter is concluded in section 3.6.

3.2 Related Work

This section provides some recent work that has focused on access network selection using some of the mathematical tools used in this chapter. The work conducted by Tran *et al.* in [64], compared the performances of SAW, Weighting product (WP also known as Multiplicative exponential weighting) and TOPSIS algorithm in selecting the best network interface in HWN. They considered three networks (UMTS, IEEE 802.11b, IEEE 802.11n and 4G) and five criteria (jitter, delay, utilization of the wireless link, packet loss and cost per byte). Their simulation results demonstrated that the SAW and WP is prone to producing inaccurate ranking of alternatives and the TOPSIS algorithm suffers from ranking abnormality. The work further proposed a modification of the TOPSIS algorithm called the Distance to ideal Alternative (DiA). The DiA algorithm managed to reduce the occurrences of ranking abnormality but fell short in completely eliminating it. Further attempts have also been made in [41, 65] to overcome the problem of ranking abnormality in the TOPSIS algorithm. Even though, these works addressed the problem, they did not succeed in completely eliminating it.

In addition, all these works assumed that information related to the parameters such as throughput, delay and jitter can be precisely obtained at any time. HWN, on the other hand is a fuzzy environment and as such getting accurate measurement of these parameters becomes a difficult task. Fuzzy logic has proved to be a useful tool in modelling the uncertainty that persists in the HWN. Prakash *et al.* in [66] proposed a multi attribute decision making process for network selection using the IF-TOPSIS framework. Their proposed scheme considered multiple decision makers running different applications. The best network was selected by considering user preferences and network parameters such as available bandwidth, delay, packet loss and cost. Their work, however, fell short in extending the IF-TOPSIS framework to incorporate multiple applications initiated by a single user. In addition, they modelled cost of service of each network to be fixed. As prices of data tend to vary during different times of the day, the scheme proposed in this chapter models cost as a dynamic parameter.

The proposed methodology in [67] focused on the use of fuzzy logic along with a MCDM technique to select the best RAT for different sets of users. Two fuzzy logic controllers, with set inference rules, were used in this work to determine the ability of two RATs (WLAN and LTE) to support the QoS requirements of two types of traffic (VoIP and File Transfer Protocol). Their results demonstrated the effectiveness of fuzzy logic in modelling the imprecise and uncertain status of the wireless environment. However, their scheme fell short in selecting the RAT most suitable for subscribers running multiple applications at the same time which is a prevalent case in HWNs.

The work conducted by authors in [68] focused on the vehicles' ability to select the best network as it moves within the coverage area of different access points (APs). The distance between an access point and a vehicle served as one of the inputs to a fuzzy logic inference system with the vehicle's speed acting as the second input. The fuzzy inference system was charged with evaluating these two inputs based on the set rules, and determine the probability of selecting the AP in question. The output of the FIS system was meant to decide the suitability of a particular candidate network for handover. However, the work fell short in placing emphasis on the role played by applications of different nature on the network selection process.

In [69], the concept of non-biogeography based optimization (NHBBO) was employed in conjunction with parallel fuzzy system (PFS) to determine the ideal network. The PFS, with set rules, served to yield the value of the probability of selecting a particular access network among WLAN and Wireless Wide Area Network (WWAN). Five criteria namely the speed of the mobile terminal, the received signal strength, network coverage, delay and data rate were considered as the inputs to the PFS. The outputs of the PFS provided the inputs to the multi-point parameter algorithms containing the weight functions. The weight functions represented the degree of importance of each of the considered criteria, with the criterion having the highest weight representing one being of utmost importance. The NHBBO algorithm was applied over the multi-point decision module with the purpose of optimizing the weight functions to determine the most suitable access network. The scheme however did not consider single or multiple applications while making a decision regarding the ideal RAT for a subscriber.

Most of the network selection schemes place their emphasis on selecting the best RAT for a user initiating single calls. However, using new generation MTs, a user now can place multiple applications on a single RAT at the same time. To address the problem of selecting a suitable RAT for a user initiating multiple calls, the schemes in [70–73] were developed. The work in [70] and [72] gave users the opportunity of choosing weights of various criteria while initiating a call consisting of multiple applications. However, as subscribers might not have the adequate technical knowledge to do so on their own, the scheme developed in this dissertation assigns weight to each application at an algorithm level that is independent of the user's interference. The work in [71] used a novel MCDM technique called MULTIplicative-form with Multi-Objective Optimization Ratio Analysis (MULTIMOORA) to model a network selection problem for subscribers running multiple applications simultaneously. This work assumed accurate information of criteria such as data rate, delay and jitter are always available which is not the case in the highly dynamic heterogeneous wireless environment. The work in [73] developed an analytical model for group calls and investigated the effect of call dynamics on RAT selection in HWN. This scheme considered only stochastic modelling and did not focus on the user preferences, application requirements and network parameters In [74], the author demonstrated the use of a MCDM technique called Complex PROportional ASsessment (COPRAS) to formulate a access network selection problem for group calls. Like other group call network selection schemes, this also assumed availability of information related to network criteria at all times. In addition, the problem of ranking abnormality common to most MCDM algorithm were not addressed in any of the work involving group calls.

After analysing the above mentioned work the following shortcomings are observed:

- Existing works are mainly focused on the vertical handover for single calls and does not place emphasis on group applications barring the work done in [70–74].

- The existing schemes assumes the availability of accurate information related to network parameters at all times while selecting the best RAT for a subscriber initiating multiple calls .

- The works done so far falls short in overcoming the ranking abnormality that affects MCDM algorithms such as TOPSIS.

In this work, the Intuitionistic Fuzzy TOPSIS method is used in the network selection process. The proposed methodology considers group of applications, user preferences and network condition while performing the selection decision in a completely fuzzy environment.

3.3 Intuitionistic Fuzzy Set with TOPSIS Algorithm

This section introduces the Intuitionistic Fuzzy set and explains its integration with the TOPSIS algorithm to formulate a MCDM algorithm capable of modelling single and group decision making processes in a fuzzy environment.

3.3.1 Intuitionistic Fuzzy Set

IF set developed by Atanassov [75], is characterized by a membership function, a non-membership function, and a hesitancy function. It is ideal for modelling the uncertainty and the vagueness associated with heterogeneous wireless networks. Assuming a fixed set X, Atanassov defined an intuitionistic fuzzy set (A-IFS) as $A = \{< x, \mu_A(x), \nu_A(x) > |x \in X\}$, which assigns to each element x, a membership degree $\mu_A(x)$ and a non-membership degree $\nu_A(x)$, with the condition $0 \leq \mu_A(x) + \nu_A(x) \leq 1$. In addition, $\Pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$ is called the indeterminacy degree of x to A.

The IF set has been used effectively in many multi criteria group decision making problems due to its greater superiority in dealing with ambiguity. In [76],[77] and [78], the authors have used the intutionistic fuzzy set alone or with other methods to evaluate the suitability of different candidate alternatives while making a group decision.

3.3.2 IF-TOPSIS Framework

The integration of IF set with TOPSIS algorithm is elaborated here. For a set of alternatives $A = \{A_1, A_2, ..., A_m\}$ and $X = \{X_1, X_2, ..., X_n\}$ criteria, the steps used in IF-TOPSIS framework can be summarised as follows:

Step 1: Determination of the weight of the decision makers.

For a decision group containing M decision makers, the importance of each decision maker is considered as linguistic terms expressed in intuitionistic fuzzy numbers.

Let $D_k = [\mu_k, \nu_k, \pi_k]$ be an intuitionistic fuzzy number for rating of kth decision maker. Then the weight of this decision maker, λ_k, can be obtained as:

$$\lambda_k = \frac{(\mu_k + \pi_k.(\frac{\mu_k}{\mu_k+\nu_k})}{\sum_{k=1}^{M}(\mu_k + \pi_k.(\frac{\mu_k}{\mu_k+\nu_k})} \tag{3.1}$$

and $\sum_{k=1}^{M} \lambda_k = 1$.

Step 2: Based on the opinions of the decision makers, the intuitionistic fuzzy decision matrix is constructed. Let $R^k = (r_{ij}^k)_{m\times n}$ be the intuitionistic fuzzy decision matrix of the k^{th} decision maker. The importances of each decision maker can be contained in a set of λ where $\lambda = \{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_M)$. In order to obtain a group decision, individual decision opinions are fused to formulate an aggregated intuitionistic fuzzy decision matrix. This is carried out using the method proposed in [79]. $R = (r_{m\times n})$, where

$$\begin{aligned} & r_{ij} = IFWA_\lambda(r_{ij}^{(1)}, r_{ij}^{(2)}, ...r_{ij}^{(M)}) = \\ & \lambda_1 r_{ij}^{(1)} \bigoplus \lambda_2 r_{ij}^{(2)} ... \lambda_M r_{ij}^{(M)} = [1 - \prod_{k=1}^{M}(1-\mu_{ij})^{\lambda_k}, \\ & \prod_{k=1}^{M}(\nu_{ij})^{\lambda_k}, \prod_{k=1}^{M}(1-\mu_{ij})^{\lambda_k} - \prod_{k=1}^{M}(\nu_{ij})^{\lambda_k}] \end{aligned} \tag{3.2}$$

Here $r_{ij} = (\mu_{A_i}(x_j), \nu_{A_i}(x_j), \pi_{A_i}(x_j))$ for $(i = 1, 2, ..., m; j = 1, 2, ..., n).r_{ij}$ represents the assessment of a decision maker of the i^{th} alternative on the j^{th} criterion. The aggregated intuitionistic fuzzy decision matrix of all the decision makers, R, can be defined as follows:

$$R = (r_{ij})_{m\times n} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{11} & r_{12} & \dots & r_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \dots & r_{mn} \end{bmatrix} \tag{3.3}$$

Each row in this matrix contains the aggregated evaluation of an alternative on all the criteria.

Step 3: This step involves determination of the weight of the criteria. As different number of criteria are used for the purpose of evaluating the ideal candidate among a group of alternative, each criterion differ in importance than the other. In order to obtain a set, W, that contains the weights of all the considered criteria, all the individual decision maker assessment of the criteria need to be merged.

Let $w_j = [\mu_k, \nu_k, \pi_k]$ be an IF number assigned to the criterion X_j by the kth decision maker. The aggregated weight of a criterion is then calculated using the IFWA operator as:

$$
\begin{aligned}
&w_j = IFWA_\lambda(w_j^{(1)}, w_j^{(2)}, ...w_j^{(M)} = \\
&\lambda_1 w_j^{(1)} \bigoplus \lambda_2 w_j^{(2)} ... \lambda_M w_j^{(M)} = [1 - \prod_{k=1}^{M}(1-\mu_j^{(k)})^{\lambda_k}, \\
&\prod_{k=1}^{M}(\nu_j^{(k)})^{\lambda_k}, \prod_{k=1}^{M}(1-\mu_j^{(k)})^{\lambda_k} - \prod_{k=1}^{M}(\nu_j^{(k)})^{\lambda_k}]
\end{aligned}
\tag{3.4}
$$

The final set of weight vector can be represented as $W = [w_1, w_2, w_3, ..., w_j]$

Step 4: This step involves constructing aggregated weighted intuitionistic fuzzy decision matrix that combines the assessments of the alternatives with the weights of the criteria. This matrix can be obtained using the following equation:

$$R \otimes W = \{x, \mu_{A_i}(x).\mu_w(x), \nu_{A_i}(x) + \nu_W(x) - \nu_{A_i}(x) \times \nu_W(x) | x \in X\} \tag{3.5}$$

and

$$\pi_{A_i.W}(x) = 1 - \nu_{A_i}(x) - \nu_W(x) - \mu_{A_i}(x).\mu_W(x) + \nu_{A_i}(x) \times \nu_W(x) \tag{3.6}$$

The aggregated decision matrix can therefore be represented as follows:

$$R' = (r'_{ij})_{m\times n} = \begin{bmatrix} r'_{11} & r'_{12} & \dots & r'_{1n} \\ r'_{11} & r'_{12} & \dots & r'_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ r'_{m1} & r'_{m2} & \dots & r'_{mn} \end{bmatrix} \tag{3.7}$$

where $r'_{ij} = (\mu'_{ij}, \nu'_{ij}, \pi'_{ij})$= $(\mu_{A_iW}(x_j), \nu_{A_iW}(x_j), \pi_{A_iW}(x_j)$ is an element of the aggregated weighted intuitionistic fuzzy decision matrix.

Step 5: In this step, the intuitionistic fuzzy positive-ideal and negative ideal solutions are obtained. The criteria are divided into J_1 and J_2 representing benefit and cost criteria respectively. A^+ denotes the intuitionistic fuzzy positive-ideal solution, while A^- denoted the IF negative ideal solution. This step is similar to the one performed in the TOPSIS algorithm. The values of A^+ and A^- are obtained as:

$$A^+ = (\mu_{A^+W}(x_j), \nu_{A^+W}(x_j)) \quad \text{and} \quad A^- = (\mu_{A^-W}(x_j), \nu_{A^-W}(x_j)) \tag{3.8}$$

where

$$\mu_{A^+W}(x_j) = (max_i \mu_{A_iW}(x_j) | j \in J_1), (min_i \mu_{A_iW}(x_j) | j \in J_2) \tag{3.9}$$

$$\nu_{A^+W}(x_j) = (min_i \nu_{A_iW}(x_j)|j \in J_1), (max_i \nu_{A_iW}(x_j)|j \in J_2) \tag{3.10}$$

$$\mu_{A^-W}(x_j) = (min_i \mu_{A_iW}(x_j)|j \in J_1), (max_i \mu_{A_iW}(x_j)|j \in J_2) \tag{3.11}$$

$$\nu_{A^-W}(x_j) = (max_i \nu_{A_iW}(x_j)|j \in J_1), (min_i \nu_{A_iW}(x_j)|j \in J_2) \tag{3.12}$$

Step 6: The separation measures are evaluated next. Using the Euclidean technique developed in [80], the positive and negative separation measures of each alternative are denoted by S^+ and S^- respectively. The separation measures of each alternative from the ideal intuitionistic fuzzy positive and negative solutions are calculated. The equations used are given below:

$$S^+ = \sqrt{\frac{1}{2n}\sum_{j=1}^{n}[(\mu_{A_iW}(x_j) - \mu_{A^+W}(x_j))^2 + (\nu_{A_iW}(x_j) - \nu_{A^+W}(x_j))^2 + (\pi_{A_iW}(x_j) - \pi_{A^+W}(x_j))^2} \tag{3.13}$$

$$S^- = \sqrt{\frac{1}{2n}\sum_{j=1}^{n}[(\mu_{A_iW}(x_j) - \mu_{A^-W}(x_j))^2 + (\nu_{A_iW}(x_j) - \nu_{A^-W}(x_j))^2 + (\pi_{A_iW}(x_j) - \pi_{A^-W}(x_j))^2} \tag{3.14}$$

Step 7: The final step of the mathematical framework includes the calculation of the relative closeness coefficient of each alternative to the intuitionistic ideal solution. For an alternative A_i, with respect to the IF positive ideal solution A^+, closeness coefficient, C_i, is given by:

$$C_i = \frac{S_{i^-}}{S_{i^+} + S_{i^-}} \tag{3.15}$$

After closeness coefficient of each of the alternatives is calculated, the alternatives are ranked according to the descending order of C_i's.

3.4 Application of IF-TOPSIS in RAT Selection for Single and Group Calls

Using the outlined steps of the IF-TOPSIS framework, the network selection problem can be modelled. For the purpose of this work, the set of decision makers, M are taken to be the user applications. Three of such applications are considered namely: Voice over IP (VoIP) calls (e.g Skype calls), non-real time (NRT) video streaming applications (e.g YouTube) and Instant messaging (IM) applications (e.g WhatsApp). Subscribers are assumed to possess multi-modal devices capable of connecting to a single interface at any given time. Furthermore, subscribers are categorised into groups based on their willingness to pay (WTP). The groups of subscribers are outlined as follows:

- Gold: This class of subscribers are assumed to have the highest WTP. The priority is mainly placed upon the bandwidth intensive application such as NRT video streaming by these subscribers.
- Silver: The members of this class are assumed to have a moderate WTP with the most emphasis placed on applications that require strict latency requirement such as VoIP calls.
- Bronze: This group of subscribers are modeled to have the least WTP who place the highest priority on applications that can be termed as 'best effort' such as instant messaging and e-mails.

Two scenarios are considered for the purpose of this work. In the first instance, each of these subscriber classes is assumed to run a single application and an access network suitable to their need is selected using the IF-TOPSIS algorithm. In the second cases, each subscriber is assumed to place a group call consisting of three different types of applications, with priority on each application varying from one subscriber class to another.

3.4.1 Single Calls

Using MATLAB simulation, the performance of the proposed IF-TOPSIS algorithm has been studied for network selection in a HWN comprising of a set of A consisting of six alternative RATs namely :LTE, UMTS, WiMAX, WLAN-I, WLAN-II AND WLAN-III. These alternatives will be assessed on a set X, consisting of five parameters in form of packet delay (PD), packet jitter (PJ), throughput (T), packet loss (PL) and security provided by an access network. The exact and accurate availability related to these parameters are assumed to be unknown as typical of HWN. The characteristics of the access networks on the considered attributes in this work are presented in Table 3.1.

