TU WIEN Informatics

Time Series Forecasting and Clustering Techniques for Cellular Network Performance for Predictive Load Management

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Masterstudium Data Science

eingereicht von

Sebastian Dolezel, BSc.
Matrikelnummer 01126006

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Markus Rupp
Mitwirkung: Dipl.-Ing. Lukas Eller

Wien, 28. August 2024

Sebastian Dolezel

Markus Rupp

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Time Series Forecasting and Clustering Techniques for Cellular Network Performance for Predictive Load Management

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Master's Programme Data Science

by

Sebastian Dolezel, BSc.
Registration Number 01126006

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Markus Rupp
Assistance: Dipl.-Ing. Lukas Eller

Vienna, August 28, 2024

Sebastian Dolezel

Markus Rupp

Erklärung zur Verfassung der Arbeit

Sebastian Dolezel, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 28. August 2024

Sebastian Dolezel

Acknowledgements

I want to thank all my dear ones who showed incredible patience and understanding during my studies. Your support means the world to me. A special thanks also goes to my supervisors for their unwavering guidance, insightful answers to my many questions, and valuable tips that greatly contributed to my progress.

Kurzfassung

Der starke Nutzungszuwachs von Mobilgeräten und damit einhergehend des Datenverkehrs stellt Telekommunikationsanbieter bei der Bereitstellung und Weiterentwicklung ihrer Leistungen vor Herausforderungen. Um eine effiziente, kostensparende Ressourcenzuweisung und zeitgleich hohe Servicequalität zu gewährleisten, ist eine genaue Vorhersage der Auslastung einzelner Funkzellen notwendig. Demzufolge müssen Mobilfunknetzbetreiber bei ihrer Kapazitätsplanung die erwartete Auslastung sowie etwaige relevante Bewegungsmuster berücksichtigen. Diese Informationen müssen dann in die Architekturplanung sowie die gezielte Auswahl von neuen Zellstandorten miteinfließen. Unter Zuhilfenahme eines Datensatzes, der von einem europäischen Telekommunikationsanbieter zur Verfügung gestellt wurde, fokussiert sich die vorliegende Diplomarbeit auf die Herstellung eines Verständnisses zur sowie die Vorhersage von Zellnutzungsmuster im Laufe der Zeit. Herkömmliche Forecasting-Methoden sind oft nicht in der Lage, die komplexen Muster zu erfassen, die in Mobilfunknetz-Datensätzen zu erkennen sind. Daher werden in dieser Arbeit verschiedene Clustering-Methoden mit traditionelleren sowie neueren Forecasting-Ansätzen – ARIMA, Holt-Winters und univariaten und multivariaten Facebook-Prophet-Modellen – kombiniert.

Die gewählte Methodik umfasst die Vorverarbeitung der Daten, um Konsistenz und Zuverlässigkeit zu gewährleisten, gefolgt von der Anwendung von Clustering-Algorithmen, zur Erkennung bestimmter Muster der Zellauslastung. Diese Cluster werden dann verwendet, um die Vorhersagefähigkeiten der Forecasting-Modelle zu verbessern.

Die Ergebnisse zeigen, dass die Integration fortschrittlicher Clustering-Techniken mit Forecasting Modellen, insbesondere einem multivariaten Prophet-Modell, die Genauigkeit von Auslastungs-Vorhersagen erheblich verbessert. Durch die Identifizierung und Nutzung von Mustern innerhalb von Clustern können die Modelle die künftige Zellenauslastung genauer vorhersagen, was eine bessere Ressourcenzuweisung und Netzoptimierung ermöglicht.

Darüber hinaus leistet diese Diplomarbeit einen Beitrag zum Bereich des Time Series Forecastings für das Anwendungsfeld der Netzoptimierung, was auf verschiedene Aspekte der Verwaltung von Mobilfunknetzen angewendet werden kann, einschließlich Kapazitätsplanung, Erkennung von Anomalien und Verbesserung der Dienstqualität.

Abstract

The exponential growth in mobile device usage and data traffic has enhanced the challenges faced by telecommunication providers in optimizing cellular network performance. Accurate prediction of cell load is essential for efficient resource allocation, cost management, and enhanced service quality. As a result, mobile network operators must consider the expected user load and movement patterns in their capacity planning, while cellular network architecture planning should include the strategic placement of cell sites to strengthen signal coverage. Using an detailed dataset provided by a European telecommunication provider, this research focuses on understanding and forecasting the patterns of cell utilization over time. Traditional forecasting methods often fall short in capturing the complex patterns inherent in cellular network data. To overcome these limitations, this thesis integrates advanced clustering methods with traditional as well as modern predictive models – namely ARIMA, Holt-Winters and the Facebook Prophet model.

The methodology involves preprocessing the data to ensure consistency and reliability, followed by the application of clustering algorithms to identify distinct patterns of cell usage. These clusters are then used to enhance the predictive capabilities of the time series models.

The findings reveal that integrating advanced clustering techniques with time series prediction models, especially a multivariate Prophet model, significantly improves the forecasting quality of the cell load. By identifying and leveraging patterns within clusters, the models can more precisely predict future cell load, facilitating better resource allocation and network optimization.

Furthermore, this thesis contributes to the field of predictive modeling and network optimization. The integration of advanced data-driven techniques offers a framework that can be applied to various aspects of cellular network management, including capacity planning, anomaly detection, and service quality enhancement.

Contents

CHAPTER 1

Introduction

This chapter aims to provide the reader with an overview of the motivation behind this thesis, problem statement and expected outcome, as well as how the thesis is structured.

1.1 Motivation and problem statement

Cellular networks increasingly face the challenge of meeting escalating demands for higher data rates, energy efficiency, and diverse user needs. [33] As both the number of mobile devices and the volume of data traffic continue to grow, optimizing network services becomes important to ensure seamless user experience. Additionally, the planning of cellular network architecture should strategically position cell sites to optimize signal coverage and minimize interference. [19, 52] To address these challenges, it is essential to predict cell load requirements in a cellular network accurately. This ability enables Mobile Network Operators (MNOs) to orchestrate resources flexibly, improving service quality while managing costs and energy consumption effectively.

The primary motivation for this thesis stems from the need to further strengthen our understanding of cell utilization patterns within cellular networks. Network optimization may be achieved by contributing to the current research on these patterns, which is critical for reducing operational costs and energy consumption. Moreover, improving the accuracy of forecasting models not only facilitates network optimization but also contributes to the broader field of predictive modeling.

