Novel Schemes for Capacity Management In Cellular Networks

22nd November 2024

Marawan. S. M. A. Azab

TUDelft

kpn

Delft University of Technology

MSc Thesis Report

Novel Schemes for Capacity Management In Cellular Networks

Author:
Marawan. S. M. A. Azab (5864887)

Thesis committee:

TU Delft supervisors:
Dr. ir. E. Smeitink
Dr. ir. Ramin Hekmat
Dr. ing. Edgar van Boven

KPN supervisors:
Ing. Hadi. S. Tabarestani
Ir. Jamal Bouhannouche
Dr. ir. Zoubir Irahhauten

22nd November 2024

Preface

This thesis represents the culmination of my journey in the Master's program in Wireless Communication and Sensing at the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. This period has been marked by remarkable growth, learning, and collaboration, profoundly shaped by the support and contributions of numerous individuals.

Foremost, I'd like to express my sincere appreciation to my KPN supervisors, Dr. ir. Zoubir Irahhauten, along with Ir. Jamal Bouhannouche and Ing. Hadi Shekarian Tabarestani. Your encouragement, insights, and technical expertise were crucial in elevating the standard and thoroughness of this thesis. I particularly appreciate your efforts in framing the research topic from KPN's viewpoint and establishing a robust foundation for this study.

Similarly, I extend my sincere gratitude to my TU Delft supervisors, Dr. ir. Eric Smeitink, Dr. ing. Edgar van Boven, and Dr. ir. Ramin Hekmat. Your critical analysis, invaluable guidance, meticulous attention to detail, and constant support were instrumental in elevating the quality of this thesis. Your combined expertise significantly shaped this work. I am especially grateful to Dr. ing. Edgar van Boven for your pivotal role in establishing this research collaboration between TU Delft and KPN, and for your ongoing support throughout the project.

I also wish to thank Dr. RangaRao Venkatesha Prasad for his valuable contribution as part of my committee.

To my wife and our newborn son, Adam, I owe a special debt of gratitude. Your unwavering support, understanding, and the quiet working environment you provided were crucial to the completion of this thesis. Adam, your arrival at the beginning of this journey has been a constant source of inspiration and motivation.

Lastly, I extend my sincere thanks to my colleagues at KPN for their invaluable assistance and cooperation. To my friends and family, your constant encouragement, understanding, and support during this challenging yet rewarding period were deeply appreciated.

Marawan Azab
Rotterdam

Abstract

The exponential growth in mobile network traffic, driven by the rapid deployment of 5G technologies and the proliferation of new services, presents significant challenges for telecommunication operators. This thesis addresses these challenges by developing a predictive capacity management solution for 4G and 5G cellular networks. The primary objective is to forecast network traffic and identify potential congestion points up to one year in advance, enabling proactive network management and optimizing resource allocation, particularly through the use of spectral efficiency as a key predictive measure.

This study utilizes data from KPN's Operations Support System (OSS), comprising 67 days of hourly data across the entire network, with a focus on predicting future traffic and network performance up to one year ahead. The methodology integrates historical data analysis, time series forecasting, and machine learning techniques. The approach combines Cumulative Distribution Function (CDF) modeling for traffic volume prediction with supervised machine learning algorithms, including Linear Regression, Lasso Regression, Random Forest, and CatBoost, to forecast Physical Resource Block (PRB) utilization and spectral efficiency at the sector level.

The detailed analysis identifies Lasso Regression as the most effective model for predicting spectral efficiency, with the lowest Mean Absolute Percentage Error (MAPE). Lasso's ability to handle extrapolation beyond observed data ranges makes it particularly well-suited for long-term capacity management when combined with CDF-based traffic prediction. The findings demonstrate significant improvements in the accuracy of congestion predictions and the efficiency of resource utilization.

The study also revealed that, without additional resources, the number of congested sectors is expected to increase as traffic demand continues to grow. This highlights the critical need for new spectrum allocation to maintain service quality. Additionally, the research evaluated the impact of deploying new spectrum resources, such as the 3.5 GHz band, in specific sectors. The results showed that the deployment of the 3.5 GHz band significantly reduced congestion and improved network performance and user experience during the forecast period.

Abbreviations and Acronyms

Acronyms

2G	Second generation
3G	Third generation
4G	Fourth generation
5G	Fifth generation
GB	Gigabyte
VR	Virtual Reality
AR	Augmented Reality
LTE	Long-Term Evolution
FDD	Frequency Division Duplex
TDD	Time Division Duplex
NR	New Radio
UE	User Equipment
BS	Base Station
PRB	Physical Resource Block
QoS	Quality of Service
KPI	Key Performance Indicator
NSA	Non-Standalone
SA	Standalone
BW	Bandwidth
OSS	Operations Support System
CA	Carrier Aggregation
DL	Downlink
UL	Uplink
CDF	Cumulative Distribution Function
MAPE	Mean Absolute Percentage Error
OFDMA	Orthogonal Frequency Division Multiple Access
FWA	Fixed Wireless Access
RAN	Radio Access Network
MNO	Mobile Network Operator
MIMO	Multiple Input Multiple Output
DBSCAN	Density-Based Spatial Clustering of Applications with Noise

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
T_{ue}	User throughput
C	Cell capacity
ρ	PRB utilization
T_{sch}	Scheduled throughput
y_i	Dependent variable
x_{i1}, x_{iD}	Explanatory variables
w_1, w_D	Regression coefficients
e_i	Observation error
$\mathbf{x_i}^T\mathbf{w}$	Matrix multiplication of explanatory variables and regression coefficients
M	Number of observations
D	Regression model order
$\mathbf{X}$	Matrix of explanatory variables
$\mathbf{y}$	Vector of dependent variables
$\hat{\mathbf{w}}$	Estimated regression coefficients
$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2$	Least squares objective function
$\hat{\beta}$	Estimated coefficients (Lasso regression)
α	Regularization parameter
$\Omega(f_k)$	Regularization term for model complexity (XGBoost and CatBoost)
$\hat{y}$	Predicted value
$\mathcal{L}(\theta)$	Loss function (XGBoost, CatBoost)
$h_i(\mathbf{x})$	Prediction from the i-th decision tree (Random Forest)
x	Original value (Min-Max normalization)
x'	Normalized value (Min-Max normalization)
Z	Z-score
μ	Mean of the dataset
σ	Standard deviation of the dataset
Y_t	Time series value
T_t	Trend component (time series decomposition)
S_t	Seasonal component (time series decomposition)
R_t	Residual component (time series decomposition)
$r_{X,Y}$	Pearson correlation coefficient
$\bar{X}, \bar{Y}$	Mean values of X and Y
$F_X(x)$	Cumulative Distribution Function (CDF)
$P(X \leq x)$	Probability that X is less than or equal to x
$F_X^{-1}(u)$	Inverse CDF of X
G_t	Dynamic growth factor
$T_m(X_i)$	Prediction from the m-th decision tree in CatBoost
γ_m	Learning rate in CatBoost
$r_i^{(m)}$	Residual for the i-th data point in CatBoost
R^2	Coefficient of determination

1

Introduction

1.1 Overview of Cellular Networks

The evolution of cellular networks from second-generation (2G) to fifth-generation (5G) represents a journey of technological advancement driven by the increasing demand for higher data rates, enhanced connectivity, and improved capacity. Each generation has addressed the limitations of its predecessor while introducing new capabilities to meet growing user demands and emerging applications.

2G technology, introduced in the early 1990s, marked the transition from analog to digital cellular systems. Global System for Mobile Communications (GSM) and its enhancement Enhanced Data rates for GSM Evolution(EDGE) provided data rates up to 384 Kbps, enabling basic data services such as text messaging [1]. However, 2G systems were primarily designed for voice communications and faced limitations in data transmission capacity [2]. Third-generation (3G) technology brought substantial improvements in data transmission. High-Speed Downlink Packet Access(HSDPA) and High-Speed Uplink Packet Access(HSUPA) technologies achieved downlink speeds up to 14.4 Mbps and uplink speeds up to 5.76 Mbps, respectively [3]. This advancement facilitated mobile internet access and video calling, though challenges in spectrum allocation and energy consumption persisted. Fourth-generation (4G) technology, introduced in 2009, represented a significant leap forward. Long-Term Evolution(LTE) and LTE-Advanced delivered peak downlink throughput exceeding 100 Mbps and uplink speeds of 50 Mbps [4]. A key innovation in 4G was the use of Orthogonal Frequency Division Multiple Access (OFDMA), which improved spectrum efficiency by dividing carrier frequency bandwidth into sub carriers allocated to different users [5].

The current deployment of 5G networks marks another revolutionary step in mobile communication technology. Figure1.1 shows that 5G is designed to support three primary use cases [6]:

- **Enhanced Mobile Broadband (eMBB)**: Supports high data rates for applications like AR, ultra-high-definition (UHD) video streaming, and cloud gaming, with average speeds of 200 Mbps [7].
- **Ultra-Reliable Low-Latency Communications (URLLC)**: Ensures 1 ms latency and 99.999% reliability, crucial for applications such as self-driving cars and industrial automation [8].
- **Massive Machine-Type Communications (mMTC)**: Facilitates connectivity for up to 1 million devices per km^2, supporting IoT applications and smart cities.

5G employs advanced technologies such as beamforming and massive Multiple Input Multiple Output(MIMO) to meet these diverse requirements. It operates across two frequency ranges: sub-6 GHz Frequency Range 1(FR1) and millimeter wave Frequency Range 2(FR2), offering both broad coverage and enhanced capacity [9].

This evolution reflects the industry's response to exponential growth in data traffic and the emergence of new applications requiring high bandwidth and low latency. Each generation has brought significant improvements in spectrum efficiency, network capacity, and user experience,

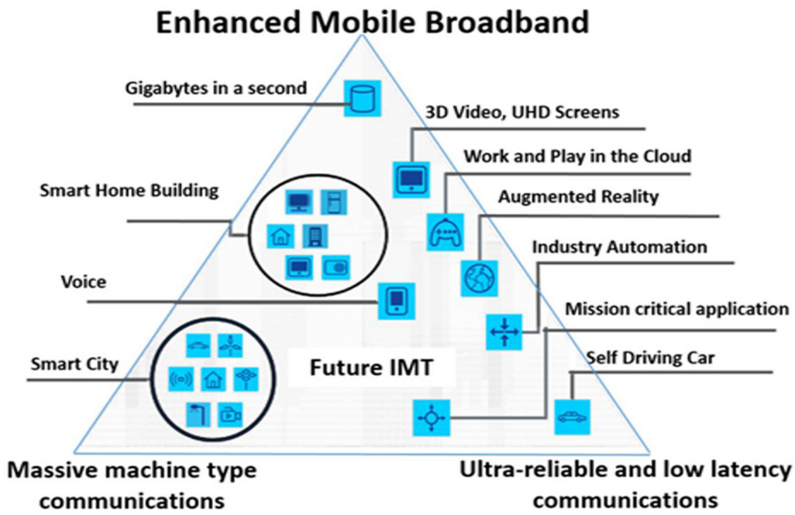


Figure 1.1: Applications relation to the three 5G categories.

setting the stage for increasingly sophisticated capacity management techniques. Understanding this progression is crucial for developing effective forecasting and congestion detection strategies in modern cellular networks, particularly as operators like KPN transition from 4G to 5G technologies.

1.2 Radio Access Network (RAN) Architecture

The Radio Access Network (RAN) is a foundational component of cellular networks that facilitates communication between user equipment (UE) and the core network. The RAN connects UEs to base stations (BSs), which in turn interface with the core network via backhaul links. These components provide services such as voice communication, video calls, and internet access, as shown in Figure 1.2 [10].

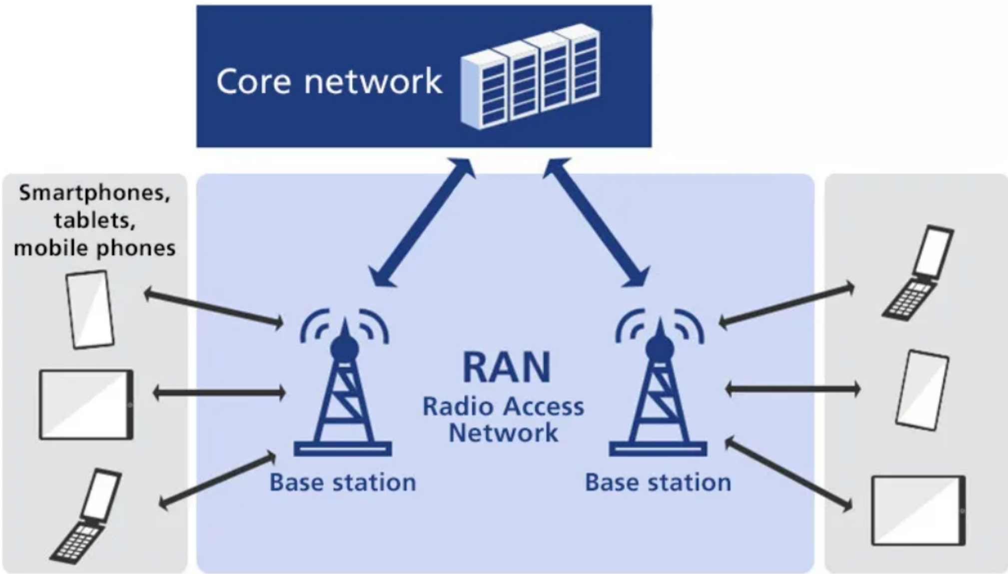


Figure 1.2: Simplified diagram of a cellular network.

The architecture of the RAN has evolved significantly from 4G to 5G to meet increasing demands for faster data rates, lower latency, and higher throughput [11]. In 4G networks, the RAN consists primarily of eNodeBs (eNBs), which handle radio communications, including signaling, control, and data transfer. A distinctive feature of the 4G RAN is its flat architecture, which reduces latency and enhances data transmission efficiency compared to previous generations [12]. Key technologies in 4G RAN include Orthogonal Frequency Division Multiple Access (OFDMA) for downlink communication, Single Carrier Frequency Division Multiple Access (SC-FDMA) for uplink, and Multiple Input Multiple Output (MIMO) technology, which increases communication capacity. Additionally, Carrier Aggregation enables operators to combine multiple frequency bands, improving overall data throughput [13]. Supporting these features is the Evolved Packet Core (EPC), which includes network components like the Mobility Management Entity (MME), Serving Gateway (SGW), and Packet Data Network Gateway (PGW) [14].

The transition to 5G introduces changes to the RAN architecture. In 5G networks, the eNodeB is replaced by the gNodeB (gNB), which serves as the central node for managing radio communications in the 5G New Radio (NR) system. The architecture is further segmented into Distributed Units (DUs), responsible for real time functions such as scheduling and transmission, and Centralized Units (CUs), which handle non real time functions like resource management [15]. This separation of functions improves scalability and performance in 5G networks. 5G RAN incorporates advanced technologies to meet the growing demand for network capacity and performance. Massive MIMO uses a large array of antennas to significantly increase capacity and coverage, while Beamforming improves signal quality by directing the signal toward specific users, reducing interference [16, 17]. Another key innovation is Network Slicing, which allows the physical network to be divided into multiple virtual networks tailored to specific use cases, such as high speed broadband or low latency industrial applications [18].

5G RAN can be deployed in two primary configurations: Non-Standalone (NSA) and Standalone (SA). NSA leverages the existing 4G LTE infrastructure for control functions while using 5G NR for higher data rates, allowing for faster 5G deployment. In contrast, the SA configuration uses a dedicated 5G core network, delivering lower latency and better overall performance [19]. Both configurations offer flexibility in deployment and support a variety of use cases, from wide area coverage with macro cells to dense urban environments served by small cells.

The evolution from 4G to 5G RAN architecture reflects the industry's focus on improving network flexibility, scalability, and performance to meet the growing demands of modern cellular communications. This progression sets the foundation for advanced capacity management techniques, which are crucial for optimizing network resources in increasingly complex and demanding cellular environments.

1.3 Data Growth and Network Challenges

The telecommunications industry is experiencing unprecedented growth and challenges, as highlighted by the Ericsson Mobility Report (November 2023) [20]. By the end of 2023, global 5G subscribers reached 1.6 billion, with projections estimating a rise to 5.3 billion by 2029. This substantial increase in mobile subscribers, coupled with the development of new applications, has significantly expanded mobile connectivity.

The surge in mobile data usage is equally striking. Individual consumption is expected to grow from 15 giga byte (GB) per month in 2022 to 75 GB per month by 2030, representing an annual growth rate of 25% [21]. This exponential increase places immense pressure on the telecommunications industry to scale its infrastructure and services rapidly. Over the past two years alone, the volume of mobile traffic has nearly doubled, with video content now constituting over 60% of the total traffic [20].

Figure 1.3 illustrates a significant shift in mobile network data traffic trends. While 2G / 3G / 4G mobile traffic is expected to peak around 2026 and then decline, 5G mobile and Fixed Wireless Access (FWA) traffic are projected to continue their upward trajectory. This trend aligns with the advancement of high resolution video applications, including virtual reality (VR) and augmented reality (AR) [20].

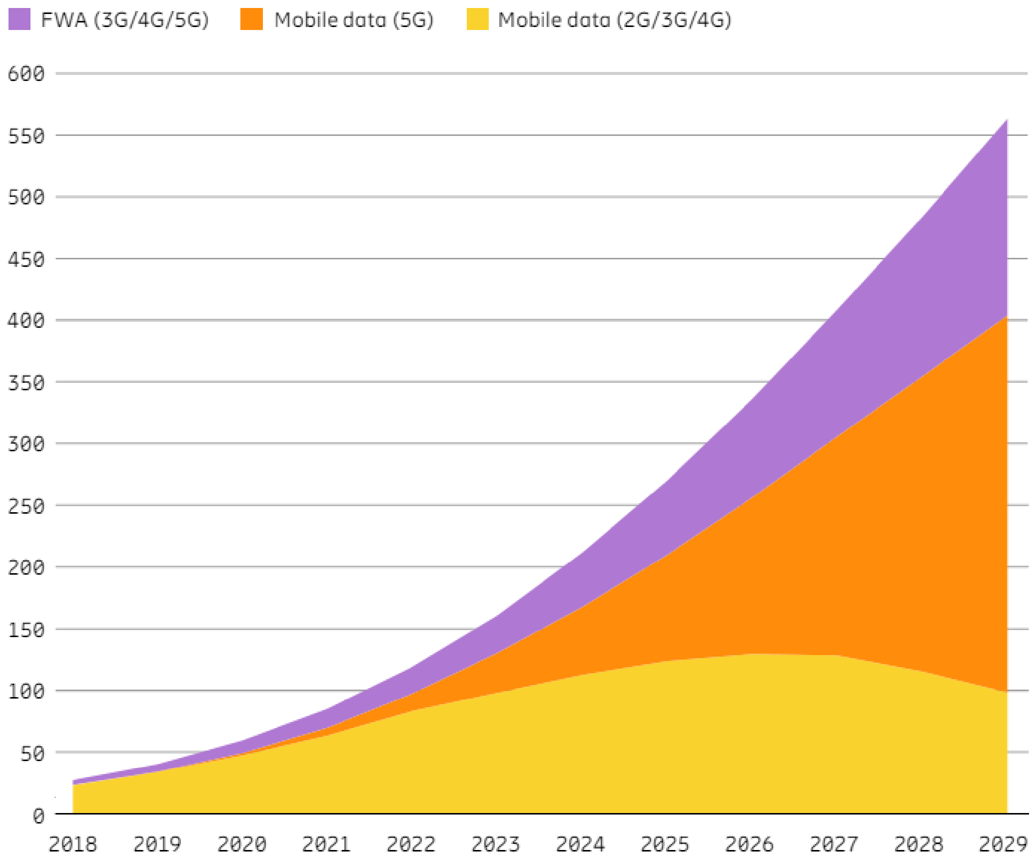


Figure 1.3: Global mobile network data traffic (exabytes per month)

To manage this increasing data traffic, Mobile Network Operators (MNOs) are compelled to adopt new solutions, techniques, and technologies. These efforts aim to bridge the gap between the growing demand for cellular network data services and the available network capacity. Strategies include acquiring additional spectrum or repurposing existing bands from older technologies. However, spectrum allocation remains a significant challenge, with frequencies being divided among MNOs and businesses through costly auctions. For instance, in 2020, the Dutch government raised 1.23 billion Euros from the auction of the 700, 1400, and 2100 MHz bands [22].

The scarcity and high costs of spectrum acquisition underscore the critical importance of effective capacity management. Capacity management, is the strategic process of optimizing network assets to meet current and projected user demands while maintaining cost effectiveness and service quality. It encompasses accurate traffic measurement, timely expansion, precise demand forecasting, and efficient congestion detection [23]. Through implementation, mobile operators can ensure sufficient capacity to handle user traffic, both presently and in the future, without overprovisioning or underutilizing their assets. Capacity planning, a key aspect of this management approach, has become essential for maintaining operational efficiency in the rapidly evolving mobile telecommunications landscape. It involves analyzing usage patterns, predicting future needs, and strategically planning expansions or upgrades. This proactive strategy prepares infrastructure for upcoming demands, prevents disruptions, and sustains service quality.

Effective capacity management maximizes spectrum utilization, ensuring this scarce resource is used to its full potential. Additionally, it reduces unnecessary expansion costs by pinpointing when and where upgrades are needed. Moreover, it enables networks to meet growing data traffic requirements while preserving the financial viability of mobile operators.

1.4 The KPN Network

1.4.1 Overview of KPN Network

KPN, a major telecom provider in the Netherlands, operates approximately 5,500 sites nationwide, ensuring near complete coverage (Figure 1.4) [24]. Each site typically comprises three sectors, though this can vary from two to nine depending on area requirements.

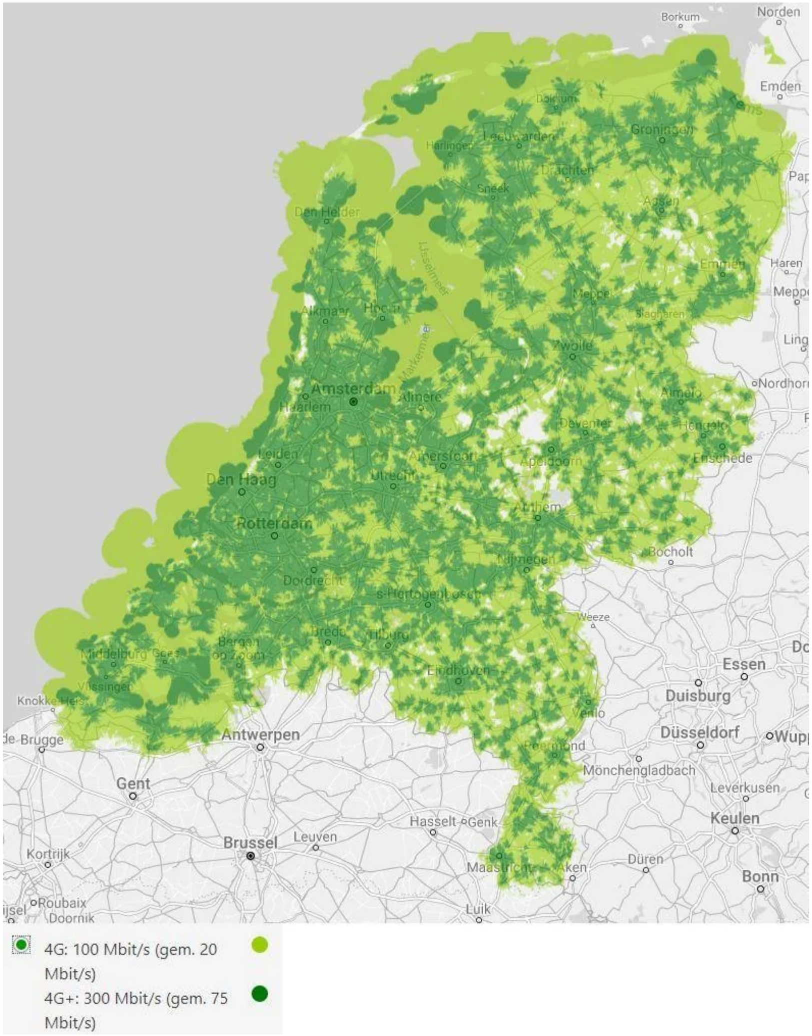


Figure 1.4: KPN 4G coverage across the Netherlands [24].

KPN's network serves over 11 million mobile subscribers and provides broadband to more than 4 million customers [25]. The company's commitment to network quality has been recognized by Umlaut, achieving the highest score in independent tests for three consecutive times [26]. These tests evaluate network quality, user experience, and performance of services like YouTube and mobile gaming, underscoring KPN's position as a leader in the Dutch telecom market.

KPN's spectrum utilization spans low and mid band frequencies (Figure 1.5) [27]. The network operates across 700, 800, 900 MHz (low-bands) and 1400, 1800, 2100, and 2600 MHz for Frequency

Division Duplex (FDD), as well as 2600 MHz for Time Division Duplex (TDD) (mid-bands). A sector refers to a specific geographic area that is served by an individual antenna within a cellular site. Each sector uses a combination of these bands, allowing for flexible network design. The 1400 MHz band is dedicated solely to downlink traffic, enhancing downlink capacity. For optimal network performance, KPN requires licenses across low bands (for basic coverage), mid bands (for higher speeds and capacity), and high bands (for extreme bandwidth at short ranges) [27]. Currently, KPN lacks high-band spectrum, as the Dutch government has not yet auctioned these frequencies (e.g., 26 GHz) for use.

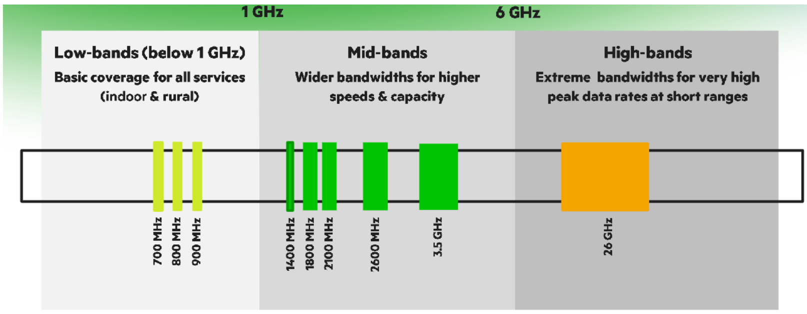


Figure 1.5: Spectrum band categorization [27].

KPN's spectrum strategy has evolved to accommodate growing demand and technological advancements. The company launched LTE services in February 2013, marking a significant improvement in spectrum efficiency and introducing a new radio interface[28]. More recently, KPN has been at the forefront of 5G deployment. The initial 5G rollout in July 2020 utilized the 700 MHz band, which continues to serve as a cornerstone of KPN's 5G network [29].

A pivotal development in KPN's network evolution occurred on May 21, 2024, when the company repurposed the 2100 MHz band from LTE to 5G New Radio (NR). This transition, carefully executed over two nights during off-peak hours, represents a major step forward in expanding KPN's 5G capacity. The repurposing of the 2100 MHz band highlights KPN's strategy to efficiently manage its spectrum resources to support the increasing number of 5G-capable devices on its network. Currently, the 700 MHz and 2100 MHz bands are fully dedicated to 5G services, while other bands, such as 800 MHz, 1800 MHz, and 2600 MHz, continue to support 4G LTE traffic. To further enhance its 5G network performance and manage the growing traffic from 5G-capable terminals, KPN plans to incorporate the 3.5 GHz band. The acquisition of this additional spectrum is critical for increasing network capacity and ensuring the delivery of high-speed 5G services as demand continues to grow. In this thesis, the impact of deploying the 3.5 GHz band on spectral efficiency will be studied, with a focus on how it can help KPN handle larger traffic volumes and maintain the high quality of service its users expect.

1.4.2 Capacity Management Process

Capacity management in cellular networks is critical to maintaining optimal network performance, especially with the rapid growth in data traffic. With the advent of 5G, MNO are facing increased pressure to efficiently manage limited resources while ensuring high quality services delivery. As data consumption continues to rise , particularly driven by services like video streaming, and virtual reality , proactive capacity management has become a key focus for operators. The deployment of 5G introduces new challenges due to higher data rates and the need for lower latency. Therefore, capacity management strategies must evolve to ensure that the network resources are allocated effectively, preventing congestion and enhancing user experience.