Following the outlined steps of IF-TOPSIS framework, the network selection problem can be modeled for both single and group calls. A crucial issue in selection of the best access network is the placement of appropriate weights to various criteria depending on

Table 3.1 Different Network Attributes

Networks	PD[ms]	PJ [ms]	T [mbps]	PL	Security
LTE	50-100	0-5	5- 50	0-2	MH
UMTS	50-100	0-3	0.5- 2	0-2	M
WiMAX	50-120	0-5	1-10	0-2	MH
WLAN-I	80- 100	0-8	0.1-5	0-4	ML
WLAN-II	80- 130	0-10	0.1-5	0-6	L
WLAN-III	100-150	5-15	0.1-3	0-7	VL

Table 3.2 QoS Requirements of Different types of Applications

Traffic Class	QoS Requirements
VoIP	1. Two-way transport
	2. Delay and Jitter are critically important
	3. PL and Throughput are relatively less important
NRT Video Streaming	1. One-way transport
	2. End-to-end delay is not important
	3. Jitter and Throughput are important
IM	1. Two-way transport that relies on request/response mechanisms
	2. Delay and PL are important
	3. Jitter and Throughput are relatively less important)

the nature of the ongoing traffic class of the user. The three types of traffic considered in this work each has unique QoS requirements. Based on specifications stated in 3GPP TS 23.203, the requirements of VoIP, NRT video streaming and IM applications can be summarised in Table 3.2.

Based on the information in table 3.2, linguistic terms can be used to set weights of importance by each applications on the considered attributes. The linguistic terms used for assignments of weights in Intuitionistic Fuzzy -TOPSIS framework are presented in table 3.3.

Table 3.3 Linguistic terms for assigning weights on criteria

Linguistic terms	Intuitionistic Fuzzy Numbers		
	μ	ν	π
Very Important (VI)	0.9	0.05	0.05
Important (I)	0.65	0.25	0.1
Medium (M)	0.5	0.4	0.1
Unimportant (U)	0.35	0.55	0.1
Very unimportant (VU)	0.15	0.8	0.05

Using the information in table 3.2 and 3.3, the following table presents the importance placed by each application on the considered criteria.

Table 3.4 Linguistic terms for rating importance of considered criteria by the applications

Applications	Criteria				
	PD	PJ	T	PL	S
VoIP	I	I	U	M	VI
Video Streaming	M	I	VI	M	U
IM	I	U	U	I	I

The Intuitionistic fuzzy-TOPSIS framework for single calls start with users assigning their priority on their preferred application. As mentioned previously, Gold class of subscribers run NRT video streaming application while the Silver and Bronze subscribers run VoIP and IM applications respectively. Using the same linguistic terms in table 3.3, the priorities assigned by different subscribers on the considered application are presented in table 3.5.

Table 3.5 Linguistic terms describing priorities for each subscriber class on different applications

Subscriber Class	Applications		
	VoIP	Video Streaming	IM
Gold	-	VI	-
Silver	VI	-	-
Bronze	-	-	VI

As these applications serve as the decision makers for the purpose of this work, the linguistic terms are converted into corresponding weights. As single calls are being considered here, the priority placed on the application by a subscriber is 1.

Each application (decision maker) also evaluates the considered alternatives on the criteria. As the applications have diverse requirements, the assessment of the individual application would vary depending on the application's requirements and each access network's performance on the considered attributes. Table 3.6 presents the linguistic terms used in the evaluation made by each application of the considered alternatives on the attributes. These linguistic terms are converted to numbers and are presented in Table 3.7.

Table 3.6 Linguistic terms for rating alternatives

Linguistic terms	Intuitionistic Fuzzy Numbers		
	μ	ν	π
Very Very High (VVH)	0.9	0.05	0.05
Very High (VH)	0.8	0.1	0.1
High (H)	0.7	0.2	0.1
Medium High(MH)	0.6	0.3	0.1
Medium (M)	0.45	0.4	0.15
Medium Low (ML)	0.4	0.5	0.1
Low (L)	0.3	0.6	0.1
Very Low (VL)	0.2	0.7	0.1
Very Very Low (VVL)	0.05	0.9	0.05

Table 3.7 Linguistic evaluation of each application against alternatives

Applications	Alternatives	Criteria				
		PD	PJ	T	PL	S
VoIP	LTE	VH	M	M	M	G
	UMTS	VH	VH	M	M	MH
	WiMAX	H	VH	M	M	H
	WLAN-I	M	MH	M	ML	ML
	WLAN-II	ML	M	M	ML	VL
	WLAN- III	VVL	MH	L	VL	VVL
NRT Video Streaming	LTE	M	H	VVH	M	MH
	UMTS	M	VH	M	M	M
	WiMAX	M	H	MG	M	M
	WLAN-I	ML	M	M	F	ML
	WLAN-II	ML	ML	ML	F	ML
	WLAN- III	VL	VVL	VVL	F	VL
Instant Messaging	LTE	VH	H	M	VH	VH
	UMTS	VH	H	M	VH	L
	WiMAX	H	H	M	VH	VH
	WLAN-I	MH	M	M	H	MH
	WLAN-II	MH	M	M	H	M
	WLAN- III	M	M	M	H	M

Each of the linguistic terms are then converted to intuitionistic fuzzy numbers using Table 3.6. The IF numbers are then aggregated using equation 3.2. Subsequently, the weight placed on each criterion by the applications are converted and aggregated into IF numbers. Using equation 3.5, the weights are then multiplied with the evaluation of the alternatives by the considered applications to form a weighted matrix, R'. The intuitionistic fuzzy positive and negative ideal solutions are then calculated, followed by the computation of the separation measure of each alternative. Finally, the closeness coefficient of each of the considered alternatives is evaluated to determine the ideal access network for the subscriber. The closeness coefficient of each alternative is considered as its IF-TOPSIS score.

3.4.2 Group Calls with Dynamic Parameter

For subscribers running group of applications at the same time, different priorities are placed on each of them. A Gold subscriber places the highest priority on video streaming application followed by VoIP and IM. A Silver subscriber prioritizes all applications equally while a Bronze subscribes places most emphasis on IM application. The linguistic priorities put on each of these applications by different classes of subscribers is presented in table 3.8.

Table 3.8 Linguistic terms describing priorities for each subscriber class on different applications

Subscriber Class	Applications		
	VoIP	Video Streaming	IM
Gold	I	VI	I
Silver	M	M	M
Bronze	M	U	VI

These priorities can be converted into IF numbers s using the information in table 3.3 and are presented in table 3.9. This produces the importance of the k^{th} application to the sth subscriber class.

Table 3.9 Priority Placed on individual application by subscriber class

Subscriber Class	Applications	Priority
Gold	VoIP	0.3952
	Video Streaming	0.5184
	IM	0.0864
Silver	VoIP	0.3333
	Video Streaming	0.3333
	IM	0.3333
Bronze	VoIP	0.2987
	Video Streaming	0.2056
	IM	0.5008

The subsequent evaluation of each access network on the considered criteria by the applications remain the same as one presented in Table 3.7. These evaluations are converted to IF numbers using Table 3.6 and are further aggregated to form a group decision matrix using equation 3.2. The linguistic terms of weights presented in Table 3.4 are also converted into IF numbers and are multiplied with the decision matrix obtained through the evaluation of the available alternatives. The remaining steps remain identical to the ones used for single calls, i.e: determination of positive and negative ideal solutions, evaluation of separation measure and closeness coefficient of each alternative.

For the purpose of group call in this work, an additional parameter in form of cost of service is introduced. With time dependent pricing gaining popularity, the price of data tends to vary during different times of the day. As a result, a subscriber's willingness to pay, while running multiple applications, need to be taken into account as it is directly related to the cost of service. The WTP score of the i^{th} access network can be modelled using the following equation:

$$WTP_i = \begin{cases} 1 - \frac{np_i}{wp_s}, & \text{if } np_i \leq wp \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

where wp_s is the willingness to pay of the s^{th} class of subscribers and np_i is the cost of accessing service from the i^{th} access network. It can be seen from equation 3.16 that whenever the cost of service of an access network exceeds WTP of a subscriber, that alternative network is assigned a score of 0 and is excluded from consideration. Using the WTP score of each network, the ultimate assessment of the i^{th} alternative can be obtained using:

$$F_i = C_i \times WTP_i \tag{3.17}$$

where F_i is the final score of the i^{th} alternative. The candidate network with the highest F score is deemed to be the ideal alternative.

3.5 Results and Discussions

In this section, the results of the IF-TOPSIS algorithm's application to the access network selection for single call and group calls of three user applications (i.e VoIP, NRT video streaming and IM) in HWN with the presence of a dynamic criteria in form of network cost are presented.

3.5.1 Single calls

The performance of the IF-TOPSIS algorithm for subscribers initiating a single application is presented in this subsection. As other MCDM algorithms fall short in operating in a completely fuzzy environment, they are unsuitable to be compared with the IF-TOPSIS framework.

IF-TOPSIS is used to rank and select the optimal network for three different subscribers each running one application. Figure 3.1, 3.2 and 3.3 present the ranking of the alternative access networks for each application user without the effect of network price. In Figure 3.1, the implemented IF-TOPSIS algorithm selected LTE as the preferred network for subscribers running VoIP calls. This result is intuitively correct given the significant advantage LTE boasts over other considered access networks in terms of parameters such as packet delay, packet jitter, throughput, loss and security offered. WiMAX possess the second best linguistic evaluation on criteria important to VoIP traffic and as such it comes second. The poorest performing alternative is seen to be WLAN-III as it has the worst rating of the considered criteria. The IF-TOPSIS scores of LTE, UMTS, WIMAX, WLAN-I, WLAN-II and WLAN-III are 0.5441, 0.4415, 0.5018, 0.4393, 0.4272 and 0.4157 respectively.

Figure 3.2 illustrates the ranking of the access networks for a subscriber running video application. Once again, LTE, due to its superior features, takes the topmost position followed by WiMAX. However, interesting trend is noticed for the order of the remaining alternatives. As throughput is the most important criteria for a video application, UMTS's poor offered data rate meant that it achieved a lower rank compared to WLAN-I and WLAN-II. WLAN-III was rated as the least favourable access network due to its

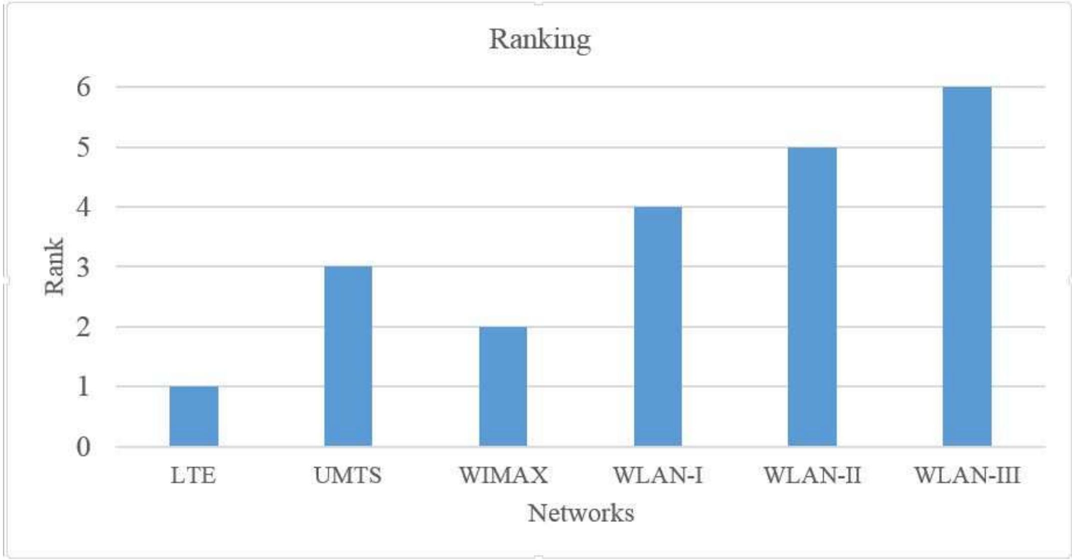

Fig. 3.1 Network selection for VoIP traffic

poor rating of packet jitter, throughput and packet loss- three parameters crucial for NRT video streaming application's performance. The IF-TOPSIS scores of the six alternatives are 0.5515, 0.4331, 0.4730, 0.4564, 0.4504 and 0.4147 for LTE, UMTS, WiMAX, WLAN-I, WLAN-II and WLAN-III respectively. This result demonstrates the effectiveness of the IF-TOPSIS framework to select the access network based on performance, an important feature of efficient RAT selection.

Figure 3.3 presents the order of the alternative access networks for a subscriber running IM applications. As IM applications are considered belonging to the best effort traffic class, where priority is placed on criteria such as packet delay, packet loss and security, LTE still emerges as the best available alternative. The remaining access networks are ranked with WiMAX, UMTS, WLAN-I, WLAN-II and WLAN-III being 2nd, 3rd, 4th and 5th. The IF-TOPSIS scores of the 3rd, 4th, 5th and 6th networks are seen to be quite close due to their similar evaluation by the application. The IF-TOPSIS scores of LTE, UMTS, WiMAX, WLAN-I, WLAN-II and WLAN-III are 0.5475, 0.5047, 0.5256, 0.4902, 0.4826 and 0.4798 respectively.

As the intuitively best access network, possessing the best features, has been selected for different types of traffic by the IF-TOPSIS algorithm, its validity in the access network selection for single call has been demonstrated. Furthermore, the change in the order of the networks depending on the attribute characteristics and application requirements further highlighted the IF-TOPSIS algorithm's ability to take under consideration various parameters while selecting the ideal solution. In the next sub-section, the

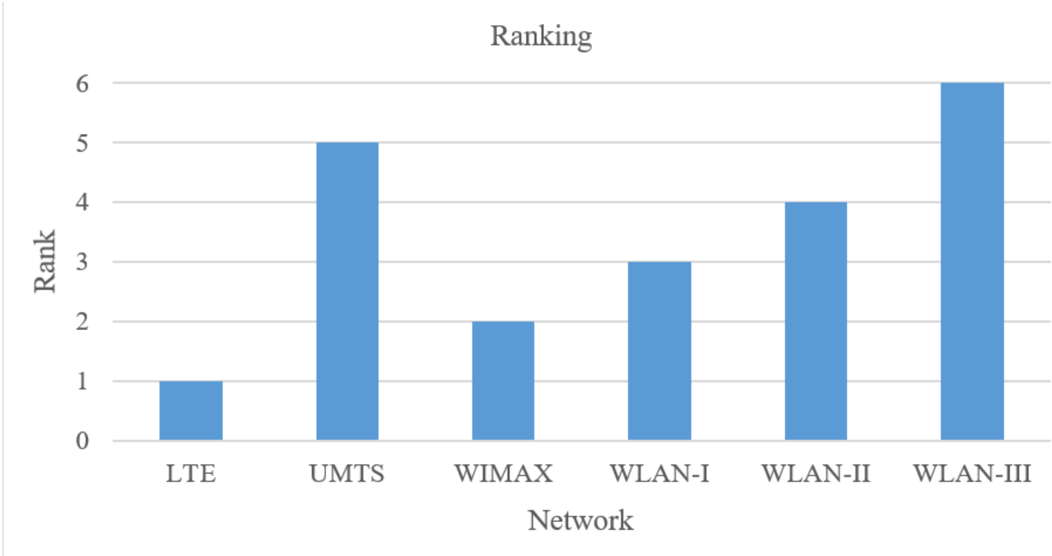

Fig. 3.2 Network selection for Video traffic

performance results of IF-TOPSIS algorithm in modelling a group decision problem for different subscribers initiating multiple calls are analysed.

3.5.2 Group calls in with fixed and varying cost of service

In this section, the performance results of the proposed IF-TOPSIS algorithm for different classes of subscribers running multiple applications are presented. Dynamic parameter in form of cost of accessing services of an access network is incorporated in the network selection problem. The cost of service is made to have direct relation on a user's ability to select a particular RAT. Two scenarios are considered for the purpose of this work, with the first scenario ensuring no effect of the cost on the access network selection and the second scenario involving its effect on the selection. Each scenario is simulated 200 times under varying priority level per subscriber class.