The problem addressed in this thesis revolves around the limitations of traditional forecasting models, such as ARIMA and Holt-Winters, which often do not fully capture the complex patterns in cellular network data. Additionally, the lack of advanced clustering techniques in these models limits their effectiveness in identifying nuanced usage patterns. This research aims to fill this gap by integrating advanced clustering

methods, specifically biclustering and spectral biclustering, with traditional but also a new predictive models, namely the Facebook Prophet model, in comparison.

1.2 Aim of the thesis

The primary aim of this thesis is to develop a deeper understanding of cell utilization patterns within cellular networks to facilitate network optimization. By analyzing a dataset provided by a local MNO, this research seeks to improve the accuracy of forecasting models, ultimately leading to reduced costs and energy consumption while enhancing network quality for consumers. To achieve this goal, the thesis will explore and evaluate various predictive modeling techniques in the context of cellular network load management. Specifically, it will examine the effectiveness of integrating advanced clustering methods with predictive models such as ARIMA, Holt-Winters, and the Prophet model by the Facebook Research Team. By analyzing temporal data and identifying patterns, the research aims to advance the understanding of time series prediction models capable of anticipating network load with high precision. This approach includes the development of ensemble models to predict cell load in the future based on the available data as described in Chapter 3.1. Those models will incorporate multiple regressors, leveraging features from both the target cell and correlated cells to enhance prediction accuracy.

The practical implications of this research are mainly focused on improving cell load prediction to assist in more efficient resource allocation, resulting in reduced cost and energy, while also adding knowledge for further predictive modeling using cellular network data – for example for cell capacity or location planning.

Resulting from that aim, the overarching research question for this thesis is:

- *How can forecasting models be improved by considering similarities and differences in the usage patterns of cellular networks within the same cluster?*

To help answer the main research question, two supporting sub-questions were defined to elaborate the goals, as well as extend on the main question:

- How can clustering techniques effectively group usage patterns of individual cell towers based on their usage characteristics?
- How can multivariate methods improve the forecast of usage patterns by capturing the trends of multiple endpoints compared to traditional univariate time series models?

By investigating these questions, this thesis aims to expand upon existing research to refine the accuracy and efficiency of network management using advanced data-driven techniques, thus improving the quality of service for the end user.

1.3 Expected results

Significant efforts of this thesis will be directed towards a new approach to clustering time series data, namely focusing on patterns in the utilization of cell towers. The results of this clustering approach will support the utilization of multiple predictive models – namely ARIMA, Holt-Winters and the Facebook Prophet model, the latter one allowing an ensemble of multiple regressors. The research will develop ensembles for features other than the target variable of the cell being predicted, as well as ensembles for regressors of the same feature in the most similar cells in terms of correlation. By identifying and utilizing homogenous subsets of data, the models are expected to focus on more relevant patterns, thereby improving prediction accuracy. Furthermore, the use of multivariate correlations and additional features is expected to provide a comprehensive understanding of the factors influencing cell load, leading to more reliable and robust forecasting models. Ultimately, this thesis aims to demonstrate that the proposed approach can outperform traditional methods – namely, the combination of clustering methods with the forecasting models ARIMA and Holt-Winters.

To summarize this, the expected results would be robust models that are able to predict time series data accurately.

1.4 Structure of the thesis

This thesis is structured into six further chapters to systematically explore the topic of cell load prediction using usage pattern clusters in cellular networks:

State of the Art (Chapter 2) shall provide background information on telecommunications, focusing on cellular networks and strategies for scheduling and planning network balancing for readers with various backgrounds.

Dataset Analysis (Chapter 3) aims to give an understanding of the dataset and its most important variables, as well as a reference to its technical specification as well as the interdependencies of features, and an analysis of selected cells.

Clustering (Chapter 4) explains how the previously explained data are analyzed to display similarities in usage patterns, the theoretical aspects of why the utilized approach was chosen, and the resulting benefits.

Timeseries Prediction (Chapter 5) offers an introduction to the algorithms utilized, the classical models they are compared to, and the approach chosen for hyperparameter optimization for all those models.

Conclusion (Chapter 6) is used to present all findings of the previous chapters in an organized manner while also offering input on future works as well as limitations of applicability on the insights gained.

CHAPTER 2

State of the art

This chapter aims to provide a short introduction into the world of telecommunications for readers and to help them gain a basic understanding how the underlying technology operates.

2.1 History of Cellular Technologies

In order to standardize cellular network communications, several consecutive standards on cellular technology have been developed.

The first generation of telecommunication technologies allowed voice calls and set the foundation for subsequent generations. There were several different first-generation systems available. A consolidation of standards started with the second generation (2G), predominantly by introducing the GSM standard. [7]

Established in 1991, 2G emerged as the first digital standard and introduced foundational mobile services, including the Short Message Service (widely known as SMS). It notably incorporated the CDMA and GSM concepts. [34] CDMA (Code Division Multiple Access), employed by the IS-95 technology, is a digital cellular technology that employs a unique code for each user, enabling multiple users to share the same frequency simultaneously. GSM (Global System for Mobile Communications) is another standard for 2G mobile networks, utilizing time division multiple access (TDMA) to allocate specific time slots to multiple users on the same frequency, allowing for efficient communication. [7] It employed Time Division Multiple Access (TDMA) for spectrum allocation and paved the way for global mobile communication. A – compared to today's standards – very limited data capacity of up to 14.4 Kbps/s was possible. [10] The Enhanced Data rates for GSM Evolution (EDGE) technology was later introduced as an upgrade to 2G, significantly improving data transmission rates. While 2G remained a reliable standard, it had slowly been replaced by the third generation (3G) standard.

3G introduced a new architecture called the Universal Mobile Telecommunications System (UMTS), offering significantly higher bandwidth than its predecessor. [35] This advancement made 3G the first mobile multimedia standard, enabling the transmission of video streams via mobile cellular networks. [10] The Long Term Evolution (LTE) standard was later introduced, enhancing the capacity and speed of mobile communications. It utilized Wideband Code Division Multiple Access (W-CDMA) technology to achieve higher data rates than it's predecessor – up to 384 Kbps/s. [7]

The subsequent standard 4G was implemented to meet rising demands due to a changed user behavior, demanding computing with much higher bandwidths than previous standards allowed. [36] 4G's primary objective is to deliver high-speed communication with enhanced security, enabling services like video conferencing. 4G LTE (Long Term Evolution) represented a significant leap in wireless technology, utilizing technologies such as Orthogonal Frequency Division Multiple Access (OFDMA) and increased Multiple Input Multiple Output (MIMO). [14, 71] Peak data rates of up to 300 Mbps/s were reached, showing a significant increase compared to 3G. [14]

The latest standard in cellular communication, 5G, is designed to facilitate communication with even higher capacities and data rates (up to 10 Gbit/s) as well as lower latency therefore meeting the high demand caused by technologies like AR or the increased use of IoT in everyday life. [37, 51, 85] Scientific research in millimeter-wave propagation, network architecture, and interference management has been fundamental to its development. [71]

In summary, the progression from GSM to 5G reflects a continuous evolution in telecommunication technologies, driven by scientific breakthroughs. Each generation has laid the groundwork for the next, contributing to the rapid expansion in the digital age.