KPN's capacity management is a continuous, cyclical process comprising several critical stages. It's important to note that this process is iterative and ongoing. Each cycle begins with predicting future network traffic and ends with implementing necessary upgrades or optimizations. However, the end of one cycle immediately triggers the start of the next, ensuring constant adaptation to changing network conditions and user demands.

This process includes:

- **Capacity Planning:** Predicting future network traffic and capacity needs based on historical data and growth trends. This will be one of the most compelling aspects of this report, offering key insights for proactive network management.
- **Dimensioning:** Establishing guidelines for air interface capacity to support expected user numbers and service quality levels.
- **Product Configuration:** Selecting and configuring hardware and software components to meet capacity and performance requirements.
- **Capacity Monitoring:** Continuously tracking key performance indicators (KPIs) to assess network performance and identify potential issues.
- **Capacity Optimization:** Analyzing performance data to improve user experience and resource utilization.
- **Network Capacity Expansion:** Implementing physical or virtual expansions, such as adding new base stations, upgrading infrastructure, or deploying new technologies when optimization reaches its limits.

The monitoring stage involves about 450 counters in the Radio Access Network (RAN), longside numerous hardware resource counters.

While the current (Excel-based) traffic prediction model of KPN provides valuable insights into overall network traffic trends, it serves primarily as a contextual input for the present research. The projection growth rate per month derived from this model will be utilized in this study.

1.5 Problem definition

KPN faces critical challenges in managing its network capacity to maintain high-quality service amidst growing demand and the transition to 5G technology. The core problem centers around two primary uncertainties: variability in traffic demand and differences in sector capacity across the network. Each network sector has distinct spectral efficiency and capacity, influenced by factors such as geography, user distribution, and local infrastructure. This variability complicates the prediction of network congestion, particularly under high traffic loads. Furthermore, traffic demand fluctuates across sectors, and the network must support both 4G-only terminals and 5G-capable terminals, increasing the complexity of resource allocation. While 5G users should benefit from reallocated 5G spectrum, the system must continue to support 4G users effectively.

Current congestion prediction methods, including KPN's existing Excel-based model, rely on multiple variables and average sector data, which are insufficient for long-term, sector-specific forecasting. The current approach does not fully account for sector-level variations in spectral efficiency or the distinct traffic patterns of 4G and 5G users, leading to inaccurate congestion predictions. Additionally, this model is limited in its use of historical data patterns, which are essential for accurately predicting long-term trends. These limitations hinder KPN's ability to proactively manage capacity and optimize resource allocation, particularly for long-term predictions and in handling large datasets. As KPN continues transitioning to 5G and faces increased data demands, existing capacity management strategies must adapt to the evolving network landscape. Expanding capacity through additional sites or spectrum allocation involves long lead times and high costs, making precise forecasting essential to avoid both overinvestment and service degradation.

However, the specific aim is to predict peak hours in each sector to detect congestion, a granular analysis that this general traffic forecast model alone cannot provide. The approach focuses on sector level dynamics during high traffic periods, complementing the broader network wide projections presented here. This more detailed analysis is necessary for identifying potential congestion issues at the sector level, which is crucial for effective capacity management.

The rapid evolution towards 5G exacerbates these challenges, requiring more management of limited network resources to meet growing demands for higher data rates and enhanced quality of service. The key challenge is to develop a predictive capacity management system that accounts for sector-specific characteristics and variability in capacity, forecasts traffic patterns, and predicts spectral efficiency under varying load conditions for both 4G and 5G users.

1.6 Research goals

To address these challenges, this research aims to assist KPN in developing an 5G radio capacity planning system with the following objectives:

1. Develop a model to forecast data volume across KPN sites for one year in advance, incorporating:
    - Pattern recognition based on historical data
    - Projected growth traffic
    - Sector-level analysis to capture localized trends
2. Create a machine learning regression model to identify and analyze congested sectors by:
    - Determining spectral efficiency for each sector individually under varying load conditions.
    - Detecting congestion based on this metric
    - Analyzing traffic distribution across all available cells within congested sectors
3. Proactively detect congestion in sectors to prevent capacity shortages before they impact service quality.
4. Ensure the delivery of the minimum required user throughput, as defined by KPN based on customer needs.
5. Develop capabilities to assess sector-specific capacity limits, evaluating whether congested sectors can accommodate additional traffic or have available capacity. Furthermore, predict how sector capacities would evolve with the implementation of future spectrum resources, determining if congestion would persist or be alleviated.

By addressing these objectives, this research aims to provide KPN with a practical tool for proactive capacity management. This approach will support both strategic network planning and tactical decision making for optimization and expansion, enhancing network efficiency, improving user experience, and optimizing resource utilization as KPN transitions to 5G technology.

1.7 Research questions

This research aims to address the challenges KPN faces in capacity management and service quality assurance as it transitions to 5G technology. The primary research question is:

How can sector capacity and spectral efficiency be predicted using historical traffic data to forecast congestion under high traffic loads in KPN's 4G/5G network up to

one year in advance?

To help answer this main research question, the following sub questions are proposed:

1. How can historical traffic data and projected growth patterns be leveraged to accurately forecast sector-specific traffic volumes for both 4G and 5G users over a one-year period?
2. How can machine learning models be used to predict sector-specific spectral efficiency under varying load conditions, and how can these predictions be applied to forecast congestion?
3. What proactive strategies can be developed to detect potential congestion points before they occur, considering the variability in sector capacities and traffic demand?

1.8 Thesis synopsis

This section describes the structure of this thesis. Chapter 2 describes the literature survey, covering RAN architecture, capacity management, and congestion management in cellular networks. Chapter 3 focuses on data analysis, detailing selection, preprocessing, and network traffic patterns, including the 4G to 5G transition. Chapter 4 outlines the methodology, explaining traffic volume forecasting models and the congestion forecasting approach using machine learning techniques. Chapter 5 presents the results, discussing traffic volume prediction, congestion forecasting, and analysis of congested sectors. Chapter 6 concludes the thesis, summarizing key findings and offering recommendations for future work in cellular network capacity management.

2

Literature Review

This chapter delves into capacity management in cellular networks, focusing on enhancing user satisfaction and preventing network issues. It examines the architecture of Radio Access Networks (RAN), explores various strategies for managing network capacity, discusses techniques for detecting congestion, and evaluates the use of predictive modeling to improve network performance.

2.1 Capacity management strategies in cellular networks

Effective capacity management in cellular networks is crucial to maintaining optimal network performance and ensuring high quality service delivery, especially as the demand for data and new services continues to grow. The primary goal of capacity management is to balance the use of network resources, such as spectrum and infrastructure, to prevent congestion, enhance user satisfaction, and optimize overall network performance. This process must address various challenges, including fluctuating user demand, limited spectrum availability, and diverse traffic patterns.

One of the core elements of capacity planning is resource allocation, which involves efficiently distributing Physical Resource Blocks (PRBs) across users and services. PRB's are the fundamental units of radio resources in LTE and 5G NR (New Radio) networks, and their effective allocation directly impacts user throughput and network efficiency [30].

In LTE networks, each PRB spans 180 kHz and 0.5 milliseconds, consisting of 12 subcarriers, each with a bandwidth of 15 kHz, and 7 OFDM symbols within a time slot, as shown in Figure 2.1. A PRB can thus be visualized as a grid, with subcarriers on one axis (frequency domain) and OFDM symbols on the other (time domain), making up a total of 84 resource elements (12 subcarriers x 7 symbols) in a time slot. These resource elements are the smallest units for data transmission, with PRBs being the blocks that the scheduler allocates to users for downlink or uplink transmission. As networks transition to 5G, PRB structures remain similar, but 5G introduces additional flexibility with various subcarrier spacings to better accommodate different deployment scenarios. This flexibility enables operators to tailor resource allocation according to specific service needs, maximizing spectrum efficiency and optimizing performance for both low-latency and high-capacity applications.

As illustrated in Table 2.1 the number of available PRBs in a 4G cell is directly related to carrier bandwidth. As bandwidth increases from 1.4 MHz to 20 MHz, the number of PRBs scales from 6 to 100 [4], enabling efficient capacity management across different deployment scenarios.

Table 2.1: 4G Carrier bandwidth and corresponding number of PRBs

Carrier Bandwidth (MHz)	**Number of PRBs**
1.4	6
3	15
5	25
10	50
15	75
20	100

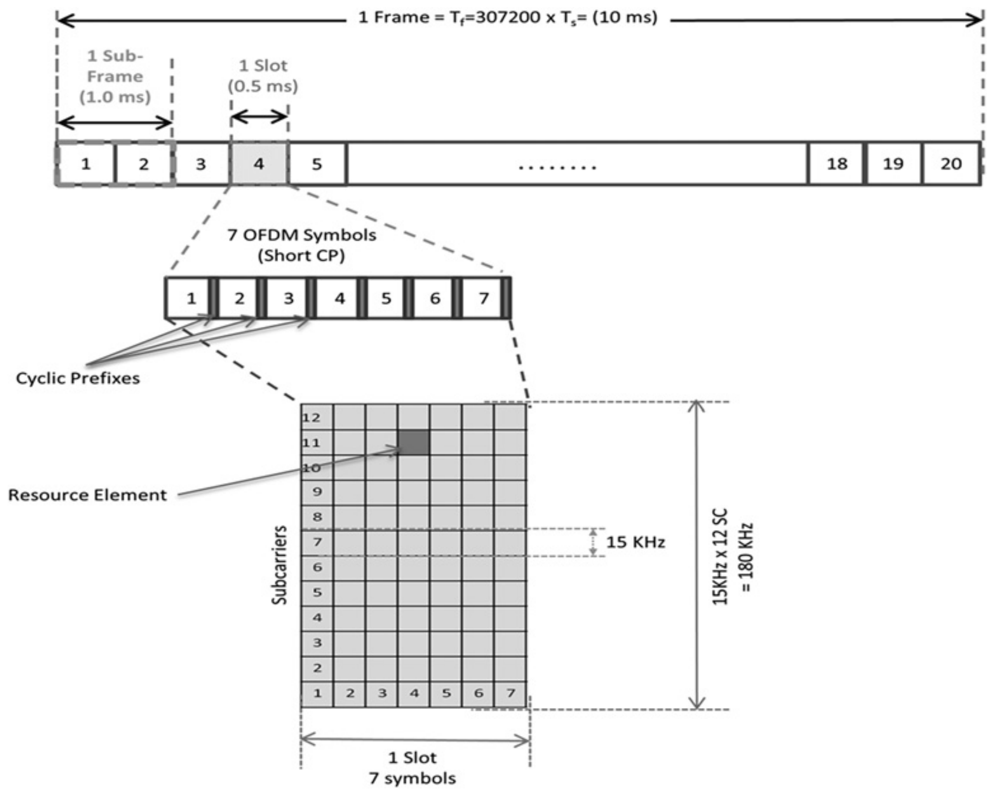


Figure 2.1: Radio resource structure in FDD mode: 4G LTE and one configuration option for 5G NR

5G's flexible numerology system, which allows for multiple subcarrier spacings (15 kHz, 30 kHz, 60 kHz, 120 kHz, and 240 kHz), further enhances spectrum efficiency and resource allocation capabilities. Table 2.2 illustrates how 5G handles various carrier bandwidths and subcarrier spacings, resulting in different numbers of PRBs [31]. For instance, a 20 MHz carrier with 15 kHz subcarrier spacing in 5G supports 106 PRBs, compared to 100 PRBs in LTE, due to reduced guard bands. This range of subcarrier spacings allows network operators to optimize their resource allocation based on specific deployment needs, whether for wider coverage or for high-capacity applications in smaller cells.

Table 2.2: 5G Maximum Carrier Bandwidth and PRBs for Different Subcarrier Spacings.

Subcarrier Spacing (Δf)	**Maximum Carrier BW**	**Maximum #PRBs**
15 kHz	20 MHz	106
15 kHz	50 MHz	270
30 kHz	100 MHz	273
60 kHz	100 MHz (FR1)	135
60 kHz	200 MHz (FR2)	264
120 kHz	400 MHz	264

Traffic forecasting plays a vital role in capacity management by predicting future network load and guiding infrastructure expansion decisions. Accurate traffic forecasting helps network operators plan for peak demand periods and preemptively address congestion, preventing network degradation. By analyzing historical traffic data, machine learning algorithms and predictive models can forecast the demand for data services, ensuring that the network is adequately prepared to handle increasing traffic loads while maintaining service quality [32].

Another critical aspect of capacity management is performance monitoring, which involves the continuous evaluation of Key Performance Indicators (KPIs). These KPIs include metrics such as PRB utilization, throughput, latency, and spectral efficiency. Monitoring these indicators allows operators to identify potential bottlenecks and take corrective measures before they impact user experience. For instance, as PRB utilization approaches a certain threshold, user throughput typically decreases, indicating that the network is nearing congestion. Spectral efficiency, defined as the ratio of data throughput to bandwidth, is another key metric used to assess how effectively the available spectrum is being utilized.

Spectrum utilization techniques, such as carrier aggregation, are essential for optimizing capacity management in cellular networks. Carrier aggregation allows network operators to combine multiple frequency bands, or component carriers, to create a larger, aggregated bandwidth for data transmission. In LTE, this aggregation can include up to five component carriers, each with a bandwidth of up to 20 MHz, resulting in a total bandwidth of up to 100 MHz, as shown in Figure 2.2 [33]. However, In 5G NR, operators can aggregate even more component carriers, up to 16 in some cases, which can combine both sub-6 GHz and mmWave frequency bands, allowing for up to 1 GHz of aggregated bandwidth. This increased bandwidth enhances both user throughput and network capacity, enabling faster data speeds and more efficient spectrum use. Carrier aggregation can be applied in both the downlink and uplink, allowing for a better user experience, higher peak data rates, and lower latency.

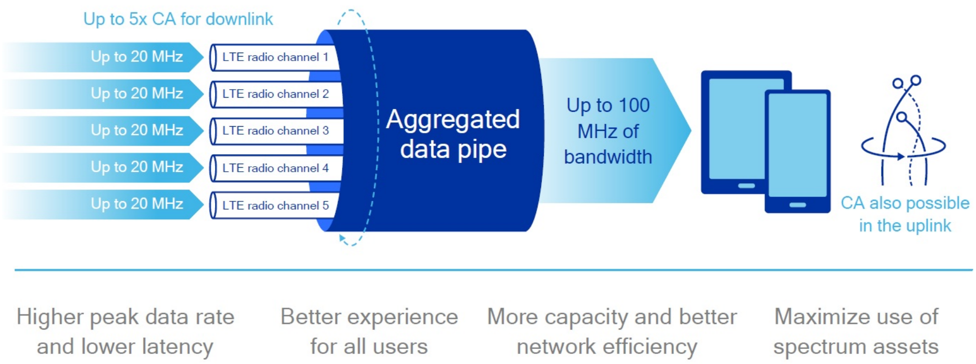


Figure 2.2: Carrier Aggregation in LTE [33]

In practice, mobile network operators, such as KPN, implement a range of strategies to manage capacity efficiently, with a key focus on selecting appropriate frequency bands for 4G and 5G to optimize data transmission and reduce latency. Initially, as part of its 5G deployment, KPN repurposed the 700 MHz band for 5G, leveraging its wide coverage capabilities to establish broad area 5G service. However, as 5G traffic and capacity demands grew, KPN strategically reallocated the 2100 MHz band from 4G to 5G, thus increasing available bandwidth for high speed data services in areas with significant user demand. This band reallocation approach underscores the importance of careful planning in capacity management, where decisions about which bands to use for 4G versus 5G are critical. Lower bands like 700 MHz are ideal for wide coverage, while mid-range bands, such as 2100 MHz, are essential for balancing coverage and capacity, especially in high-density urban environments. To further enhance 5G capacity in these high-demand areas, KPN is planning to acquire spectrum in the 3.5 GHz band. This high-frequency band will support significantly greater data rates and capacity, making it well-suited for densely populated urban centers where network demand is highest.

The selection and deployment of 4G and 5G bands is time-intensive, with new 4G or 5G layers typically taking six months to implement and new base stations up to two years. This means operators must plan proactively to ensure that infrastructure is in place to support expected future demands, which requires a forecasting model that can accurately predict network congestion based on PRB utilization, throughput, and spectral efficiency in both 4G and 5G settings [34].

In conclusion, capacity management in cellular networks involves a combination of capacity planning, dimensioning, product configuration, capacity monitoring, capacity optimization, and network capacity expansion. These strategies work together to ensure that the network operates efficiently, balancing growing demand with available resources. Key metrics like PRB utilization, throughput, and spectral efficiency provide insight into network performance and help operators optimize service delivery while planning for future growth.

2.2 Predictive Modeling for Network Capacity and Congestion

Accurate traffic forecasting and capacity management in cellular networks heavily rely on predictive modeling techniques that utilize historical data and machine learning algorithms. These models are essential for anticipating network demand and optimizing resource allocation, ensuring the network is prepared to handle future traffic loads efficiently. With the evolution of networks from 4G to 5G, predictive modeling becomes increasingly vital due to the growing complexity of user behaviors, service requirements, and network technologies.

Time Series Analysis is a widely used approach in traffic forecasting. Time series models, such as Auto-Regressive Integrated Moving Average (ARIMA), are employed to identify patterns and trends in historical data to predict future values. By analyzing past traffic, ARIMA models forecast network load and guide capacity planning decisions. ARIMA's structure relies on three components: the Auto-Regressive (AR) part, which uses past values for predictions; the Integrated (I) component, which ensures stationarity by differencing data points; and the Moving Average (MA) component, which models the residual error from lagged observations [35] [36].

However, while ARIMA and other traditional statistical models offer a flexible framework for network traffic modeling, they often fall short in handling complex, non-linear relationships in data. Recent studies indicate varying effectiveness of ARIMA and machine learning models across different domains. For example, Kontopoulou et al. (2023) [37] demonstrated that ARIMA models perform well for linear pattern forecasting with smaller datasets, such as in short-term financial predictions and certain environmental and healthcare metrics. Conversely, machine learning models like Long Short-Term Memory (LSTM) networks and XGBoost excel in handling larger datasets and complex, non-linear patterns, making them particularly effective in areas such as network traffic and COVID-19 forecasting. This limitation has driven a growing interest in developing estimation models that leverage known traffic growth projections for adaptability and accuracy, especially in dynamic, heterogeneous environments like cellular networks. These models are further discussed in the Methodology chapter.

To ensure optimal Quality of Service (QoS), Mobile Network Operators (MNOs) continuously monitor their infrastructure through network counters at base stations. These counters provide raw data that is aggregated into specific Key Performance Indicators (KPIs), which summarize network performance. According to the European Telecommunications Standards Institute (ETSI), the critical KPI categories for LTE and NR include [38] [39]:

- **Accessibility**: Measures the ability of users to access network services, e.g., RRC Setup Success Rate.
- **Retainability**: Reflects the network's capacity to maintain ongoing sessions, e.g., Call Drop Rate.
- **Integrity**: Concerns the quality of service experienced by the user, focusing on performance aspects like latency.
- **Mobility**: Evaluates the network's effectiveness in maintaining service continuity as the user moves, e.g., Handover Success Rate.

- **Availability**: Indicates the proportion of time network elements are operational, e.g., eNodeB uptime.
- **Utilization**: Measures how efficiently the network's resources, such as Physical Resource Blocks (PRBs), are used.
- **Quality of Service**: Assesses how well the network supports various service requirements, e.g., QCI performance.

In [40], the authors focused on analyzing KPIs like Accessibility, Mobility, Retainability, and Traffic KPIs. They developed a traffic forecasting platform using big data analytics and machine learning algorithms to predict traffic patterns for GSM, 3G, and 4G cells. Although their approach is valuable for general traffic prediction, it does not directly address spectral efficiency or user-experienced throughput, critical factors in 4G and 5G networks. Similarly, authors in [41] focused on forecasting average downlink throughput using LTE probes and predicted congestion events up to 30 hours in advance. However, this study did not account for the long-term evolution of spectral efficiency over time. While Channel Quality Indicator (CQI) prediction has been explored, the relationship between CQI, spectral efficiency, and user throughput remains underexplored in dynamic environments with changing network loads [34].

Congestion in cellular networks occurs when resource demand exceeds availability, leading to degraded service quality. This significantly impacts average user throughput and overall network performance. As network demands increase, congestion detection becomes critical for maintaining performance and ensuring user satisfaction. Several factors influence average user throughput in cellular networks [23]:

1. Number of Scheduled PRBs.
2. Signal to Interference and Noise Ratio (SINR), reported as Channel Quality Indicator (CQI).
3. MIMO Usage Performance

These factors interact in complex ways, creating dynamic relationships between individual user throughput and overall cell traffic. In LTE, the eNodeB determines the Modulation and Coding Scheme (MCS) based on UE reports. The combination of MCS and scheduled PRBs defines the transport block size (TBS) for each Transmission Time Interval (TTI) [42]. For example, scheduling 100 PRBs with MCS 28 on a 20 MHz LTE cell can yield around 150 Mbit/s for MIMO 2x2 QAM64 users. However, the same 100 PRBs with MCS 0 and transmit diversity would result in only 2792 kbit/s [42]. In NR networks, this variability increases due to wider bandwidths and advanced modulation schemes. Allocating 273 PRBs on a 100 MHz NR cell with MCS index 28 can achieve 2 Gbps per MIMO stream with 256-QAM modulation [43].

Given this complexity, PRB Utilization defined as the proportion of PRBs in use relative to the total available PRBs, becomes a critical measure for congestion detection and performance assessment. Here, PRB utilization, denoted by ρ, is calculated at the sector level, representing the degree of resource consumption within a sector based on its specific user demands. This metric reflects both the number of PRBs in use and the impact of varying MCS levels on each PRB's effective throughput.

The relationship between PRB utilization and user throughput can be modeled using queuing theory, specifically the M/G/1 Processor Sharing (PS) approach, which expresses user throughput (T_{ue}) as:

$$T_{ue} = \frac{C(1-\rho)}{\rho} \ln\left(\frac{1}{1-\rho}\right) \tag{2.1}$$

Where C is the sector capacity, adjusted for MCS, and ρ is the sector-level PRB utilization. A simplified expression for scheduled throughput is:

$$T_{\text{sch}} = C(1 - \rho) \tag{2.2}$$

This model enables congestion detection by monitoring PRB utilization, providing a consistent reflection of sector capacity and usage dynamics under varying MCS and traffic conditions.

It's crucial to distinguish between user throughput (T_{ue}) and scheduled throughput (T_{sch}):

- User throughput: Actual data rate experienced by individual users.
- Scheduled throughput: Network-level metric representing overall cell throughput.

The scheduled throughput is typically lower due to resource sharing and network overhead. This relationship enables congestion detection and performance assessment by focusing on PRB utilization, offering a consistent measure of cell capacity and utilization.

Congestion Thresholds are typically determined based on specific commercial requirements, which can vary depending on the network configuration and operator policies [23]. For example:

- An LTE cell on the 800MHz band with 10MHz bandwidth might be deemed congested at 60% utilization (20 PRBs unused).
- An LTE cell on the 1800MHz band with 20MHz bandwidth might have a higher congestion threshold, perhaps 80% utilization (20 PRBs remaining available).

These thresholds help operators optimize resource management based on traffic load.

2.3 Predictive modeling for spectral effeciency

The relationship between PRB load and traffic enables the calculation of spectral efficiency, a critical factor in assessing how much traffic a cell can handle before reaching its PRB load threshold. Spectral efficiency, in turn, informs how much data can be transmitted over a given amount of spectrum under current network conditions. As traffic increases, predicting spectral efficiency becomes vital for forecasting network congestion and planning capacity expansions.

Given the complex interaction between spectral efficiency, PRB load, and congestion, machine learning models are increasingly being used to predict congestion and PRB utilization more accurately. These models can dynamically adjust predictions based on current traffic patterns and network performance, providing more effective tools for congestion prevention and capacity management. By forecasting PRB utilization and congestion, machine learning techniques enable operators to proactively address network issues and optimize resource allocation.

Recent studies have explored advanced techniques like Linear Regression (LR), XGBoost, Lasso Regression, Random Forest (RF), and CatBoost to predict network performance and manage capacity effectively. For example, the study by Tomic et al. (2022) demonstrated that XGBoost excelled in modeling spectral efficiency for network performance prediction, significantly outperforming linear models in handling large-scale data and incorporating complex features such as modulation schemes and PRB utilization from surrounding cells [34]. This study underscored the model's adaptability to dynamic load conditions, a factor that is highly relevant for cellular networks facing variable user demand.

Additionally, Chmieliauskas et al. (2019) used Facebook's fbProphet algorithm for LTE cell traffic forecasting, demonstrating its effectiveness in anticipating traffic growth and congestion across a limited number of LTE cells (100 sites). Their approach allowed for proactive planning by accurately predicting busy hour trends and congestion risks over a short-term period [23].

2.4 Research gaps

Despite numerous studies focusing on network traffic prediction, significant gaps remain in addressing the challenges of capacity planning, particularly in terms of predicting user experience. Most current research emphasizes traffic forecasting and network-centric Key Performance Indicators (KPIs) but often neglects the end-user experience, especially regarding data throughput. The majority of studies [[44] [45] [46] [47]] concentrate on general traffic prediction or short-term forecasting of network performance metrics. For example, authors in [40] developed a traffic forecasting platform using big data analytics and machine learning algorithms to predict traffic patterns in GSM, 3G, and 4G networks. While this approach is valuable for general traffic forecasting, it does not adequately address spectral efficiency or user-experienced throughput.

Similarly, [41] demonstrated the ability to predict cell congestion events up to 30 hours in advance, but this work was limited to short-term forecasting and did not consider the long-term evolution of spectral efficiency. A critical limitation in existing research is the absence of comprehensive models that integrate traffic prediction with spectral efficiency forecasting to estimate user-experienced throughput over longer periods. While studies such as [34] have focused on predicting Channel Quality Indicator (CQI), they have not fully explored how CQI translates to actual user throughput or how it evolves under changing network loads.

These research gaps have significant implications for capacity planning. Inefficient planning often stems from the lack of accurate long-term predictions of user-experienced throughput, leading network operators to mismanage resources. This mismanagement can result in inefficient use of capital and potential declines in service quality. The predominant focus on short-term predictions has led to reactive network management, which may result in suboptimal user experiences during peak usage periods. Furthermore, the absence of comprehensive models that consider both 4G and 5G technologies has hindered effective planning for the transition to 5G, potentially slowing adoption and optimization.

This research addresses these gaps. First, it proposes an integrated approach to spectral efficiency and resource block modeling, combining traffic prediction with network performance metrics to offer a more complete view of network capacity and user experience. Second, it introduces a methodology for predicting the long-term evolution of spectral efficiency under changing network loads, which is essential for strategic capacity planning. Finally, it addresses the challenges of predicting performance in hybrid 4G/5G environments, filling a gap in current literature that often focuses on single-generation networks without considering their coexistence.

3

Data Analysis and Network Traffic Characterization

This chapter provides an analysis of KPN's LTE and NR network data, forming the foundation for research on network capacity management. It covers the selection of datasets, including their granularity and scope, as well as data preprocessing techniques to ensure quality and consistency, such as handling missing values, normalization, and outlier detection. The analysis examines network wide traffic patterns, trends, and distributions, followed by a detailed sector level analysis, introducing concepts like spectral efficiency. Additionally, a correlation analysis explores the relationships between key performance metrics.

3.1 Data Selection

The time granularity of the dataset is crucial in traffic and spectral efficiency predictions, influencing two key factors. First, the length of the time series affects the forecast horizon, determining the prediction period. It is important that the number of observations exceeds the model parameters during training [48]. Second, data aggregation can reduce variations and limit the available data points for model training, impacting prediction performance. Therefore, selecting an appropriate granularity is essential for deriving accurate insights from the forecast results.

MNO employs an Operations Support System (OSS) for performance monitoring. This system utilizes network counters placed at base stations, which serve as network access points. These counters provide raw measurements that are aggregated to form specific KPIs. This research employs two primary datasets to analyze mobile network performance and traffic patterns:

- The dataset includes hourly data over 67 days, exceeding the typical one month OSS collection window to capture weekly and monthly variations, resulting in a more robust and representative dataset. It contains around 159 million records from 5,500 sites. This hourly granularity is essential for studying peak usage congestion and detailed sector level analysis [23].