Scenario 1: No effect of cost

Here, the price of service of all available access networks are assumed to be within the willingness to pay for all classes of subscribers for 200 calls. The priorities of each class of subscribers during the simulation steps are varied. As stated before, Gold subscribers placed highest priority on video streaming application and predominantly chose LTE

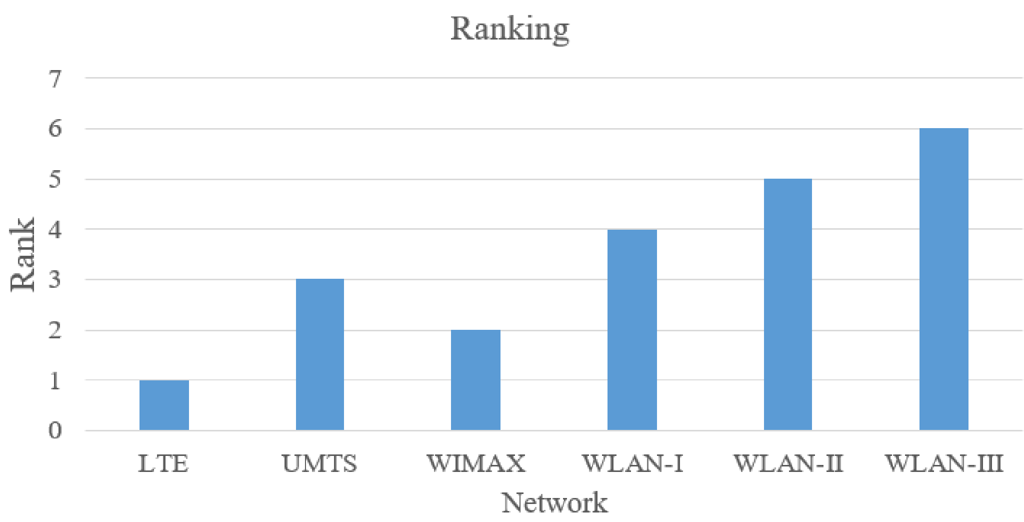

Fig. 3.3 Network selection for IM traffic

network for its superior throughput and jitter ratings. As priority value on the video streaming application increased, more calls were admitted to the LTE network since it offered significantly higher throughput as required by the application.

Similar trend is noticed for Silver and Bronze subscribers as most of their calls were admitted to the LTE network. As LTE boasts the superior features for the considered parameters, it shows that the proposed scheme selects LTE more times as priority on any application increases. Figure 3.4 shows the number of times each network was selected under this scenario.

Scenario 2: Effect of varying cost

The effect of varying network prices on the selection procedure are investigated in this sub-section. With the popular use of time dependent pricing by the network operators, price per resource often varies during different times of the day. However, it is ensured that if the difference of IF-TOPSIS score between the best performer and the worst exceed a certain threshold value (0.1 in this simulations), the worst network is not considered during the selection process. This is to ensure that a worst performing network element is not selected solely on its low resource price. Table 3.10 presents the set price of accessing services (cents/MB) from each of the considered network elements used in the simulation.

The WTP of the Gold subscribers were varied between 40c/MB and 100c/MB. The WTP of the Bronze subscribers were set between 5c/MB and 20c/MB. While the WTF of silver class was maintained between 5c/MB and 55c/MB. The LTE prices are set as

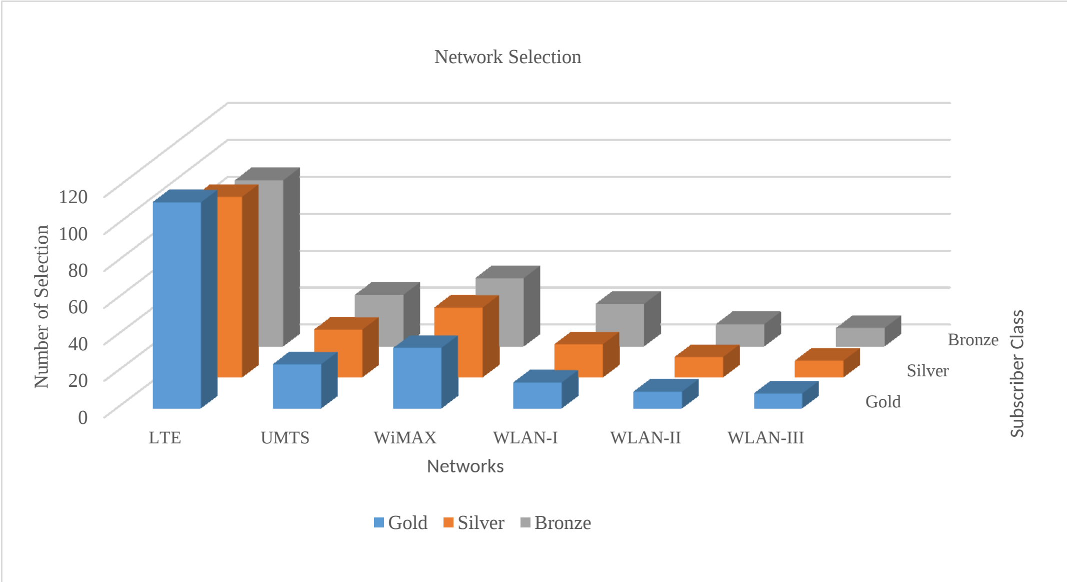

Fig. 3.4 Network selection for all classes with constant cost

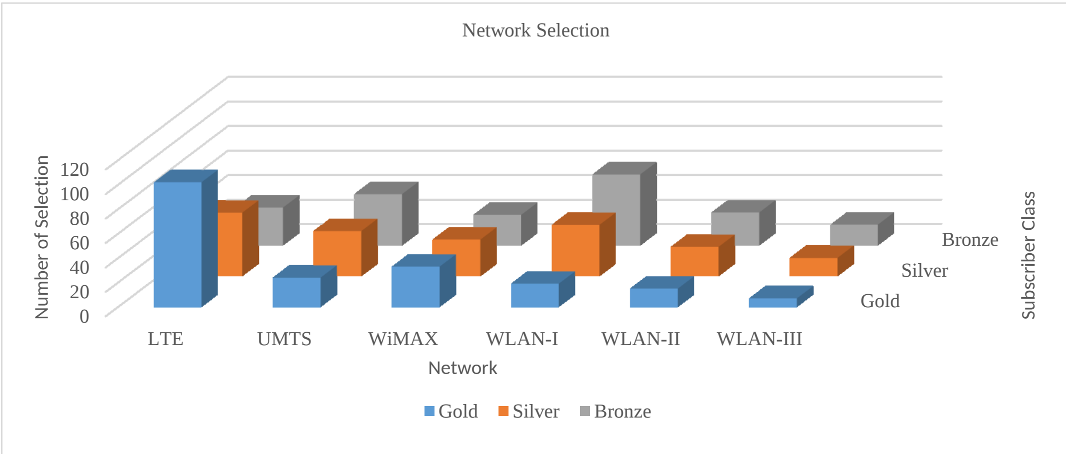

Fig. 3.5 Network selection for all classes with varying cost

Table 3.10 Network Prices

Networks	Price [c/MB]
LTE	45
UMTS	25
WiMAX	30
WLAN-I	15
WLAN-II	10
WLAN-III	5

the highest which fell below the WTP of most Gold subscribers .As such, most calls from this subscriber group maintained their attachment to LTE. The network cost of UMTS, WLAN-I, WLAN-II and WLAN-III were kept within the WTP of the Bronze subscribers. It was seen that most of the calls from Bronze subscribers were placed on WLAN-I followed by UMTS. This is primarily because of these network elements' good delay and jitter limits as well as their network prices being within the WTP of the subscribers. Silver subscribers maintained fairly even selection of all networks due to their even priority assignment on each application. This scheme demonstrates its ability to select the best network under dynamic network condition. Figure 3.5 shows the number of times each network was selected under this scenario.

Investigation on occurrence of rank reversal

The effect of removal of the worst network in the network selection process under scenario 1 are investigated here. This is performed by taking the average IF-TOPSIS score of each network after 200 simulations before and after the elimination of the WLAN-III . The Silver class subscribers are considered here since they place equal priority on each application.

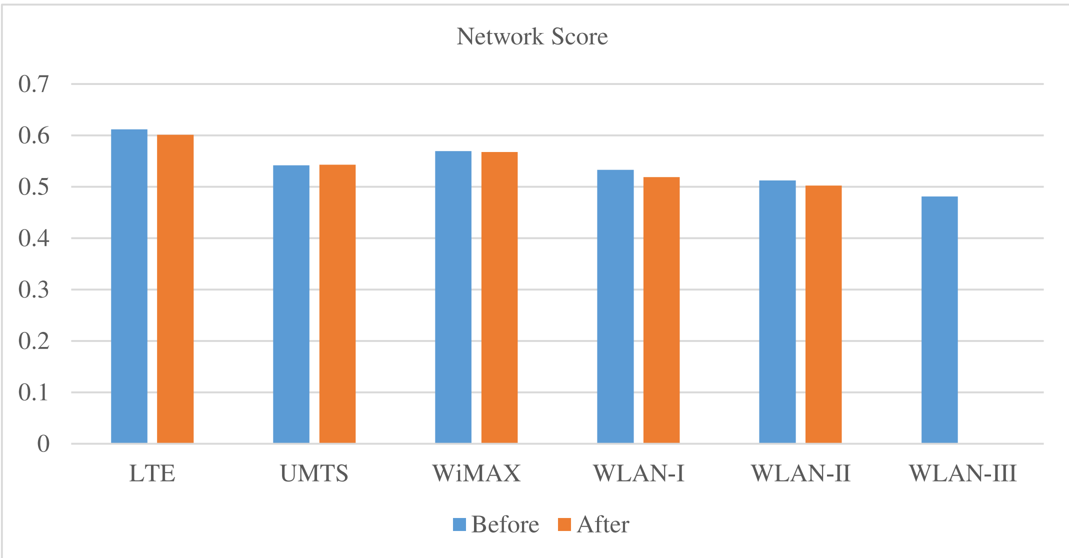

Fig. 3.6 IF-TOPSIS score before and after removal of WLAN-III

As associated with traditional TOPSIS as well as other commonly used MCDM techniques such as VIKOR, such elimination of the worst performing alternative leads to an uncanny rank reversal which leads to abnormal ranking of available networks. As can be seen from the network score, the proposed scheme is robust enough to eliminate rank reversal and maintain almost the same score for other alternatives as WLAN-III is removed. As such, this scheme is fit to be used in the highly dynamic heterogeneous wireless environment where a network element might fail suddenly. Figure 3.6 illustrates the IF-TOPSIS network scores of each RAT before and after the worst performing network (WLAN-III) is eliminated.

3.6 Chapter Summary

In this chapter, the problem of access network selection in a fuzzy environment mirroring the heterogeneous wireless networks was investigated. While many work has thus far been done in the area of access network selection, none has been able to accurately model the impact of ambiguity on the selection problem. A thorough analysis of the proposed scheme has been conducted using multiple wireless interfaces and three different types of traffic namely VoIP, NRT video streaming and IM.

The algorithm has been tested for subscribers initiating single and multiple calls at the same time. For single calls, LTE network was chosen by all classes of applications. As the LTE network is seen to possess the best attribute features, its selection confirms the algorithm's ability to accurately pick the intuitively best alternative for different types of traffic.

In addition to single calls, the proposed algorithm can also select the optimal network for a subscriber running multiple applications under varying network condition (price per resource) . Furthermore, the proposed algorithm is robust as demonstrated by the absence of ranking abnormality- a shortcoming of other MCDM algorithms.

Chapter 4

Data Demand Prediction for Data Centers in 5G Networks

4.1 Introduction

The next generation of the wireless network will be the 5G. Its deployment has been necessitated due to the immense surge in data demand that the wireless communication sector has witnessed in recent years. Demand for data is forecast to reach close to 50 Exabytes per month in the year 2021 [1]. This demand would be a 7-fold growth from the data that was consumed per month in 2016. The expenditure (capital and operational) associated with present wireless service providing infrastructure would simply be astronomical if this forecast demand is to be met. Along with traditional users of the mobile broadband internet, 5G will involve consumers from other vertical industries such as automation, health, energy and industries. To meet demands of this diverse clientele using a monolithic approach would be infeasible and as such 5G is proposed to have at least three logically separated portions of the network, with each portion called a 'network slice'.

Network slicing would enable operators to separate a physical network into multiple virtual networks tailored to meet requirements of different user groups [3]. As the current mobile core network mainly consists of hardware dependent network functions, flexible and scalable way of providing service becomes an issue. Two main concepts in form of NFV and SDN have come to the forefront to aid the transition of the core architecture to a more softwarized domain. NFV allows network functions such as service gateway to be deployed as Virtual Network Functions (VNFs) implemented in form of software on commercial off the shelf (COTS) hardware (such as servers in data centers) [4]. This allows service providers to be less dependent on hardware and aids seamless scaling

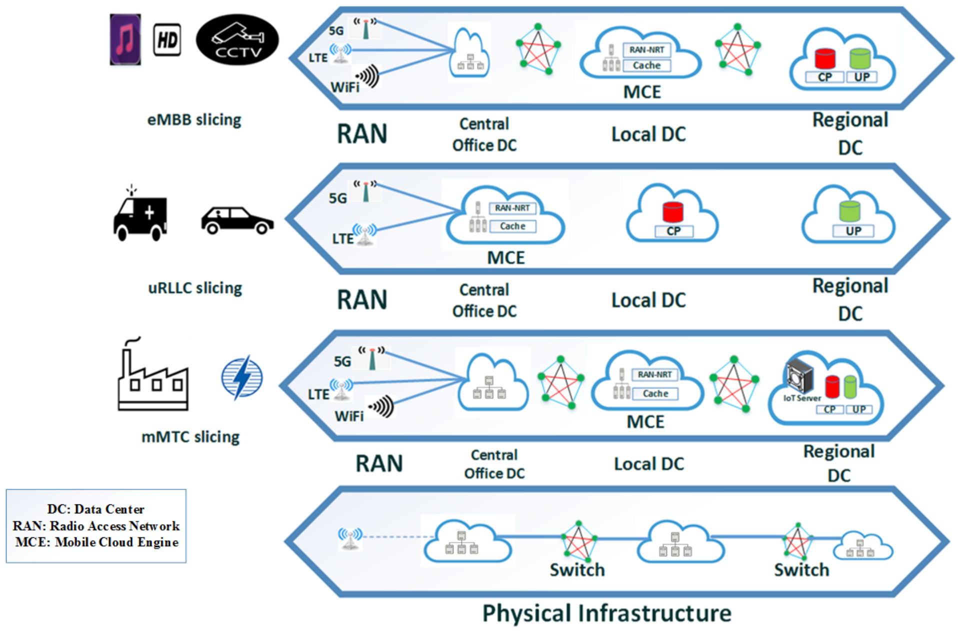

Fig. 4.1 Data Centers in 5G Networks

in case changes are needed to be made. This further reduces the operator's expenses associated with establishment of the core architecture. SDN [5], on the other hand splits the control and data plane of the network functionalities thereby producing a programmable environment that greatly simplifies network management. SDN allows service providers to have more control over the various functionalities of a software based network.

Data centers hosting various VNFs are likely going to be placed at the edge of the current networks to supplement those deployed in traditional cloud data centers [80]. Figure 4.1 shows the higher level physical infrastructure of 5G networks [6]. Figure 4.1 illustrates the vital role regional data centers will play in processing the traffic in different zones. Adequate planning, therefore, needs to take place to determine the optimal positioning and dimensioning of these data centers. Existing research works either focus on the problem of determining resource requirements and placement of individual VNFs [81–83], or architectural design of data centers [84–86]. The dimensioning and placement problem of data centers were addressed in [87] with the aid of optimization models. However, a key element in form of the cellular traffic that originates from users is not considered in the design of data centers in the literature. As 5G will involve different types of subscribers, each data center needs to be designed based on the volume of traffic it will be required to deal with to guarantee optimal performance. In addition, to ensure service requests are met within the delay constraints, the data centers need to be also optimally located from the base stations. Furthermore, to guarantee dynamic utilization of data center resources, accurate prediction models need to be employed to forecast future traffic demands.

In this work, the open telecommunications data set of Telecom Italia is utilized to first study the data traffic profile of different regions of the city of Milan. The big data set is processed to seperate the city into twenty regions and establishment of a data center in each region is proposed to provide computational power to various NFV and SDN functionalities of one or multiple network operators. The optimal positions of each data center within the regions are determined so as to be best placed from all the base stations that will be delivering services to the end users. This ensures that the data centers remain capable to adequately handle the traffic with minimal delay from any part within the coverage area. Furthermore, the hourly traffic pattern within each region under consideration is analysed to heuristically determine the dimension of the data centers in terms of the number of CPU cores needed to handle the traffic within their coverage. This analytic approach towards determining the size of data centers would enable the infrastructure providers to understand the nature of the data center that would be needed to establish in a certain area. Additionally, based on the traffic pattern analysis and previous information, machine learning algorithms are employed to forecast the next day's traffic within each region under consideration. With the aid of forecast values, data center's resource manager (usually a SDN controller) would be able to determine the amount of computational power that will be needed to handle the future traffic from different wireless service providers. This would greatly reduce the operational cost for the infrastructure providers and ensure optimal utilization of resources.

The main contributions of this work can therefore be summarised as follows:

- Conduct a detailed analysis on the cellular traffic dataset of Telecom Italia (TIM).
- Determine the optimal placement of the regional data center according to the minimal aggregated distance from all the base stations within the zone.
- Determine the redundant capacity of a data center based on a heuristic approach.
- Apply recurrent neural network (RNN) models to the TIM data set to predict future data demand that occurs in different regions within a city.
- Validate and compare the performances of different RNN models in predicting the next-day's data demand.