2.2 A primer on Telecommunications

Telecommunications describes the science of transmitting information of various types over long distances in a cellular network. A cellular network is comprised of segmented areas called cells or cell sites, which are strategically distributed over a defined geographic region. Each cell site incorporates one or more receivers to provide coverage for their area. [19, 59]

As is explained in more detail in Chapter 2.1 below, contemporary cellular network architectures, eNodeBs (E-UTRAN NodeBs) are central to the LTE Radio Access Network (RAN), facilitating direct communication between the network's core and User Equipment (UE) via the S1 interface. eNodeBs manage crucial functions such as radio resource management, mobility handling, and the initiation of security protocols. They employ MIMO (Multiple Input, Multiple Output) technology and OFDMA to optimize spectrum use and increase throughput, enabling support for a high volume of simultaneous user connections. [56]

The architecture integrates eNodeBs with core network entities like the Mobility Management Entity (MME) and the Serving Gateway (SGW) through defined interfaces, ensuring seamless data and signaling flow. The MME orchestrates session management and security processes, while the SGW facilitates the routing of user data packets. This infrastructure is pivotal in achieving efficient, high-speed data transmission and robust mobility support across the network, laying the groundwork for the transition to more advanced network technologies, including 5G NR (New Radio) systems, which introduce gNodeBs (Next-Generation NodeBs) to accommodate broader bandwidths, lower latency, and increased connectivity demands. [71]

MNOs are the primary entities responsible for delivering telecommunication capabilities to end users by establishing cellular networks that interconnect individuals within specified regions, ranging from local entities such as villages and cities to broader scales encompassing provinces, countries, or even global connectivity. [12] The dimensions of the cells and the concentration of cell towers within them fluctuate based on factors like population density and the network's capacity demands for each region. They are placed close to each other to provide sufficient coverage and seamless transfer of signals and communications with devices in motion.

2.2.1 Long Term Evolution (LTE)

This section shall provide a comprehensive overview of the 3rd Generation Partnership Project's (in short: 3GPP) Long Term Evolution (LTE), starting with the key requirements set for LTE to achieve its performance goals. In Release 8, the 3GPP established LTE as a better replacement to the Universal Mobile Telecommunications System (UMTS) standard. [41]

These requirements include supporting data rates within a 20 MHz spectrum allocation up to 100 Mb/s downlink and 50 Mb/s uplink, corresponding to spectral efficiency values of 5 bps/Hz and 2.5 bps/Hz, respectively. In terms of throughput, the downlink average throughput per MHz was approximately 3 to 4 times higher than what could be achieved prior, with the uplink average throughput per MHz about 2 to 3 times higher than its predecessor. LTE's Radio Access Network (RAN), or Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), surpasses initial goals, supporting up to 300 Mbit/s downlink and 75 Mbit/s uplink rates. [86]

Mobility optimization within LTE is geared towards low terminal speeds, only up to 15 km/h, while also ensuring that high speeds, up to to 350 (and potentially up to 500) km/h, can be reached. Set targets for coverage must be met within 5 km cell ranges, with acceptable levels of throughput and spectrum efficiency degradation for cells extending up to 30 km. LTE also optimized frequency scheduling and minimized inter-cell interference, enhanced by the X2-interface and MIMO technologies to support a range of User Equipment (UE) capabilities, from 10 Mbit/s to 300 Mbit/s downlink throughputs. [42]

2.2.2 LTE Handover Management

When a user initiates a call or data activity, the mobile device connects to the nearest cell tower (base station). When the user, and thereby the user's device, is on the move, a handover process as illustrated in Figure 2.1 is taking place. [12] This means that the dedicated radio connection between the device and the network is shifted from one cell, the current transmission point (TP), to another. This is managed by the eNodeB base station, ensuring that the device maintains a connection with the optimal communications link as the user moves. [17]

In general, the handover is being executed based on the reference signal received power (RSRP). While the device is being commanded to switch to a new channel, at the same time, the session is being switched, too. This process can also be utilized to distribute the load among several cell sites. [71] The handover is, if not earlier due to maxed out capacity in the current cell, initiated when the signal strength of the cellular device is weaker than the signal strength in the cell it is approaching.

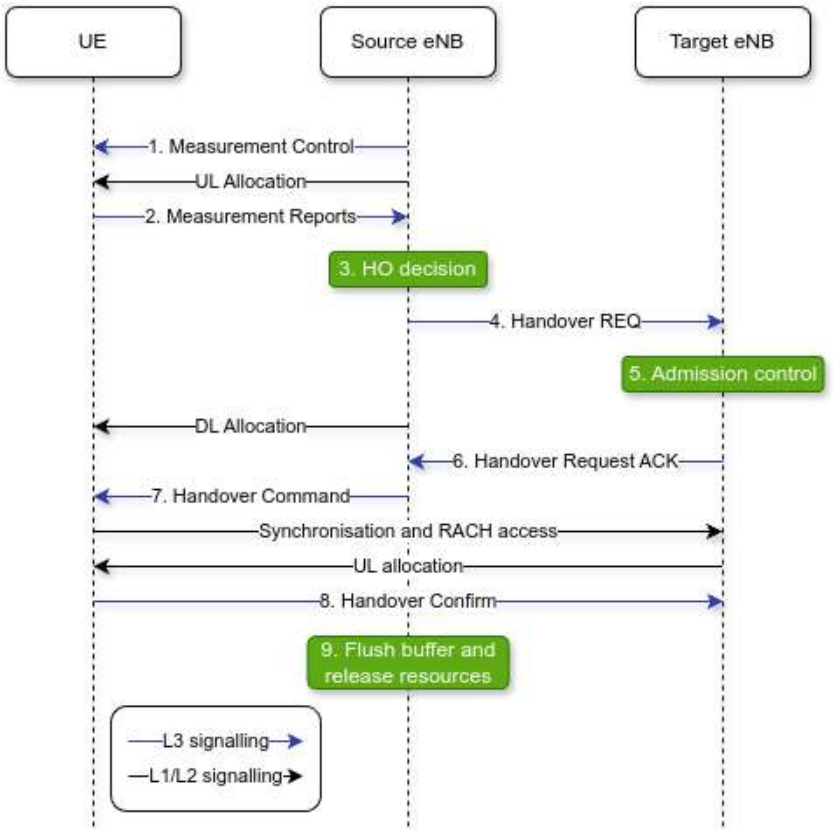


Figure 2.1: The Handover Management Procedure (based on [17])