- Additionally, The research uses a year long daily dataset, representing the maximum historical data available from the OSS system. This dataset is used to analyze long term trends, including the transition from 4G to 5G, traffic distribution patterns, and changes in downlink and uplink traffic.

This research is based on an analysis of KPN's LTE and NR network data, encompassing approximately 5,500 sites across the Netherlands. These sites cover diverse geographical areas with varying traffic patterns, including urban and suburban areas, as well as wide rural coverage zones with low traffic density.

Figure 3.1 illustrates the total data traffic volume over the year long study period. It shows an increasing trend in data volume traffic throughout the year, signifying growing demand for data services. Periodic fluctuations are observable, likely due to user activity patterns and network management activities. A notable dip in traffic during the last week of December 2023 and the first week of January 2024 is attributable to reduced data usage during the Christmas holiday and New Year in the Netherlands.

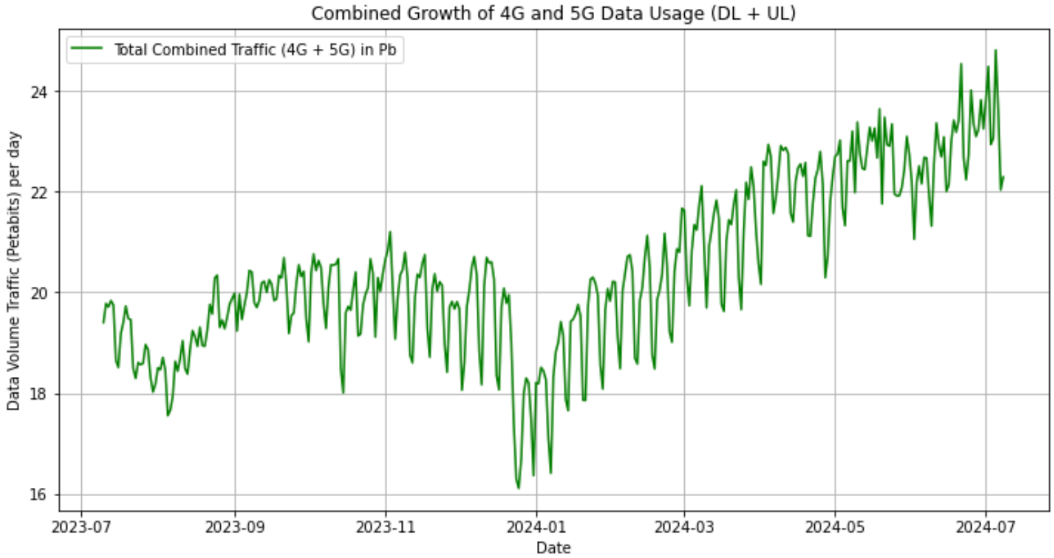


Figure 3.1: Growth Trend in KPN Data Traffic Volume (4G and 5G) combined

The primary input variables for the predictive models are shown in the appendix A in Table A.1.

3.2 Data Preprocessing

Data preprocessing is a crucial step in any data analysis or machine learning project, particularly in the context of network capacity management. Raw data from network systems often contains inconsistencies, missing values, and varying scales that can impact the accuracy and reliability of subsequent analyses. The preprocessing phase involved two main steps: handling missing values and data normalization. These steps are essential to create a clean, consistent dataset that could provide meaningful insights into network traffic patterns and facilitate accurate capacity management.

3.2.1 Handling Missing Values

Missing data in the network traffic records is addressed using linear interpolation. This method estimates unknown values within a set of known data points, providing a continuous and smooth estimation of the missing data. Linear interpolation is chosen over other methods, such as mean imputation or more complex algorithms, because it preserves the overall trend and pattern of the data, which is essential for capacity management. The interpolation formula used is given in Equation 3.1:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)} \tag{3.1}$$

where y is the interpolated value, x is the estimation point, and (x_1, y_1) and (x_2, y_2) are known data points.

3.2.2 Data normalization

Data normalization is a crucial preprocessing technique for the congestion prediction model, which utilizes LR, RF, Lasso, and CatBoost. This process adjusts the scale of features in the dataset, ensuring they fall within a specific range, typically $[0, 1]$. Normalization is particularly important in this context, where features like traffic volume, time of day, and bandwidth usage are measured on different scales. By applying Min-Max normalization, consistency in feature evaluation is achieved, preventing any single feature from disproportionately influencing the model and improving overall model performance.

The Min-Max normalization technique, chosen for its advantages in this scenario, scales the data to a fixed range using the following equation 3.2:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.2}$$

where x is the original value, $\min(x)$ and $\max(x)$ are the minimum and maximum values in the feature, and x' is the normalized value.

This normalization technique leads to more stable gradients during training. Figure 3.2 illustrates the result of applying Min-Max normalization to the traffic data from daily dataset , scaling values between [0,1] range. The plot demonstrates the preservation of the overall upward trend in traffic volume, clear visibility of daily and weekly fluctuations, and easy identification of anomalies.

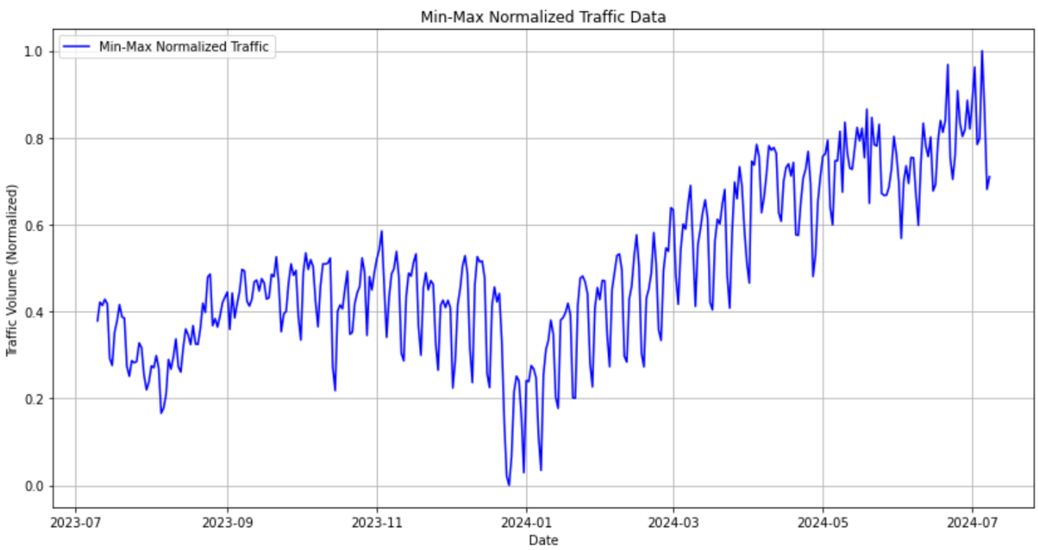


Figure 3.2: Daily traffic data after Min-Max normalization

3.2.3 Outlier Detection

In the analysis of network traffic data, particularly at the granular hourly level, the presence of outliers, the data points that deviate from the normal traffic patterns can negatively affect the accuracy of predictive models and skew the interpretation of network behavior. Outliers in this context may represent anomalous network events, measurement errors, or extreme but valid traffic spikes. Handling outliers is crucial because they can distort key calculations, such as spectral efficiency and congestion predictions, leading to inaccurate assessments of network performance. Several outlier detection methods are evaluated for this task:

Initially, the Interquartile Range (IQR) method is a statistical approach that focuses on the central 50% of the data distribution. It identifies outliers based on the spread between the first quartile (Q1) and the third quartile (Q3), known as the Interquartile Range. Data points are classified as outliers if they fall below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$. This method is effective for detecting extreme values in symmetrically distributed data, but it only detects outliers in a single dimension at a time. For example, when applied separately to PRB utilization and traffic data as illustrated in Figure 3.3c, the IQR method can overlook multidimensional anomalies where unusual combinations of PRB utilization and traffic might signal congestion events.

Secondly, the Z-Score method standardizes data points based on their deviation from the mean, measured in standard deviations, and is also limited to univariate outlier detection. Each data point is assigned a Z-score calculated as:

$$Z = \frac{x - \mu}{\sigma} \tag{3.3}$$

where x is the data point, μ is the mean, and σ is the standard deviation of the dataset. Typically, points exceeding a predetermined threshold (e.g., a Z-score of ± 3) are flagged as outliers. While this method is effective for detecting anomalies in normally distributed data, it does not account for interactions between multiple dimensions. In complex, real-world network traffic, single-dimensional outlier detection fails to capture unusual combinations of PRB utilization and traffic. This limitation is evident in Figure 3.3b where the Z-score method struggles to identify multidimensional anomalies.

Given the complexity of the dataset, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) emerges as the optimal choice for outlier detection. DBSCAN's density-based approach can identify isolated points from high-density regions regardless of their absolute values, making it particularly effective for network traffic analysis [49]. While traditional methods like IQR and Z-score examine variables independently, DBSCAN analyzes data in multidimensional space, enabling the detection of unusual combinations of PRB utilization and traffic metrics that may indicate network anomalies. This multidimensional analysis captures atypical patterns and relationships between variables that single-dimensional methods would miss, providing a more assessment of potential congestion.

Figure 3.3 illustrates a comparison of original data and outlier cleaning methods applied to PRB utilization and traffic volume for sector 75271-4 over an hourly data period from May 2 to July 8, 2024. During this time, the 2100 MHz band is reconfigured for 5G to accommodate increased traffic, but the 3.5 GHz spectrum had not yet been introduced (its deployment occurred at the end of week 28 in 2024). The reason for selecting this specific sector for analysis will be discussed in the upcoming section. In Figure 3.3a, the dataset contains outliers that may not be extreme in terms of absolute value but are spatially isolated from the denser data clusters. DBSCAN, a density-based algorithm, is chosen for its ability to identify such outliers by focusing on data density rather than just value extremity. This approach is particularly useful in identifying points with unusual combinations of PRB utilization and traffic volume, indicating atypical network conditions.

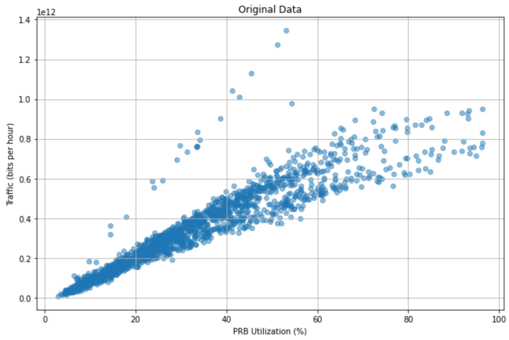
The DBSCAN algorithm relies on two primary parameters:

- ϵ (eps): This parameter sets the maximum distance between two samples for them to be considered in the same neighborhood. Through iterative tuning, an optimal ϵ value is determined by observing the formation of data clusters, ensuring that only closely packed points are grouped while isolated points remained detectable as potential outliers.

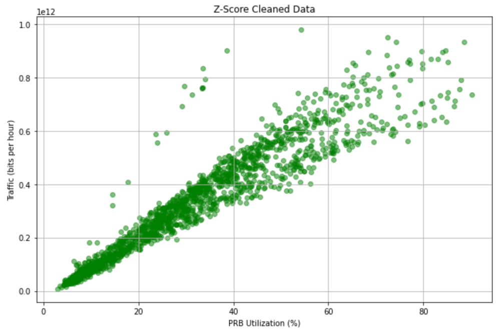
- MinPts: The minimum number of samples require to form a dense region. By setting MinPts based on the natural density of typical data points in the distribution, the model ensured that sparsely populated areas with low PRB utilization, where favorable propagation conditions are likely more effectively flagged as outliers.

After tuning these parameters, DBSCAN is applied to the hourly dataset, focusing on key metrics such as traffic volume and PRB utilization. The algorithm specifically flagged data points with unusually low PRB utilization relative to high traffic levels.

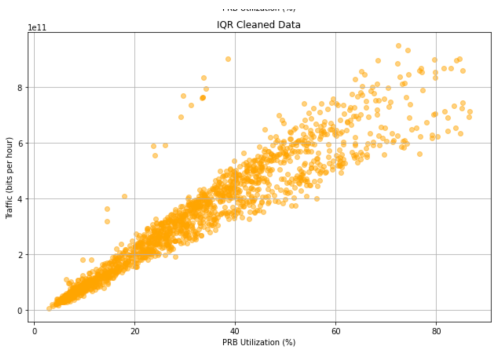
The outliers removed by DBSCAN were primarily those with unusually low PRB utilization relative to the observed traffic volume. These cases often reflect favorable propagation conditions or optimal user distribution within a sector, where high data throughput is achieved with minimal PRB usage. By excluding such points, the goal was to obtain conservative spectral efficiency estimates that would inform a cautious planning approach, as shown in Figure 3.3d. A key consideration was the potential removal of data points that exhibit high PRB utilization for the given traffic, as these could indicate challenging propagation conditions or other network limitations. To prevent the erroneous exclusion of such critical cases, the fine-tuning process included a threshold that preserves clusters near the higher bounds of PRB utilization for each traffic volume level. This approach ensures that high utilization outliers remain in the dataset, capturing instances where suboptimal spectral efficiency could signal capacity constraints that should be addressed in the planning process.



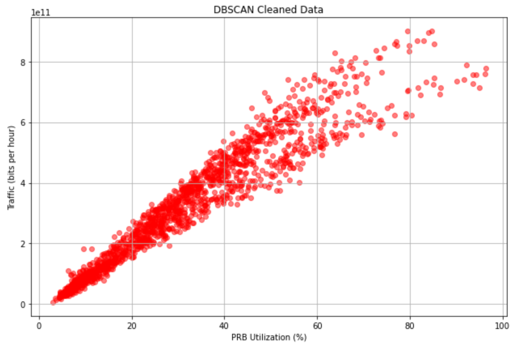
(a) Original Data



(b) Z-Score Cleaned Data



(c) IQR Cleaned Data



(d) DBSCAN Cleaned Data

Figure 3.3: Comparison of Original Data and Outlier Cleaning Methods on sector 75271-4

This cleaned dataset provides the foundation for subsequent analyses, including spectral efficiency calculations and the development of the congestion prediction model, which will be discussed in detail in the Methodology chapter. By applying these preprocessing techniques, a robust data pipeline is ensured.

To further analyze the development of spectrum efficiency over time and understand its dependency on daily patterns, Figure 3.4 presents the average traffic and spectral efficiency by hour of the day for sector 75271-4. Here, the traffic and spectral efficiency averages were calculated for each hour of the day. The figure shows that as traffic load decreases during early morning hours, spectral efficiency also drops to its lowest levels. As the day progresses and traffic load increases, spectral efficiency rises correspondingly, demonstrating an upward trend with higher traffic volumes, up to a certain threshold. During peak traffic hours, spectral efficiency tends to stabilize or slightly decrease, likely due to high PRB utilization approaching the sector's capacity limits. This pattern suggests that at higher traffic loads, the sector manages demand by efficiently utilizing available resources.

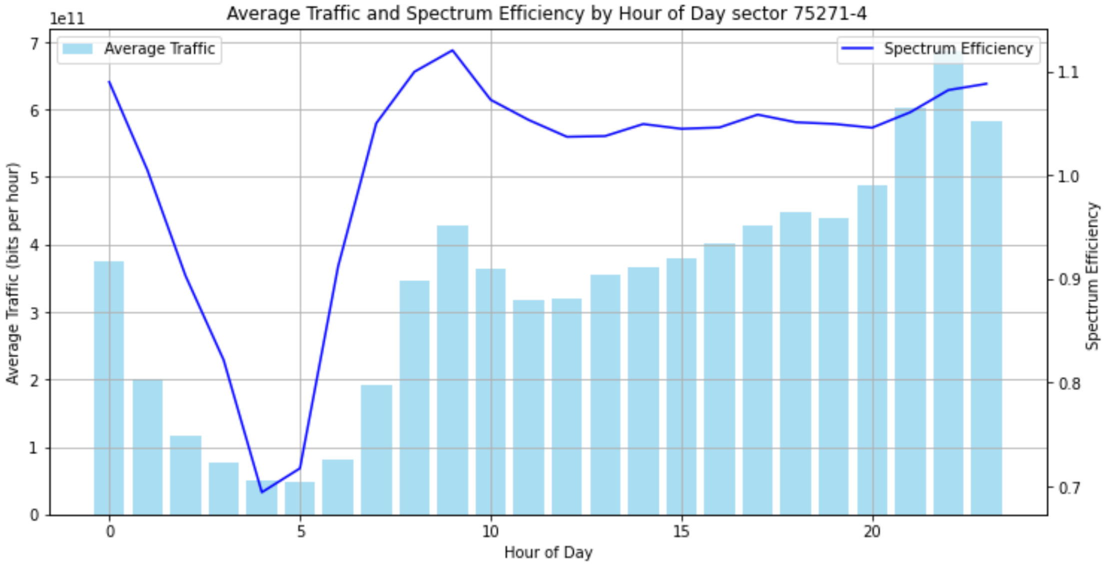


Figure 3.4: Hourly patterns of traffic and spectral efficiency for sector 75271-4.

3.3 Time series analysis

This research is predicting network traffic for the upcoming year. To achieve this, analyzing the daily dataset spanning one year of 4G and 5G traffic is crucial. This time series analysis aims to assess the data characteristics and extract insights critical for the prediction model described in Chapter 6. Time series data can be decomposed into three distinct components: trend, seasonality, and noise [50]. The trend represents long term changes (increasing, constant, or decreasing) over time. Seasonality captures recurrent patterns within specific periods (e.g., weekly or monthly), influenced by seasonal factors. The residual component, also known as noise, accounts for unexplained variability not attributed to trend or seasonality. Empirical analysis of the dataset reveals an increasing trend in traffic volume over time, with a consistent weekly pattern, likely influenced by seasonal changes in user behavior. To better understand these underlying patterns, time series decomposition is performed using Python [51], employing a multiplicative model [52]. This model combines various components to reconstruct the historical time series, as expressed in Equation 3.4:

$$Y_t = T_t \times S_t \times R_t \tag{3.4}$$

Where T_t represents the trend component, S_t the seasonal component, and R_t the residual component.

Given the daily data with weekly seasonality, a 7 day season length is used for decomposition. Figure 3.5 illustrates the results of this multiplicative decomposition for both 4G and 5G data time series. The trend component is defined by traffic volume in bits per day, while the seasonality exhibits a weekly pattern. The y-axes for seasonality and residual components represent factors to be multiplied by the trend to reconstruct the historical time series.

Figure 3.5: The multiplicative decomposition of the time series.

Figure 3.6 provides a detailed view of the trend and seasonality components. The observed increasing trend indicates consistent traffic growth, with a steep slope reflecting an annual increase of approximately 50%. This aligns with the expanding demand for network services. The weekly seasonality is evident, showing recurrent patterns influenced by seasonal factors. Notable features include an upward trend from the start of the calendar year, significant traffic drops during Christmas and New Year's, followed by a drastic increase. As traffic grows, the magnitude of the seasonal component increases proportionally.

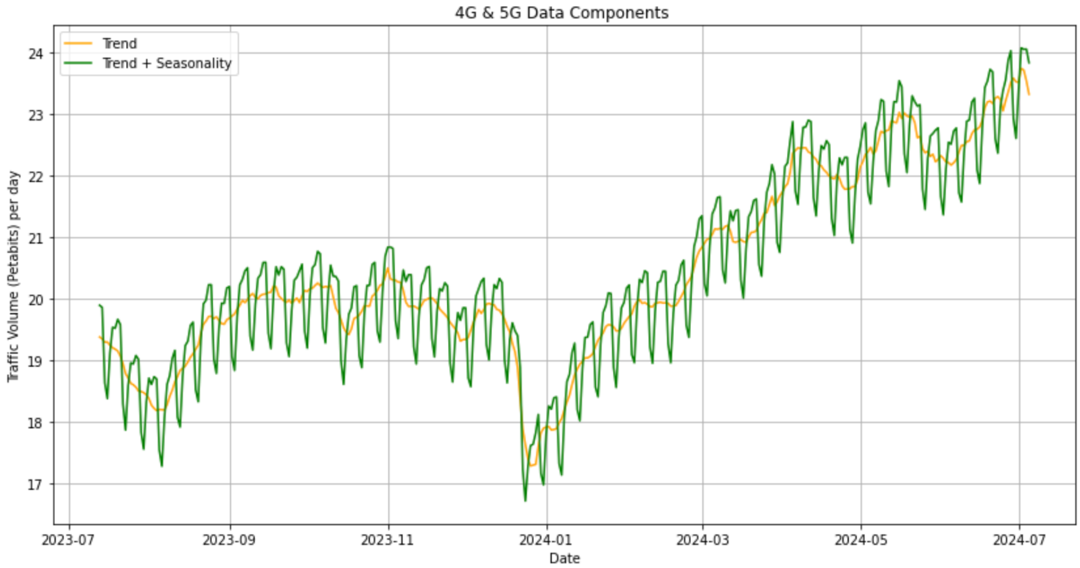


Figure 3.6: The trend and seasonality components of 4G & 5G data.

Further investigation is conducted to examine the traffic patterns in DL and UL directions. This analysis is crucial for understanding network usage and informing the focus of the congestion prediction model. Figure 3.7 illustrates the comparison between DL and UL traffic volumes over time.

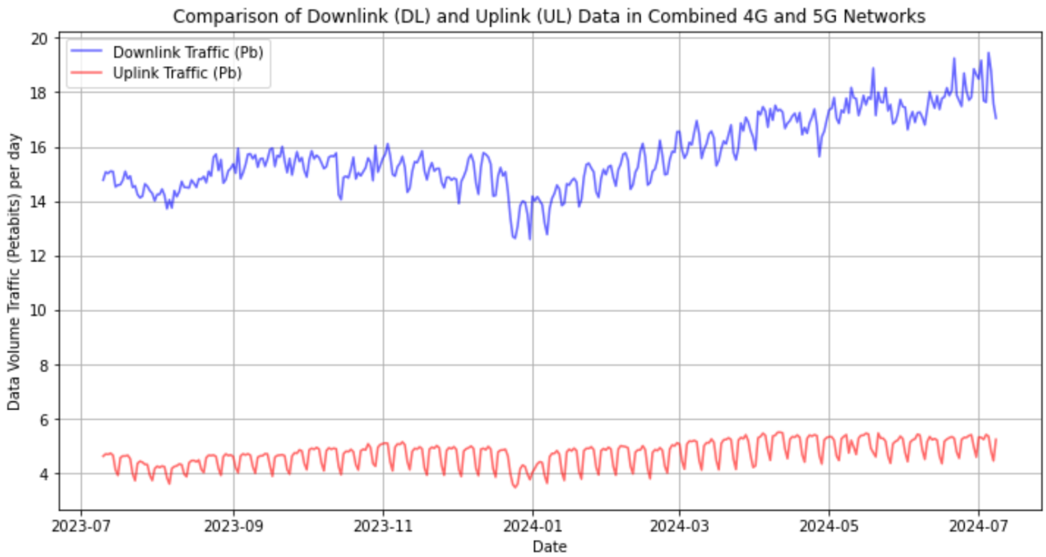


Figure 3.7: Comparison of Data Volumes in DL and Uplink UL.

As shown in the figure, DL traffic consistently exceeds UL traffic, with DL volumes approximately 3-4 times greater than UL volumes. The observed asymmetry between DL and UL traffic aligns with typical network usage patterns, driven by download-centric activities such as video streaming, browsing media-rich websites, and downloading large files. While uplink capacity remains a crucial consideration, many network technologies are optimized for greater DL capacity to handle the heavier load associated with these high-bandwidth activities.

Although DL traffic is significantly higher, this alone does not directly imply that DL is the sole cause of congestion, as congestion is also influenced by the capacity of each traffic direction. However, because DL traffic generally represents a larger portion of the total network load and is more likely to contribute to congestion under high usage conditions, subsequent analyses and model development will primarily focus on DL traffic patterns.

Figure 3.8 presents a heatmap of combined LTE and NR bands' traffic distribution across the entire network. This visualization reveals that the 1800 MHz and 2100 MHz bands consistently carry the highest percentage of total downlink traffic, with the 1800 MHz band handling 35-41% and the 2100 MHz band managing 25-29% before its repurposing. This concentration stems from the network's traffic steering algorithm, which preferentially directs traffic to these bands due to their higher capacity.

A notable transition occurs around May 21, 2024, when the 2100 MHz band is repurposed for 5G deployment. This shift redistributes traffic, with the 1800 MHz band increasing to over 41% and the 2100 MHz band, now operating as 5G, decreasing to about 17%. The decrease in 2100 MHz band traffic results from limited 5G technology support among client end-user equipment. Meanwhile, other bands such as 2600 MHz experience slight increases in traffic share.

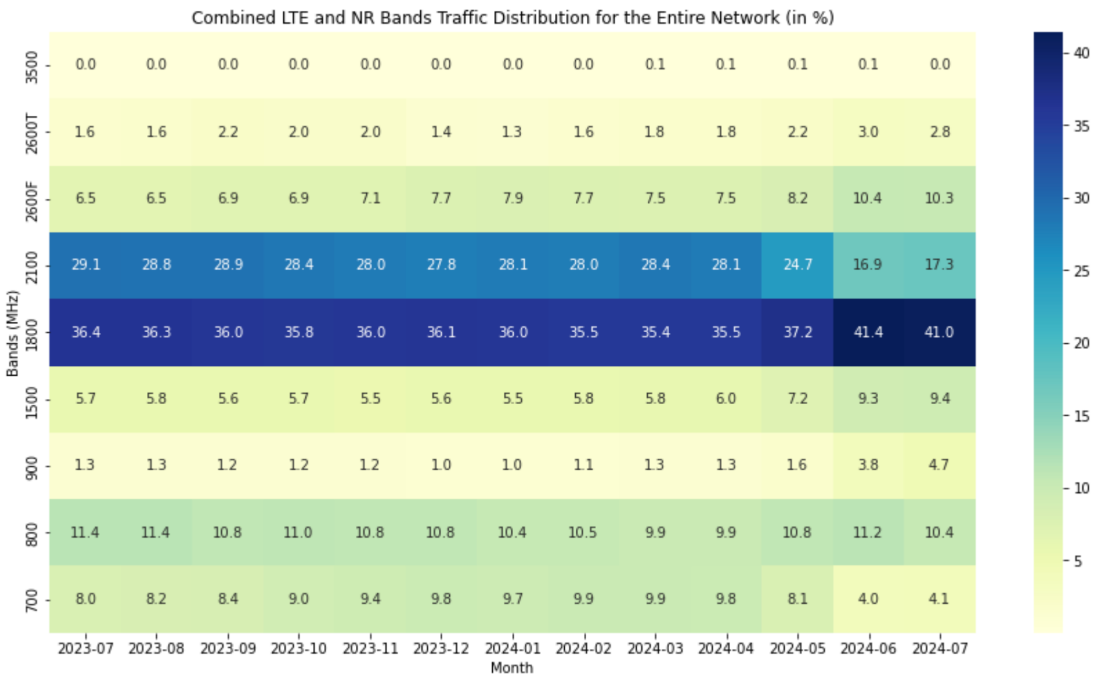


Figure 3.8: Traffic Distribution Across Different Frequency Bands

This traffic distribution analysis provides valuable insights into how different frequency bands handle traffic and where potential congestion may occur. Higher-frequency bands like 1800 MHz and 2100 MHz have more available capacity and are suitable for managing increased traffic volumes. However, they also tend to experience higher congestion, with higher data demand.

Moreover, by conducting a sector-level analysis that considers all frequency bands within a sector, the congestion prediction model aligns with real-world network operations, which utilize carrier aggregation a technique that combines multiple frequency bands to increase capacity and improve data rates—to manage traffic more efficiently.