The rest of the chapter is organized as follows. Section 4.2 provides an overview of the background and work related to this field. In Section 4.3, the data set that has been used in this work is analysed. Section 4.4 presents a detailed analysis of traffic profile of three chosen zones in the data set. This is followed by Section 4.5 that presents the design of the data centers. In section 4.6, machine learning algorithms are presented to predict the future traffic profiles. Finally, conclusions are presented in section 4.7.

4.2 Related Work

The literature related to this work can be classified into two areas. The first area focuses on the traffic analysis and prediction models that have been proposed in the literature. The second area is related to the design of various components of the cloud infrastructure in cellular and wired network.

4.2.1 Traffic Analysis and Prediction in Cellular Network

A lot of work has thus far been conducted to determine the traffic profile that a single or group of base stations experience over a period of time. Given the proliferation of bandwidth intensive applications, cellular data trace reveals more related to network behaviours than traditional voice traffic. Furno *et al.* [88] developed a cognitive framework to analyse traffic profiles using the TIM and Orange open data sets. The idea behind this work was to identify anomalies that usually occur while considering a network-wide usages. The authors however did not utilize various traffic profiles to design data centers. Wang *et al.*[89] studied the traffic experienced by multiple geographically seperated base stations to demonstrate the strong spatial-temporal relationship that exists in the domain of the cellular traffic. Sinusoid superposition and lognormal distribution methods were employed to describe the temporal and spatial traffic variations respectively. They however, did not use any prediction model to determine future traffic demand. The learning and prediction aspects were considered in the model proposed by Li *et al.* in [90] used a big data set to study the parametric differences that exists between different types of cellular network applications. They further proposed a predictive algorithm to make forecasts related to application-level traffic. The same authors in [91] used entropy theory to demonstrate the inherent pattern that exists in cellular traffic.They also concluded that the data traffic prediction is solely reliant on temporal and spatial relevancies. These works however fell short in using traffic information to design key network infrastructure such as data centers.

To predict traffic patterns in cellular networks, some algorithms have been commonly employed in the literature. Linear models such as autoregressive integrated moving average (ARIMA) and its modified version Fractionally- ARIMA (FARIMA) have been used in [92, 93]. Kalman filtering has been used in [94, 95] with mobility and network traffic model utilized in [96, 97]. These shallow learning linear architectures however, have proven incapable of accurately modelling deep and complex non-linear relationships that are usually present in cellular traffic traces. Recently, a deep of learning-based predictive algorithms have emerged and proven effective in the prediction of traffic pattern. Oliveira *et al.* in [98] compared the performances of recurrent neural network (RNN) with stacked autoencoder to forecast internet usage. Their analysis showed that RNN is superior to autoencoder in making accurate prediction in this use case. The work in [99] used two state of the art neural network models: RNN and Convolutional Neural Network (CNN), alone and in conjunction with each other to forecast maximum, average and minimum traffic volume of different regions in the TIM dataset. Their evaluation presented prediction accuracies between 70 to 90 percent. The work in [100] compared the performances of neural network models and linear models. Their results showed that RNN model is more suitable in modelling complex data when compared to linear models.

The existing literature in the domain of traffic prediction and analysis either used models that have proven to be less effective with complex data or have not performed hourly time series analysis with state of the art RNN models. RNN in its architecture contains a recurring loop which allows it to combine present information with the past. This makes RNN and its models suitable for analysis of time series sequence. As cellular traffic is highly dependent upon time, RNN's ability to recognize patterns in such data is highly desirable. In addition, the more recent models of RNN have significantly enhanced the performance of the simple RNN model. However, the effect of activation functions on the prediction accuracy of the RNN models has not gained much attention in the literature. In this work, the performances of different RNN models with activation functions to obtain future traffic demands are analysed.

4.2.2 Design and operation of Cloud Infrastructure

Cristina *et al.* in [101] designed a NFV oriented architecture for edge data centers. They implemented a server centric data center architecture that produced better results when compared to traditional network-centric architectures. The authors in [86] proposed optimization models to reduce the total energy consumption in both data centers and data center networks. Gebert *et al.* proposed potential solutions that accommodates a sudden increase in traffic demand with the aid of dynamic and optimal placements of various VNFs [102]. The work in [103] proposed modification on the existing CU algorithms to determine characteristics of different traffic that a data center network experiences. They validated the proposed algorithm by using a real data set and achieved improvement on error performance, space cost and time complexity. The problem of optimal placement of several VNFs was explored further in [104]. In this work, they proposed mathematical frameworks that determined the ideal location for several VNFs within the boundaries of network capacity and latency limits. Shi *et al.* in [105] combined a decision making approach with Bayesian learning to dynamically allocate computing cloud resource in data centers to various NFV components. In [87], the authors proposed an optimization model that addressed the issue of optimal placement of both SDN and NFV components along with ideal size of data centers. Their work considered network cost with load to determine the best location for various data centers in the cases of Germany and USA.

After reviewing the existing research works so far conducted in this area, one can observe that they do not focus on the determination of the dimensions of the data centers using real world cellular traffic. As suggested in the literature, adaptive utilization and placement of cloud computing resources and functionalities are essential to provide optimal services for the end users. As traffic pattern changes with space and time, it is therefore crucial to determine the traffic profile that exists in various regions within a geographical area. With proper analysis and forecast of traffic patterns in different locations, data centers can be optimally placed and their resources can be properly utilized. A through review of the existing literature shows that none has studied the TIM data set in this context, which is the main focus of this work.

4.3 Dataset Analysis

In this section, the dataset utilized for the purpose of this work is presented. The data set is processed and analysed to facilitate the design.

The data used in this work was released by Telecom Italia in 2015 and has been made available for public use [106]. It contains call detail records (CDRs) of different areas over the period of November 1, 2013 to January 1, 2014 within the city of Milan and the Province of Trento. As illustrated in Figure 1, the data set breaks down the considered area into 10,000 square cells, with each cell representing an area of 235m by 235m. The CDRs contain information related to each cell's different telecommunication activities such as number of calls, sms and internet activity within a ten minute period. As internet activities certainly demand more resources in today's wireless communication, for the purpose of this work, the amount of internet activities that occur within a cell in a given time frame is considered. The cells of the Milan grid are presented in Figure 4.2.

9901	9902	9903	...	...	...	...	...	...	10000
9801	9802	9803	...	...	...	...	...	...	9900
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
101	102	103	...	...	...	...	...	...	200
1	2	3	...	...	...	...	...	...	100

Fig. 4.2 The Milan Grid

Even though the dataset is normalized and anonymized, it is still possible to extract certain features from it. In order to analyse the internet traffic that different areas in the grid experience, the geographically collocated cells within an area of 27 km^2 are clustered to form a single zone. This is due to the fact that data centers in 5G will serve to host VNFs of different base stations and network operators. As such, from the infrastructure provider's point of view, it would be important to have an understanding of the traffic profiles that different regions exhibit before establishing the regional data centers. For example, cells numbered from 1 to 499 constitute zone 1 in this work and demonstrate a significantly different traffic profile when compared to cells 1500 to 1999 that form zone 4.

Given that the original data set consists of records of internet activity that occurs over 10 minute period, further aggregation of these values are performed to have an hourly breakdown of internet traffic that each zone experiences over a day. As the data set also contains 62 days worth of records, the holidays (22 days) are separated from the working days (40 days). Another important attribute of this data set is in its recording of the CDRs. Each entry of the internet activity in the dataset represents the number of times a connection is initiated or terminated. A new CDR is also registered every time a previous connection exceeds 5 Megabytes (MB) of traffic. Since it is neither possible to identify exact number of new and terminated internet connections nor the exact volume of traffic exchanged within a connection, it is therefore assumed that half of the records are new connections with each connection having a volume of 5 MB of internet traffic. After performing these tasks, the traffic loads that different zones experience every hour within a day are obtained. For the sake of brevity, the results of three zones are presented in this chapter. Interested readers are suggested to go through the detailed analysis presented in [107]. The rest of the data analysis included mapping the square cell ids into coordinates to obtain their geographical locations. Furthermore, location of the base stations present within the entire area of the dataset was also provided. These pieces of information were made available by the authors who conducted the work in [108].

4.4 Traffic Profile Analysis

In this section, three zones are identified to investigate their traffic profiles. The three chosen zones are Zone 1, Zone 10 and Zone 14. The reason behind doing so is that each of these zones have different volumes of traffic that can be classified as low level, medium level and high level. Among all the considered zones, Zone 1 experiences the least amount of peak and average traffic. Zone 14, on the other hand, experiences what can be termed as the medium amount of peak as well as average traffic when compared to others in the data set. The maximum peak traffic observed in Zone 10 is among the highest in all the zones and can be classified as a high level traffic zone. This work proposes the establishment of a data center in each of these zones to meet the various level of traffic experienced in them. Figure 4.3 presents the architecture of the proposed model. In Figure 4.4 and Figure 4.5, the geographical zones selected for the purpose of this work and the base station distributions in each of these zones are demonstrated respectively.

Based on the aggregated hourly data obtained in the previous section, the maximum and average hourly traffic volumes of these zones during holidays and workdays are determined. In addition, the sum of the mean and one standard deviation is also presented to illustrate the amount of variation in each hour of traffic. Despite the differences in the traffic profiles of these zones, there exists some common phenomena among them. The detailed analysis of the traffic profiles of these zones are presented below.

Common Characteristics of the Zones: Figure 4.6 (a), (c) and (e) represent the holiday traffic profiles of Zone 1, 14 and 10 respectively. Conversely, (b), (d) and (f) of the same figure present the workday traffic profiles of these zones. Table 4.1, 4.2 and 4.3 further provide additional information that aid the analysis of traffic profiles of these zones. For both holiday and workday traffic, the presence of 'tidal effect' is evident. Traffic is gradually seen to decline during late night hours (starting from 9 pm) and the minimum is reached around 3 am. For holidays however, during midnight hours (12 am to 3 am), more traffic volumes are experienced compared to workdays. This can be attributed to people's tendency to stay awake late at night during holidays as opposed to sleeping early in working days. Furthermore, traffic consumption is seen to gradually increase from early morning hours (5 am to 8 am) in both types of days. However, the rate of increase can be seen to be higher in the case of working days. This

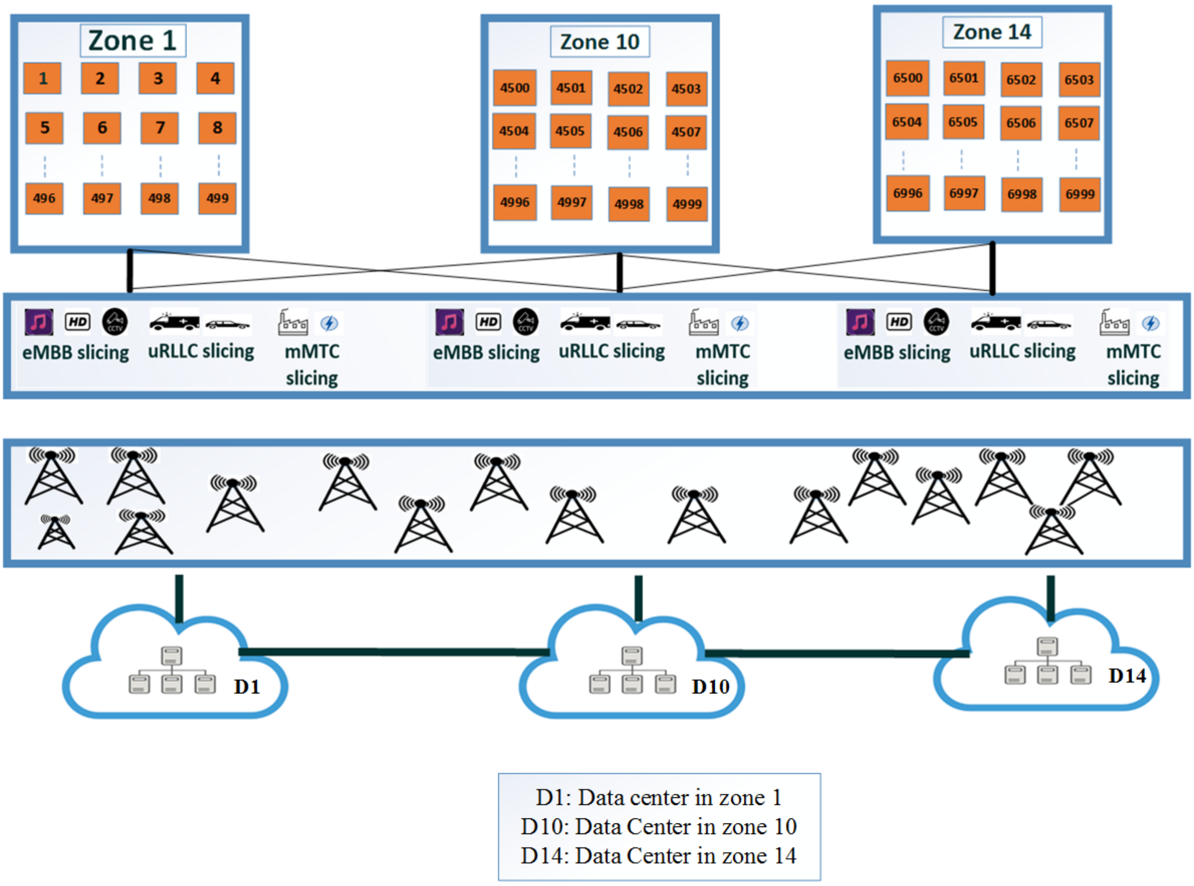

Fig. 4.3 Proposed Model

could be as a result of people waking up early to get to work on work days which is in contrast to people sleeping longer during mornings of holidays. In addition, workdays commonly experience higher volumes of traffic than holidays. The physical quantity of the standard deviation observed in these traffic volumes is also seen to increase as the traffic volume increases. However, it is observed that the ratio between the standard deviation and mean is not constant and it is larger in periods of low traffic. Another important quantity of interest is the probability with which the traffic volumes of past days tend to fall within the sum of the mean and deviations. It is observed that, in the cases of both holidays and workdays, the hourly traffic of past days mostly fall within the sum of the mean and one standard deviation with a probability of 75 percent and more. However, this probability increases to the range of 90 to 100 percent when the sum of the mean with two standard deviations is considered. This probability analysis enables us to understand the hourly variation in traffic volume which would further aid in the subsequent design of the data centers. The inherent traffic characteristics of each of the three zones are explained below.

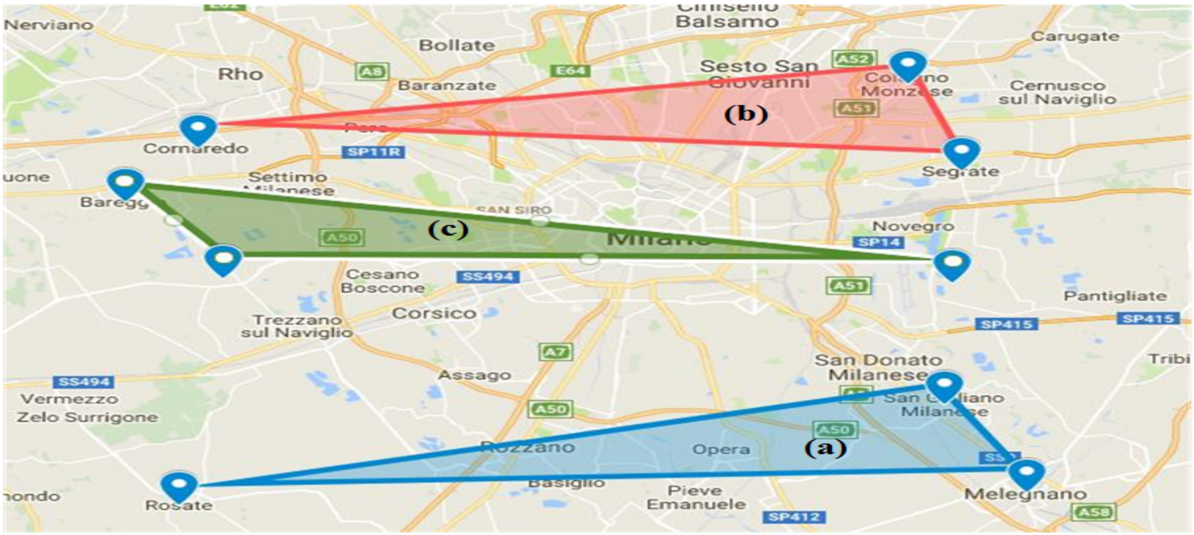

Fig. 4.4 Zone 1, 14 and 10 in Milan.

Zone 1: This zone experiences the lowest amount of traffic among all others and comprises of cells 1 to 499 in the original data set. Figure 4.4 (a) shows the geographical mapping of these cells. This zone consists of sparsely populated areas outskirts of the central part of the Milan city . Figure 4.5 maps the geographical positions of the base stations within this zone. There are a total of 17 TIM base stations (BSs) within this area. This relatively low number of BSs highlight the low level of traffic this zone experiences for both holidays and workdays as seen from Figure 4.6 (a) and 4.6 (b). Some key features associated with the traffic volumes in this zone are presented in Table I.