To ensure a smooth process, a slight overlapping of signal coverage of cells and a strategic

placement of those cells by the MNO is necessary. [56] Proper planning and optimization of cell tower locations help minimize interference and maximize the efficiency of the handover process. [64]

Consequently, MNOs must account for the anticipated load and movement patterns of users in capacity planning, while cellular network architecture planning must strategically position cell sites to optimize signal coverage and minimize interference. [17]

LTE Downlink Architecture

LTE downlink architecture utilizes OFDMA to allocate the spectral resource across multiple users, optimizing the transmission by adapting to channel conditions in real-time. Central to its design is the use of a time-frequency grid, where the allocation of Physical Resource Blocks (PRBs) to users is dynamically managed by a scheduler. This allows LTE to respond faster to user demands and additional factors, ensuring optimal performance. [40]

In this context, PRB utilization is a critical metric for understanding cell load. PRBs represent the smallest unit of resource allocation in LTE, consisting of a specific number of subcarriers over a certain number of time slots. High PRB utilization indicates that a cell is experiencing heavy traffic, as more resource blocks are being allocated to active users. Conversely, low PRB utilization suggests lighter traffic and potentially more available capacity. Monitoring PRB utilization helps network operators manage and optimize the load on each cell, balancing user demands and maintaining service quality. [20, 56]

This system also leverages advanced MIMO (Multiple Input Multiple Output) technologies. LTE's ability to dynamically adjust its MIMO transmission modes is essential since its flexibility allows for real-time adaptation to current channel conditions, ensuring optimal performance. [5]

The design of the LTE downlink supports a range of bandwidths from 1.25 MHz to 20 MHz, catering to various deployment scenarios and spectrum availability. The spacing between subcarriers is maintained at 15 kHz, which is a key parameter for OFDMA and ensures efficient utilization of the available spectrum. This flexibility in bandwidth support allows LTE to be deployed in diverse environments, from rural areas with limited spectrum to urban areas with high user density and spectrum demand. [86]

2.2.3 Mobility Prediction in Cellular Networks

Research and development regarding planning and executing the extension of cellular networks has been significantly influenced by user mobility prediction studies. [29, 31]Predicting user mobility is considered as a practical approach to manage connections between user behavior and mobile networks – by leveraging signaling data from users moving in or between cell areas. [46]

Predicting a users next cell area or general trajectory across cells is crucial for mobile network planning. Zhang et al. [84] identified three key areas in which mobility prediction

might contribute to promising advancements: improving handover management (for example handover failure or latency), dynamic resource management and location based service planning.

The first area, handover management, describes how the aforementioned handover process may have users experiencing service instability due to RSRP fluctuations. Mobility prediction offers insight into users' long-term movements and can assist in improving handover prioritization and reduction of overall handovers. The second research field, resource management, focuses on the prediction and availability of the capacities necessary to accommodate all incoming user requests. TPs are able to reserve resources for users to reduce resource collision and interferences. This area of resesarch might also profit from gaining insight into the dwell time a user spends in a particular TP, also allowing more precise managmenet of resources, improving energy efficiency in the course of doing so. The third area of application, location based services, moves into the direction of providing users with location-specific information, for example by assisting vehicular network application to allow real-time updates to adapt a suggested route to the respective traffic conditions. [48, 78].

Earlier research regarding mobility prediction in cellular networks focused more on analyzing user mobility characteristics. User movement features have been the subject of research since the late 1990s. Back then, the focus was already on optimizing capacity planning for future large-scale mobile network infrastructures. It has been argued that the more locations a user visits, the closer they are on average, forming a Gaussian distribution. [72] This comparatively low mobility was also reconfirmed by other studies over the following years. [61] Furthermore, this generally low mobility is not only reflected in the distance traveled, but also in the places visited. [31] Research into these spatiotemporal regularities in user mobility showed that strong regularities in a subjects' mobility pattern can be identified, suggesting that people often revisit places within a 24-hour cycle – even go so far as to say that 93% of user mobility is theoretically predictable. [28, 70]

2.2.4 Load Capacity Planning in Cellular Networks

As explained before, capacity planning for mobile networks has become a more and more significant challenge for network planners over the past decade. [76] Mobile network traffic has seen exponential growth, with the rate varying by market but approximately doubling every two years. [21, 22] As traffic loads grow, network performance tends to degrade unless there is continuous investment in additional capacity. Concurrently, user demand for higher throughput and lower latency has risen. Adding capacity to mobile networks is also complicated by long implementation periods, adding another reason to rely on predictive cellular network planning to estimate future network performance. [76]

There are three essential parameters for capacity planning in a network: Traffic estimates, average antenna height, and frequency (re-)usage [57, 73]:

1. **Traffic Estimates:** Traffic estimation is based on theoretical assumptions and studies of existing networks. It depends on user communication rates and movement within the network. User mobility impacts handover rates and, subsequently, capacity planning.

2. **Average Antenna Height:** Average antenna height influences the frequency re-use pattern and capacity calculations. Lower antenna heights in urban environments lead to smaller coverage areas, necessitating more cells and increasing frequency re-use.

3. **Frequency Usage:** Frequency re-use is the frequency allocation within the network and is calculated based on the number of transceivers per base station and the total number of available frequencies.

Load capacity planning shall ensure that a cellular network can handle the anticipated traffic load without compromising service quality. This involves predicting future traffic patterns, identifying potential bottlenecks, and making informed decisions about network upgrades and resource allocation. [6] Inadequate capacity planning can lead to users experiencing a low-quality network. [55, 63]

Previous applications of Machine Learning in this field have focused on system improvements such as beamforming or coding/decoding schemes as well as network anomaly detection, spectrum allocation, and handover prediction, to name a few. [67] However, regarding capacity planning and deep learning techniques, most studies have concentrated on traffic forecasting, focusing, as mentioned above, on user movement streams. [83]

Previous studies have utilized forecasting models such as ARIMA, LSTM and MLP for performance evaluation and predicting cell loads in mobile networks, contributing to a better understanding of dynamic resource optimization and cellular network planing, though it is reiterated that more research is needed in this field. [55, 58] While methods like exponential smoothing and ARIMA have proved to be feasible and suitable for time-series forecasting, a focus on improved accuracy is still needed. One study from 2019 [6] highlights the relevance of simple models for network management and concludes that further evaluation and testing of new methods are necessary, especially extending to 4G and 5G networks. [20]

Various additional sources underscore the necessity of time series forecasting techniques for accurately predicting future demand and adjusting capacity levels. The incorporation of historical data and error rates into forecasting models allows for continuous improvement and more accurate capacity planning. [50]

CHAPTER 3

Dataset analysis

This part of the research involves a detailed look at the dataset to familiarize oneself with what information it holds and how it can be utilized to aid the research goal.