Figure 3.9 illustrates the comparison between downlink traffic for permanent macro sites and the total downlink traffic, including indoor sites. The graph shows that macro sites handle the majority of network traffic, which aligns with KPN's infrastructure of approximately 5,000 macro sites compared to only 500 indoor sites. Although the difference between the two lines in the graph suggests that indoor sites contribute a smaller percentage of the overall network traffic. Both macro and indoor sites can experience congestion depending on their spectral efficiency and capacity, especially as more 5G devices enter the network. While this study focuses on macro sites due to their role as primary load bearers and their higher likelihood of experiencing congestion, the developed model is capable of analyzing traffic patterns and predicting congestion for indoor sites as well. However, the macro sites will remain the primary focus, given their larger contribution to overall network traffic and their critical role in network-wide congestion management.

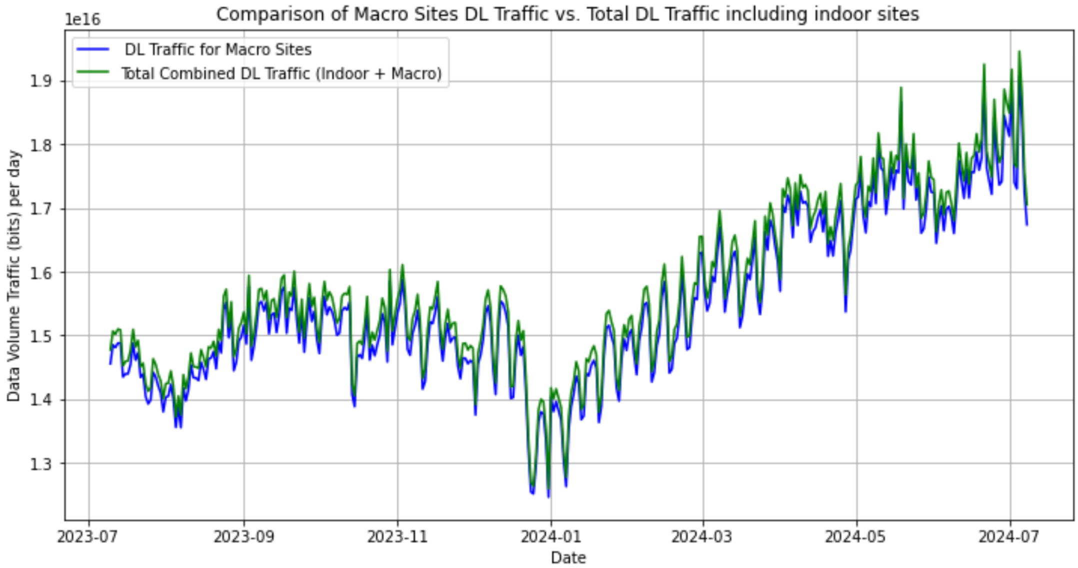


Figure 3.9: Comparison of Downlink traffic in Macro Sites and All sites.

Conclusion

Based on the above analysis of traffic patterns and network infrastructure, the congestion prediction model will focus on downlink traffic at the sector level for permanent macro sites. This approach aligns with the predominant traffic flow, accounts for dynamic frequency allocation within sectors, and targets the network elements handling the majority of data traffic. By concentrating on these key aspects, the model aims to effectively predict and address potential congestion issues in the most critical areas of the network.

3.4 Sector level Analysis

While a network wide analysis provides a broad view of traffic patterns and trends, a more granular approach is essential for understanding network performance and accurately identifying congestion points. Sector-level analysis delves deeper into individual sectors, assessing traffic patterns, resource utilization, and, critically, spectral efficiency to offer insights that would be obscured in an aggregated, network wide view. This focus on sectors rather than individual cells is particularly valuable due to the unique structure of cellular networks. Each sector comprises multiple cells operating on different frequency bands, where higher frequency cells provide higher capacity but more limited coverage than lower frequency cells. The goal of sector-level planning is to maintain a consistent Quality of Service (QoS) across the entire sector, even in areas with lower capacity.

This shift to sector-level analysis is also driven by variations in user density, geographical features, and local events that affect each sector differently. By examining each sector individually, network planners can optimize resource allocation, address localized congestion points, and improve spectral efficiency where it is most needed. This approach enables a tailored strategy for capacity management, ensuring that resources are allocated effectively to maintain network performance and meet user demands across all areas of the network.

For this analysis, The focus on Sector 75271-4, which consistently experiences high traffic volumes, aims to address its unique challenges in traffic and spectral efficiency.

Spectral efficiency, measured in bits per second per Hz, reflects how effectively each sector uses its allocated spectrum. Figure 3.10 compares the spectral efficiency of Sector 75271-4 and Sector 1001-1. The plot clearly shows that Sector 75271-4 (represented by green circles) exhibits a wider spread in both traffic volume and PRB usage, reflecting higher spectral efficiency compared to Sector 1001-1 (represented by black triangles). This supports the observation that Sector 75271-4 experiences higher traffic and utilizes its spectrum more efficiently, particularly during peak hours. Sector 75271-4, located in a high-demand recreational area, sees a high concentration of users engaged in data-intensive activities, such as media streaming and navigation. The combination of modern devices and an open environment further enhances signal quality and data rates, driving the sector's spectral efficiency. In contrast, Sector 1001-1, which experiences lower traffic volumes, has a more concentrated, narrower distribution of PRB usage and traffic, reflecting lower spectral efficiency.

Figure 3.10: Comparison of spectral efficiency between two network sectors.

These observations from the sector level analysis provide crucial insights that will guide the subsequent modeling efforts, particularly in developing predictive models for congestion and spectral efficiency.

3.5 Correlation Analysis

This section explores the relationships between key performance metrics at the sector level, focusing on PRB utilization, traffic volume, and average user throughput. Understanding these relationships is essential for predicting network performance and managing congestion effectively. Correlation measures the degree to which two variables are linearly related, and the strength of this relationship is quantified using the Pearson correlation coefficient (r), which ranges from -1 to 1 [53]. The Pearson correlation coefficient is calculated as:

$$r_{X,Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \cdot \sqrt{\sum(Y - \bar{Y})^2}} \tag{3.5}$$

Where X and Y represent the variables being analyzed, such as PRB utilization and traffic volume. $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y, respectively. The summation $\sum$ is calculated over all paired data points. The denominator consists of the product of the standard deviations of X and Y. A positive correlation indicates that as one variable increases, the other also rises, showing a direct relationship. Conversely, a negative correlation suggests that as one variable increases, the other decreases, highlighting an inverse relationship. A correlation close to zero implies no significant linear relationship between the variables.

The correlation analysis is performed on daily data over a one year period for Sector 75271-4, revealing strong relationships between the key metrics, as illustrated in Figure 3.11:

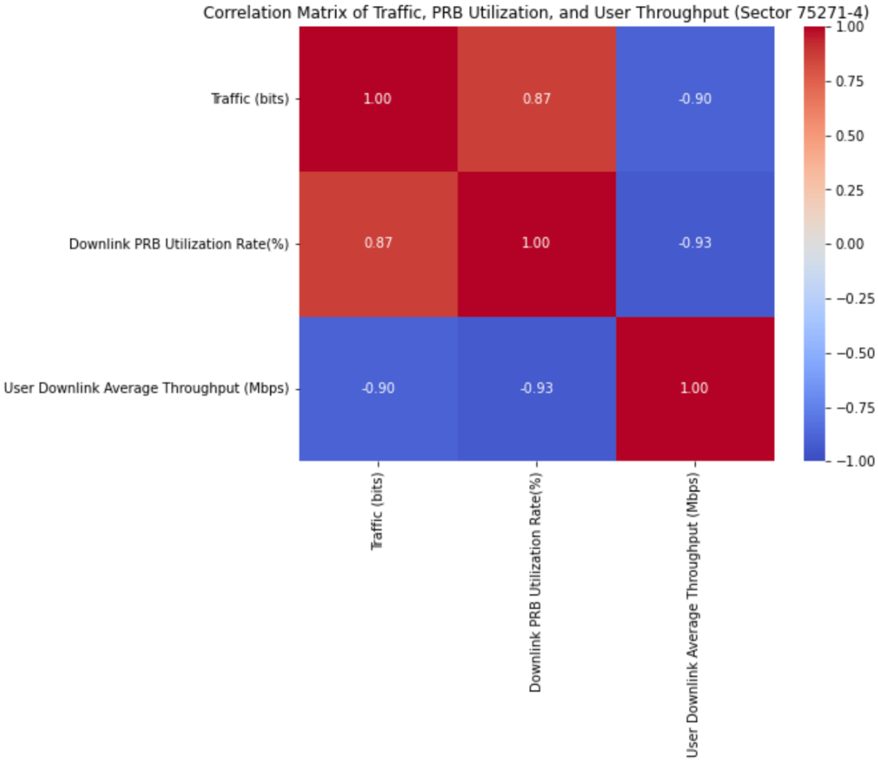


Figure 3.11: Correlation matrix for sector 75271-4

As expected, the analysis revealed a strong positive correlation (0.87) between traffic volume and PRB utilization, indicating that higher traffic leads to increased resource usage. However, this correlation is intuitive and commonly observed in network performance analysis. The primary insight comes from the negative correlation between traffic volume and user throughput (-0.90) and between PRB utilization and user throughput (-0.93). These findings indicate that as traffic and PRB usage increase, service quality, reflected by user throughput, degrades significantly.

While the correlation between traffic and PRB utilization is predictable, the strong negative correlations involving user throughput underscore the critical importance of managing PRB resources effectively to avoid congestion and ensure consistent service quality. The results also underscore the role of spectral efficiency as a key measure, integrating these variables to assess sector performance and identify areas where capacity issues are most likely to affect service quality, offering key insights for network management improvements.

4
Forecast Methodology

This chapter details the methodology for developing a prediction model using time series data for Cumulative Distribution Functions (CDF) and machine learning regression models. The goal is to produce weekly forecasts of congested sectors for up to one year, providing insights into future traffic patterns and potential congestion areas.

4.1 Overview of the Forecasting Process

The forecasting process integrates several key components to predict traffic demand, resource utilization, and potential congestion points in the network:

- **Time Series Models**: Utilizing both projection growth rates and hourly historical time series data to predict traffic volume.
- **CDF Model for Traffic Prediction**: Employing a Cumulative Distribution Function approach to identify sector specific traffic distribution patterns and integrate known projection growth rates. The CDF model allows for detailed analysis of peak traffic loads and assists in determining the likelihood of congestion occurring in each sector. .
- **Machine Learning Regression Models**: Applying various supervised learning techniques, such as Linear Regression, Lasso, Random Forest, and CatBoost to model the relationship between traffic demand, PRB utilization, and spectral efficiency.
- **Spectral Efficiency and Congestion Forecasting**: Spectral efficiency, is a primary measure used to assess each sector's ability to handle increasing traffic. By analyzing spectral efficiency alongside traffic capacity, the model can predict potential congestion points.

In radio networks, spectral efficiency is closely related to factors such as system design, modulation schemes, and radio channel propagation characteristics. These aspects influence the model's predictions of spectral efficiency, as each sector's radio environment is unique. This thesis approaches spectral efficiency predictions on a sector-by-sector basis, using historical traffic and PRB utilization data specific to each sector. This methodology inherently captures the local propagation conditions and design factors, indirectly accounting for fluctuations in radio environments. While system design and propagation conditions significantly impact spectral efficiency, the sector-specific approach ensures that these factors are incorporated within the predictions, supporting a more accurate assessment of each sector's capacity to manage traffic growth.

4.2 Time series models

While mobile network operators store daily performance indicators, shorter time periods (e.g., 1 hour) are needed to evaluate peak congestion. This study uses hourly historical time series data and projection growth rates to predict traffic volumes. Growth projections consider factors such as subscriber increases and new service rollouts, providing a comprehensive view of potential future congestion. The predicted traffic volumes are then compared against each sector's spectral efficiency to evaluate whether the sector can handle the anticipated load. Spectral efficiency is a critical measure in this comparison because it reflects how effectively each sector utilizes its allocated spectrum.

By doing so, the model identifies sectors where the load threshold may be exceeded, providing insight into potential congestion points during peak hours.

As illustrated in Figure 4.1, the forecasting workflow begins with historical network traffic measurements and projected growth rates. These inputs are processed through a CDF to generate a one-year traffic prediction. Machine learning models, such as LR, CatBoost, Lasso, and RF, are then employed to predict each sector's spectral efficiency, which plays a pivotal role in identifying capacity limitations. Based on the predicted spectral efficiency and the sector's traffic capacity, the model generates weekly forecasts of congested sectors.

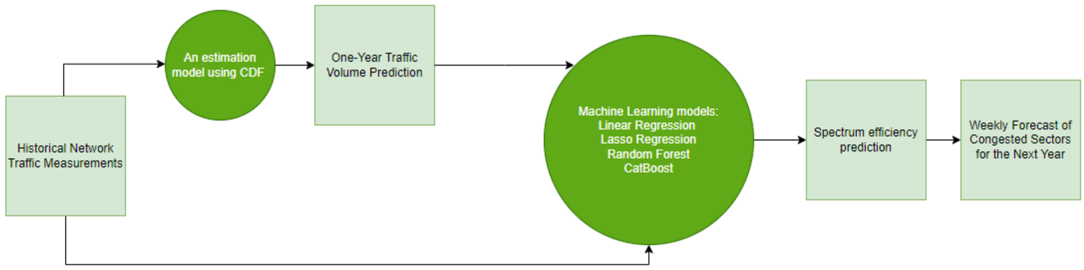


Figure 4.1: Integrated Process for Traffic Volume Prediction and Congestion Forecasting

To ensure efficiency in the methodology development, The initial focus is on Sector 75271-4, which experiences high peak hour traffic due to its location in a popular recreation area. As shown in Figure 4.2 spectral efficiency in this sector approaches its capacity during peak times, making it an ideal candidate for model development. Once validated in Sector 75271-4, the model will be extended to other sectors across the network.

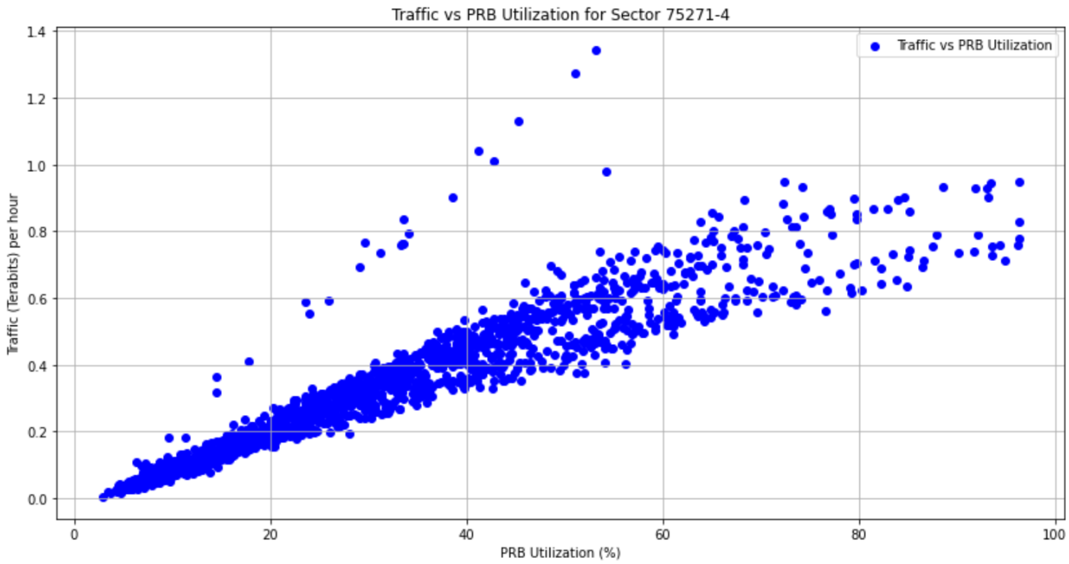


Figure 4.2: Spectral efficiency for sector 75271-4.

4.3 Cumulative Distribution Function (CDF) Model for Traffic Prediction:

The CDF model is preferred over other commonly used models for mobile network traffic prediction due to its ability to capture sector-specific traffic patterns and provide a probabilistic view of traffic distribution. While many forecasting models rely solely on historical data, the CDF model incorporates known growth projections, allowing it to account for future changes in traffic demand. Additionally, the CDF model is particularly effective in handling bursty and peak traffic patterns and is adaptable to sector-specific variability, making it better suited for mobile network traffic forecasting.

The CDF model predicts future traffic volumes in mobile networks. It represents the probability that traffic volume will be less than or equal to a given amount. This is expressed as:

$$F_X(x) = P(X \leq x) \tag{4.1}$$

where $F_X(x)$ is the CDF and X is the traffic volume.

The implementation of this model for traffic prediction begins with the collection and preprocessing of historical traffic data, organized into time series for each network sector and categorized by day of the week and hour of the day.

To capture temporal traffic patterns, the model constructs 168 distinct CDFs for each sector, corresponding to each hour of the day across every day of the week. These unique CDFs allow the model to capture sector-specific traffic distributions that vary by hour and day, ensuring that each sector's forecast reflects its own traffic patterns. The traffic data for each hour and day is sorted in ascending order, which allows for the calculation of cumulative probabilities for any given traffic volume and provides insights into the likelihood of different traffic levels, especially during peak usage periods. A key metric derived from each CDF is the 95th percentile:

$$P_{95} = F_X^{-1}(0.95) \tag{4.2}$$

which serves as a threshold for identifying peak traffic volumes and helps in classifying hours as peak or non-peak.

The model assumes independence in traffic values generated per hour, meaning that sequential correlations (e.g., traffic increases or decreases across consecutive hours) are not directly modeled. However, by creating unique CDFs per hour and day, the model partially captures natural daily and weekly traffic patterns. This level of granularity reflects typical traffic variations across different times of the week.

The model incorporates a sector specific weighting system to adapt the network wide projection growth rate. This weighting system is designed to more accurately apply the overall network growth rate to individual sectors based on each sector's unique capacity and traffic characteristics. Since the projected growth rate data is provided monthly for the entire network (spanning all sites and sectors), the weight is calculated to adjust this network-wide growth rate to fit the specific traffic profile of each sector. Each sector's weight is determined by calculating the ratio of its total traffic volume to the total network traffic volume. Specifically:

$$\text{Weight}_i = \frac{\text{Total Traffic}_i}{\sum_{j=1}^{n} \text{Total Traffic}_j} \tag{4.3}$$

where n is the total number of sectors. This weight allows us to proportionally allocate the projected growth rate across sectors, reflecting sector-specific traffic demand relative to the network as a whole.

Traffic predictions are generated through a process of inverse CDF sampling. For each hour, a new random value (u) is drawn from a uniform distribution (0,1), and the corresponding traffic volume is determined by applying the inverse CDF ($F_X^{-1}(u)$). This process, called inverse CDF sampling, reflects the probabilistic nature of historical traffic patterns and enables realistic variation in traffic forecasts, as each prediction is based on a unique random draw from the traffic distribution.

The base prediction is then scaled by a dynamic growth factor:

$$G_t = \text{Base Growth} \times \left(1 + \frac{\text{Week}_t}{52}\right) \tag{4.4}$$

where:

- **Base Growth** represents the initial growth factor specific to either peak or non-peak hours, accounting for different rates of increase during high and low usage times.
- **Weekly Adjustment**: The term $\left(1 + \frac{\text{Week}_t}{52}\right)$ scales the growth factor incrementally over the weeks in a year, adjusting the base growth upward as time progresses. By dividing by 52 (the number of weeks in a year), the growth factor increases smoothly, reflecting cumulative traffic growth throughout the forecast period.

The final traffic prediction for a given time t (T_t) is calculated as:

$$T_t = F_X^{-1}(u) \times G_t \times \text{Weight}_i \tag{4.5}$$

where:

- $F_X^{-1}(u)$ Provides the base traffic prediction using a unique random value u from the CDF, simulating the inherent variability in traffic demand.
- G_t: The dynamic growth factor, which adjusts the base prediction based on expected network-wide growth over time.
- Weight_i: This sector-specific weight (Weight_i) is a sector-specific weight that scales the growth rate according to each sector's unique traffic characteristics. This is calculated as each sector's proportion of the total network traffic volume, ensuring that growth adjustments align with each sector's actual traffic profile.

By implementing the CDF model in this way, MNOs generate sector-specific traffic predictions that incorporate historical patterns, future growth, and sector-specific characteristics. This approach allows each sector's forecast to reflect both the expected network-wide traffic increase and the individual traffic characteristics of that sector.

4.4 Spectral efficiency and congestion forecasting

In mobile networks, predicting sector congestion and managing user throughput are essential for effective capacity management. Spectral efficiency is a key metric that reflects how efficiently a sector uses its spectrum resources, measured up to the PRB load threshold. It directly quantifies the relationship between traffic volume and PRB utilization. Strong correlations between these factors indicate that as traffic increases, PRB utilization also rises, impacting user throughput. Therefore, spectral efficiency is the primary measure used to assess a sector's ability to manage traffic growth and avoid congestion.

Spectral efficiency is calculated using the following equation:

$$\text{SE}_{\text{sector}} = \frac{\text{Data Volume}_{\text{sector}}}{\text{PRB Util}_{\text{sector}} \times \text{Total Bandwidth}_{\text{sector}} \times \text{Measurement Period}} \tag{4.6}$$

where:

- $\text{SE}_{\text{sector}}$ represents the spectral efficiency at the sector level in bits/Hz/second,
- $\text{Data Volume}_{\text{sector}}$ is the total data volume handled by the sector during the measurement period (in bits),

- PRB $\text{Util}_{\text{sector}}$ is the average PRB utilization across all cells in the sector,
- Total $\text{Bandwidth}_{\text{sector}}$ is the sum of the bandwidths of all carriers (frequency bands) operating within the sector (in Hz),
- Measurement Period is the duration over which the data is collected (e.g., 1 hour).

PRB is the smallest unit of bandwidth and time allocated for data transmission in cellular networks. The number of PRBs available in a cell depends on the channel bandwidth and subcarrier spacing used as discussed in Section 2.1.

Given that cells within a sector may operate on different bandwidths and subcarrier spacings, each cell may inherently have a different number of PRBs thus, a different spectral efficiency. To incorporate this variation, historical PRB utilization data is collected for each cell individually. The sector-level PRB utilization (PRB $\text{Util}_{\text{sector}}$) is then calculated as the average PRB utilization across all cells within the sector. This averaged utilization reflects the sector's performance as a whole, accounting for the varying PRB numbers and spectral efficiency across different cells.

The congestion forecasting process involves a detailed analysis of each sector's spectral efficiency and traffic handling capacity before reaching the PRB utilization threshold. This approach captures sector-specific traffic patterns and user distributions. The forecasting process consists of four key steps:

1. Collect and analyze historical PRB utilization and traffic volume data at the cell level, then aggregate these metrics by summing traffic volume and averaging PRB utilization across all cells within each sector. This aggregation creates a sector-level performance model reflecting typical traffic behavior and resource usage.
2. Apply machine learning regression techniques to model the relationship between traffic demand and PRB utilization. This helps in understanding how traffic growth affects spectrum utilization and PRB load thresholds at the sector level.
3. Use the regression model to calculate spectral efficiency by determining the maximum traffic volume each sector can handle before exceeding the PRB utilization threshold. This sector-level approach assumes that cell-specific variations in spectral efficiency, such as those due to differences in frequency bands, are minimal or effectively averaged out in the aggregation process. Spectral efficiency serves as the benchmark for congestion prediction.
4. Forecast future traffic patterns and identify potential congestion points by monitoring spectral efficiency trends at the sector level. This enables proactive network management and allows operators to take targeted interventions before congestion affects user throughput.

Machine learning techniques are categorized into unsupervised, supervised, and reinforcement learning. Unsupervised learning groups data based on similar characteristics, while supervised learning builds a mapping function between input and output data to estimate output features. Reinforcement learning involves training an agent through interaction with an environment. For this study, supervised learning is chosen due to the clear relationship between PRB utilization and traffic volume in the historical data as shown above. With extensive labeled data pairs of traffic volumes (input) and PRB utilization (output), supervised learning is well suited to meet the study's objectives. Therefore, this chapter introduces supervised learning techniques and their implementation challenges to achieve accurate predictions.

4.4.1 Supervised machine learning Regression approach

The primary reason for using machine learning algorithms in this project is to estimate and predict PRB utilization and spectral efficiency in mobile network sectors. By utilizing sector level metrics such as traffic volume and PRB usage from large historical datasets, machine learning can model complex relationships between these variables, enabling more accurate and adaptive predictions. This approach is particularly effective for managing the vast amounts of data generated by mobile networks, which traditional methods would struggle to handle. Machine learning enhances resource allocation and network performance by improving the forecasting of potential congestion points. LR and Lasso are used to explore the linear relationship between traffic volume and resource utilization, while CatBoost is applied to capture non linear relationships. RF combines both linear and non linear boundaries through tree based modeling. More information about the machine learning techniques used in this project can be found in Appendix B.

The methodology for predicting network congestion and traffic volume integrates a systematic data flow and operations pipeline, as illustrated in Figure 4.3. The process begins with Data Collection and Preparation, where raw data is collected from the OSS and stored in a local database. This data includes historical hourly traffic data and PRB utilization for all network sectors. Feature Engineering is then applied to add time based features such as hour, day, and weekday, which are crucial for capturing the temporal variations in traffic patterns and accounting for the dynamic and nonlinear nature of network traffic.

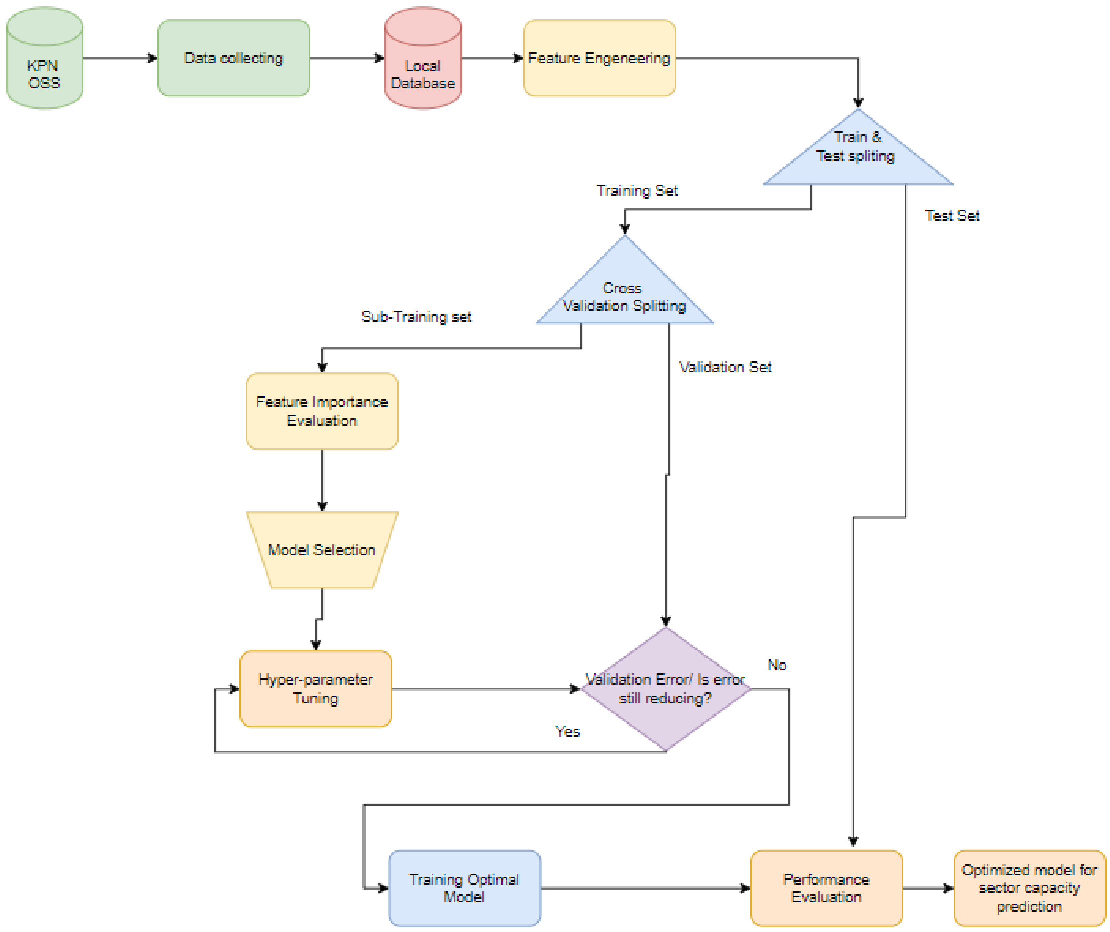


Figure 4.3: Data flow in the pipeline

In the Data Splitting phase, the dataset is divided into two parts: the training set and the test set. The training set is used to train the machine learning model, allowing the model to learn patterns from the data. This set forms the foundation for building the predictive model. The test set, on the other hand, is kept separate and unseen by the model during training. It is used later in the process to evaluate the model's performance, ensuring that the model can generalize well to new, unseen data and doesn't overfit to the training data. To ensure robust model performance, the training set is further split into a sub-training set and a validation set using Cross-Validation. Cross-validation helps assess the model's performance on different subsets of the training data, allowing for better hyper-parameter tuning and model selection. The validation set is specifically used to fine-tune the model by evaluating it during training, while the test set remains reserved for final model evaluation.