Zone 14: This zone experiences medium amount of traffic and comprises of cells 6500 to 6999 in the original data set. Figure 4.4 (b) shows the geographical mapping of these cells. It covers areas that are just outside the central of the city and consists of populated areas. There are 127 BSs in this zone which is significantly more when compared to Zone 1. This is expected given the larger volume of traffic that is experienced in these areas over time. Figure 4.5 presents the geographical locations of the BSs within zone 14. Table III presents some of the attributes noticed for the traffic in zone 14.

Zone 10: This zone experiences one of the highest volumes of traffic in comparison to others. This zone, as seen from figure 4.4 (c), is in the heart of Milan (cells 4500 to 4999) and experiences heavy traffic volumes throughout the day. To meet up with the traffic demands in these areas, there are 171 TIM base stations located in this area. Figure 4.5 also illustrate the locations of these BSs within zone 10 with Table II presenting some key features of this zone's traffic volumes.

Table 4.1 Zone 1 Traffic Profile

Hour	Holiday						Workday					
	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)
0	24544.17	63903.00	11175.28	35719.45	0.95	0.95	10750.78	21149.00	4962.67	15713.45	0.83	0.95
1	10083.26	32141.00	6911.84	16995.10	0.91	0.95	2718.80	10416.00	1963.81	4682.61	0.90	0.95
2	4412.61	10587.00	2968.77	7381.38	0.86	0.91	1186.11	5253.60	1065.70	2251.81	0.93	0.95
3	1565.15	4695.90	1122.35	2687.51	0.91	0.95	1043.66	4900.60	1077.14	2120.79	0.90	0.93
4	1977.62	6766.60	1439.68	3417.30	0.86	0.95	3868.69	8773.40	1880.27	5748.96	0.83	0.95
5	8670.61	20939.00	5653.95	14324.57	0.77	0.95	33163.59	50757.00	9220.51	42384.10	0.95	1.00
6	32944.12	62204.00	15615.34	48559.46	0.77	1.00	88662.81	104510.00	18288.70	106951.52	1.00	1.00
7	67550.60	91372.00	15448.05	82998.65	0.82	1.00	105756.15	126860.00	16010.90	121767.05	0.98	1.00
8	90923.48	109000.00	13600.95	104524.43	0.95	1.00	102776.98	120300.00	11238.11	114015.09	0.93	1.00
9	101792.75	124000.00	13068.78	114861.52	0.95	1.00	98767.57	111380.00	7981.26	106748.84	0.95	1.00
10	106492.81	126000.00	12956.57	119449.38	0.95	1.00	96463.24	105970.00	6735.01	103198.25	0.95	1.00
11	106867.63	131000.00	13284.02	120151.64	0.95	1.00	98191.21	107060.00	7514.03	105705.24	0.98	1.00
12	102338.26	109000.00	11236.42	113574.68	1.00	1.00	101494.13	114000.00	8487.77	109981.90	0.90	1.00
13	105473.01	121000.00	12556.62	118029.63	0.95	1.00	104512.33	118360.00	8733.90	113246.23	0.93	1.00
14	106319.82	118000.00	11948.20	118268.02	1.00	1.00	105652.76	117180.00	8607.07	114259.83	0.93	1.00
15	108625.08	122000.00	13362.39	121987.48	0.95	1.00	109990.14	123410.00	10284.28	120274.42	0.98	1.00
16	112540.64	132000.00	14447.60	126988.24	0.95	1.00	118021.28	128740.00	11994.19	130015.47	1.00	1.00
17	115954.26	132000.00	15054.93	131009.19	0.95	1.00	123291.81	136620.00	13313.68	136605.49	0.98	1.00
18	113807.83	125000.00	13505.42	127313.25	1.00	1.00	121636.20	136760.00	13236.05	134872.25	0.95	1.00
19	108910.49	121000.00	12900.16	121810.65	1.00	1.00	115708.50	127520.00	11669.87	127378.38	0.98	1.00
20	105082.34	114000.00	11883.19	116965.53	1.00	1.00	110256.61	118400.00	10172.70	120429.31	1.00	1.00
21	93726.61	100000.00	9151.17	102877.78	1.00	1.00	101049.12	109020.00	9049.93	110099.05	1.00	1.00
22	77204.35	96757.00	9874.22	87078.57	0.95	1.00	79700.18	90174.00	7213.97	86914.15	0.90	1.00
23	54307.38	83133.00	12859.25	67166.63	0.91	0.95	40405.45	51530.00	5411.29	45816.74	0.93	0.98

Table 4.2 Zone 10 Traffic Profile

Hour	Holiday						Workday					
	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)
0	540909.09	800000.00	180138.06	721047.15	0.82	1.00	329675.00	477000.00	82191.05	411866.05	0.85	1.00
1	357897.23	559000.00	131816.59	489713.82	0.86	1.00	130731.18	245000.00	46100.37	176831.55	0.85	0.95
2	202176.19	406000.00	94289.78	296465.97	0.86	0.95	49220.28	115000.00	23149.44	72369.72	0.88	0.95
3	115826.83	315000.00	68606.31	184433.14	0.91	0.95	37992.08	71009.00	14788.22	52780.30	0.88	0.95
4	95131.67	275000.00	54559.96	149691.63	0.95	0.95	79531.73	121000.00	20374.99	99906.72	0.88	0.98
5	147603.30	297000.00	70395.47	217998.77	0.86	0.95	330700.29	417000.00	73423.14	404123.43	0.95	1.00
6	337300.15	573000.00	128696.14	465996.29	0.82	1.00	741192.51	924000.00	159133.97	900326.47	0.95	1.00
7	551740.92	822000.00	142234.65	693975.57	0.86	1.00	1101879.81	1360000.00	233557.44	1335437.25	0.93	1.00
8	713079.13	1050000.00	175390.73	888469.87	0.86	1.00	1291522.00	1540000.00	260117.42	1551639.42	1.00	1.00
9	844867.23	1220000.00	196999.48	1041866.71	0.86	1.00	1370688.05	1630000.00	260135.91	1630823.96	1.00	1.00
10	927403.06	1300000.00	208527.30	1135930.36	0.86	1.00	1418442.20	1670000.00	260525.71	1678967.91	1.00	1.00
11	960791.05	1310000.00	209899.01	1170690.06	0.86	1.00	1440261.06	1710000.00	268739.67	1709000.73	0.98	1.00
12	944081.41	1270000.00	211097.06	1155178.47	0.86	1.00	1479631.53	1740000.00	288542.34	1768173.87	1.00	1.00
13	964367.34	1250000.00	212485.02	1176852.36	0.86	1.00	1447640.79	1700000.00	277554.82	1725195.61	1.00	1.00
14	977153.06	1240000.00	213767.59	1190920.65	0.91	1.00	1387241.02	1630000.00	256319.91	1643560.93	1.00	1.00
15	992506.96	1260000.00	216963.95	1209470.91	0.91	1.00	1372081.03	1600000.00	251954.74	1624035.76	1.00	1.00
16	998659.41	1280000.00	213621.74	1212281.15	0.95	1.00	1344377.03	1560000.00	246201.89	1590578.91	1.00	1.00
17	1001620.88	1300000.00	218509.19	1220130.07	0.95	1.00	1307909.43	1540000.00	245963.64	1553873.06	1.00	1.00
18	990528.22	1270000.00	224543.38	1215071.60	0.95	1.00	1226322.74	1440000.00	229883.41	1456206.15	1.00	1.00
19	987569.46	1260000.00	228799.98	1216369.44	0.91	1.00	1160983.07	1370000.00	213498.87	1374481.94	1.00	1.00
20	988798.61	1310000.00	239227.15	1228025.76	0.86	1.00	1116324.58	1340000.00	203711.83	1320036.41	0.98	1.00
21	956218.12	1270000.00	233621.91	1189840.03	0.91	1.00	1056308.11	1270000.00	188890.55	1245198.67	0.95	1.00
22	867646.28	1170000.00	212134.53	1079780.81	0.86	1.00	952882.70	1130000.00	165144.78	1118027.49	0.95	1.00
23	791347.56	1020000.00	211289.31	1002636.87	0.95	1.00	676797.07	853000.00	124692.12	801489.19	0.95	1.00

Table 4.3 Zone 14 Traffic Profile

Hour	Holiday						Workday					
	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)	Mean	Max	Std	Mean+1std	P(Mean+1std)	P(Mean+2std)
0	273764.55	403850.00	71594.46	345359.00	0.82	1.00	184732.00	299390.00	36670.19	221402.19	0.88	0.95
1	153598.71	246150.00	49455.05	203053.76	0.86	1.00	68832.03	117080.00	18258.49	87090.51	0.88	0.98
2	88882.53	173360.00	38107.25	126989.78	0.95	0.95	27868.50	41473.00	8406.77	36275.27	0.83	1.00
3	53735.02	107930.00	23749.66	77484.69	0.86	0.95	27330.26	44046.00	8655.06	35985.32	0.85	1.00
4	49381.73	99621.00	16424.86	65806.58	0.91	0.95	59331.48	88070.00	13642.78	72974.26	0.90	0.98
5	105660.99	183210.00	38065.17	143726.16	0.82	0.95	229218.41	296580.00	53212.90	282431.31	0.93	1.00
6	239913.04	407580.00	88184.08	328097.13	0.82	1.00	524643.71	661780.00	115872.63	640516.34	0.90	1.00
7	393281.05	563690.00	93385.86	486666.91	0.86	1.00	742546.09	918310.00	158206.12	900752.21	0.93	1.00
8	495317.32	663840.00	101409.39	596726.71	0.86	1.00	779188.15	936650.00	151748.54	930936.69	0.95	1.00
9	563360.33	739710.00	107666.67	671027.00	0.86	1.00	784446.95	946960.00	144771.51	929218.46	0.90	1.00
10	595031.38	753550.00	104085.03	699116.41	0.95	1.00	782239.92	934800.00	140627.70	922867.63	0.88	1.00
11	599775.52	751510.00	100944.96	700720.47	0.91	1.00	790106.00	964080.00	146498.86	936604.86	0.90	1.00
12	584609.80	727050.00	99051.56	683661.35	0.86	1.00	824517.15	1025200.00	164818.79	989335.94	0.85	1.00
13	595426.35	719640.00	99924.82	695351.18	0.86	1.00	804994.43	984830.00	154326.83	959321.25	0.88	1.00
14	607654.83	745550.00	102738.98	710393.82	0.91	1.00	787910.11	949010.00	142992.98	930903.09	0.90	1.00
15	612763.86	755380.00	102939.27	715703.13	0.86	1.00	788851.00	949430.00	141734.02	930585.03	0.88	1.00
16	625453.81	766010.00	107201.54	732655.35	0.86	1.00	801789.53	970200.00	145565.39	947354.92	0.93	1.00
17	631854.26	780810.00	110654.65	742508.91	0.91	1.00	801204.99	972460.00	149241.95	950446.94	0.90	1.00
18	630209.28	763040.00	111743.88	741953.17	0.91	1.00	750385.87	910640.00	132894.10	883279.97	0.90	1.00
19	622242.24	763340.00	113050.46	735292.70	0.86	1.00	700596.40	851230.00	119767.86	820364.26	0.93	1.00
20	613426.47	755340.00	114429.48	727855.95	0.86	1.00	672963.16	809760.00	113314.70	786277.86	0.90	1.00
21	588956.20	742990.00	112248.61	701204.81	0.86	1.00	634689.58	777680.00	104973.21	739662.79	0.88	1.00
22	529723.46	663400.00	94946.92	624670.38	0.91	1.00	567888.49	686040.00	89462.32	657350.81	0.88	1.00
23	442051.98	540870.00	87128.28	529180.25	0.91	1.00	400390.96	509120.00	60218.54	460609.51	0.93	1.00

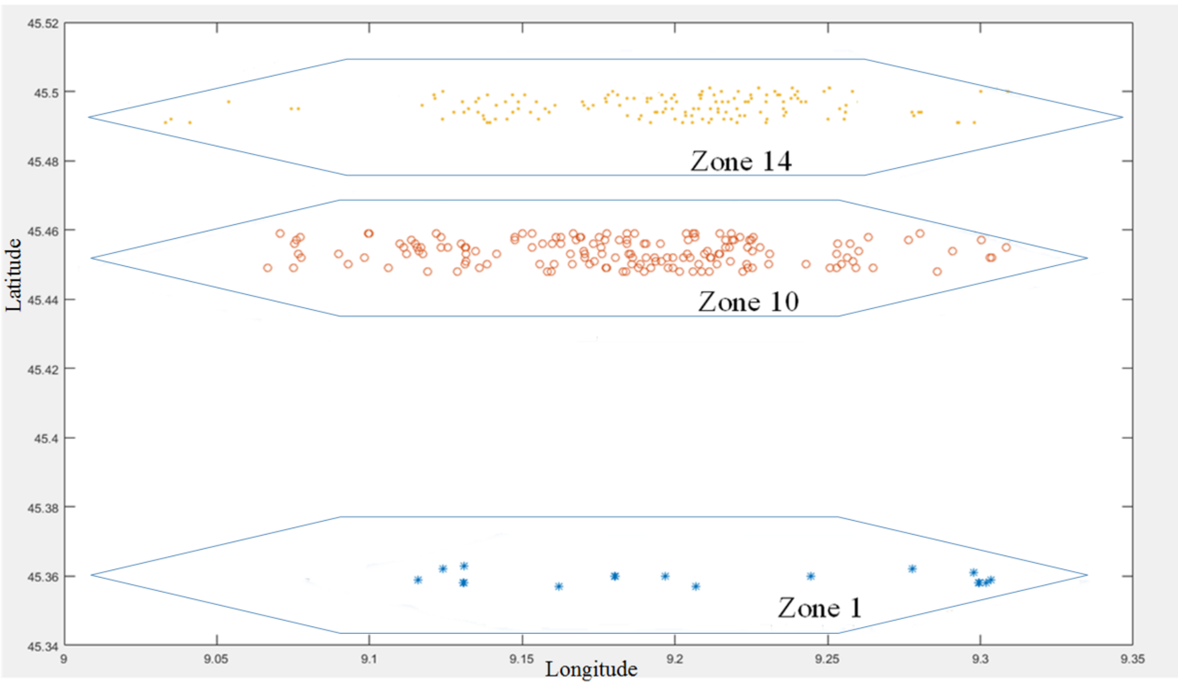

Fig. 4.5 Base stations in Zone 1, 10 and 14

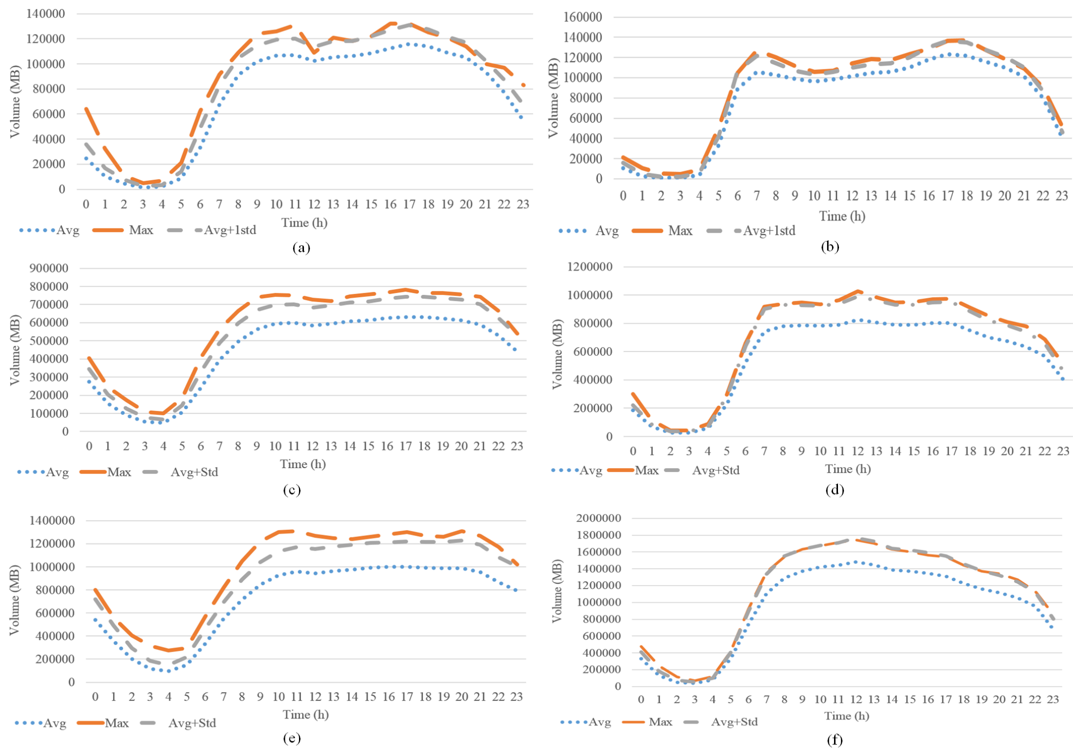

Fig. 4.6 Traffic profiles in Zones 1, 14 and 10.