3.1 Description of the available data set and its features

The MNO provided two data sets featuring ≈ 90 features to describe the cell towers as well as their observed utilization:

Info dataset contains meta information about a subset of the LTE cell towers operated by the MNO. This subset comprises over 150 cell towers, each typically configured with a three-sector deployment, and includes approximately 1500 cells operating across three different frequency bands. The area was chosen as it contains all of the cell towers in an Austrian city, its industrial district, and the surrounding areas to reflect a representative mix of urban, suburban, and rural areas. Commuting is thus expected to be included in the data set which makes the results of this work applicable to scaling out on other areas as well.

Load dataset contains measurements of the cell towers within in the chosen subset in regular intervals. Multiple features are included that give an understanding of the cell load in various ways to analyze the capacity, efficiency and efficacy of the towers. The measurements were obtained within a time frame of six months. Each row represents the observed variables with tags when they were recorded in which cell. To understand the meaning of the observed variables, one can refer to documents published by the European Telecommunications Standards Institute (ETSI) for detailed technical specification. [2]

Both datasets encompass information on the unique cell sites, the cellular technology used indicated by a frequency band, the sectors of cell coverage and the unique eNodeBs.

3.1.1 Info dataset

The info dataset primarily includes variables related to the infrastructure, thereby providing details about the physical and technological setup of the cellular network in question. It provides information on the 1473 unique cells that are distributed over 164 physical locations with cell towers that make up this dataset.

This dataset covers different technologies across multiple sites, with some additional specific details about the cell sites including the antenna types and frequency band in use.

It also includes specific technical details such as the name of each cell, the direction an antenna is pointing (azimuth) as well as key technical identifiers like the Physical Cell ID (PCI) of the cell.

In addition, the dataset provides detailed information about the network's technical setup, including the identifiers for each cell, the eNodeB at which it is located and the frequency it is set up for.

For each observation of those unique cells, the eNodeBs, Mobile Country Codes, the UARFCN (UTRA Absolute Radio Frequency Channel Number), antenna height or gain, beam width, their geographical location postal codes, longitude and longitude coordinates are given.

In the analysis for this work, the info data set was used to perform general analysis on the meta information available. The assigned name for each eNodeB, information about the location in the form of latitude and longitude, as well as a classification into one of the three given frequency bands ($\approx$ 800MHz, $\approx$ 1800, and $\approx$ 2600 MHz with 50%, 30% and 20% of all cells) was provided for interpretation.

3.1.2 Load dataset

The load dataset focuses on performance metrics and shows the operational status and efficiency of the network. It captures detailed measurements of the cell sites over 24 weeks. It includes metrics such as the average and maximum number of active users (downlink), the downlink spectral efficiency as well as the PRB utilization.

The data also offer a view into how the network operates, which is the main source of information this thesis draws from, with observations of how many users are active, how efficiently data is transmitted (downlink spectral efficiency), and the average number of users connected to the network for every observation.

The features of the data set that were most relevant for the prediction in this thesis are explained below, including a brief explanation and a reference to further documentation of the European Telecommunications Standards Institute. [2]

Selected features in the load data set include:

Average Users E-UTRAN The average number of users served by the cell in an E-UTRAN network. In contrast to the PRB utilization, this metric gives an insight into the number of user equipment in the serving cell at a given time. [43]

`e_utran_average_number_of_ues_in_serving_cell`

Average Users RRC Average number of Users with an active RRC (Radio Resource Control) connection. This number is highly correlated with the number of average user equipment in the serving cell, is an indication of how many devices might request data and result in active network utilization. [39]

`avg_rrc_conn_ue`

Average Active Users Downlink Average active users in the downlink. The more users populate a cell and actively use the resources, the fewer resources each user will have, therefore this is an important KPI when assessing the load in an LTE cell. This becomes evident during large social events such as demonstrations, soccer games, or concerts. [38]

`avg_act_ues_dl`

Average PRB per TTI Usage Average usage per Transmission Time Interval (TTI) of PRBs in the E-UTRAN downlink. This metric describes the ratio of resource blocks assigned to UEs in a given TTI. [76]

`e_utran_avg_prb_usage_per_tti_dl`

Figure 3.1 depicts the pairwise relationships between those four variables and gives an idea how they interact with each other. It is logical to see that all of the variables have a direct proportionality and thus should be well-fit to train the multivariate model that incorporates the other features to explain the target variable.

The graph gives insight that the PRB utilization follows an exponential trend with the number of active users in the downlink before saturating at 100. Notably, cells operating in lower frequency bands tend to support more average active UEs in the downlink, leading to higher PRB utilization. This is likely due to lower frequency bands covering larger areas, where other cell towers may not provide comparable service quality. This discrepancy is evident in the top right of the third graph in the second row, showing fewer cells with this specific combination.

A similar observation can be made in the third subgraph in the first row, which indicates predominantly low number of connected user equipment in the downlink with some exceptions in the 800 MHz frequency range that show many equipment in the downlink associated with a large number of connected user equipment. This makes sense, as lower frequency bands offer larger coverage and thus the number of equipment in the downlink has a higher potential to increase, while still offering marginally slower speeds.