During Feature Importance Evaluation, the significance of various features is assessed to identify those that most influence the model's predictions. Afterward, the Model Selection phase compares different machine learning algorithms, such as Linear Regression (LR), Lasso, Random Forest, and CatBoost.

In the Hyper-parameter Tuning phase, GridSearchCV with cross-validation is used to optimize model parameters. If the validation error is still reducing, the process loops back to hyper-parameter tuning for further fine-tuning. The pipeline continues this iterative process until the validation error reaches a constant or acceptable value, ensuring the model is well-optimized. Once the validation error has stabilized or is minimized, the pipeline proceeds to Training the Optimal Model. The model is trained on the full training set with the best-tuned hyperparameters.

The final step involves Performance Evaluation, where the trained model is tested on the test set to assess its predictive accuracy. This process produces the Optimized Model for Sector Capacity Prediction, which can be used to forecast network congestion and manage traffic across different network sectors.

4.5 Hyperparameter Tuning

Hyperparameter tuning is essential for optimizing the performance of machine learning models used for PRB utilization prediction. This study uses the GridSearchCV tool to systematically explore the hyperparameter space for models including LR, Lasso Regression, RF, and CatBoost.

For each model, a set of hyperparameters and their potential values are defined, as shown in Table 4.1. GridSearchCV evaluates all combinations by training the model on subsets of the data and validating performance using k-fold cross validation. The average performance metric (negative mean squared error) is calculated for each fold, and the hyperparameter set with the best overall performance is selected.

The study uses 5-fold cross validation to ensure reliable performance estimation by testing the model on different data subsets. This method reduces the risk of overfitting and ensures the model generalizes well to unseen data.

Table 4.1: Hyperparameters and Their Ranges for Lasso Regression, Random Forest, and CatBoost Models

Model	Parameter	Values
Lasso Regression	`alpha`	[0.1, 0.5, 1.0, 10.0]
Random Forest	`n_estimators`	[100, 200, 500]
	`max_depth`	[10, 20, None]
	`min_samples_split`	[2, 5, 10]
	`min_samples_leaf`	[1, 2, 4]
CatBoost	`iterations`	[100, 500]
	`depth`	[4, 6, 8]
	`learning_rate`	[0.01, 0.1, 0.3]

Negative mean squared error is used as the scoring metric for GridSearchCV, aiming to minimize prediction errors in PRB utilization. After completing the grid search, the best hyperparameters for each model are selected based on the highest average cross validation score. These optimal hyperparameters are then used to train the final model on the entire training dataset.

By carefully tuning the hyperparameters as outlined in Table 4.1, this approach ensures that the best possible version of each algorithm is compared in the evaluation of PRB utilization prediction performance. This process contributes to the robustness and reliability of the congestion forecasting methodology.

4.6 Model Evaluation

The results from the forecasting methods are evaluated and compared to determine the necessary steps for the final design of the prediction model. To assess the performance of the models, the Mean Absolute Percentage Error (MAPE) is used, a widely recognized metric for evaluating prediction accuracy. The model with the lowest MAPE is considered the most optimal, as it indicates closer alignment between the predicted values and the actual observed data. The formula for calculating MAPE is shown in:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100 \tag{4.7}$$

where t represents a specific time period, n is the total number of observations, and Y_t is the observed throughput at time t. The MAPE is calculated by taking the absolute difference between the predicted $\hat{Y}_t$ and observed values, relative to the observed values.

In addition to MAPE, the R^2 score, also known as the coefficient of determination, was used to measure the proportion of the variance in the observed data that is predictable from the independent variables. The formula for calculating R^2 is shown in Equation 4.8:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{4.8}$$

where Y_t represents the observed throughput at time t, $\hat{Y}_t$ represents the predicted throughput, and $\bar{Y}$ is the mean of the observed data. The R^2 score provides insight into how well the model explains the variability of the observed data. A value of $R^2 = 1$ indicates a perfect fit, meaning that the model explains all the variance in the data, while an R^2 of 0 indicates that the model explains none of the variability.

Together, MAPE and R^2 provide a assessment of model performance. While MAPE quantifies the percentage error of predictions, R^2 indicates how much of the variation in the data can be explained by the model.

4.7 Congestion threshold and Forecasting

To maintain a minimum service requirement of 6.0 Mbit/s for 90% of user connections, it is crucial that a sector's PRB load does not reach full capacity. KPN has set the PRB load threshold at 89% of the utilization, leveraging carrier aggregation across its sites [54]. Congestion in a sector is defined by two key criteria [55]:

- **Short-term**: PRB utilization exceeds the established threshold, resulting in service degradation (below 6.0 Mbit/s for 90% of connections) over a 1-hour period.
- **Long-term (Structural)**: Weekly congestion is identified when the total traffic during congested hours exceeds 5% of the total weekly traffic volume (including both congested and non-congested hours). This criterion must be met for at least 6 out of the last 13 weeks, including one of the most recent two weeks.

The congestion threshold of 89% PRB utilization in the downlink is based on KPN's satisfaction criteria, which indicates when the network is considered congested from an average user experience perspective, aligning with KPN's service quality standards. This threshold signifies the point at which average user satisfaction significantly decreases, irrespective of variations in active user count during high PRB utilization.

In this study, congestion is assessed weekly. The final output is a weekly report identifying congested sectors across the network. These reports help network operators quickly detect problem areas, monitor congestion trends, and prioritize interventions for capacity improvement or traffic management. This approach supports proactive network management, enabling timely responses to emerging and ongoing congestion issues.

5

Results

This chapter presents the results of traffic volume prediction, PRB utilization modeling, and congestion forecasting for the network. The analyses were conducted using the methodologies outlined in the previous chapter, which involved both time series forecasting with the Cumulative Distribution Function (CDF) model and machine learning models for PRB utilization. The results are structured into the following sections.

5.1 Time series forecasting results

The time series forecasting approach uses the CDF model, as outlined in the Methodology chapter. This model predicts traffic volumes for Sector 75271-4, selected due to its higher frequency of peak hours and elevated traffic compared to other sectors.

Figure 5.1 illustrates the hourly forecasted traffic for Sector 75271-4, providing a granular view of traffic fluctuations essential for identifying short term congestion periods.

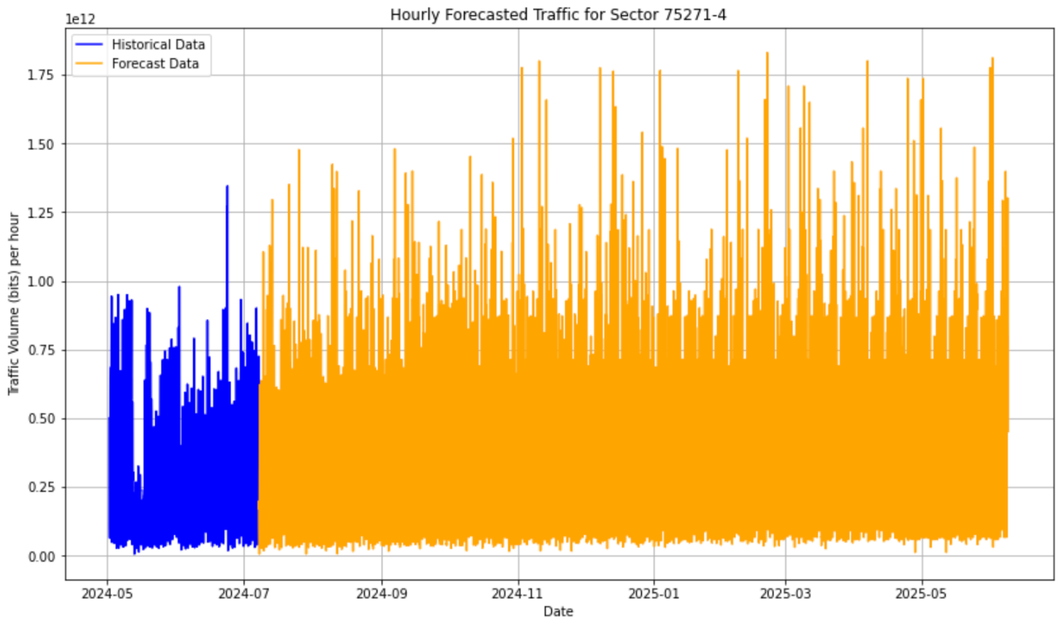


Figure 5.1: Hourly traffic forecasting .

The hourly forecast reveals clear daily patterns in both historical and predicted data, with regular peaks corresponding to high usage hours. Over time, there is a noticeable increase in peak traffic magnitudes, aligning with the overall growth trend. Additionally, the baseline traffic shows a rise, indicating growth in off peak usage. Despite the inherent variability at the hourly level, the forecast maintains consistent patterns throughout the prediction period, suggesting model stability.

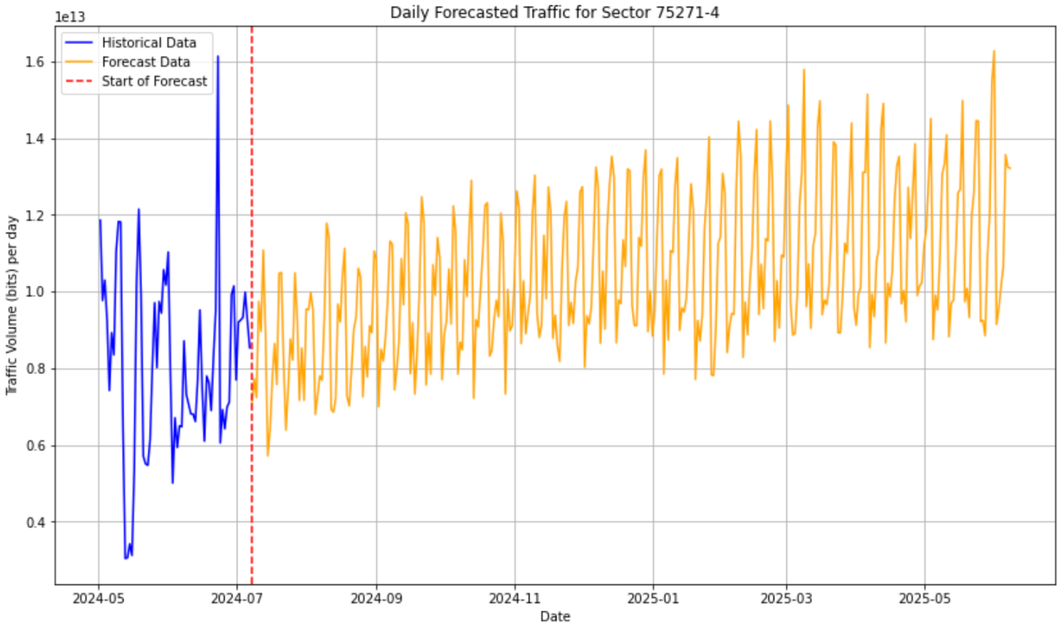


Figure 5.2: Daily traffic forecasting.

Figure 5.2 presents the daily forecasted traffic for the same sector over approximately one year. The daily forecast demonstrates a smooth continuation from historical data, indicating good model fit. An upward trend is evident, suggesting increasing data usage over time. The model successfully preserves cyclical patterns, capturing weekly and monthly usage trends. It also maintains realistic day to day variability, avoiding overly smoothed predictions. Several high peaks are forecasted, potentially indicating periods of exceptionally high demand and possible congestion.

The CDF model effectively captures long-term growth trends and seasonal traffic patterns while maintaining realistic variability, as demonstrated in Figure 5.3, which compares historical and forecasted traffic distributions. The close alignment of the curves highlights the model's ability to learn from past patterns, while slight divergences at higher volumes reflect the inherent variability in traffic projections.

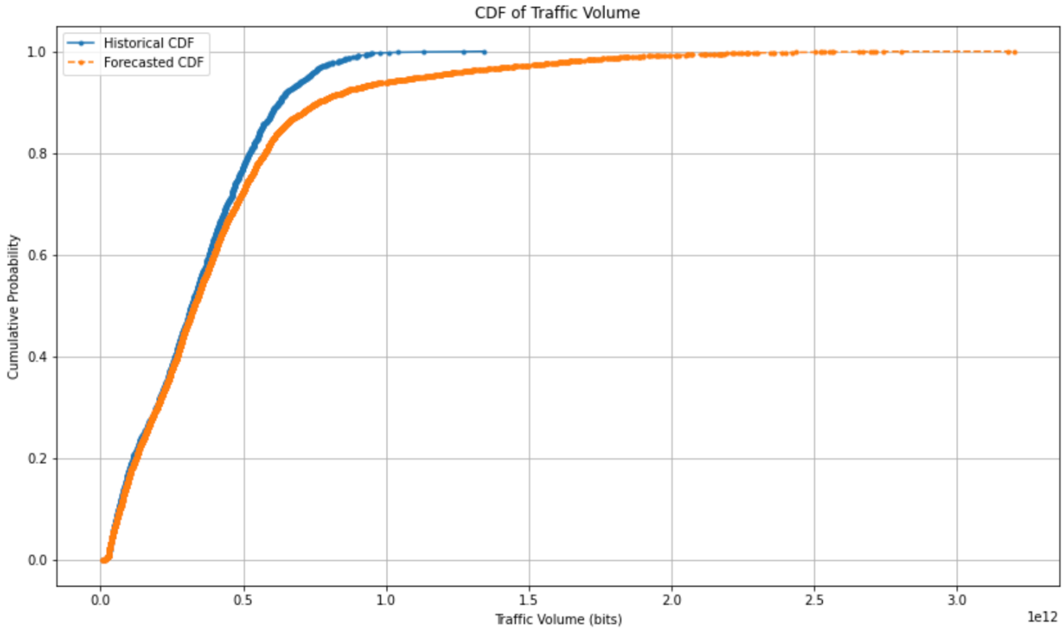


Figure 5.3: CDF of Traffic Volume – Historical vs. Forecasted

Quantitatively, the model achieved a MAPE of 13.6%, demonstrating strong predictive accuracy. However, the model's assumption of independence between hourly predictions may limit its ability to capture sequential patterns or extreme traffic spikes.

The model's hourly granularity provides actionable insights for short-term network management, while daily forecasts support strategic long-term capacity planning. These results serve as crucial inputs for subsequent analyses, including PRB utilization and congestion forecasting, forming a solid foundation for optimizing network performance.

5.2 Machine learning model performance

Machine learning models are developed to predict spectrum efficiency using four different algorithms: LR, Lasso Regression, RF, and CatBoost. Each model is trained on the same dataset and evaluated using consistent metrics to ensure a fair comparison. The primary metrics for evaluating the models are R^2 and MAPE, which provide insights into each model's goodness of fit and prediction accuracy.

The prediction line generated by each model is essential for understanding its performance. By plotting the predicted PRB utilization against the true PRB utilization, the prediction line visually represents the model's accuracy and reliability. A strong alignment of points along the prediction line indicates how well the model captures underlying patterns in the data, reflecting its effectiveness in making accurate predictions.

Table 5.1 summarizes the performance metrics for each model, both before and after outlier removal.

As shown in Figures 5.4 till 5.7, the prediction lines for each model before outlier detection demonstrate that both Random Forest and CatBoost achieve strong alignment with the actual SE. The actual SE illustrates the observed efficiency of resource usage relative to the traffic load within the network, providing insight into how effectively resources are allocated at varying levels of demand. In contrast, the predicted SE refers to the values estimated by the models based on the input features. Both models exhibit a natural ability to handle outliers effectively. CatBoost, in particular, shows minimal deviation from the true values, as indicated by its high R^2 value of 0.9592 and low MAPE of 8.76% on the original dataset (see Table 5.1 for detailed performance metrics).

Table 5.1: Performance Comparison of Machine Learning Models Before and After Outlier Removal

Model	Dataset	R^2	MAPE (%)
Linear Regression	Original Data	0.8628	15.19%
Linear Regression	After Outlier Removal	0.9255	11.39%
Lasso Regression	Original Data	0.8609	17.80%
Lasso Regression	After Outlier Removal	0.9251	11.25%
Random Forest Regressor	Original Data	0.9504	10.15%
Random Forest Regressor	After Outlier Removal	0.9586	8.77%
CatBoost Regressor	Original Data	0.9592	8.76%
CatBoost Regressor	After Outlier Removal	0.9783	6.02%

Similarly, the Random Forest model achieves an R^2 of 0.9504 and a MAPE of 10.15%. These results indicate that both models manage outliers effectively while maintaining a high level of prediction accuracy. In contrast, the linear models LR and Lasso Regression struggle to handle outliers in the original data. As depicted in Figures 5.4 and 5.5, the points deviate more noticeably from the prediction line, particularly for higher PRB utilization values. This deviation is reflected in the higher MAPE values of 15.19% for Linear Regression and 17.80% for Lasso Regression, indicating larger prediction errors when outliers are present.

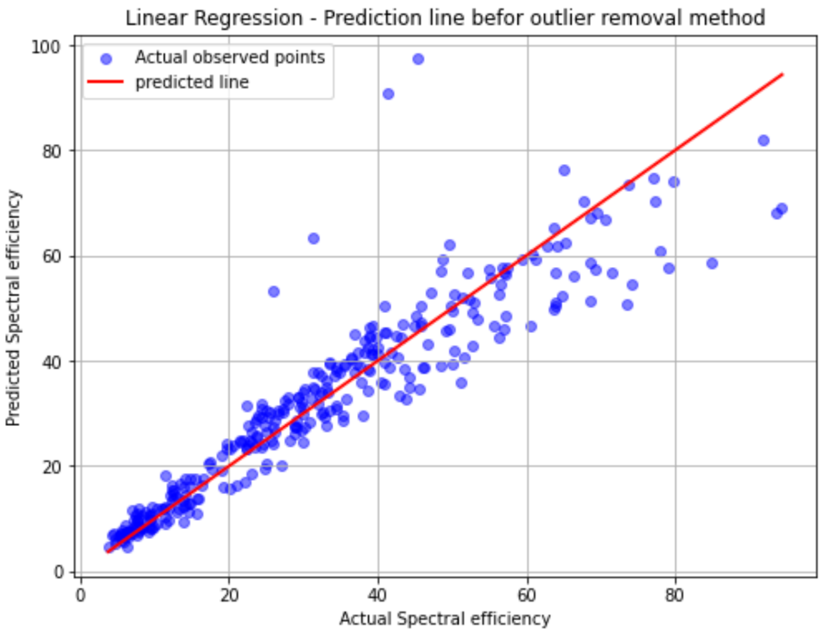


Figure 5.4: Linear Regression - Before Outlier Removal

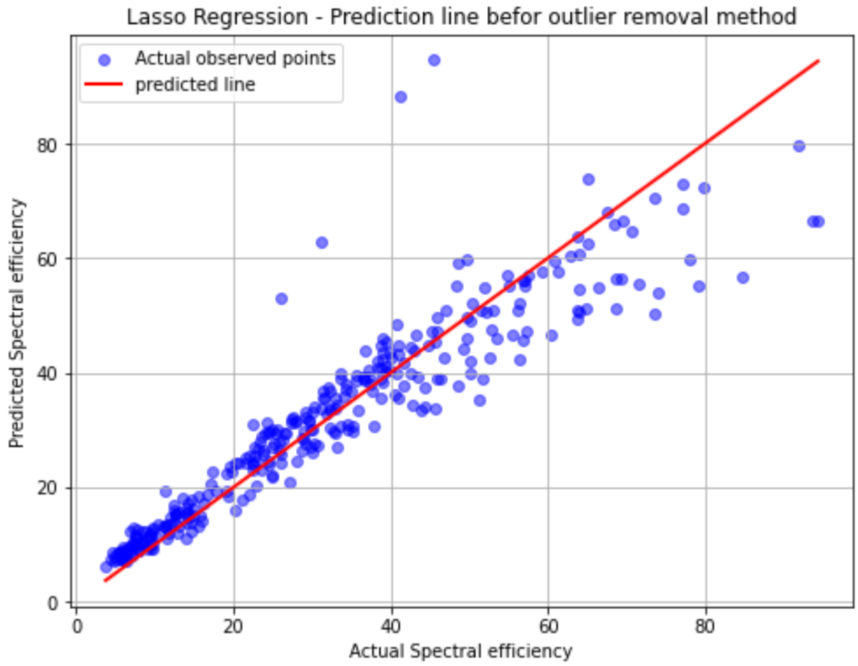


Figure 5.5: Lasso Regression - Before Outlier Removal

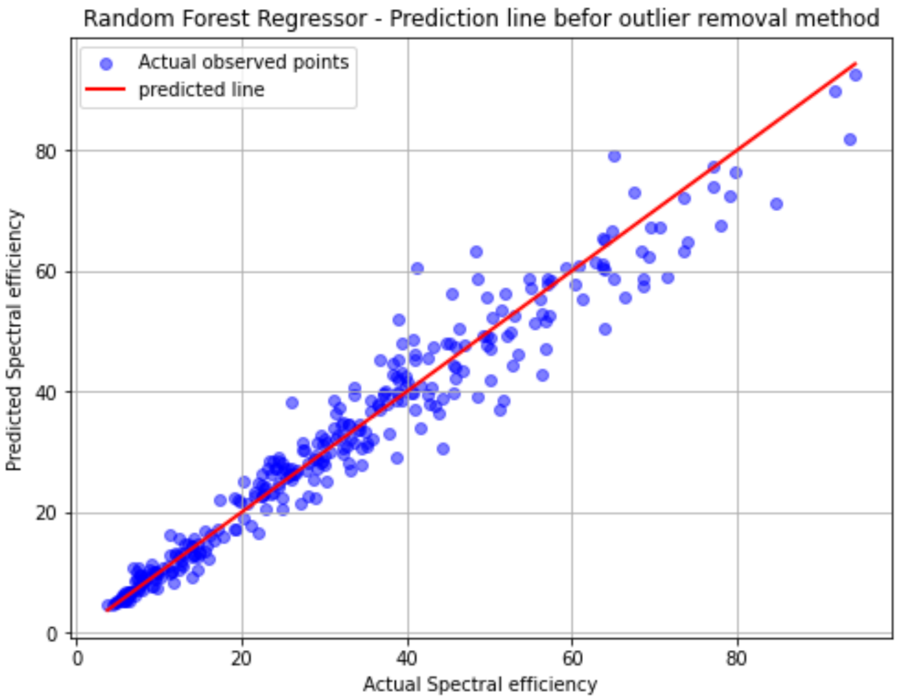


Figure 5.6: Random Forest - Before Outlier Removal

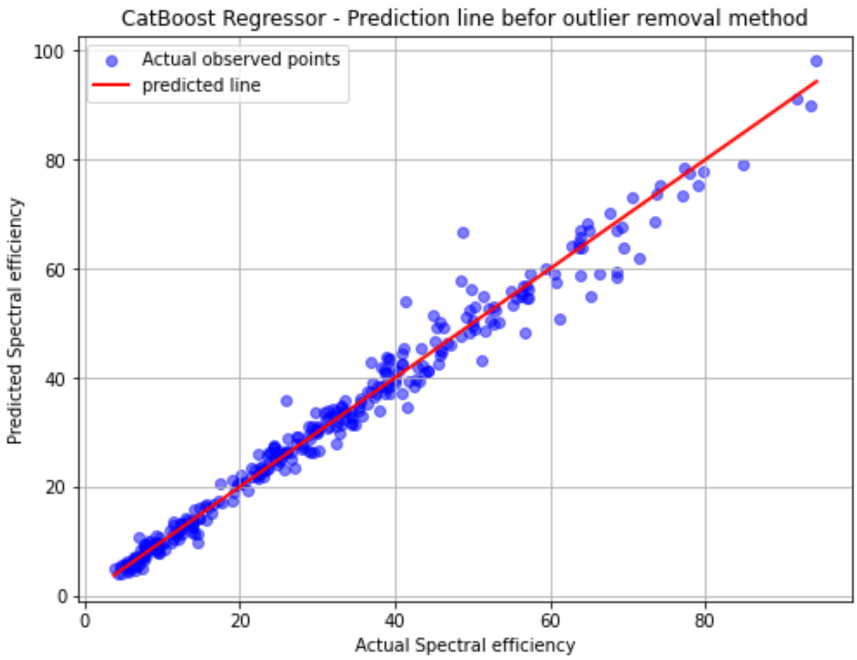


Figure 5.7: CatBoost - Before Outlier Removal

In contrast, the linear models LR and Lasso Regression struggle to handle outliers in the original data. As depicted in Figures 5.4 and 5.5, the points deviate more noticeably from the prediction line, particularly for higher PRB utilization values. This deviation is reflected in the higher MAPE values of 15.19% for Linear Regression and 17.80% for Lasso Regression, indicating larger prediction errors when outliers are present.

To improve prediction accuracy, DBSCAN is applied to detect and remove outliers from the dataset. The impact of this preprocessing step is clear in the improved prediction lines shown in Figures **??**, where the models show closer alignment with the true PRB utilization values after outlier removal.

The removal of outliers has a particularly positive effect on the performance of the Linear Regression and Lasso Regression models. After outlier removal, the Linear Regression model's R^2 improved from 0.8628 to 0.9255, and its MAPE decreased from 15.19% to 11.39%, as shown in Table 5.1. Similarly, Lasso Regression's R^2 increased from 0.8609 to 0.9251, and its MAPE dropped from 17.80% to 11.25%. These improvements, depicted in Figures from 5.8 till 5.11, demonstrate that linear models benefit from outlier detection.

RF and CatBoost, which already performed well on the original dataset, also shows further improvement after outlier removal. CatBoost's R^2 increased from 0.9592 to 0.9783, and its MAPE dropped from 8.76% to 6.02%, reflecting an even stronger predictive capability. Random Forest's R^2 rose to 0.9586, while its MAPE reduced to 8.77%. These enhancements, depicted in Figures from 5.8 till 5.11, underscore the effectiveness of these models on both the original and cleaned datasets.