4.5 Determination of Data Center Placement and Dimension

In this section, the analysis conducted in the previous section is used to first optimally place a data center in each of the considered zones. This is followed by the determination of the dimension of the data centers to meet up with the traffic demands from the zones.

4.5.1 Placement of the Data Centers

One consideration while determining the position of a data center is its distance from the base stations. Minimizing the aggregate distances between the data centers and the base stations would reduce the cost of front haul links and also would lower the delay in propagation. The problem of optimally determining the best location for such a facility can be identified as the *Weber's probelm* which also is a special case of the *Facility Location Problem* [109].

The aim of any facility location problem is to determine the most suitable place to establish one or multiple facilities in the presence of many candidate locations. The facilities are usually required to provide services to meet demands that are imposed by their customers (whose locations are known). In the model proposed in this chapter, the facilities are the data centers which provide the base stations (the customers) with computational power to process the traffic experienced in each base station. The Weber algorithm reaches a point that ensures that the weighted sum from the point to the known customers' locations reaches its minimum [110]. Since the distance of each base station from the data center are considered as the only parameter to minimize, the following mathematical model can be employed to represent it:

$$\arg\min_{Y \in R^2} f(Y, \theta) = \sum_{i=1}^{m} w_i \theta(C_i, Y) \tag{4.1}$$

Here w_i denotes the weights assigned to the i^{th} customer (base station), among m customers, belonging to the customer group C. The function θ changes the distance $\theta(C_i, Y)$ into a serving cost per customer depending on the established location of the facility Y. In this work, all BSs are judged to be equally important and are assigned with the weight of 1.

A convenient method for solving such a problem is the Weiszfled procedure [111] which is based on the gradient descent algorithm. Weiszfeld algorithm do not require differentiation or line search to obtain the minimal aggregated distance. As a result of its significantly low complexity, this algorithm converges faster in comparison to other similar algorithms. Since all the BSs are weighed equally in this case, the problem of determining the optimal location for the data center facility simplifies to a *geometric median* problem. Geometric median essentially minimizes the sum of l_2 norm of each element of the customer group C_i and iterates to obtain the best possible location for the establishment of the facility.

As three zones are considered, each with its own boundaries in terms of latitude and longitude, the possible location for a data center in z^{th} zone has to be contained in a 2-dimensional vector space, J, which also includes the location of all the BSs within that zone . The algorithm begins at a random coordinate point having latitude (x_{1z}) and longitude (y_{1z}) and attempts to locate the optimal point within the set J to minimize the sum of Euclidean distances from the BSs within that zone. The x and y values are calculated using the formula:

$$x = \frac{\sum_{j \in J} (w_{jz} x_{jz}) / d_{oj}}{\sum_{j \in J} (w_{jz} x_j) / d_{oj}}, \quad y = \frac{\sum_{j \in J} (w_{jz} y_{jz}) / d_{oj}}{\sum_{j \in J} (w_{jz} y_j) / d_{oj}} \tag{4.2}$$

where x_{jz} and y_{jz} represent the j^{th} candidate's location for the data centre in zone z with d_{oj} representing the distance between the candidate location for the data center to

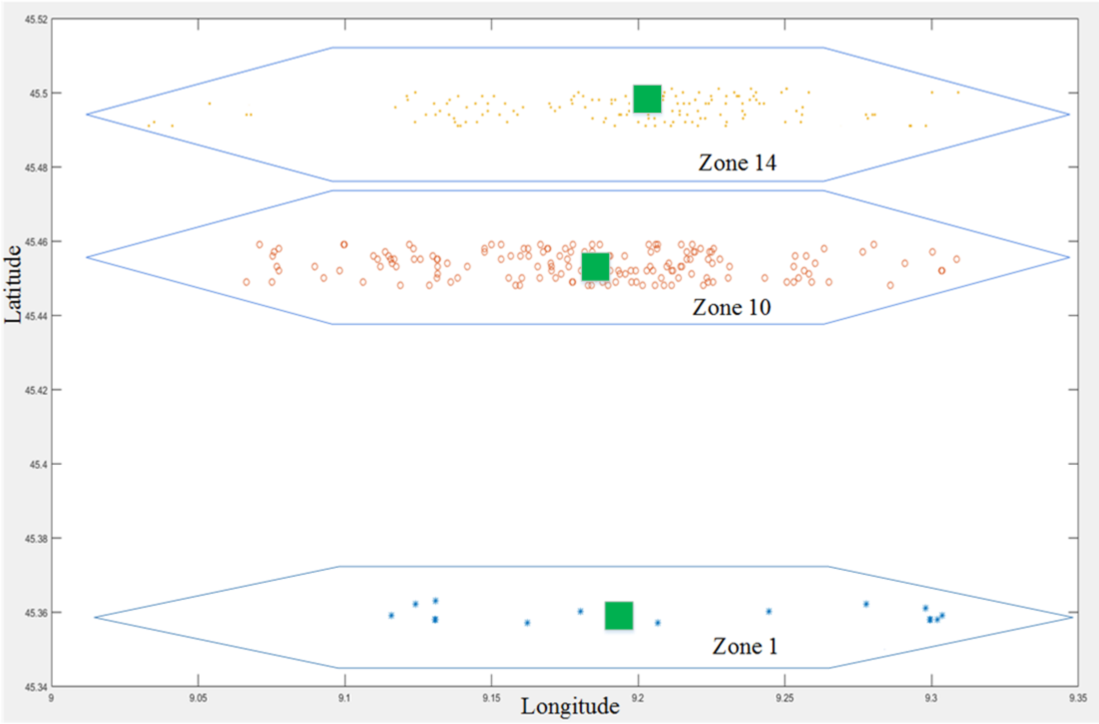

Fig. 4.7 Data Centers in Zone 1, 14 and 10

a point in set J. w_{jz} represents the weight assigned to the j^{th} candidate which is set to be 1 in this model. The iterations are continued until either a convergence is reached or if the maximum number of evaluations is completed.

With the aid of this algorithm, the ideal location for each data center is determined in Zones 1, 10 and 14. The location of the data centers among the base stations are illustrated in boxes in Figure 4.7.

4.5.2 Dimension of the Data Centers

Once the data center is established, it becomes critically important to determine its dimension. This largely depends upon the traffic demand that the data center is expected to cater for. For the purpose of this work, as the information of the single mobile operator (TIM) is only available, the size of the data centers are determined based on the traffic volumes experienced by its BSs in the zones under consideration. In a real life design case, it is expected that the infrastructure providers would lease their services to multiple operators and as such would require relevant information from other operators as well. Also, the focus is placed on the resources in terms of the computational power required to process the traffic in these zones.

The computational power is provided by servers within a data center. An area with large number of BSs would require higher amount of computational resources (CPU cores provided by the servers) to serve the traffic demands as well as to host various VNFs such as SDN controllers and virtual gateways. It is assumed that the VNFs for a particular zone are all hosted in a single centralized data center rather than being distributed all over. This approach requires less number of servers, and subsequently cores, as opposed to having a distributed VNF architecture [87]. VNFs that are intended to serve both data and control plane functionalities require more computational power than SDN controllers that deal with only control plane functionalities. The authors in [112] demonstrated that 20 cores of CPU processing power are required to handle 1 unit of data traffic demand (i.e. 1 Gbps). Conversely, only 6 cores are required by the SDN controllers to process the same amount of traffic. Based on the current practice in data centers, the authors also suggested that each server should possess 48 CPU cores to provide computational power. In the model proposed in this chapter, these specifications are adopted to design the dimensions of each zone's data center based on the traffic profile analysis made in the previous section.

While determining the processing power required by a data center (data center's capacity), the traffic profiles that the base stations under its coverage experiences need to be carefully evaluated. As 62 days worth of data is available, certain traffic characteristics might not have been captured within this time frame. One characteristic that has been observed is that the traffic volume experienced in workdays surpasses that of holidays'. However, allocating resources to meet just the maximum of peak-hour traffic would lead to over provisioning of resources that would remain underutilized most of the times. Similarly, having enough servers to meet only the average demand would lead to shortage of resources during peak demand hours thereby resulting in poor quality of services. Referring to Table 4.1, it can be seen that a good metric to determine the volume of traffic that a data center needs to be designed for is the sum of the average and standard deviations of the traffic volume. The hourly traffic volume of zones had surpassed the sum of the mean and one standard deviation in considerable number of occasions (more pronounced in the entire dataset). However, most of these volumes fall well within the sum of the mean and two standard deviations. The ideal traffic volume that a data center need to cater for, therefore, lies somewhere in the range between these two. Therefore the ideal design capacity, V_z, of the data center in zone z is heuristically determined to be:

$$V_z = D_z \times (f(std)_z) + g(mean)_z) \tag{4.3}$$

$f(std)_z$ and $g(mean)_z$ are the corresponding values of the standard deviation and mean for the maximum hourly sum of mean and one standard deviation for a particular zone z. $D_z \in [1, 2]$ is the multiplier which aids in determining the maximum traffic volume , V_z, that the z^{th} data center would be capable of serving at any given time. This multiplier is inversely proportional to the probability of the h^{th} hour's traffic volume to fall within the sum of the mean and one standard deviation in zone z, $P_h(mean + 1std)_z$. In this work, the D_z value is 1.6 when the $P_h(mean+1std)_z$ is 0.6 and D_z is 1 when $P_h(mean+1std)_z$ is 1. Then, D_z can be obtained using the following heuristically obtained mathematical relationship:

$$D_z = b + \left(\frac{(max(P_h(mean + 1std))_z - (b - 1))) \times (1 - b)}{1 - P_s} \right) \tag{4.4}$$

where b for this dataset is 1.6 and P_s is 0.6 as mentioned above. $max(P_h(mean{+}1std))_z$ denotes the probability value for the hour that demonstrates the highest sum of mean with one standard deviation ($mean + 1std$) of traffic volume in that zone. For example, in zone 1, it can be seen from Table 4.1 that the highest value of $mean + 1std$ is observed for the hour 17 (between 5 pm and 6 pm) workday traffic which corresponds to a $P_{17}(mean + 1std)_1$ value of 0.98. Therefore, to determine the deviation factor in this zone, this value is used as the $max(P_h(mean + 1std))_1$. With this, the traffic volume that the data center of the z^{th} zone needs to be designed for can be determined using equation 4.3.

As mentioned previously, based on the specification adopted this work, to process one unit of traffic i.e 1 Gbps, it is assumed that 26 cores of processing power is required. Given the aggregated nature of the data set, it is not possible to evaluate the demand volume that is experienced per second for each zone. Therefore, in this work, a linear relationship between the time and the demand is assumed i.e at every second, same amount of demand is generated resulting in a cumulative volume of v_z in the z^{th} zone per hour. Note also that the capacity of the fronthaul links between the base stations and the data center also play a crucial role in the processing of the traffic demand. This, however, is beyond the scope of this work. Using the above equations and the specifications, the capacity of each data center can be designed.

Zone 1 Data Center: This light traffic volume zone, as mentioned above, possesses $max(P_h(mean + 1std))_1$ value of 0.98 based on the maximum of sum and 1 std that is noticed in the 17^{th} hour. Using equation 4.4, the multiplier of zone 1, D_1, is obtained to be 1.03 . The corresponding mean and standard deviation values of this hour's of traffic are $12,3291.8$ MB and $13,313.68$ MB respectively. Therefore, using equation 4.3, the

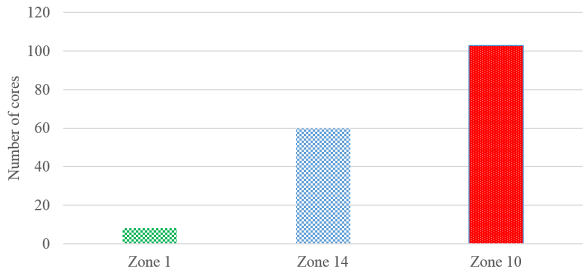

Fig. 4.8 Capacity of Data Centers

value of $137,0105$ MB (1,070 Gb) is determined as the maximum volume of traffic that this data center would be required to handle at any given hour. Note that this value is greater than any of the peak hourly traffic for both working days and holidays based on the available data. This value is also much below than the sum of mean and two standard deviation value. As such, it is a value with ample tolerance to meet the highest traffic demand that might be encountered in this zone. Using the assumptions and specifications, it is evaluated that the data center for this zone would require maximum of 8 cores to process the traffic volume for the TIM subscribers.

Zone 14 Data Center: Zone 14's medium level traffic has a $max(P_h(mean+1std))_{14}$ value of 0.85 corresponding to the 12^{th} hour. The deviation factor of this zone, D_{14} is evaluated to be 1.2250 using equation 2. With corresponding values of mean and standard deviation of 824517.20 MB and 164818.80 MB respectively, the maximum volume of traffic that the data center of this zone would have to handle at any given hour is determined to be $1,026,400$ MB ($8,019$Gb). To fulfil this volume of traffic demand, the servers in the data center in this zone need to have 60 cores of CPU power.

Zone 10 Data Center: Traffic level of this area surpasses others and possess a $max(P_h(mean + 1std))_{10}$ value and deviation factor, D_{10}, of 1 for the 12^{th} hour. The mean and standard deviation value corresponding to this hour are $1,479,632$ MB and $288,542.30$ MB respectively. As expected, the maximum volume of traffic that the data center in this zone needs to cater for, $1,768,200$ MB ($13,814$ Gb) is also the highest among all others. To satisfy this level of demand, the capacity of this data center would also have to be higher than others. This zone's data center would require 103 cores to process the traffic demands from the subscribers.

Figure 4.8 shows the number of cores that the data centers in each zone would require to process their respective traffic demands.

4.6 Machine Learning based Traffic Demand Prediction

In this section, several state-of-the-art recurrent neural network (RNN) models are employed to forecast next day's traffic of each zone based on previously collected data. The idea is to have the future traffic demand of these areas in hand to determine utilization of data center resources. Dynamic utilization of the CPU cores in a data center can reduce operational cost as well as ensure proper quality of services for end users.

RNN has proved to be an effective tool to perform prediction on time-series data. Given the inherent seasonality present in the data of the hourly aggregated traffic demand of different zones, RNN models can be used to make forecasts of future demand. Two RNN models: Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) are used. The fitness of two activation functions : Rectified Linear Unit (ReLU) and hyperbolic tangent (tanh) are also examined to determine the combination that produces the result with highest accuracy. As the holiday demand is different from the workday one, these models are tested on each type of day for the considered zones. For holidays, as 22 days of data is available, 21 days demand values are used as as input to the RNN to predict the 22^{nd} day's demand. Similarly for workdays, the RNN is fed with 39 days of data to generate the values for the 40^{th} day.

The performance of the simple RNN model is evaluated along with LSTM and GRU models on the zonal data to predict future demands. These models are used with relu and tanh activation functions. As mentioned previously, 21 days of holiday data was used to train each of these RNN models and the 22^{nd} day's data was used as a test. Similarly, the networks were trained with 39 days of workday data and the 40^{th} day's data was used for testing the prediction performances of these models. Google's open source machine learning platform- Tensorflow is used on a 2.6 GHz, 4 cores and NVDIA GTX 970 graphics card enabled computer to analyse the performances of these algorithms on holiday and workday data. The obtained results are explained below:

Holiday Prediction: Figure 4.9 (a), (c) and (e) demonstrate the performances of the considered algorithms on holiday data of zone 1, 14 and 10 respectively. Note that this zone consists of less amount of data compared to workdays (528 data points compared to 960). LSTM and GRU models generally perform well with accuracy of 90 percent and more across all regions with both activation functions. Simple RNN models however struggle to capture and maintain good performance over the span of the data set. Figure 4.10 (a) and (c) show the average of root mean square errors (RMSE) and mean absolute percentile error (MAPE) of these algorithms on holiday data set. RNN-tanh is seen to perform the worst while GRU-tanh performing marginally better than other LSTM and GRU models.

Workday Prediction: The performances of the considered machine learning algorithms on a slightly larger workday dataset are presented in Figure 4.9 (b), (d) and (f). Once more, the GRU and LSTM algorithms performed similarly to each other and better than the simple RNN algorithms. An interesting trend is noticed here for simple RNN model with tanh activation function. This particular model failed to follow the trend at a very early stage of the training and subsequently generated a steady result that did not rise or fall with the actual test data (workday demand). This behavior is consistently observed for this model on all the zones. RNN-relu model however followed the seasonality present within the data but failed to capture it entirely. Figure 4.10 (b) and (d) shows the RMSE and MAPE values of each algorithm on workday data set. The GRU and LSTM models predict with least error while maintaining accuracy of greater than 88 percent on this data set. LSTM-relu model marginally outperforms other LSTM and GRU models in this case . Notice that on both these metrics, RNN- relu perfrom worst than the RNN-tanh model. This is due to the fact that RNN-tanh model's output was somewhat close to the actual demand values as compared to the RNN-relu model which attempted to follow the trend.