When taking a look at the average number of UEs in the serving cell, it becomes evident that the number of connected user equipment is directly proportional with the PRB

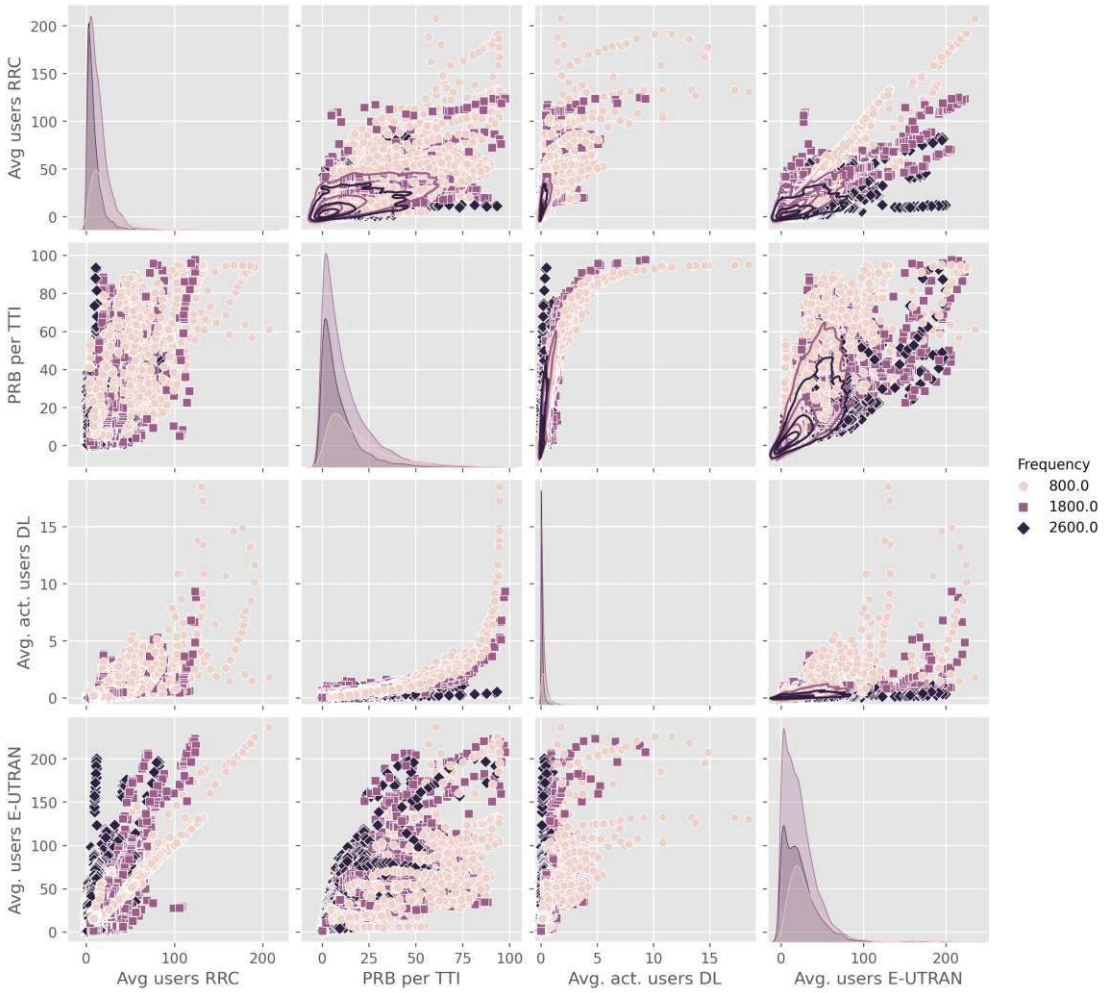


Figure 3.1: Pairwise relationships between the four most influential features

utilization and the number of active users in the downlink are each lower than the purely connected users. The decreasing trend of those variables with a rising number of UEs in the serving cell is more evident when looking at the cells with higher frequencies, as is indicated by the colorcoded plots in the rightmost column. This means that, especially for higher data traffic demands, the cells offering service at a larger frequency, less and less user equipment is prone to request large amounts of downlink than on lower frequencies.

3.2 Data Cleansing and Preparation

While the dataset was received quite clean already, some steps were taken to ensure consistency and reliability for the procedures to follow. First, various notations that represented the same information but were displayed with minor differences, space-

trimming of various fields, or that showed small discrepancies in naming conventions, were standardized and redundancies removed. Additionally, the format of the given timestamps was unified, ensuring data integrity and correcting any anomalies or inconsistencies. Uniform representations of cell names, eNodeB identifiers and frequency bands were adopted for the purpose of enabling systematic analyses. Missing data for the observed variables were no issue as only a few values were missing, and measurements were averaged per hour (as well as other intervals for different analyses where it was required, as explained in Chapter 3.2). There were a few features that included more NA values, however those were negligible parts of the info dataset and did not pose danger of inconclusive analyses later on and thus were disregarded, because the interval based average grouping accounted for those missing values.

The focus for preparing the dataset for the subsequent analysis lies in the conversion and segmentation of data into a structured format. For the facilitation of the spatial analysis, the geographical coordinates, initially provided in latitude and longitude, were converted into a two-coordinate system using the EPSG 31287 (WGS 84) standard [1] to transform the geographic locations into coordinates compatible with a coordinate system that supports distance and area calculations for easier handling. This conversion also ensured the dataset's independence and scalability across different geographic locations as it removed the influence of the Earth's curvature and allowed distance calculation between towers with its standard coordinate system for the central European area. The upcoming analyses were carried out for multiple time intervals, which were each created by grouping the actually recorded observations and taking the average value for that group, resulting in, for example, 24 hourly observations per day (instead of the originally provided measurements in 15 minute intervals). This step was conducted to be able to discern more usage patterns free of random stutters or noisy data since the research in this thesis shall include analysis steps that rely heavily on identifying patterns within those time intervals selected prior.

3.3 First insights into the Data set

3.3.1 Proximity vs. correlation

As the frequency that a cell operates at also dictates the range and number of users the MNO can provide connectivity for, this leads to the conclusion that the intermediate frequency band is the most popular among users, as it balances coverage area and the number of users served. [81] As the medium of those three served distances (which are, by nature, affected by the height of the antennas) is the most common one according to the dataset, the question arises if the correlation between multiple cells correlates to the distance between those cells.