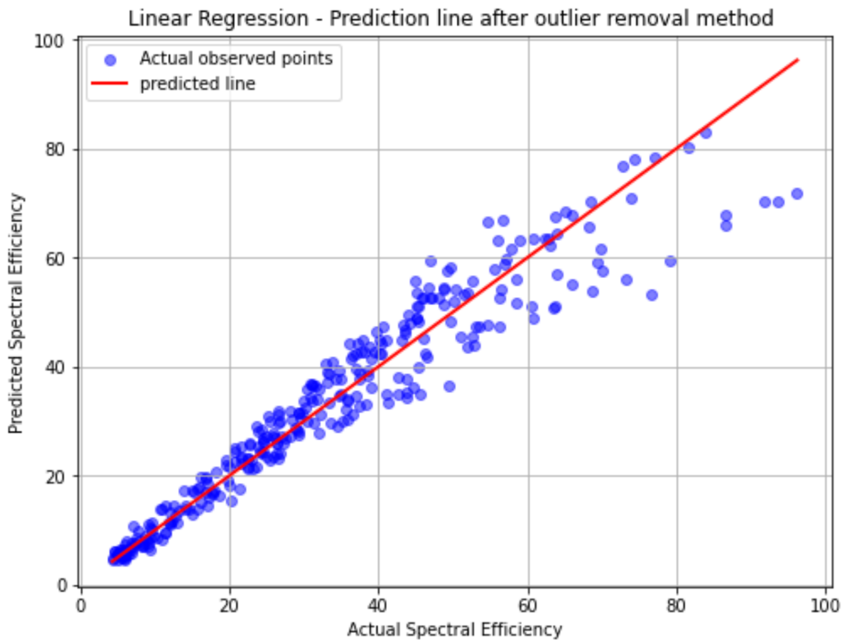


Figure 5.8: Linear Regression - After Outlier Removal

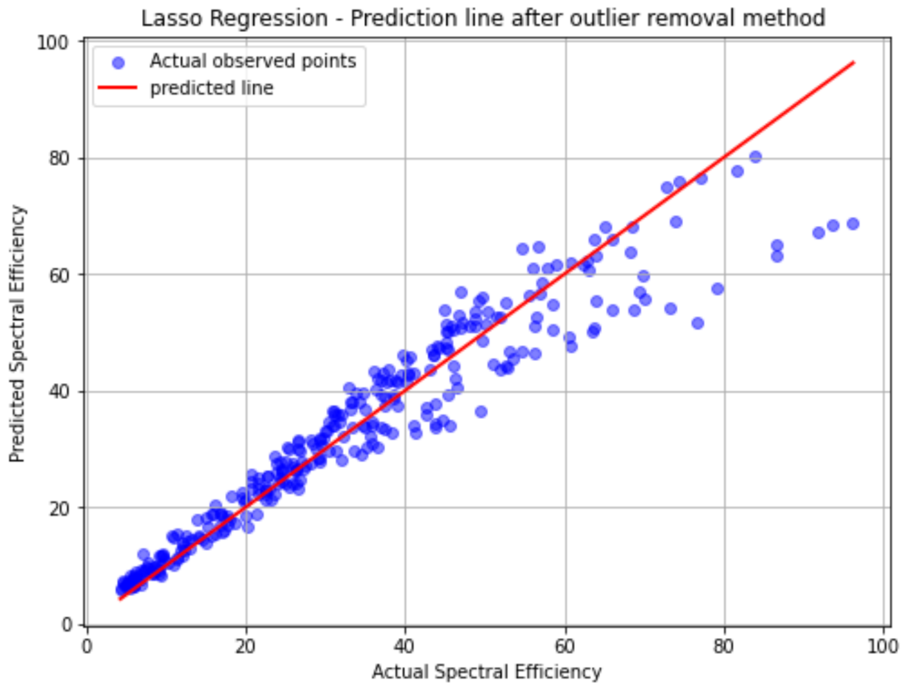


Figure 5.9: Lasso Regression - After Outlier Removal

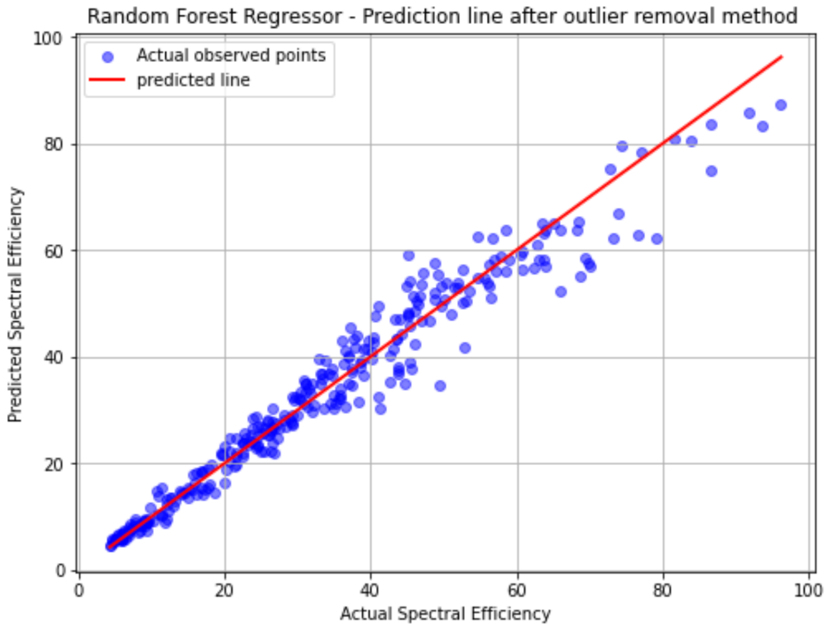


Figure 5.10: Random Forest - After Outlier Removal

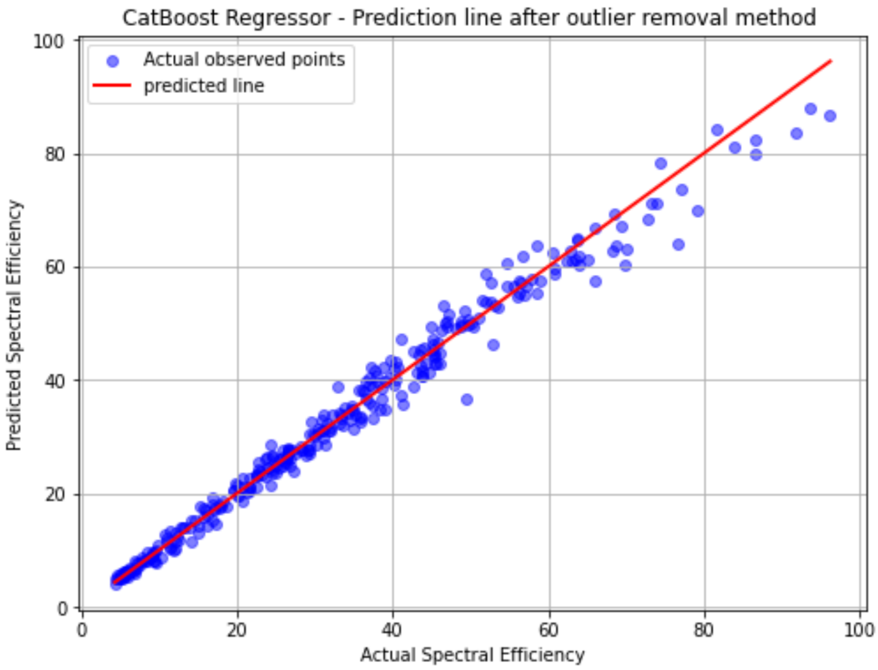


Figure 5.11: CatBoost - After Outlier Removal

5.3 Forecasting spectral efficiency

A grid search identifies the optimal hyperparameters for LR, Lasso Regression, RF, and CatBoost models. These models then predict spectral efficiency, a key metric related to PRB utilization in a network sector. Performance metrics, including R^2 and MAPE, show how effectively each model predicts traffic and PRB utilization. Accurate spectral efficiency predictions help operators anticipate high PRB utilization periods and take proactive steps to prevent congestion.

As shown in Figures from 5.12 till 5.15 LR and Lasso regression models exhibit a clear linear trend, indicating a reliable relationship between traffic and resource block utilization. In these models, PRBs scale proportionally with traffic demand, ensuring effective resource management, even during peak hours. In contrast, while Random Forest and CatBoost models perform well within the training data range, they struggle to extrapolate beyond it. Random Forest achieves an R^2 of 0.9592 and a MAPE of 8.76%, while CatBoost reaches an R^2 of 0.9783 and a MAPE of 6.02%. Despite their accuracy under normal conditions, these models have limitations when predicting traffic surges beyond typical levels.

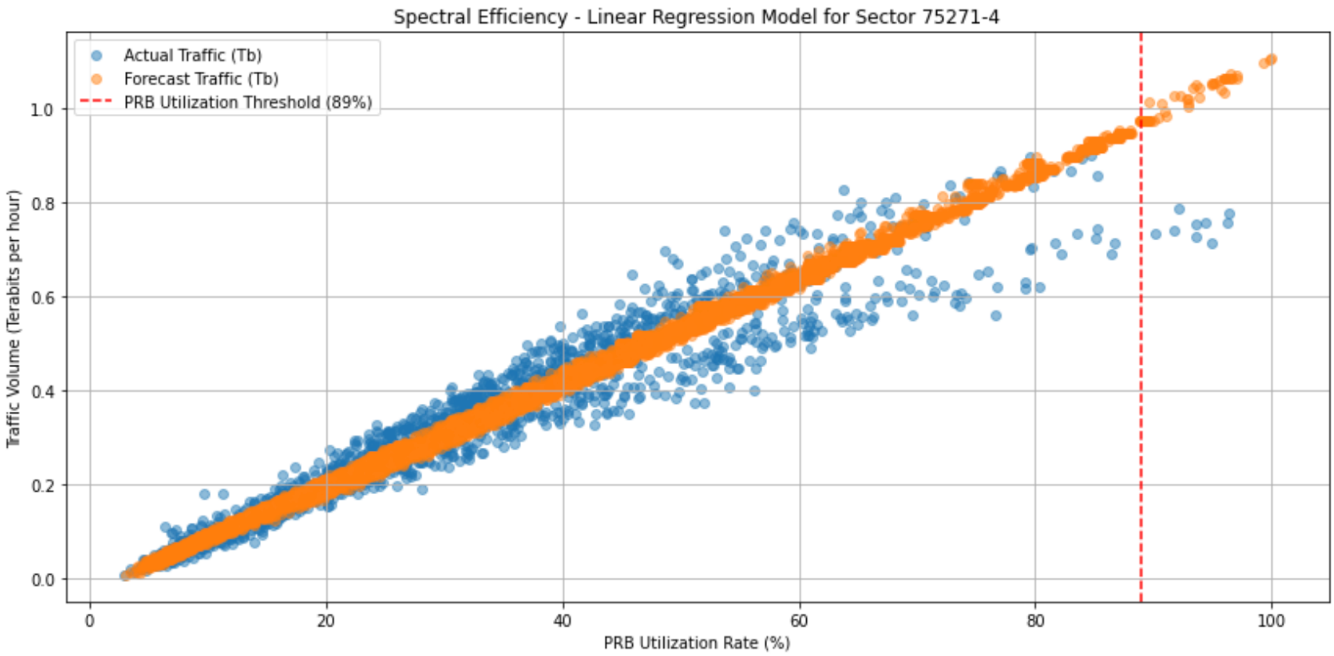


Figure 5.12: Linear Regression

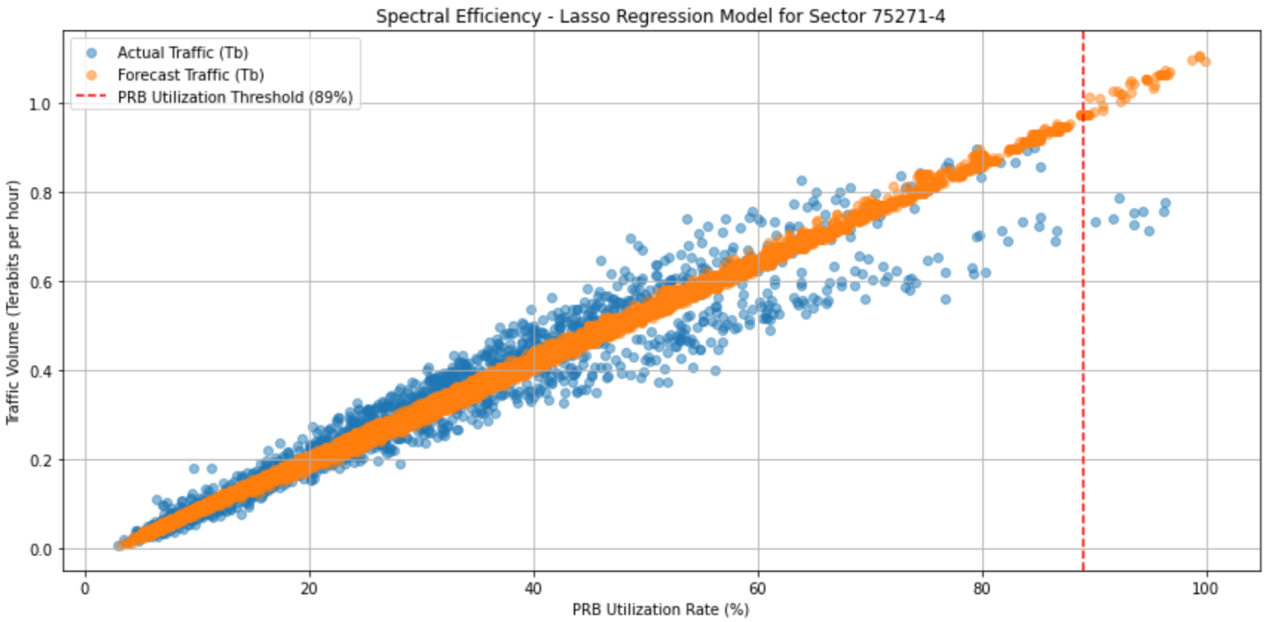


Figure 5.13: Lasso Regression

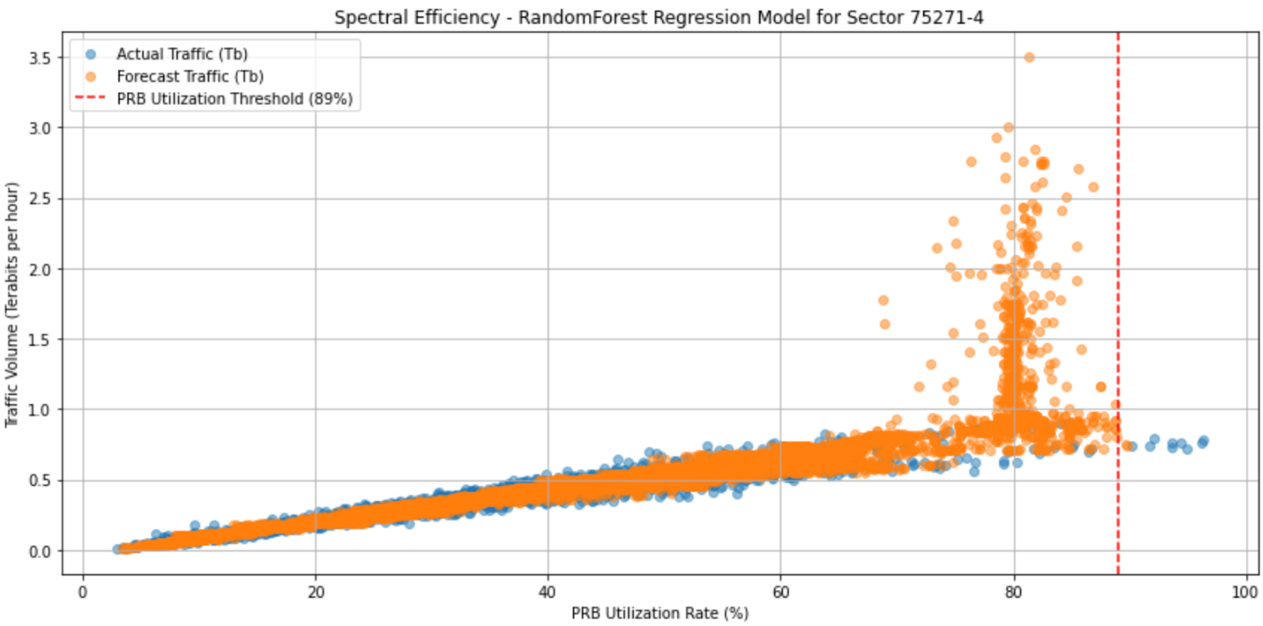


Figure 5.14: Random Forest

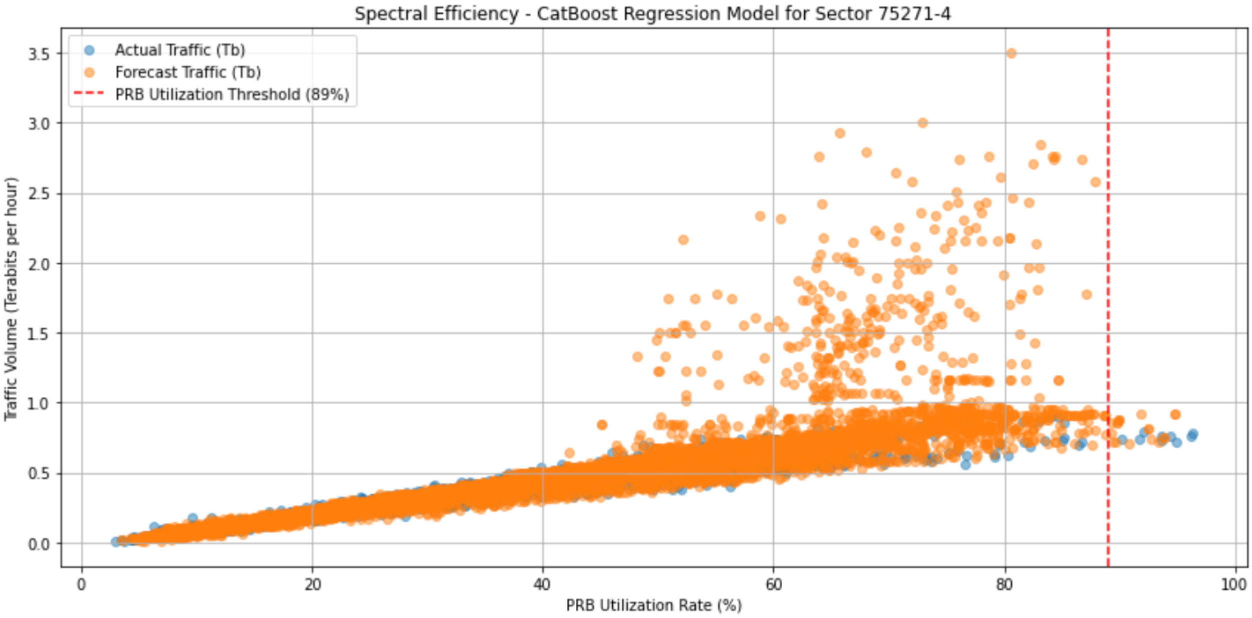


Figure 5.15: CatBoost

LR and Lasso regression models consistently show a linear relationship between PRB utilization and traffic, even when predictions extend beyond observed data. This aligns with the expectation that increased traffic requires more resource blocks. In contrast, CatBoost and Random Forest models falter at higher PRB utilization rates, particularly with CatBoost, where forecasts plateau rather than rise with increasing traffic. While these ensemble models perform well within the observed data range, they struggle to predict traffic surges, which is crucial for effective spectral efficiency forecasting. Linear and Lasso models are more reliable for long-term capacity management, as they better capture the relationship between traffic and resource blocks at higher utilization rates.

The consistency of LR and Lasso regression models in maintaining predictable behavior across the full range of PRB utilization rates reinforces their suitability for forecasting PRB utilization rates at higher traffic volumes. On the other hand, CatBoost and Random Forest models flatten out as traffic rises beyond average utilization, leading to potential inefficiencies and missed congestion indicators. This limitation is visually evident in the plots, where ensemble models fail to predict a proportional rise in PRB usage during high traffic periods, potentially leading to underestimations of resource needs. As a result, Linear and Lasso models, with their consistent relationship between traffic and PRB utilization, are better suited for forecasting peak traffic and planning for congestion. Although CatBoost and Random Forest excel in moderate traffic conditions, their inability to handle traffic surges makes them less effective for predicting resource demands during peak hours.

Accurate spectral efficiency forecasting enables proactive network management, ensuring optimal performance during congestion prone periods. By predicting traffic surges and corresponding PRB utilization, operators can allocate resources effectively, reduce the risk of network saturation, and maintain consistent service quality.

5.3.1 Weekly Congestion Analysis

Building upon the spectral efficiency predictions, the focus now shifts to forecasting and analyzing network congestion for Sector 75271-4. This analysis is critical for proactive network management and capacity planning.

The Lasso regression model was utilized for spectral efficiency prediction due to its consistent performance in extrapolation and the lowest MAPE of 11.25%. From this model, the relationship between PRB utilization and traffic volume is derived, enabling the determination of the sector's traffic capacity threshold.

Figure 5.16 illustrates the forecasted traffic volume for Sector 75271-4 during Week 8 of 2025. This specific week was selected from a one-year forecast period due to its higher congestion ratio, indicating a greater number of congested hours compared to other weeks. The forecast is generated using the CDF model, which leverages historical traffic distributions to produce a probabilistic view of future traffic loads. In the CDF model, traffic predictions for each hour are generated by sampling from the cumulative distribution of historical traffic data. A random variable u, drawn from a uniform distribution, maps to a corresponding traffic volume in the inverse CDF, allowing the model to reflect both typical and peak traffic patterns. By applying this probabilistic sampling over the full forecast period, the model identifies periods where traffic is expected to approach or exceed the congestion threshold. The choice of Week 8, 2025, allows for a clear visualization of expected peak points and highlights critical periods for network management, where proactive measures may be necessary to mitigate congestion. The green dashed line in the figure represents the traffic capacity threshold, which is the maximum traffic volume the sector can handle before congestion occurs, approximately 9.74×10^{11} bits. Congestion is defined as periods when PRB utilization exceeds 89%.

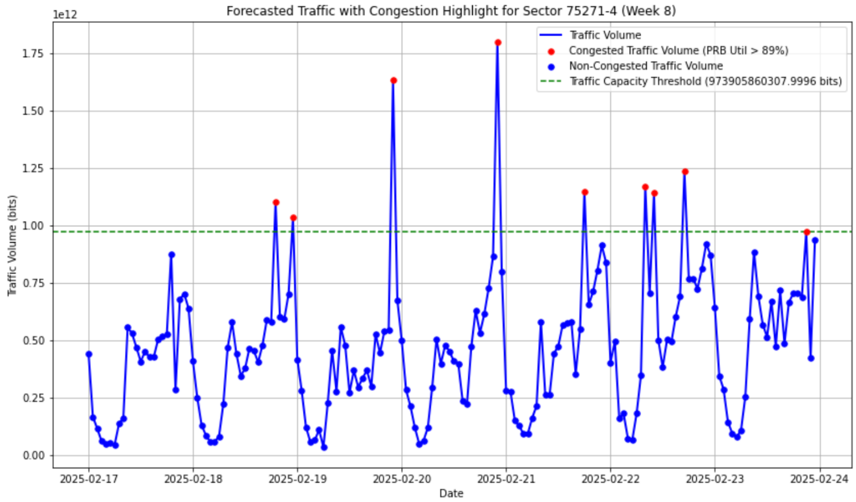


Figure 5.16: Forecasted Traffic with Congestion hours for Sector 75271-4 for week 8, 2025

This analysis suggests that Sector 75271-4 can efficiently handle traffic up to 1×10^{12} bits before experiencing congestion. During congested periods, the sector operates beyond its optimal capacity, potentially leading to degraded service quality, slower data transmission rates, and increased risk of dropped connections. The recurring congestion peaks indicate a need for increased capacity or traffic offloading strategies, particularly during identified high traffic periods.

By anticipating these congestion points, network operators can implement targeted interventions such as dynamic resource allocation during predicted peak times, load balancing to nearby sectors, and temporary capacity enhancements during critical periods. The analysis of Sector 75271-4 reveals a pattern of recurring congestion across multiple weeks in 2024 and early 2025.

Sector 75271-4's current configuration utilizes multiple frequency bands (700, 800, 900, 1400, 1800, 2100, 2600 FDD, and 2600 TDD MHz), with a total of approximately 550 PRBs allocated across these bands. Despite this substantial resource allocation, the persistent congestion issues revealed in the forecast suggest that the current capacity is insufficient to meet future demand. This indicates the need for capacity expansion or optimization strategies to address the increasing traffic volume and prevent service degradation.

To address these capacity constraints, KPN has taken a proactive approach by acquiring a new 3.5 GHz spectrum with 100 MHz bandwidth. This strategic acquisition will provide an additional 273 PRBs per site, increasing the total available PRBs for Sector 75271-4 by more than 50%. This enhancement in resources is expected to accommodate the projected traffic growth and substantially mitigate congestion occurrences.

5.4 Impact of 3.5 GHz Spectrum Implementation on Sector 75271-4

The implementation of the 3.5 GHz spectrum in Sector 75271-4 has demonstrated significant improvements in network performance, as evidenced by updated forecasts and spectral efficiency analysis. Figure 5.17 highlights the traffic forecast for Week 8, revealing no red points, which previously indicated congestion. This demonstrates that adding 3.5 GHz resources has effectively resolved congestion in the sector for the forecasted period. The absence of congestion points, even during peak times, indicates that the 3.5 GHz spectrum provides an effective buffer against traffic fluctuations.

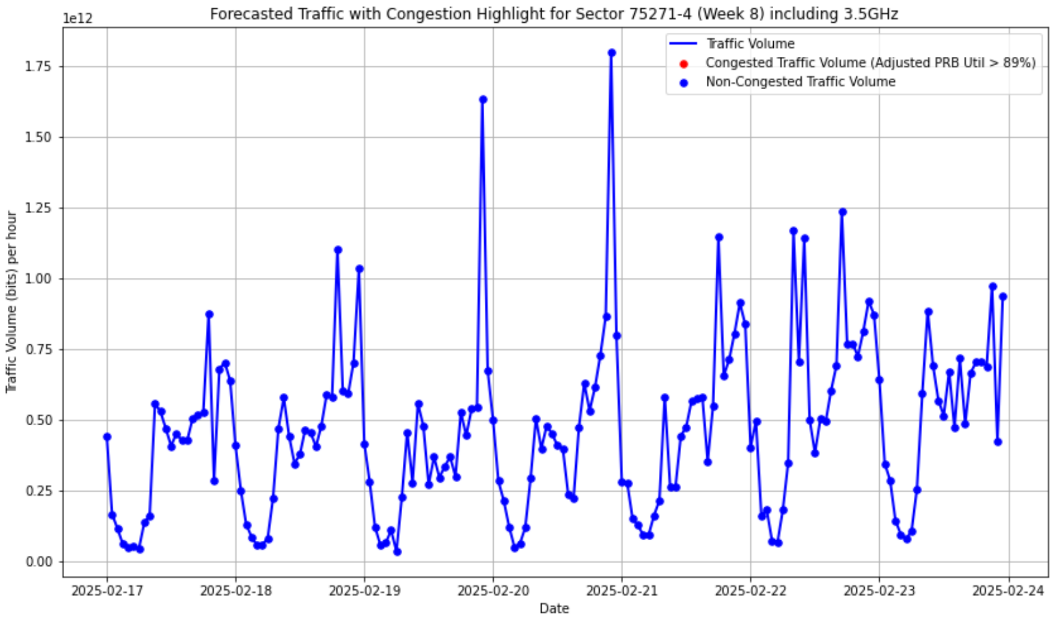


Figure 5.17: Forecasted Traffic with Congestion Highlight for Sector 75271-4 (Week 8) including 3.5GHz

The quantification and prediction of these improvements are achieved through a machine learning approach using a Lasso regression model. This model is selected for its proven accuracy in minimizing MAPE during initial testing phases, as detailed in the Methodology section. The model's training incorporate the substantial changes introduced by the 3.5 GHz spectrum implementation, particularly focusing on the additional PRBs. With the 3.5 GHz band operating at 100 MHz bandwidth and 15 kHz subcarrier spacing, the theoretical maximum of 273 additional PRBs is calculated. This parameter is designed to be adjustable in the model, allowing for future recalibration if bandwidth or subcarrier spacing configurations change.

For practical implementation, the model assumes uniform 3.5 GHz coverage across the sector. While this simplification helps in strategic capacity planning, it's worth noting that high-frequency bands (3.5 GHz) typically have more limited coverage compared to lower frequencies (700, 800 MHz). Despite this limitation, the model provides valuable insights for capacity forecasting without requiring complex cell-level analysis.

The impact of these additional resources is clearly demonstrated in Figure 5.18, which compares original PRB utilization with the model's adjusted predictions. The steeper slope of the green points (adjusted prediction) compared to the blue points (original utilization) indicates a significant improvement in the sector's ability to handle higher traffic volumes before approaching critical utilization levels.