Runtime: Figure 4.11 (a) and (b) demonstrates the time it takes for each of these algorithms to complete training and make prediction for both holiday and workday data respectively. Due to RNN's simpler architecture, it takes significantly less time to train and predict when compared to LSTM and GRU models. As expected, due to the presence of three gates in the LSTM architecture, it takes slightly longer runtime when compared to a simpler GRU architecture.

With the aid of machine learning algorithms, it would be possible for infrastructure providers to determine the hourly demand that will be encountered from a particular MNO within a zone. For example, data center 1 has been designed to have 8 CPU cores to handle the traffic demand within zone 1 from TIM subscribers. However, to meet the demand that occurs at about 5 am, only a fraction of these cores would be required. With hourly predicted demand values in hand, the data center resources can be optimally and dynamically allocated leading to efficient utilization of resources.

4.7 Chapter Summary

In this chapter, the open Big data set of Telecom Italia was analysed to determine the traffic profiles that exist in different zones within the city of Milan. The data set was

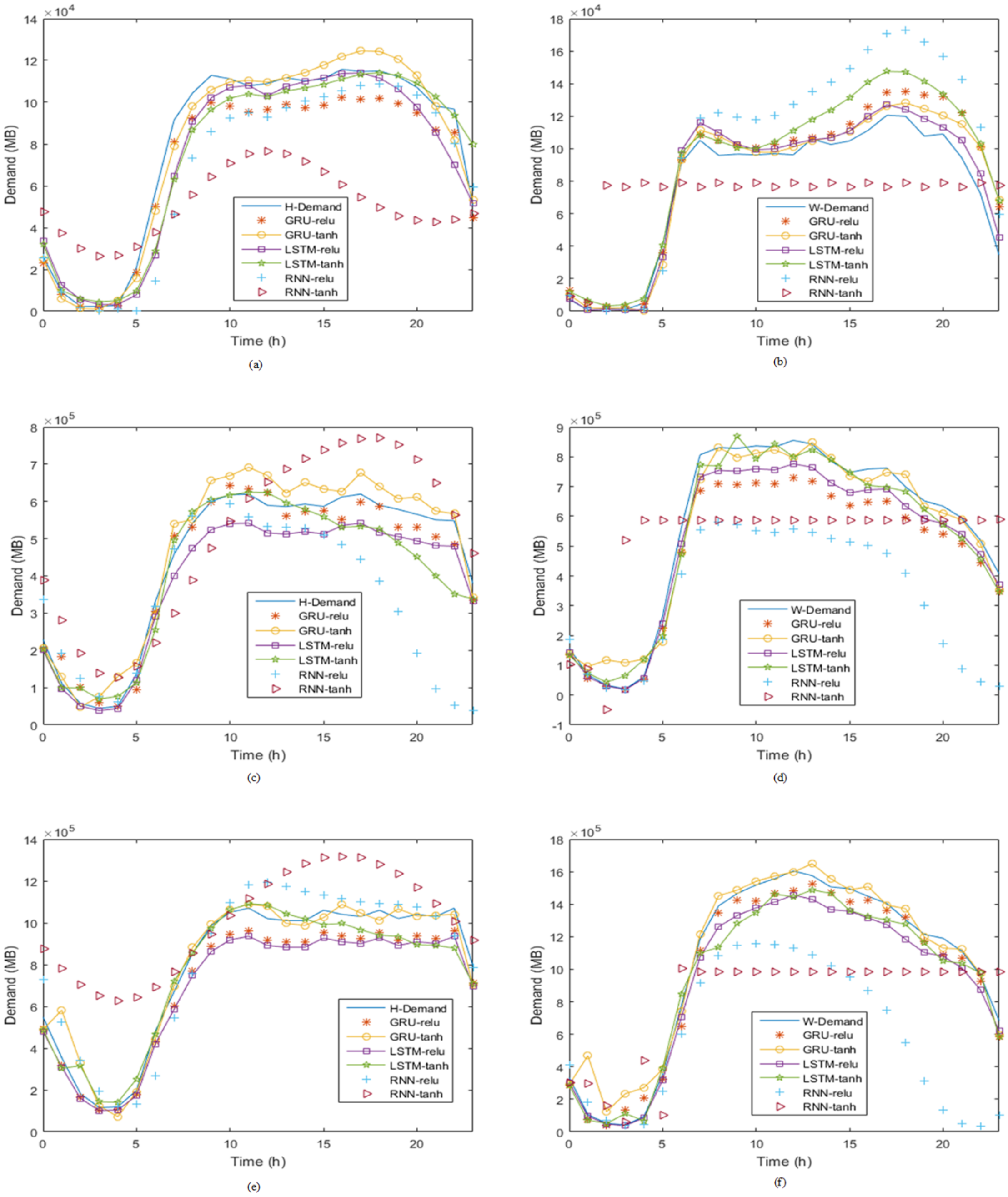

Fig. 4.9 Machine Learning Algorithms' Predictions

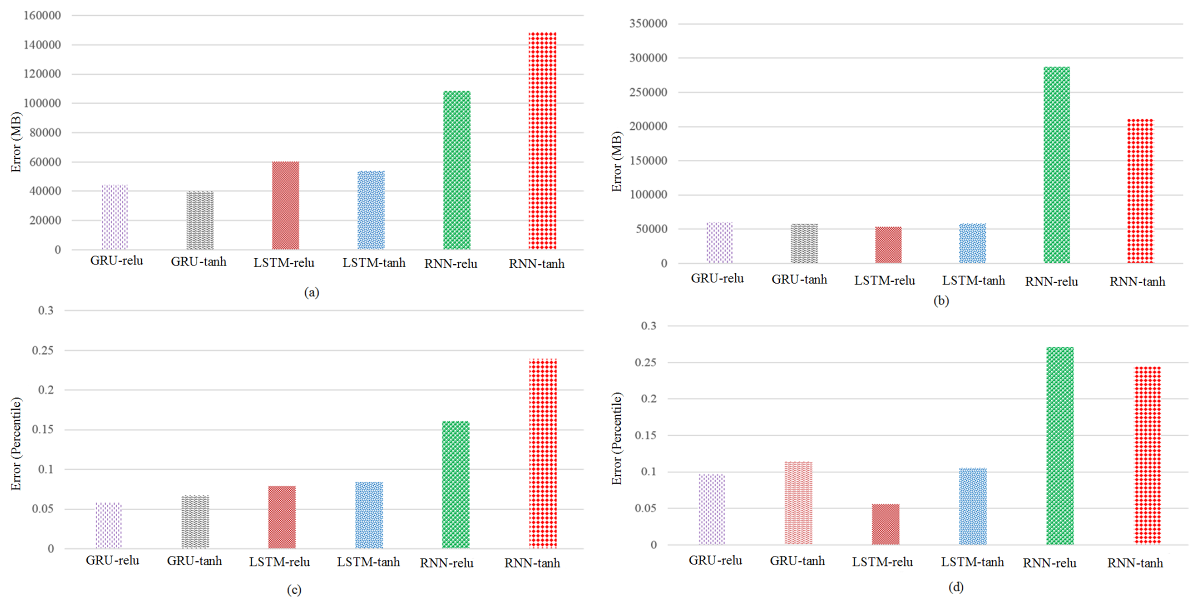

Fig. 4.10 RMSE and MAPE

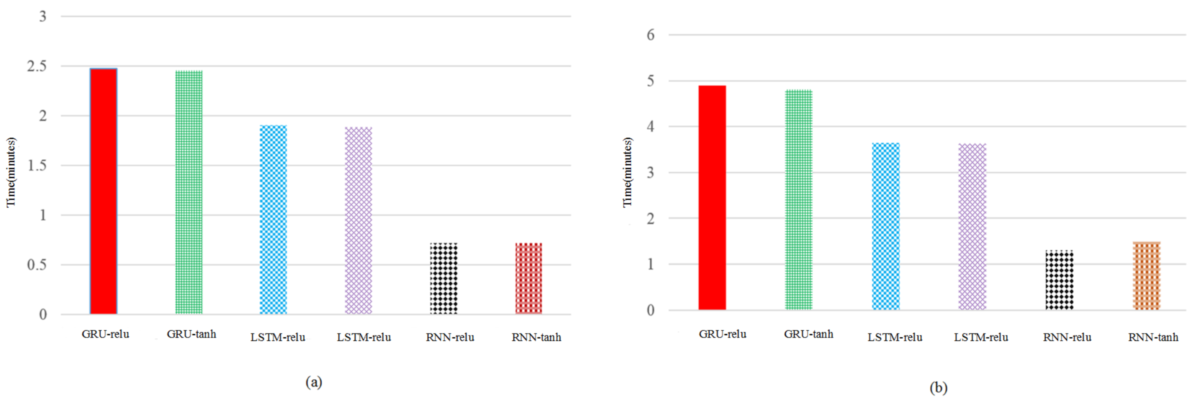

Fig. 4.11 Runtime

processed to have a hourly cellular traffic demand that arises in different parts of the city during the course of the day. The city of Milan was split into 20 zones and from that three zones (Zone 1, Zone 14 and Zone 10) were isolated as they experienced the least, medium and most volumes of traffic respectively.

Based on the location of the base stations in a zone, the establishment of a data center was proposed to host the VNFs and SDN controllers in each zone. The problem of optimal placement of data center was identified as a facility location problem which was solved using the Weiszfeld's algorithm. Furthermore, based on the traffic profile of each zone, the ideal dimension of a data center that will be capable of handling the traffic within that zone was heuristically determined.

Finally, machine learning algorithms were employed to predict the future data demand that occurs in holidays and workdays in each of the considered zones. Results showed the ability of LSTM and GRU models to predict future demand values with considerably high accuracy as opposed to simple RNN models.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

With the constant introduction of new applications and powerful mobile terminals, the wireless industry is experiencing a massive surge in the demand for data. In addition to the users of the mobile broadband applications such as video streaming and instant messaging, the 5G network is set to provide services to multiple vertical industries such as the health, automation and energy sectors. To meet up with this high demand, various changes need to be made for both the users' equipment and for service providing infrastructures. In this dissertation, a user problem in terms of selecting the best available access network among many available in a heterogeneous wireless network is first solved. Then the dissertation addressed a key challenge that the infrastructure providers face when establishing a key network facility such as data center. Subsequently, to ensure proper quality of service for users and aid dynamic utilization of data center resources, several data demand prediction models were employed. Section 5.1 presents a summary of the contributions made by this dissertation. Finally section 5.2 highlights some important directions for future work.

5.2 Summary of Contributions

In chapter 1, an overview of the current challenges that both users and service providers in the wireless industry face today was provided. This chapter highlighted some of the research problems and questions that were addressed in this work. Additionally, the key research objectives of this dissertation were also provided. Chapter 2 presented elaborate information on topics associated with the dissertation.

Chapter 3 analysed the problem of selecting the ideal access network for different types of user, under different network conditions, in the dynamic heterogeneous wireless network (HWN) environment. This dissertation proposed the integration of a branch of fuzzy set called the Intuitionuistic fuzzy set with a famous multi-criteria decision making algorithm called the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) to formulate an access network selection scheme. Traditional MCDM access network selection schemes were observed to either focus on single calls or not consider the dynamic nature of the HWN. To address these issues, IF-TOPSIS algorithm was introduced in this work for users initiating single or multiple applications simultaneously. The performance of the proposed algorithm was evaluated under various conditions. Results demonstrated the ability of the proposed algorithm to accurately select the best access network for different users. In addition, the proposed algorithm overcame the problem of ranking abnormality that is a common problem for MCDM algorithms.

Chapter 4 presented a detailed analysis of the Telecom Italia's (TIM) real world big data set. The data was studied to determine the volume of traffic that different geographical locations within the city of Milan experiences over the period of a day. It was observed that more volume of traffic is experienced in workdays than holidays. The data set was divided into small zones (consisting of 500 square cell) and establishment of a data center in each zone to cater for the experienced traffic was proposed. As data centers in 5G will be hosting various key network functionalities, their accurate location and dimension (in terms of computational power) need to be carefully determined. The work determined the optimal positioning of a data center by minimizing the aggregate distance between the facility and all the base stations within its coverage. Additionally, based on the traffic volume, the optimal dimension of different data centers were heuristically determined. Finally, several machine learning algorithms to obtain the predictions of future data demands to aid the dynamic utilization of the resources of the data centers were implemented.

5.3 Future Work

5.3.1 Access Network Selection

The computational complexity of fuzzy set could be high with increasing number of alternatives and criteria. Further work needs to be conducted with more available access networks and network parameters to assess the computational complexity of the IF-TOPSIS scheme. Additionally, other MCDM schemes such as VIKOR and GRA could be incorporated with the IF set to determine the best MCDM technique for access network selection problem.

5.3.2 Data Center Design and Data Demand Prediction

In this work the cellular trace of a single mobile network operator in Italy has been used to design the location and dimension of regional data centers. Further work needs to be conducted with multiple service providers' data to analyse the overall demand that a zone experiences from all the subscribers. This could further impact the dimension and location of a data center. Additionally, future work should attempt to devise novel forecast schemes to better predict future demand for data.

5.4 Concluding Remarks

To address various problems that the wireless communications industry face, this dissertation has proposed a RAT selection scheme to solve a user level problem and a data center design and data demand prediction framework to solve a service provider's problem. This dissertation has also provided some possible areas that still need further exploration. This dissertation, therefore, can serve as a benchmark to solve other relevant problems.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Feb. 2016.

[2] T. Han, G. Mao, Q. Li, L. Wang, and J. Zhang, "Interference Minimization in 5G Heterogeneous Networks," *ArXiv e-prints*, Jan. 2017.

[3] N. G. M. N. Alliance, "Ngmn 5g initiative white paper," Feb. 2015.

[4] ETSI and S. Antipolis, "Network functions virtualisation (nfv); management and orchestration," Feb. 2014.

[5] O. N. Foundation, "Sdn architecture," Feb. 2014.

[6] Huaweii, "5g network architecture: A high-level perspective," Feb. 2016.

[7] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5g: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, May 2017.

[8] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous lte/lte-a networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2142–2180, Fourthquarter 2014.

[9] J. Churi, T. Surendran, S. Tigdi, and S. Yewale, "Evolution of networks (2g-5g)," in *International Conference on Advacnes in Communication and Computing Technologies*, 2012, pp. 8–13.

[10] 3GPP, *Architecture Enhancements for Non-3gpp Accesses*, 2007-2012.

[11] P. S. Khodashenas, J. Aznar, A. Legarrea, C. Ruiz, M. S. Siddiqui, E. Escalona, and S. Figuerola, "5g network challenges and realization insights," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, July 2016, pp. 1–4.

[12] R. Chávez-Santiago, M. Szydełko, A. Kliks, F. Foukalas, Y. Haddad, K. E. Nolan, M. Y. Kelly, M. T. Masonta, and I. Balasingham, "5g: The convergence of wireless communications," *Wireless Personal Communications*, vol. 83, no. 3, pp. 1617–1642, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11277-015-2467-2

[13] O. T. Eluwole and M. Lohi, "Coordinated multipoint power consumption modeling for energy efficiency assessment in lte/lte-advanced cellular networks," in *2012 19th International Conference on Telecommunications (ICT)*, April 2012, pp. 1–6.

[14] R. Pries, H. J. Morper, N. Galambosi, and M. Jarschel, "Network as a service - a demo on 5g network slicing," in *2016 28th International Teletraffic Congress (ITC 28)*, vol. 01, Sept 2016, pp. 209–211.

[15] A. M. Miyim, M. Ismail, and R. Nordin, "Vertical handover solutions over lte-advanced wireless networks: An overview," *Wireless Personal Communications*, vol. 77, no. 4, pp. 3051–3079, Aug 2014. [Online]. Available: http://dx.doi.org/10.1007/s11277-014-1695-1

[16] F. Kaleem, "Vhits: Vertical handoff initiation and target selection in a heterogeneous wireless network," Ph.D. dissertation, Florida International University, 2012.

[17] F. Zhu and J. McNair, "Multiservice vertical handoff decision algorithms," *EURASIP J. Wirel. Commun. Netw.*, vol. 2006, no. 2, pp. 52–52, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1155/WCN/2006/25861

[18] F. Siddiqui and S. Zeadally, "Mobility management across hybrid wireless networks: Trends and challenges," *Comput. Commun.*, vol. 29, no. 2, pp. 1363–1385, May 2006.

[19] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Communications Magazine*, vol. 44, no. 10, pp. 96–103, Oct 2006.

[20] S. Maaloul, M. Afif, and S. Tabbane, "Vertical handover decision policy based on the end user's perceived quality of service," in *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, March 2013, pp. 493–498.

[21] P. TalebiFard and V. C. M. Leung, "A dynamic context-aware access network selection for handover in heterogeneous network environments," in *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2011, pp. 385–390.

[22] V. Gupta, "Network selection in 3g-wlan interworking environment using topsis," in *2016 11th International Conference on Industrial and Information Systems (ICIIS)*, Dec 2016, pp. 512–517.

[23] D. Zhang, Y. Zhang, N. Lv, and Y. He, "An access selection algorithm based on gra integrated with fahp and entropy weight in hybrid wireless environment," in *2013 7th International Conference on Application of Information and Communication Technologies*, Oct 2013, pp. 1–5.