In order to answer that question, the author has produced a graph-based solution: an undirected graph was created in which nodes represent the eNodeBs in the dataset, and each eNodeB is connected to their three strongest correlation partners, creating a simple, easy to understand visual representation. For the creation of that graph, two

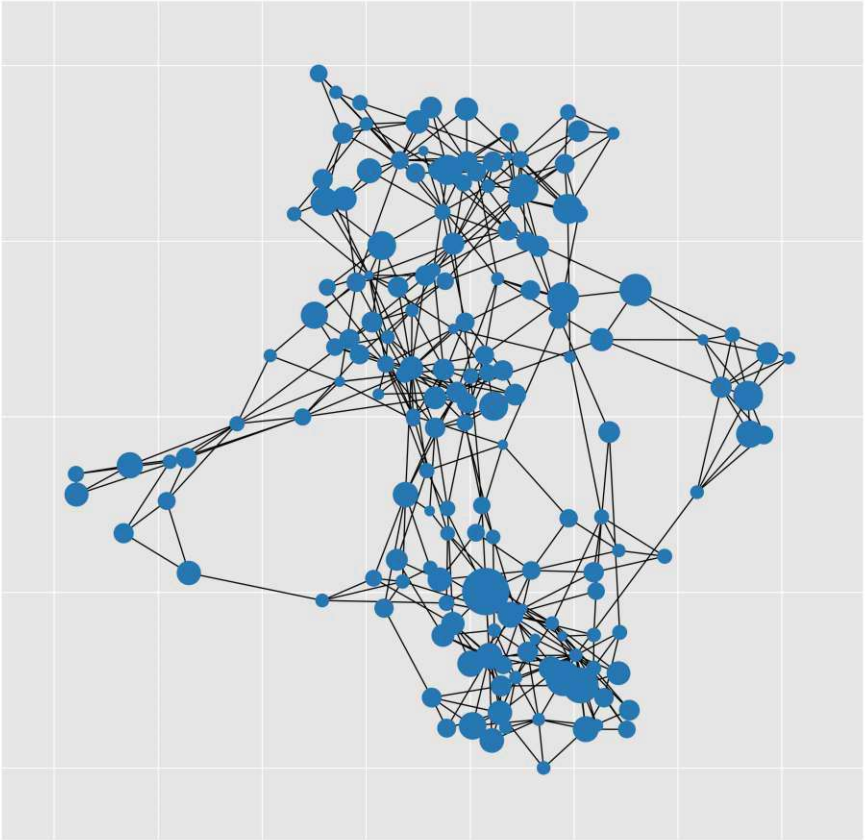


Figure 3.2: Visual representation of the eNodeBs within the data set and the connection between correlation and geographic distance.

matrices were created: a distance matrix (from each tower of every frequency band to every other) as well as a correlation matrix (based on every cell of the corresponding towers). The number of connected user equipment is depicted by the size of the nodes, whereas the distance between the nodes is representative of the actual distance between the towers. It is important to note, that, while the distances are accurately presented, the locations within the graph are not indicative of the geographic position. Thus, lots of edges between nodes located far from each other are indicative for user masses spread out geographically, whereas lacking edges are indicative of users gravitating towards a

single geographic location.

When looking at Figure 3.2, the presence of five clusters is noticeable. On the one side, each cluster is of different size, but with each cluster separated spatially, the inter-cluster connections are very limited resembling the usual travel trajectories between city centre and rural areas. The clusters obtained from the graph are similar in the fact that nodes with large user numbers are centrally aligned with other nodes surrounding in a star-shaped form. Notable is a number of eNodeBs with average load that act as a buffer between the correlation clusters, which seems logical considering the fact that stronger correlations between the neighbors of the „buffer towers“ would mean that the farther apart nodes were correlating even stronger. The main conclusion that can be drawn from the graph is that the distance based clustering leads to similar results as the correlation analysis such that the cells are strongly correlated with other cells in the neighborhood.

While it is not in the scope of this work to identify the common location properties of each cluster, this is an important aspect to keep in mind as well as a possible outlook to a future work as part of the overarching topic how cells can be classified and the MNO can improve its service based on a cell's classification in, for example, shopping malls, nightlife districts, office parks, industrial territory and logistic centers such as airports, ports or railway stations. This way a small set of tagged eNodeBs can be extended via cluster assignments for the utilization of supervised learning approaches to identify meaningful patterns in cell load.

3.3.2 Timeseries

To provide an understanding how the utilization of a cell develops over time, Figure 3.3 gives a short overview and sets a comparison between a single cell to other cells within the same enodeB, the same frequency as well as the average of all cells within the entire provided dataset.

The red line gives an impression of the observed values during the time the data was available for this specific cell of the number of active users. Specifically the peaks in week 37 and week 46 of the observed time frame as well as the lows in week 52 (Christmas time with a peak on new year's eve), one and two are specifics that might be easier to predict if data over an entire year were in the data set, as long seasonality cannot be predicted from holidays specifically, which are an option to improve predictions for some models. As the peaks are possibly explained by shopping through holiday seasons before Christmas holidays, this is not directly influenced by the holidays itself and thus would not be captured by the models. This is evident by observing the aforementioned peaks on weekends that are caused by shopping weekends before the festive season.

Specifically on Tuesdays and the weekends, the chosen cell is apparently influencing the average pattern of its enodeB as well as the overall average positively, as the utilization is repeatedly above those averaged values on those weekdays with the averaged trends of the enodeB and all of the cells even displaying internal lows for those days.

All those observations show that, in this case, the selected cell is in a suburban area where a few small shops and a train station are located and social events as well as holidays influence the trend strongly (as is evident by inspecting the holiday season as well as peaks on the turn of the year).

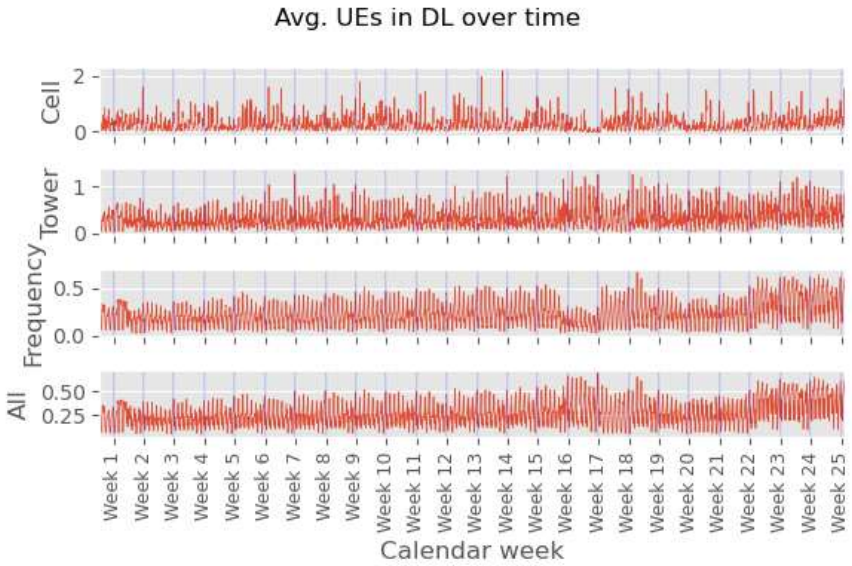


Figure 3.3: Utilization of a cell over time

As previously mentioned during the inspection of the single cell, it is also important to have a look how other cells of the same enodeB develop in the course of the day (figure 3.4). Aside from the six cells within the medium frequency band, there are also three in the lowest as well as three more in the highest frequency band.