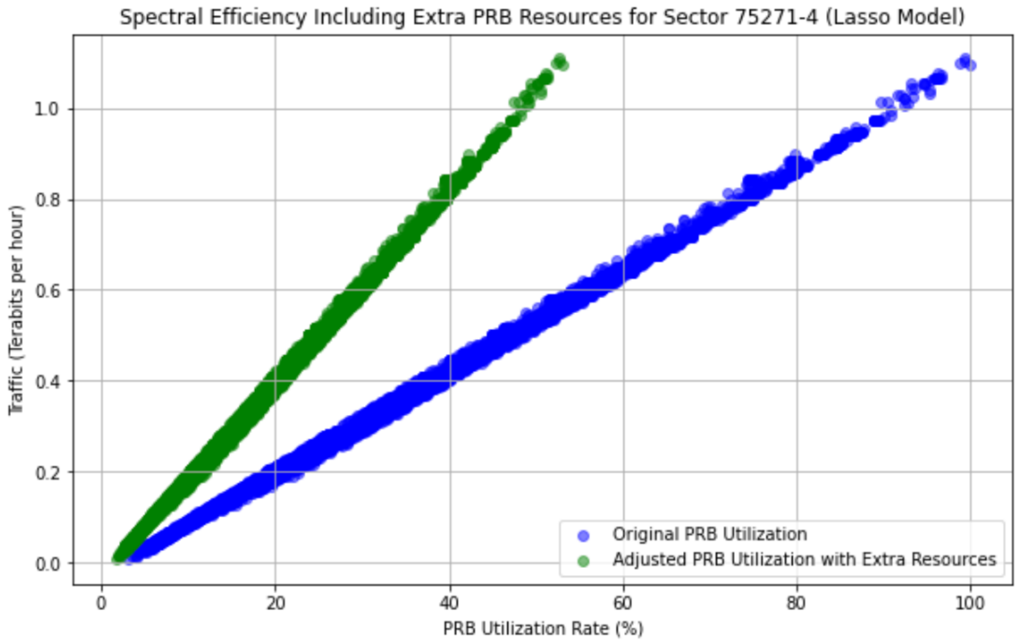


Figure 5.18: Spectral Efficiency Including Extra PRB Resources for Sector 75271-4 (Lasso Model)

The enhanced spectral efficiency revealed by the model translates directly to tangible user benefits. The sector can now manage traffic volumes that previously would have caused severe congestion, resulting in improved data rates, reduced latency, and more consistent service quality during peak periods. This improvement is particularly significant given the increasing demands on network resources. While the model and forecasts suggest complete congestion elimination for the upcoming year, actual performance may vary due to external factors such as 5G device adoption rates, changes in usage patterns, or unexpected demand spikes. Continuous monitoring will be essential to verify that real-world performance aligns with the predictions. Nevertheless, the implementation of 3.5 GHz spectrum in Sector 75271-4 has not only resolved current congestion issues but also created a robust foundation for accommodating future data growth, demonstrating the effectiveness of strategic spectrum deployment in maintaining high-quality service in the face of increasing demand.

5.5 Network Wide Congestion Analysis

The study initially focused on Sector 75271-4, selected as a representative test case due to its high traffic demand and frequent congestion. This sector was used to develop and validate machine learning models. The CDF model for traffic prediction and the Lasso regression model for PRB utilization demonstrated the best performance based on MAPE metrics. Subsequently, the methodology was scaled to encompass all macro sectors across the network, enabling network-wide congestion forecasting and illustrating that models developed for a single sector could be applied across a larger network with effective results. To further validate the scalability and reliability of this approach, the models were tested on three additional sectors chosen randomly across the network. Detailed validation results are provided in Appendix C.

The network wide analysis replicated the steps of data collection, feature engineering, model training, and evaluation for each macro sector. This approach captured the unique traffic patterns and PRB utilization characteristics of individual sectors, facilitating effective network wide capacity management. The results reflect the application of these predictive models across the entire network.

Figure 5.19 illustrates the total number of congested sectors steadily increases over time. This is because traffic demand is expected to grow, while the available resources remain the same. With increasing demand and unchanged resources, congestion naturally rises.There are fluctuations in congestion levels, with occasional dips in certain weeks. These variations are expected and can be attributed to the fact that different sectors experience peak usage at different times. Some sectors may experience temporary relief or heightened congestion due to shifts in traffic demand. This fluctuation highlights the dynamic nature of network traffic and reflects the capacity and operational limits of individual sectors.

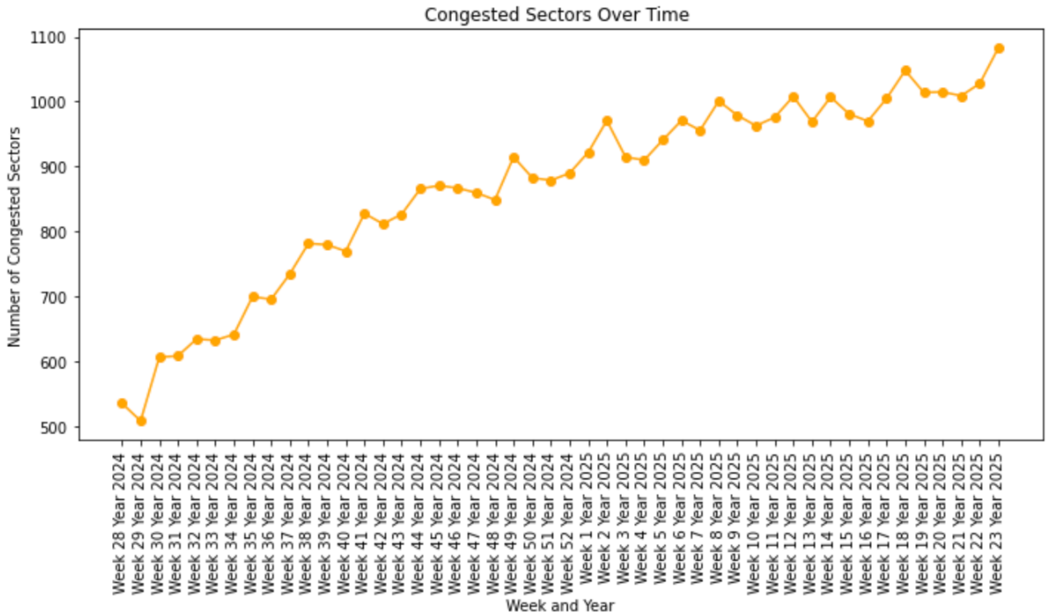


Figure 5.19: Total number of congested sectors over time across the network.

Figure 5.20 presents a comparison between predicted congested sectors and the actual number of congested sectors for an eight weeks period. The prediction line shows an increase in the number of congested sectors, suggesting worsening congestion over time without additional resources. In contrast, the actual number of congested sectors exhibits a decreasing trend, likely due to the 3.5GHz band deployment. The deployment of new 3.5GHz band has been done in Week 28 of 2024. This additional spectrum provided more resources to handle increased traffic, reducing congestion. The actual trend mirrors the pattern of the predicted trend with a negative correlation. Both trends display a drop from Week 28 to Week 29, followed by a slight increase from Week 30 onwards. This similarity in pattern, despite the opposite directions, indicates the model's ability to capture underlying traffic dynamics, even without accounting for new spectrum deployment.

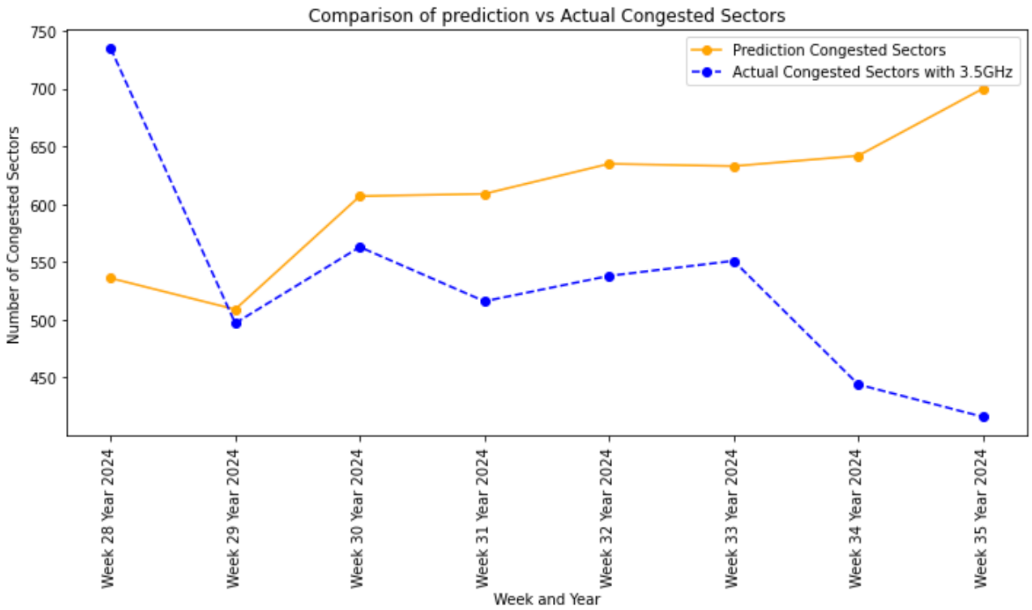


Figure 5.20: Comparison of predicted and actual congested sectors.

6 Conclusion

6.1 Conclusion

This study aimed to develop an effective capacity management model to proactively address network congestion and maintain Quality of Service for both 4G and 5G networks. The primary objective was to create a model capable of accurately forecasting sector-specific traffic volumes by leveraging historical traffic data and projected growth patterns. Using these traffic forecasts, the study then applied machine learning models to predict spectral efficiency and identify potential congestion points within the network. Additionally, the model was designed to be adaptable across multiple sectors, allowing for accurate, network-wide forecasting to manage diverse traffic demands.

To forecast future traffic loads, the study implemented a CDF model for traffic forecasting, achieving effective results in capturing network behaviors. By incorporating historical data and applying inverse CDF sampling, the model generated accurate sector-specific traffic predictions up to a year in advance. The model's performance highlighted its strength in handling diverse traffic patterns, allowing for realistic congestion forecasting at peak times. This model is therefore valuable for capacity planning, with the potential to optimize proactive resource allocation. However, while effective on a sector level.

Four distinct machine learning models were evaluated for spectral efficiency prediction: Linear Regression, Lasso Regression, Random Forest, and CatBoost. While Random Forest and CatBoost performed well on training data, they struggled with extrapolation beyond the observed data ranges. On the other hand, Linear Regression and Lasso Regression demonstrated superior performance in long-term predictions. Lasso emerged as the most effective model, due to its ability to handle extrapolation with high accuracy and low Mean Absolute Percentage Error (MAPE), making it particularly well-suited for predicting spectral efficiency and, consequently, identifying potential congestion.

A significant finding of this study was the effect of deploying the 3.5 GHz spectrum on network performance in Sector 75271-4. By simulating the model under scenarios that incorporated this new spectrum, we observed a reduction in PRB utilization rates due to the additional bandwidth, confirming improved congestion management. However, while spectral efficiency appeared higher due to reduced PRB strain, this improvement is nuanced. Since spectral efficiency is influenced by both bandwidth and traffic load, the observed gains may partly result from the increased bandwidth rather than an intrinsic efficiency gain.

After validating the models in Sector 75271-4, the methodology was extended to all macro sectors across the network, the study assessed the accuracy and scalability of the developed model across multiple macro sectors. By replicating the traffic forecasting and congestion prediction process across sectors with varying demand patterns, the model maintained a good accuracy in predicting spectral efficiency and congestion. Validation tests across different sectors revealed consistent MAPE scores, confirming that the model can be adapted for network-wide forecasting and resource planning. This scalability highlights the model's utility for broader network capacity management, enabling MNO's to anticipate congestion points across diverse sector profiles.

In conclusion, this research provided an effective approach to proactive capacity management in 4G/5G networks. By combining a CDF model for traffic forecasting with Lasso Regression for spectral efficiency prediction, the study delivered a scalable method for long-term network planning and congestion management.

6.2 Recommendation for future work

Several recommendations are proposed to enhance capacity management strategies in cellular networks.

An extension of this work could involve refining spectral efficiency predictions by calculating and forecasting at the individual cell level rather than at the sector level used in this analysis. This approach would enable a more granular understanding of traffic distribution and spectral efficiency across cells within each sector. Additionally, incorporating detailed system design parameters, such as antenna configurations, MIMO technology, and localized propagation effects, would improve the accuracy of spectral efficiency predictions, capturing the unique radio environment of each cell. By explicitly modeling these factors, future studies could enhance congestion predictions and enable more targeted capacity enhancement strategies.

Additionally, it is recommended to enhance the data measurement pipeline by retaining more detailed hourly traffic data. This would offer deeper insights into network performance across seasonal, yearly, and quarterly patterns, and enable dynamic prediction models for resource optimization. Such models could result in cost savings and greater network efficiency by using daily traffic profiles for more informed capacity planning and management. Expanding data storage capabilities would allow for more advanced analyses, leading to more informed decision making regarding data retention and operational automation in the evolving landscape of AI and machine learning.

To better capture long term trends, seasonal variations, and the impact of major events, future studies should extend the data collection period to span multiple years. Real time adaptation is another priority, where systems should continuously update predictions based on live network data, allowing for more dynamic and responsive capacity management. Developing systems that can suggest the optimal timing and locations for capacity upgrades based on predictive analytics would also improve resource allocation and network planning. Finally, enhancing the interpret ability of machine learning models is crucial for providing network operators with clear insights into the factors driving congestion predictions, which would improve decision making and operational strategies.

Bibliography

[1] B. Lindoff. "Using a direct conversion receiver in EDGE terminals-a new DC offset compensation algorithm". In: *11th IEEE International Symposium on Personal Indoor and Mobile Radio Communications. PIMRC 2000. Proceedings (Cat. No.00TH8525)*. Volume 2. 2000, 959–963 vol.2. DOI: 10.1109/PIMRC.2000.881564 (cited on page 1).

[2] Baha Mkheimer. "Evaluation and optimisation of GSM network in Jenien City, Palestine". In: (2012). DOI: 10.1504/IJMNDI.2012.054458 (cited on page 1).

[3] Syed Ismail Shah. "UMTS: High Speed Packet Access (HSPA) Technology". In: *2008 IEEE International Networking and Communications Conference*. 2008, pages 2–2. DOI: 10.1109/INCC.2008.4562669 (cited on page 1).

[4] 3GPP. *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)*. 3GPP TS 36.300. Version 0.3.3. 2014. URL: https://www.3gpp.org/specifications-technologies/releases/release-8 (cited on pages 1, 11).

[5] Kan Zheng, Bin Fan, Jianhua Liu, Yicheng Lin, and Wenbo Wang. "Interference coordination for OFDM-based multihop LTE-advanced networks". In: *IEEE Wireless Communications* 18.1 (2011), pages 54–63. DOI: 10.1109/MWC.2011.5714026 (cited on page 1).

[6] International Telecommunication Union. *IMT Vision: Framework and overall objectives of the future development of IMT for 2020 and beyond*. Recommendation ITU-R M.2083-0. RECOMMENDATION ITU-R M.2083-0. 2015. URL: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf (cited on page 1).

[7] 3GPP. *Service requirements for cyber-physical control applications in vertical domains*. ETSI TS 122 104. Version 16.5.0. 2020. URL: https://www.etsi.org/deliver/etsi_ts/122100_122199/122104/16.05.00_60/ts_122104v160500p.pdf (cited on page 1).

[8] 3GPP. *Digital Enhanced Cordless Telecommunications (DECT); Study on URLLC use cases of vertical industries for DECT evolution and DECT-2020*. ETSI TR 103 515. Version 1.1.1. 2018. URL: https://www.etsi.org/deliver/etsi_tr/103500_103599/103515/01.01.01_60/tr_103515v010101p.pdf (cited on page 1).

[9] Ijaz Mansoor Quershi Qazi Kamal Ud Din Arshad A. Kashif. "A Review on the Evolution of Cellular Technologies". In: (2019). DOI: 10.1109/IBCAST.2019.8667173 (cited on page 1).

[10] SoftBank News. *RAN (Radio Access Network)*. Technical report. 2022. URL: https://www.softbank.jp/en/sbnews/entry/20220630_01 (cited on page 2).

[11] Mohammad Asif Habibi, Meysam Nasimi, Bin Han, and Hans D. Schotten. "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System". In: *IEEE Access* 7 (2019), pages 70371–70421. DOI: 10.1109/ACCESS.2019.2919657 (cited on page 3).

[12] Mianxiong Dong Zhenyu Zhou. *Energy Efficiency and Spectral Efficiency Tradeoff in Device-to-Device (D2D) Communications*. Technical report. 2014. URL: https://arxiv.org/abs/1407.1556 (cited on page 3).

[13] Shribala Nagul. "A review on 5G modulation schemes and their comparisons for future wireless communications". In: *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*. 2018, pages 72–76. DOI: 10.1109/SPACES.2018.8316319 (cited on page 3).

[14] Tara Ali Yahiya. *Understanding LTE and its Performance*. Technical report. 2011. DOI: 10.1007/978-1-4419-6457-1 (cited on page 3).

[15] Alberto Martínez Alba, Shakthivelu Janardhanan, and Wolfgang Kellerer. "Enabling Dynamically Centralized RAN Architectures in 5G and Beyond". In: *IEEE Transactions on Network and Service Management* 18.3 (2021), pages 3509–3526. DOI: 10.1109/TNSM.2021.3071975 (cited on page 3).

[16] Wonil Roh, Ji-Yun Seol, Jeongho Park, Byunghwan Lee, Jaekon Lee, Yungsoo Kim, Jaeweon Cho, Kyungwhoon Cheun, and Farshid Aryanfar. "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results". In: *IEEE Communications Magazine* 52.2 (2014), pages 106–113. DOI: 10.1109/MCOM.2014.6736750 (cited on page 3).

[17] Long Jiao, Ning Wang, Pu Wang, Amir Alipour-Fanid, Jie Tang, and Kai Zeng. "Physical Layer Key Generation in 5G Wireless Networks". In: *IEEE Wireless Communications* 26.5 (2019), pages 48–54. DOI: 10.1109/MWC.001.1900061 (cited on page 3).

[18] Jose Ordonez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J. Ramos-Munoz, Javier Lorca, and Jesus Folgueira. "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges". In: *IEEE Communications Magazine* 55.5 (2017), pages 80–87. DOI: 10.1109/MCOM.2017.1600935 (cited on page 3).

[19] Ramy Mohamed, Sofiane Zemouri, and Christos Verikoukis. "Performance Evaluation and Comparison between SA and NSA 5G Networks in Indoor Environment". In: *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. 2021, pages 112–116. DOI: 10.1109/MeditCom49071.2021.9647621 (cited on page 3).

[20] Tech. Rep Ericsson. *Ericsson mobility report*. Technical report. 2023. URL: https://www.ericsson.com/4ae12c/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-november-2023.pdf (cited on page 4).

[21] Arthur D. Little. *Data growth in Europe*. 2023. URL: https://www.adlittle.com/sites/default/files/reports/ADL_Data_growth_Europe_2023.pdf (cited on page 4).

[22] *KPN, T-Mobile and VodafoneZiggo acquire frequencies in Dutch mobile communications auction"*. Technical report. 2020. URL: https://www.government.nl/latest/news/2020/07/21/kpn-t-mobile-and-vodafoneziggo-acquire-frequencies-in-dutch-mobile-communications-auction (cited on page 5).

[23] Darius Chmieliauskas and Darius Guršnys. "LTE Cell Traffic Grow and Congestion Forecasting". In: *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*. 2019, pages 1–5. DOI: 10.1109/eStream.2019.8732145 (cited on pages 5, 15, 16, 18).

[24] KPN. *"4G cellphone coverage in the Netherlands"*. Technical report. 2018. URL: https://www.reddit.com/r/MapPorn/comments/9ripcy/4g_cellphone_coverage_in_the_netherlands_kpn_2018/?rdt=49654 (cited on page 6).

[25] KPN. *KPN at a Glance*. Technical report. 2023. URL: https://www.overons.kpn/en/the-company/kpn-at-a-glance (cited on page 6).

[26] KPN. *KPN mobile network tested as the best with the highest score worldwide*. Technical report. Feb,2024. URL: https://www.overons.kpn/nieuws/en/umlaut-kpn-mobile-network-tested-as-the-best-with-the-highest-score-worldwide/ (cited on page 6).

[27] Eric Smeitink. *Strategy, Spectrum and 5G*. Technical report. 2024. URL: https://brightspace.tudelft.nl/d2l/le/content/595916/viewContent/3653045/View (cited on pages 6, 7).

[28] KPN. *KPN to Roll out LTE in February 2013*. Technical report. 2013. URL: https://telecomgrid.com/news/kpn-to-roll-out-lte-in-february-2013/ (cited on page 7).

[29] KPN. *KPN to Roll out 5G in July 2020*. Technical report. 2020. URL: https://nltimes.nl/2020/07/28/kpn-t-mobile-launching-5g-networks-today (cited on page 7).

[30] Jari Salo. *Analysis of LTE Radio Load and User Throughput*. Technical report. 2017. DOI: 10.5121/ijcnc.2017.9603 (cited on page 11).

[31] 3GPP. *5G; NR; Physical channels and modulation.* Technical report. TS 38.211 version 16.3.0 Release 16. 2020. URL: https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/16.03.00_60/ts_138211v160300p.pdf (cited on page 12).

[32] M. Raftopoulou. *Resource management in wireless networks.* Technical report. 2024. URL: https://research.tudelft.nl/en/publications/resource-management-in-wireless-networks (cited on page 13).

[33] Qualcomm Technologies. "Delivering on the LTE Advanced promise". In: (2016). URL: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/delivering_on_the_lte_advanced_promise.pdf (cited on page 13).

[34] Tomic and Predrag Ivanis. *"Predictive Capacity Planning for Mobile Networks—ML Supported Prediction of Network Performance and User Experience Evolution.* Technical report. Electronics 11, no. 4: 626. 2022. URL: https://doi.org/10.3390/electronics11040626 (cited on pages 13, 15–17).

[35] Neda Tavakoli Sima Siami Namini and Akbar Siami Namin. *Comparison of ARIMA and LSTM in Forecasting Time Series.* Technical report. 2018. URL: DOI:%2010.1109/ICMLA.2018.00227 (cited on page 14).

[36] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. "A Comparison of ARIMA and LSTM in Forecasting Time Series". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA).* 2018, pages 1394–1401. DOI: 10.1109/ICMLA.2018.00227 (cited on page 14).

[37] Vaia I. Kontopoulou, Athanasios D. Panagopoulos, Ioannis Kakkos, and George K. Matsopoulos. "A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks". In: *Future Internet* 15.8 (2023). ISSN: 1999-5903. DOI: 10.3390/fi15080255. URL: https://www.mdpi.com/1999-5903/15/8/255 (cited on page 14).

[38] 3GPP. *Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN).* Technical report. TS 32.450 version 15.2.0 Release 15. 2021. URL: https://www.etsi.org/deliver/etsi_ts/132400_132499/132450/15.02.00_60/ts_132450v150200p.pdf (cited on page 14).

[39] 3GPP. *5G; Management and orchestration; 5G end to end Key Performance Indicators (KPI).* Technical report. TS 128 554 Version 16.7.0. 2021. URL: https://www.etsi.org/deliver/etsi_ts/128500_128599/128554/16.07.00_60/ts_128554v160700p.pdf (cited on page 14).

[40] Luong-Vy Le, Do Sinh, Li-Ping Tung, and Bao-Shuh Paul Lin. "A practical model for traffic forecasting based on big data, machine-learning, and network KPIs". In: *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC).* 2018, pages 1–4. DOI: 10.1109/CCNC.2018.8319255 (cited on pages 15, 17).

[41] Pedro Torres, Paulo Marques, Hugo Marques, Rogério Dionísio, Tiago Alves, Luis Pereira, and Jorge Ribeiro. "Data analytics for forecasting cell congestion on LTE networks". In: *2017 Network Traffic Measurement and Analysis Conference (TMA).* 2017, pages 1–6. DOI: 10.23919/TMA.2017.8002917 (cited on pages 15, 17).

[42] 3GPP. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures.* Technical report. TS 136 213 version 8.3.0 Release 8. 2008. URL: https://www.etsi.org/deliver/etsi_ts/136200_136299/136213/08.03.00_60/ts_136213v080300p.pdf (cited on page 15).

[43] 3GPP. *5G; NR; Physical layer procedures for data.* Technical report. TS 38.214 version 15.9.0 Release 15. 2020. URL: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/15.09.00_60/ts_138214v150900p.pdf (cited on page 15).

[44] C. Gijón, M. Toril, S. Luna-Ramírez, M. L. Marí Altozano, and J. M. Ruiz-Avilés. *Long-Term Data Traffic Forecasting for Network Dimensioning in LTE with Short Time Series.* Technical report. 2021. URL: https://doi.org/10.3390/electronics10101151 (cited on page 17).

[45] J. A. Bastos. *Forecasting the capacity of mobile networks. Telecommunication Systems.* Technical report. 2019. URL: https://doi.org/10.1007/s11235-019-00556-w (cited on page 17).

[46] Rongpeng Li, Zhifeng Zhao, Jianchao Zheng, Chengli Mei, Yueming Cai, and Honggang Zhang. "The Learning and Prediction of Application-Level Traffic Data in Cellular Networks". In: *IEEE Transactions on Wireless Communications* 16.6 (2017), pages 3899–3912. DOI: 10.1109/TWC.2017.2689772 (cited on page 17).

[47] Yuxiu Hua, Zhifeng Zhao, Zhiming Liu, Xianfu Chen, Rongpeng Li, and Honggang Zhang. "Traffic Prediction Based on Random Connectivity in Deep Learning with Long Short-Term Memory". In: *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. 2018, pages 1–6. DOI: 10.1109/VTCFall.2018.8690851 (cited on page 17).

[48] Rob Hyndman and Andrey Kostenko. "Minimum Sample Size Requirements for Seasonal Forecasting Models". In: *Foresight: The International Journal of Applied Forecasting* 6 (Feb. 2007), pages 12–15 (cited on page 18).

[49] Omkaresh Kulkarni and Adnan Burhanpurwala. "A Survey of Advancements in DBSCAN Clustering Algorithms for Big Data". In: *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*. 2024, pages 106–111. DOI: 10.1109/PARC59193.2024.10486339 (cited on page 21).

[50] "Introduction". In: *Introduction to Time Series and Forecasting*. Edited by Peter J. Brockwell and Richard A. Davis. New York, NY: Springer New York, 2002, pages 1–44. ISBN: 978-0-387-21657-7. DOI: 10.1007/0-387-21657-X_1. URL: https://doi.org/10.1007/0-387-21657-X_1 (cited on page 23).

[51] Guido Van Rossum and Fred L. Drake. "Python 3 Reference Manual". In: (2009) (cited on page 23).

[52] Rob J. Hyndman and George Athanasopoulos. "Forecasting: Principles and Practice". In: (2018). 2nd ed (cited on page 23).

[53] Bima Wahyu Maulana, Satryo Budi Utomo, Khairul Anam, and Iwanah Bilfaqih. "Analysis of Movement Patterns Based on Electroencephalograph Signals Using Pearson Correlation Coefficient". In: *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. 2023, pages 593–597. DOI: 10.1109/ISITIA59021.2023.10220971 (cited on page 28).

[54] Umlaut. *The 2019 Mobile Network Test in the Netherlands.* Technical report. 2019. DOI: https://www.simpel.nl/files/PDF/2019-network-test-netherlands-umlaut.pdf (cited on page 37).

[55] KPN. *Specification congestion calculation 4G and 5G.* Technical report. 2023 (cited on page 37).

[56] Johnson Kolluri, Vinay Kumar Kotte, M.S.B. Phridviraj, and Shaik Razia. "Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method". In: *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. 2020, pages 934–938. DOI: 10.1109/ICOEI48184.2020.9142992 (cited on page 63).

[57] Gengjian Xue, Li Song, Jun Sun, and Meng Wu. "Foreground estimation based on robust linear regression model". In: *2011 18th IEEE International Conference on Image Processing*. 2011, pages 3269–3272. DOI: 10.1109/ICIP.2011.6116368 (cited on page 64).

[58] P. Suresh Kumar, Anisha Kumari K, Subhashree Mohapatra, Bighnaraj Naik, Janmenjoy Nayak, and Manohar Mishra. "CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages". In: *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*. 2021, pages 1–6. DOI: 10.1109/ODICON50556.2021.9428943 (cited on page 66).