[24] J. D. Martínez-Morales, U. Pineda-Rico, and E. Stevens-Navarro, "Performance comparison between madm algorithms for vertical handoff in 4g networks," in *2010 7th International Conference on Electrical Engineering Computing Science and Automatic Control*, Sept 2010, pp. 309–314.

[25] A. Ahmad, M. M. Rathore, A. Paul, S. Rho, M. Imran, and M. Guizani, "A multi-parameter based vertical handover decision scheme for m2m communications in hetmanet," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–8.

[26] S. Liang, Y. Zhang, B. Fan, and H. Tian, "Multi-attribute vertical handover decision-making algorithm in a hybrid vlc-femto system," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1521–1524, July 2017.

[27] A. F. Almutairi, M. A. Landolsi, and A. O. Al-Hawaj, "Weighting selection in gra-based madm for vertical handover in wireless networks," in *2016 UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim)*, April 2016, pp. 331–336.

[28] Meenakshi and N. P. Singh, "A comparative study of cooperative and non-cooperative game theory in network selection," in *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, March 2016, pp. 612–617.

[29] T. Bi, Z. Yuan, R. Trestian, and G. M. Muntean, "Uran: Utility-based reputation-oriented access network selection strategy for hetnets," in *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, June 2015, pp. 1–6.

[30] N. Abbas and J. J. Saade, "A fuzzy logic based approach for network selection in wlan/3g heterogeneous network," in *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, Jan 2015, pp. 631–636.

[31] L. Wang and G. S. G. S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks x2014; a tutorial," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 271–292, First 2013.

[32] Y. Yeryomin and J. Seitz, "Enhanced multi-criteria-based path selection algorithm for heterogeneous networks," in *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2016, pp. 804–809.

[33] A. Rafik and H. Oleg, *Decision theory with Imperfect Information.* World Scientific Publishing Company, 2014, an optional note.

[34] Y. Cao, D. Duan, X. Cheng, L. Yang, and J. Wei, "Dynamic network selection in hetnets: A social-behavioral (sobe) approach," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 4653–4658.

[35] M. E. Helou, S. Lahoud, M. Ibrahim, and K. Khawam, "A hybrid approach for radio access technology selection in heterogeneous wireless networks," in *European Wireless 2013; 19th European Wireless Conference*, April 2013, pp. 1–6.

[36] D. Todinca and C. Cernăzanu-Glăvan, "Novel network selection algorithm based on fuzzy inference," in *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, May 2013, pp. 467–472.

[37] C. Hwang and K. Yoon, *Multiple Attribute Decision Making.* Springer-Verlag Berlin Heidelberg, 1981.

[38] S. Maaloul, M. Afif, and S. Tabbane, "Vertical handover decision policy based on the end user's perceived quality of service," in *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, March 2013, pp. 493–498.

[39] R. Tawil, G. Pujolle, and O. Salazar, "A vertical handoff decision scheme in heterogeneous wireless systems," in *VTC Spring 2008 - IEEE Vehicular Technology Conference*, May 2008, pp. 2626–2630.

[40] L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications.* New York: Springer, 1981.

[41] M. Lahby, L. Cherkaoui, and A. Adib, "An enhanced-topsis based network selection technique for next generation wireless networks," in *ICT 2013*, May 2013, pp. 1–5.

[42] I. Chamodrakas and D. Martakos, "A utility-based fuzzy topsis method for energy efficient network selection in heterogeneous wireless networks," *Applied Soft Computing*, vol. 11, no. 4, pp. 3734 – 3743, 2011.

[43] F. Bari and V. Leung, "Multi-attribute network selection by iterative topsis for heterogeneous wireless access," in *2007 4th IEEE Consumer Communications and Networking Conference*, Jan 2007, pp. 808–812.

[44] M. A. Senouci, S. Hoceini, and A. Mellouk, "Utility function-based topsis for network interface selection in heterogeneous wireless networks," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[45] G. H. Opricovic, "Multiple criteria planning of postearthquake sustainable reconstruction," *Journal of Computer- Aided Civil and Infrastructure Engineering*, vol. 17, no. 3, pp. 211–220, 2002.

[46] S. Baghla and S. Bansal, "Effect of normalization techniques in vikor method for network selection in heterogeneous networks," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, Dec 2014, pp. 1–6.

[47] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 2015.

[48] Á. J. Sánchez-García, H. V. Ríos-Figueroa, M. de Lourdes Velasco Vasquez, G. C. Vega, and A. Marín-Hernández, "Motion prediction of regions through the statistical temporal analysis using an autoregressive moving average (arma) model," *Research in Computing Science*, vol. 77, pp. 9–20, 2014.

[49] D. Miao, W. Sun, X. Qin, and W. Wang, "Msfs: Multiple spatio-temporal scales traffic forecasting in mobile cellular network," in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and*

Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Aug 2016, pp. 787–794.

[50] C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averagesl,” *International Journal of Forecasting*, vol. 20, pp. 5–10, 2004.

[51] D. Tikunov and T. Nishimura, “Traffic prediction for mobile network using holt-winter’s exponential smoothing,” in *2007 15th International Conference on Software, Telecommunications and Computer Networks*, Sept 2007, pp. 1–5.

[52] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, Dec 1943.

[53] S. Y. King and J. N. Hwang, “Neural network architectures for robotic applications,” *IEEE Transactions on Robotics and Automation*, vol. 5, no. 5, pp. 641–657, Oct 1989.

[54] R. Akmeliawati, M. P. L. Ooi, and Y. C. Kuang, “Real-time malaysian sign language translation using colour segmentation and neural network,” in *2007 IEEE Instrumentation Measurement Technology Conference IMTC 2007*, May 2007, pp. 1–6.

[55] A. Graves, “Supervised sequence labelling with recurrent neural networks,” 2012.

[56] C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2553–2561. [Online]. Available: http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf

[57] R. Rojas, *Neural Networks*, 1996.

[58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[59] “An Intuitive Explanation of Convolutional Neural Networks,” https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/, 2016, [Online; accessed 19-December-2018].

[60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 11, pp. 533–536, September 1986.

[61] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML’13. JMLR.org, 2013, pp. III–1310–III–1318. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043083

[62] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[63] J. Chung, Ç. Gülçehre, K. Choand, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv e-prints*, vol. abs/1412.3555, 2014, presented at the Deep Learning workshop at NIPS2014. [Online]. Available: https://arxiv.org/abs/1412.3555

[64] P. N. Tran and N. Boukhatem, "The distance to the ideal alternative (dia) algorithm for interface selection in heterogeneous wireless networks," in *Proceedings of the 6th ACM International Symposium on Mobility Management and Wireless Access*, ser. MobiWac '08. New York, NY, USA: ACM, 2008, pp. 61–68. [Online]. Available: http://doi.acm.org/10.1145/1454659.1454671

[65] M. Lahby, L. Cherkaoui, and A. Adib, "A novel ranking algorithm based network selection for heterogeneous wireless access," *Journal of Networks*, vol. 8, 2013.

[66] S. Prakash, R. B. Patel, and V. K. Jain, "A multi-attribute intuitionistic fuzzy group decision method for network selection in heterogeneous wireless networks using topsis," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 11, pp. 5229–5252, 2016.

[67] F. Bouali, K. Moessner, and M. Fitch, "A context-aware user-driven framework for network selection in 5g multi-rat environments," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Sept 2016, pp. 1–7.

[68] E. Ndashimye, N. I. Sarkar, and S. K. Ray, "A novel network selection mechanism for vehicle-to-infrastructure communication," in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2016, pp. 483–488.

[69] S. Sangeetha. and T. A. A. Victoire, "Radio access technology selection in heterogeneous wireless networks using a hybrid fuzzy-biogeography based optimization technique," *Wirel. Pers. Commun.*, vol. 87, no. 2, pp. 399–417, Mar. 2016. [Online]. Available: http://dx.doi.org/10.1007/s11277-015-2890-4

[70] O. E. Falowo and H. A. Chan, "Rat selection for multiple calls in heterogeneous wireless networks using modified topsis group decision making technique," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2011, pp. 1371–1375.

[71] E. Obayiuwana, "Efficient radio resource management in next generation wireless networks," Ph.D. dissertation, University of Cape Town, 2017.

[72] O. E. Falowo and H. A. Chan, "Dynamic rat selection for multiple calls in heterogeneous wireless networks using group decision-making technique," *Computer Networks*, vol. 56, no. 4, pp. 1390–1401, 2012.

[73] ——, "Effect of call dynamics of a multiservice multimode terminal on rat selection in heterogeneous wireless networks," in *2012 IEEE Global Communications Conference (GLOBECOM)*, Dec 2012, pp. 5249–5253.

[74] J. F. Orimolade, "Access network selection schemes for multiple calls in next generation wireless networks," Ph.D. dissertation, University of Cape Town, 2017.

[75] K. T. Atanassov, "Intuitionistic fuzzy sets," *Fuzzy Sets Syst.*, vol. 20, no. 1, pp. 87–96, Aug. 1986. [Online]. Available: http://dx.doi.org/10.1016/S0165-0114(86)80034-3

[76] C. Kahraman, M. K. Ghorabaee, E. K. Zavadskas, S. C. Onar, M. Yazdani, and B. Oztaysi, "Intuitionistic fuzzy edas method: an application to solid waste disposal site selection," *Journal of Environmental Engineering and Landscape Management*, vol. 25, no. 1, pp. 1–12, 2017. [Online]. Available: http://dx.doi.org/10.3846/16486897.2017.1281139

[77] H. Liao, Z. Xu, X.-J. Zeng, and D.-L. Xu, "An enhanced consensus reaching process in group decision making with intuitionistic fuzzy preference relations," *Information Sciences*, vol. 329, no. Supplement C, pp. 274 – 286, 2016, special issue on Discovery Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025515006787

[78] Kavita, S. P. Yadav, and S. Kumar, *A Multi-criteria Interval-Valued Intuitionistic Fuzzy Group Decision Making for Supplier Selection with TOPSIS Method.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 303–312. [Online]. Available: https://doi.org/10.1007/978-3-642-10646-0_37

[79] Z. Xu, "Intuitionistic fuzzy multiattribute decision making: An interactive method," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 3, pp. 514–525, June 2012.

[80] B. Martini, F. Paganelli, P. Cappanera, S. Turchi, and P. Castoldi, "Latency-aware composition of virtual functions in 5g," in *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, April 2015, pp. 1–6.

[81] W. Rankothge, F. Le, A. Russo, and J. Lobo, "Optimizing resource allocation for virtualized network functions in a cloud center using genetic algorithms," *IEEE Transactions on Network and Service Management*, vol. 14, no. 2, pp. 343–356, June 2017.

[82] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, "Topology-aware prediction of virtual network function resource requirements," *IEEE Transactions on Network and Service Management*, vol. 14, no. 1, pp. 106–120, March 2017.

[83] K. Suksomboon, M. Fukushima, M. Hayashi, R. Chawuthai, and H. Takeda, "Lawnfo: A decision framework for optimal location-aware network function outsourcing," in *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, April 2015, pp. 1–9.

[84] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck, "Towards edge slicing: Vnf placement algorithms for a dynamic realistic edge cloud environment," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[85] R. Tripathi, S. Vignesh, V. Tamarapalli, and D. Medhi, "Cost efficient design of fault tolerant geo-distributed data centers," *IEEE Transactions on Network and Service Management*, vol. 14, no. 2, pp. 289–301, June 2017.

[86] K. Zheng, W. Zheng, L. Li, and X. Wang, "Powernets: Coordinating data center network with servers and cooling for power optimization," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 661–675, Sept 2017.

[87] A. Basta, A. Blenk, K. Hoffmann, H. J. Morper, M. Hoffmann, and W. Kellerer, "Towards a cost optimal design for a 5g mobile core network based on sdn and nfv," *IEEE Transactions on Network and Service Management*, vol. 14, no. 4, pp. 1061–1075, Dec 2017.

[88] A. Furno, D. Naboulsi, R. Stanica, and M. Fiore, "Mobile demand profiling for cellular cognitive networking," *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 772–786, March 2017.

[89] S. Wang, X. Zhang, J. Zhang, J. Feng, W. Wang, and K. Xin, "An approach for spatial-temporal traffic modeling in mobile cellular networks," in *2015 27th International Teletraffic Congress*, Sept 2015, pp. 203–209.

[90] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3899–3912, June 2017.

[91] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 234–240, June 2014.

[92] J. Dai and J. Li, "Vbr mpeg video traffic dynamic prediction based on the modeling and forecast of time series," in *2009 Fifth International Joint Conference on INC, IMS and IDC*, Aug 2009, pp. 1752–1757.

[93] O. Cappe, E. Moulines, J. C. Pesquet, A. P. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 14–27, May 2002.

[94] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, "Traffic matrices: Balancing measurements, inference and modeling," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '05. New York, NY, USA: ACM, 2005, pp. 362–373. [Online]. Available: http://doi.acm.org/10.1145/1064212.1064259

[95] M. C. Falvo, M. Gastaldi, A. Nardecchia, and A. Prudenzi, "Kalman filter for short-term load forecasting: An hourly predictor of municipal load," in *The 16th IASTED International Conference on Applied Simulation and Modelling*, ser. ASM '07. Anaheim, CA, USA: ACTA Press, 2007, pp. 364–369. [Online]. Available: http://dl.acm.org/citation.cfm?id=1659042.1659112

[96] F. Ashtiani, J. A. Salehi, and M. R. Aref, "Mobility modeling and analytical solution for spatial traffic distribution in wireless multimedia networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1699–1709, Dec 2003.

[97] K. Tutschku and P. Tran-Gia, "Spatial traffic estimation and characterization for mobile communication network design," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 804–811, Jun 1998.

[98] T. Oliveira, J. S. Barbar, and A. S. Soares, "Computer network traffic prediction: a comparison between traditional and deep learning neural networks," *International Journal of Big Data Intelligence*, vol. 3, no. 1, pp. 28–37, 2016, pMID: 73903. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJBDI.2016.073903

[99] C. W. Huang, C. T. Chiang, and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–6.

[100] M. Barabas, G. Boanea, A. B. Rus, V. Dobrota, and J. Domingo-Pascual, "Evaluation of network traffic prediction based on neural networks with multi-task learning and multiresolution decomposition," in *2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing*, Aug 2011, pp. 95–102.

[101] C. K. Dominicini, G. L. Vassoler, L. F. Meneses, R. S. Villaca, M. R. N. Ribeiro, and M. Martinello, "Virtphy: Fully programmable nfv orchestration architecture for edge data centers," *IEEE Transactions on Network and Service Management*, vol. 14, no. 4, pp. 817–830, Dec 2017.

[102] S. Gebert, D. Hock, T. Zinner, P. Tran-Gia, M. Hoffman, M. Jarschel, E. Schmidt, R. Braun, C. Banse, and A. Kopsel, "Demonstrating the optimal placement of virtualized cellular network functions in case of large crowd events," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 359–360, Aug. 2014. [Online]. Available: http://doi.acm.org/10.1145/2740070.2631428

[103] C. H. Liu and J. Fan, "Scalable and efficient diagnosis for 5g data center network traffic," *IEEE Access*, vol. 2, pp. 841–855, 2014.

[104] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the nfv provisioning puzzle: Efficient placement and chaining of virtual network functions," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2015, pp. 98–106.

[105] R. Shi, J. Zhang, W. Chu, Q. Bao, X. Jin, C. Gong, Q. Zhu, C. Yu, and S. Rosenberg, "Mdp and machine learning-based cost-optimization of dynamic resource allocation for network function virtualization," in *2015 IEEE International Conference on Services Computing*, June 2015, pp. 65–73.

[106] G. Barlacchi, M. Nadai, R. Larcher, A.Casella, C. . Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of milan and the province of trentino," *Scientific Data*, vol. 2, October 2015.

[107] "Telecom Italia Dataset Traffic Profiles," https://www.scribd.com/document/372629127/Telecom-Italia-Dataset-Traffic-Profile, 2018, [Online; accessed 3-March-2018].

[108] S. Troia, "Machine learning-based traffic prediction and pattern extraction for dynamic optical routing in sdn mobile metro networks," 2016.

[109] P. V. Heiningen, E. Reehuis, and T. Bäck, "Comparing a weiszfeld's-based procedure and (1+1)-es for solving the planar single-facility location-routing problem," in *2015 IEEE Symposium Series on Computational Intelligence*, Dec 2015, pp. 1743–1750.

[110] R. Z. Farahani and M. Hekmatfar, *Facility Location*, 2009.

[111] E. Weiszfeld. and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Annals of Operations Research*, vol. 167, no. 1, pp. 7–41, March 2009. [Online]. Available: https://doi.org/10.1007/s10479-008-0352-z

[112] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K.Hoffmann, "Applying nfv and sdn to lte mobile core gateways, the functions placement problem," in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, & Challenges*, New York, NY, USA: ACM, 2014, pp. 33–38. [Online]. Available: http://doi.acm.org/10.1145/2627585.2627592