A

Data analysis

Table A.1: Primary input variables for predictive models

<table>
<tr><th>Input Variable</th><th>Definition</th></tr>
<tr><td>Time stamp</td><td>A specific point in time, typically recorded in a standardized format (e.g., YYYY-MM-DD HH:MM:SS), indicating when the data was collected or the event occurred.</td></tr>
<tr><td>Site number</td><td>A unique identifier assigned to each physical location where network equipment (such as base stations) is installed, allowing for geographical tracking and management of network resources.</td></tr>
<tr><td>Sector number</td><td>An identifier for a specific directional antenna or set of antennas within a site, typically covering a 120-degree arc. Most sites have three sectors for complete 360-degree coverage.</td></tr>
<tr><td>Cell (frequency band)</td><td>Refers to the specific frequency range used for communication in a particular sector. The bands include 700, 800, 900, 1800, 2100, 2600 FDD, and 2600 TDD MHz.</td></tr>
<tr><td>Downlink Traffic Volume</td><td>The total amount of data transmitted from the network to end-user devices, calculated as follows:
• In LTE:
$$\begin{aligned}\text{Total DL Traffic} = &\sum \text{L.Traffic.DL.SCH.QPSK.TB.bits}\\ &+ \sum \text{L.Traffic.DL.SCH.16QAM.TB.bits}\\ &+ \sum \text{L.Traffic.DL.SCH.64QAM.TB.bits}\\ &+ \sum \text{L.Traffic.DL.SCH.256QAM.TB.bits}\end{aligned}$$
• In NR:
$$\text{Total DL Traffic} = \sum \text{N.ThpVol.DL.Cell(kbit)}$$</td></tr>
<tr><td>PRB Utilization Rate</td><td>Indicating the efficiency of resource usage, calculated as follows:
• In LTE:
$$\text{PRB Utilization Rate} = \frac{\sum \text{L.ChMeas.PRB.DL.Used.Avg}}{\sum \text{L.ChMeas.PRB.DL.Avail}}$$
• In NR:
$$\text{PRB Utilization Rate} = \frac{\sum \text{N.PRB.DL.Used.Avg}}{\sum \text{N.PRB.DL.Avail.Avg}}$$</td></tr>
<tr><td>User DL average throughput</td><td>Represents the average data transfer rate experienced by users in the downlink direction. It's typically measured in bits per second (bps) or a variation thereof (Kbps, Mbps, etc.).</td></tr>
</table>

The counters mean:

- **L.ChMeas.PRB.DL.Avail:** Number of available downlink PRBs.
- **L.ChMeas.PRB.DL.Used.Avg:** The average number of used PDSCH PRBs.
- **L.Traffic.DL.SCH.QPSK.TB.bits:** Number of bits of TBs initially transmitted on the downlink SCH in QPSK modulation scheme.
- **L.Traffic.DL.SCH.16QAM.TB.bits:** Number of bits of TBs initially transmitted on the downlink SCH in 16QAM modulation scheme.
- **L.Traffic.DL.SCH.64QAM.TB.bits:** Number of bits of TBs initially transmitted on the downlink SCH in 64QAM modulation scheme.
- **L.Traffic.DL.SCH.256QAM.TB.bits:** Number of bits of TBs initially transmitted on the downlink SCH in 256QAM modulation scheme.
- **N.ThpVol.DL.Cell(kbit):** Total volume of downlink data sent at the MAC layer in a cell.
- **N.PRB.DL.Avail.Avg:** Number of available downlink PRBs.
- **N.PRB.DL.Used.Avg:** The average number of used PDSCH PRBs.

B

Machine Learning models

Machine learning, a subset of artificial intelligence, enables systems to learn and adapt without explicit programming. It focuses on developing computer programs that evolve when exposed to new data. These algorithms use data to detect patterns and adjust actions accordingly. Machine learning techniques can be broadly categorized into supervised, unsupervised, and reinforcement learning.

- Supervised Learning: Algorithms apply learned knowledge from labeled data to predict outcomes on new, unseen data.
- Unsupervised Learning: Algorithms discover hidden patterns or intrinsic structures in input data without labeled outcomes.
- Reinforcement Learning: Algorithms learn optimal policies by interacting with the environment, using feedback signals from actions to guide learning.

In the context of forecasting, the primary goal is to identify a model that can accurately fit past measured data from the network and generalize well to future observations. Models can vary from linear time series models to state-space and non-linear models. This study focuses on well-known regression algorithms, including Linear Regression (LR), Random Forest (RF), Lasso Regression, and CatBoost.

The choice of models is driven by their ability to capture different relationships between features:

- Linear and Lasso Regression are used to explore linearity between traffic and PRB utilization.
- Random Forest (RF) can capture both linear and non-linear boundaries through its ensemble of decision trees.
- CatBoost is advanced ensemble techniques capable of capturing complex patterns, often enhancing prediction accuracy.

The Machine Learning Technique is trained on a given dataset and tested on new samples to achieve good estimation performance, measured by low MAPE values in both phases. This process aims for generalization, where a well-generalized model performs well on new data within the same domain, addressing challenges like over-fitting and under-fitting FigureB.1. Fitting refers to how well the model approximates the mapping between input and output variables [56]:

- Over-fitting occurs when the model fits the training data too well, including noise, leading to poor performance on new data.
- Under-fitting occurs when the model fails to fit the training data or generalize to new data, resulting in poor performance even during training.
- The optimal model strikes a balance between over-fitting and under-fitting, learning the data pattern without memorizing noise.

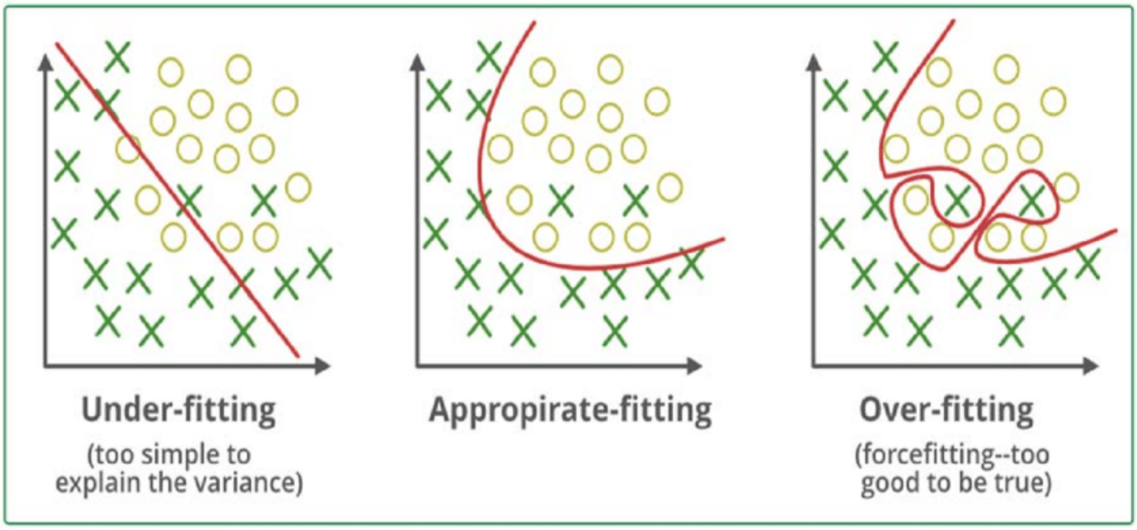


Figure B.1: Under fitting and overfitting cases in ML.

Linear Regression (LR):

Linear Regression identifies the statistical relationship between variables by fitting a linear equation to observed data. The prediction of PRB utilization ($\hat{y}_i$) is based on input features X_i, such as traffic volume, hour of the day, day of the month, and weekday [57]. In linear regression, the relationship between the input features and the target variable is modeled as a linear combination of the input features:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot \text{traffic}_i + \beta_2 \cdot \text{hour}_i + \beta_3 \cdot \text{day}_i + \beta_4 \cdot \text{weekday}_i \tag{B.1}$$

Where:

- $\hat{y}_i$ is the predicted PRB utilization for the i-th data point.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients (weights) for the corresponding features.
- $\text{traffic}_i, \text{hour}_i, \text{day}_i, \text{weekday}_i$ are the feature values for the i-th data point.

Linear regression estimates the coefficients by minimizing the sum of squared errors (SSE). The objective is to find the values of $\beta_0, \beta_1, \ldots, \beta_4$ that minimize the following cost function:

$$J(\beta_0, \beta_1, \ldots, \beta_4) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{B.2}$$

Where:

- y_i is the actual PRB utilization for the i-th data point.
- $\hat{y}_i$ is the predicted PRB utilization as defined in Equation (B.1).
- n is the total number of data points.

The solution to this optimization problem provides the coefficients that best fit the data.

LR will help understanding if there's a straightforward linear relationship between traffic volume and PRB utilization. It serves as a baseline model and can reveal if more complex models are necessary.

Lasso Regression (L1 Regularization):
Lasso regression is a type of linear model that adds an L_1 regularization term to the cost function. This helps prevent overfitting by shrinking some of the coefficients towards zero, effectively performing feature selection. The Lasso regression model predicts the PRB utilization using the same linear form as in linear regression:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot \text{traffic}_i + \beta_2 \cdot \text{hour}_i + \beta_3 \cdot \text{day}_i + \beta_4 \cdot \text{weekday}_i \tag{B.3}$$

However, unlike linear regression, Lasso regression introduces an additional regularization term to the cost function. In Lasso regression, the objective is to minimize the sum of squared errors, with an additional L_1 regularization term that penalizes the absolute size of the coefficients:

$$J(\beta_0, \beta_1, \ldots, \beta_4) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{B.4}$$

Where:

- λ is the regularization parameter that controls the strength of the penalty. A higher λ leads to stronger regularization.
- $\sum_{j=1}^{p} |\beta_j|$ is the L_1 norm of the coefficients, which shrinks the coefficients towards zero.
- p is the number of features (in this case, $p = 4$).

The regularization term encourages sparsity in the model by forcing some coefficients to become exactly zero, which can be useful for feature selection.

The main advantage of Lasso Regression is its ability to perform automatic feature selection by shrinking some coefficients to exactly zero, effectively removing less important features from the model. This can lead to simpler, more interpretable models, especially when there are many irrelevant or redundant features. However, Lasso can be sensitive to the choice of λ and may struggle when features are highly correlated.

Random Forest (RF)
RF technique is an ensemble learning model that improves estimation performance by combining predictions from multiple learning algorithms. RF is based on building multiple trees, called Decision tree (DT). The DT models the problem in a tree structure, as illustrated on Figure B.2. from a root node the algorithm takes several decisions to reach the leaf nodes. Each leaf node is an estimation value. And, each internal node split refers to the taken decision. The individual decision tree is sensitive to data and prone to over fitting. For that, RF combines at training phase many individual decision trees by adding randomness in order to select the training subset and the feature in each node.

The RF algorithm includes the following steps:

1. Draw multiple trees using bootstrap sampling. The number of trees (*n_estimators*) is a user defined hyper-parameter.
2. For each node split during tree construction, select the best split among a random subset of predictors instead of choosing from all predictors. The number of features (*max_features*) controls the randomization strength.
3. Estimate new data by averaging the estimations of the constructed trees.

Each tree in the forest incorporates traffic volume. Each root contains a random subset of metrics, and each internal node represents a decision about these. Leaf nodes hold the PRB utilization estimation. The final PRB utilization estimation is obtained by averaging the estimations from all trees. RF is known for its robustness against over fitting due to the introduced randomness, its simplicity, and its efficiency on large datasets with numerous input features. Hyper-parameters such as *n_estimators*, *max_features*, *max_depth*, *min_samples_split*, and *min_samples_leaf* are critical

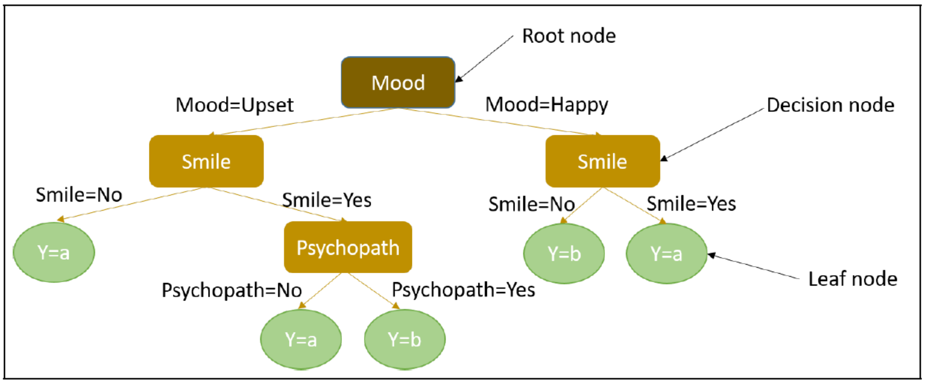


Figure B.2: Example of decision tree (DT).

for optimal model performance.

Gradient Boosting (CatBoost) The aim is to predict the PRB utilization ($\hat{y}_i$) based on several features such as traffic volume, hour of the day, day of the month, weekday. CatBoost is an ensemble of decision trees that are trained sequentially with each new model correcting the errors of the previous ones, to improve overall prediction accuracy [58]. The final prediction, $\hat{y}_i$ (PRB utilization) for the i-th data point, is given by the sum of predictions from all decision trees in the ensemble:

$$\hat{y}_i = F_M(X_i) = \sum_{m=1}^{M} \gamma_m T_m(X_i) \tag{B.5}$$

Where:

- $F_M(X_i)$ is the final predicted PRB utilization for the i-th data point after M trees.
- $T_m(X_i)$ is the prediction from the m-th decision tree for the input features $X_i = (\text{traffic}, \text{hour}, \text{day}, \text{weekday})$.
- γ_m is the learning rate (step size) for the m-th tree.
- M is the total number of trees in the ensemble.

Each tree $T_m(X_i)$ represents a non linear mapping from the feature space to the target variable.
The training process in CatBoost follows a sequential approach.
Initially, CatBoost starts with a constant prediction, which is often the mean PRB utilization across the training data:

$$\hat{y}_i^{(0)} = \text{mean}(y) \tag{B.6}$$

For each boosting iteration m, the algorithm calculates the residuals, which represent the difference between the actual PRB utilization y_i and the predicted PRB utilization $\hat{y}_i^{(m-1)}$ from the previous iteration:

$$r_i^{(m)} = y_i - \hat{y}_i^{(m-1)} \tag{B.7}$$

Where $r_i^{(m)}$ is the residual for the i-th data point at iteration m, and y_i is the actual PRB utilization for that data point.

The next decision tree $T_m(X_i)$ is trained to predict these residuals, effectively learning from the errors made by the previous trees:

$$T_m(X_i) \approx r_i^{(m)} \tag{B.8}$$

Here, $T_m(X_i)$ is the prediction of the residual $r_i^{(m)}$ based on the input features X_i = (traffic,hour, day,and weekday).

After training the new tree, the model's prediction is updated by adding the prediction from the current tree, weighted by the learning rate γ:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \gamma T_m(X_i) \tag{B.9}$$

Where γ controls how much the new tree's prediction affects the overall model update.

After all M trees have been trained, the final prediction for PRB utilization for the i-th data point is the sum of all individual tree predictions:

$$\hat{y}_i = \sum_{m=1}^{M} \gamma T_m(X_i) \tag{B.10}$$

This equation represents the overall CatBoost model, where each tree $T_m(X_i)$ contributes to the final prediction based on the input features $X_i = (\text{traffic}, \text{hour}, \text{day}, \text{weekday})$.

The model handles non linear features efficiently, often achieves high accuracy. However, it can be prone to overfitting if not properly tuned, may be computationally intensive, and is less interpretable than simpler models.

Model performance is evaluated by comparing estimated values with actual measurements. Optimal performance depends on selecting the appropriate hyperparameters, which is detailed in the hyperparameter tuning section. To ensure robustness, the models' ability to generalize estimation error to unseen data is assessed, providing a reliable measure of predictive accuracy on new data. This approach ensures the models' effectiveness in real world scenarios.

C

Validation of Congestion Forecasting Models Across Multiple Sectors

The purpose of this appendix is to validate the scalability and reliability of the congestion forecasting models developed in this study. While the initial model development and validation are conducted using data from Sector 75271-4, the methodology is subsequently expanded to a network-wide application. To verify that the models provide consistent and reliable results across different sectors, three additional sectors—1001-2, 9981-3, and 10161-1 are randomly selected for further testing. Each of these sectors presents distinct traffic characteristics:

- Sector 1001-2 and Sector 9981-3: These sectors have relatively low average traffic loads and do not exhibit frequent congestion. Testing in these sectors assesses the model's accuracy under typical, low-demand conditions.
- Sector 10161-1: This sector experiences higher average traffic loads and exhibits peak congestion hours. Testing in this sector evaluates the model's ability to accurately forecast congestion in high demand environments.

This selection of sectors provides a balanced validation across both low- and high-demand conditions, allowing for a thorough assessment of the models' performance and adaptability across varying traffic patterns.

Table C.1 summarizes the performance metrics for each sector, comparing results for both the Lasso regression model (predicting PRB utilization) and the CDF model (forecasting traffic). These metrics demonstrate consistency across sectors, with MAPE values and R^2 scores close to those achieved in the initial test case, Sector 75271-4, which had a MAPE of 13.6% for the CDF model and 11.25% for the Lasso regression model after outlier removal.

Table C.1: Validation Results for Additional Sectors

Sector	Model	R^2	MAPE (%)	CDF Model MAPE (%)
1001-2	Lasso Regression	0.9229	9.69	12.13
9981-3	Lasso Regression	0.9203	8.83	14.0
10161-1	Lasso Regression	0.9439	10.68	11.68

The validation results demonstrate effective traffic forecasting through the CDF model, which maintains MAPE values within a comparable range across all tested sectors. This consistency confirms that the CDF model effectively captures variations in traffic patterns throughout the selected sectors. Additionally, the Lasso regression model achieves similar R^2 values and low MAPE percentages across all sectors, with performance closely aligning with the initial results from Sector 75271-4. This alignment supports the scalability and reliability of the model for predicting PRB utilization across various sectors. The validation in both low-demand sectors (1001-2 and 9981-3) and a high-demand sector (10161-1) further indicates that the models are adaptable to different traffic conditions and congestion patterns, which is essential for effective network-wide congestion forecasting.

The following figures illustrate the forecasted traffic volumes and predicted PRB utilization for each of the selected sectors (1001-2, 9981-3, and 10161-1), showing both the CDF model's traffic forecast and the Lasso regression model's PRB utilization predictions.

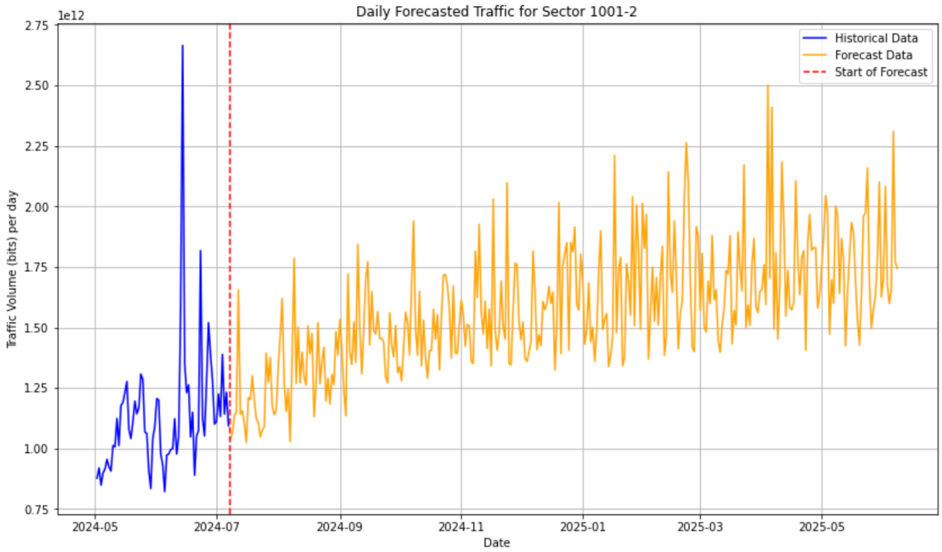


Figure C.1: Daily Forecasted Traffic for Sector 1001-2

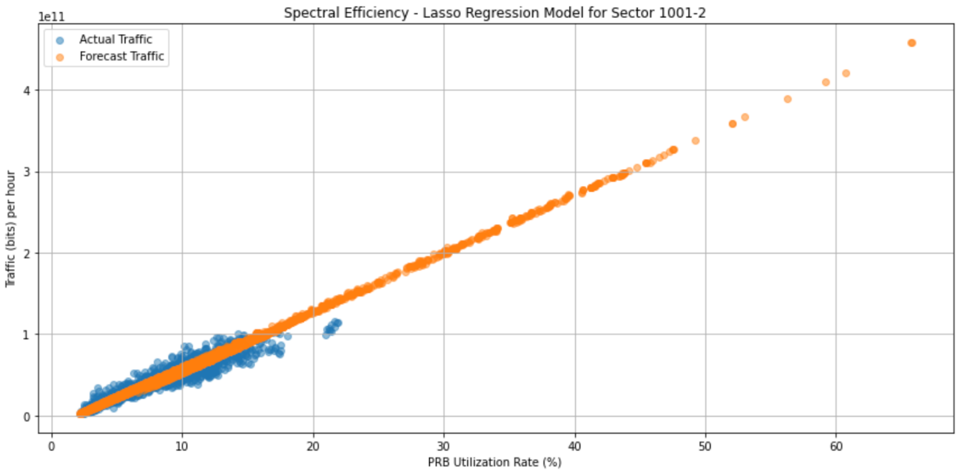


Figure C.2: Spectral Efficiency - Lasso Regression Model for Sector 1001-2

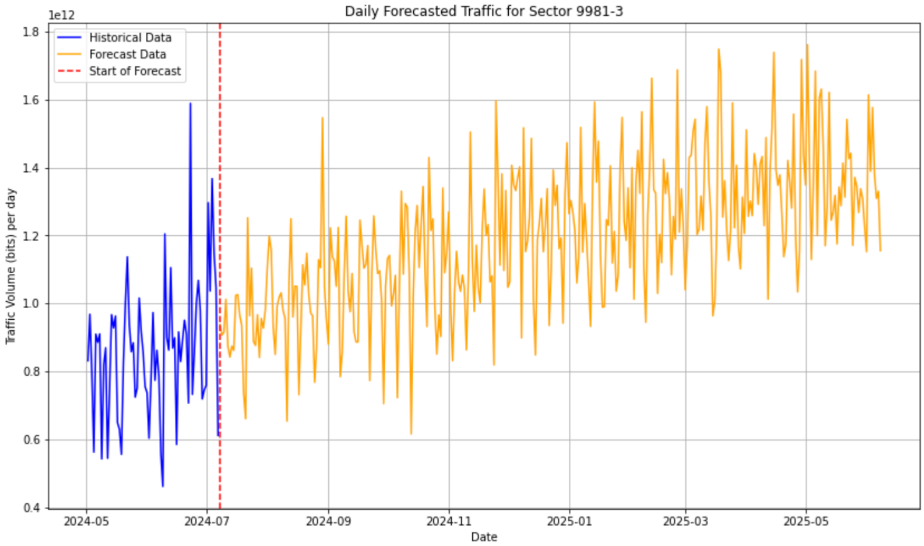


Figure C.3: Daily Forecasted Traffic for Sector 9981-3

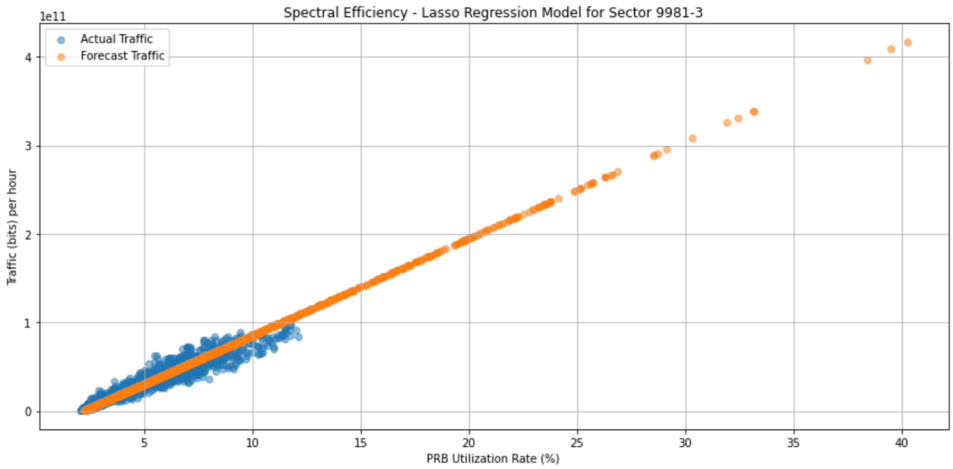


Figure C.4: Spectral Efficiency - Lasso Regression Model for Sector 9981-3

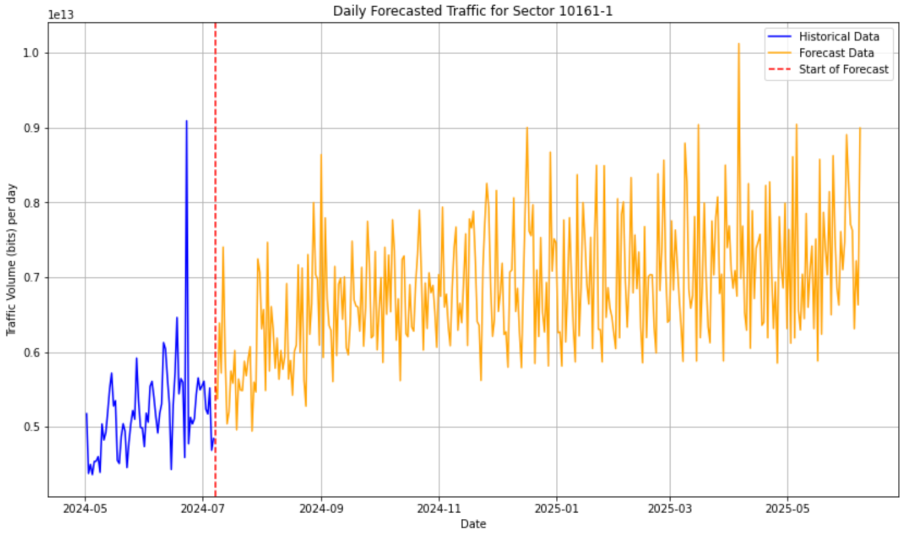


Figure C.5: Daily Forecasted Traffic for Sector 10161-1

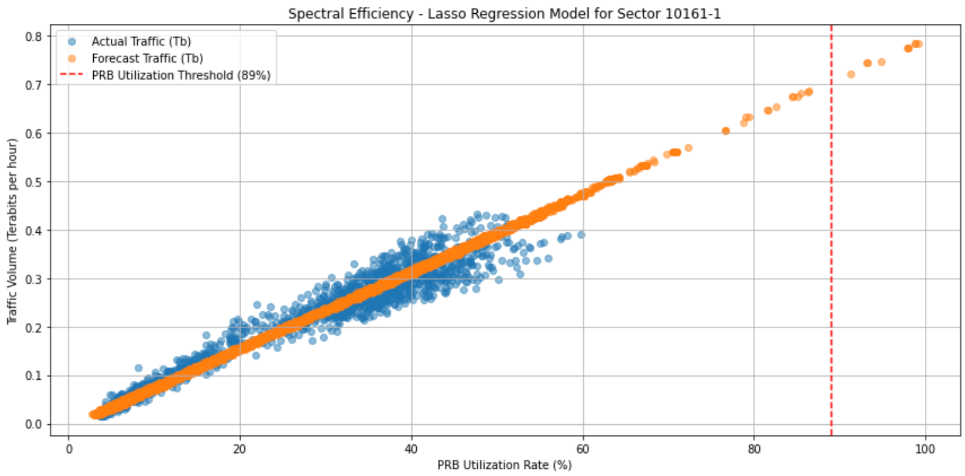


Figure C.6: Spectral Efficiency - Lasso Regression Model for Sector 10161-1