In the course of a single day, on average the selected cell (denoted as "Cell 0") only corrects the trend downwards between 06:00 and 08:00, but upwards for the remaining hours of the day until the first half of the night. Even among the other cells of its enodeB in the same frequency band, the selected cell is above average consistently (observed in hourly trends).

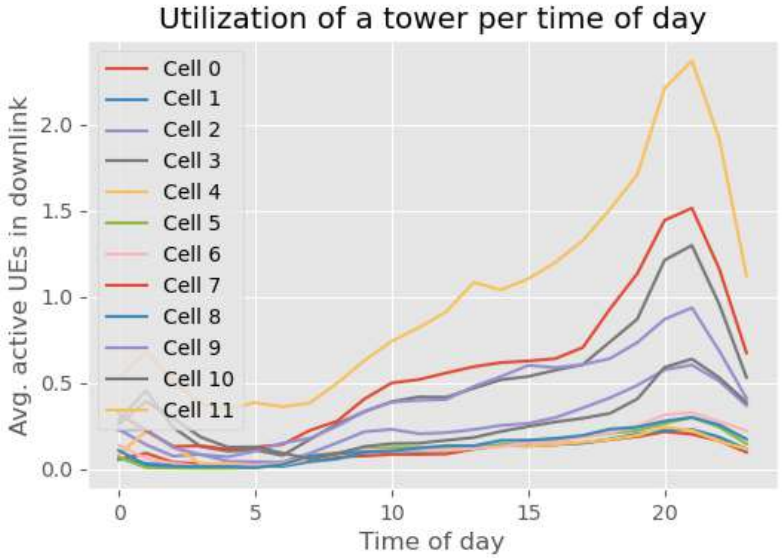


Figure 3.4: Utilization over time in comparison to the average utilization of the corresponding enodeB and the entire data set overall per hour

Similar behaviour is depicted in the trends per weekday in the observed enodeB (figure 3.5): our chosen cell (denoted as "Cell 0") shows higher utilization in terms of active user equipment than most other cells within its cell tower.

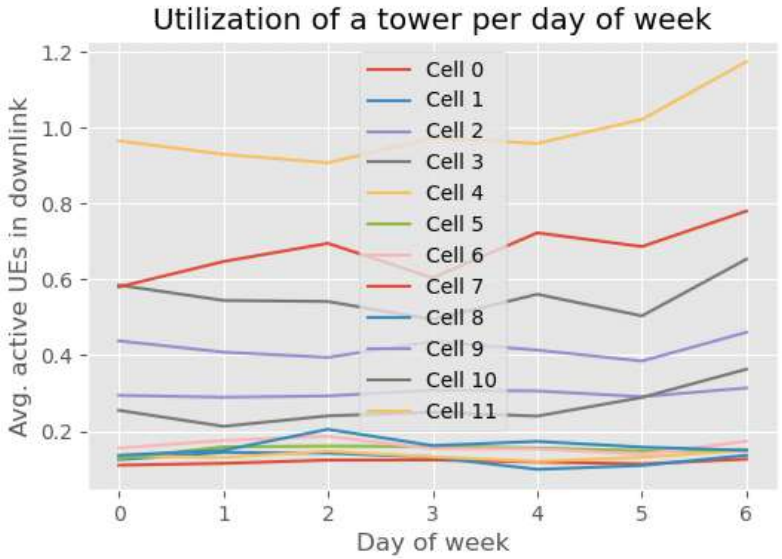


Figure 3.5: Utilization over time in comparison to the average utilization of the corresponding enodeB and the entire data set overall per weekday

CHAPTER 4

Clustering

The goal of this thesis is to identify different cell patterns and relate them to geographic features. Specifically, it aims to predict the area of a town in which a given cell is located, distinguishing between office areas, suburban housing, shopping districts, nightlife venues and such. To achieve this, the research focuses on clustering techniques that group cells within the dataset based on their usage characteristics. These clusters are expected to reflect differing utilization patterns corresponding to the specific functions of the geographic areas they serve. For that reason, the first experiments were aimed at clustering the cells within the dataset into distinct buckets that promised to display differing utilization based on the usage pattern of each of those dedicated areas.

In prospect of the second part of this thesis, the question if those patterns can be better predicted with the aid of other cells within the same cluster in comparison with univariate predictions based on a single cell.

Clustering structures are typically represented as a set of clusters $C = \{C_1, \ldots, C_k\}$ from a data set X, where the union of all clusters equals X, and each cluster is distinct from the others. Clustering methods vary based on input data types, criteria for measuring similarity or dissimilarity, and underlying theoretical concepts, leading to a wide array of proposed algorithms in the literature.[27] Clustering algorithms can be classified into several main categories – most notably hierarchical and partitioning algorithms, as well as grid-based, density-based and graph-based algorithms. [47]

The first of the aforementioned algorithm categories builds a hierarchy of clusters wherein data is either divided into smaller or agglomerated into larger clusters by separating (top-down) or merging (bottom-up) the neighboring pairs of clusters. [24, 47, 80] Partitioning Clustering Algorithms divide a data set into a specific number of clusters without utilizing a hierarchical structure. The next category mentioned above, grid-based algorithms, clusters the original data in a grid-like structure based on a pre-defined number of cells. [24, 47, 80] Density-based algorithms identify areas of various data point density and

separates them into cells, allowing for arbitrarily shaped areas. [24, 80] Graph-based Algorithms present data as a graph wherein data points are presented as nodes while distances between those points are shown as edges. The clustering itself is achieved by identifying densely connected sub-graphs. [47]

For the problem defined in this thesis, partitioning methods are used that utilize an iterative way to determine clusters based on moving data points that can move from one cluster to the next. These clustering methods usually require a predefined number of clusters (k). A further developed clustering method, the K-Means, is one of the most frequently applied in research and shall be examined in the next section.

4.1 K-Means for time series

The K-Means algorithm is a method designed to divide a collection of numerical vectors into k distinct clusters. It was initially proposed by Ball and Hall in 1965. [4] Despite its age, the theory is still a widely used and highly regarded method for clustering. [44] Its applications span a wide range of possible applications, including medical imaging and pattern recognition, among many others. At its core, the K-Means approach seeks to identify k centroids corresponding to k clusters, with the goal of minimizing the cumulative distance between every data point and its nearest cluster centroid.

This is achieved by the solution for $x_i \in X$:

$$\underset{\{C_1,\ldots,C_k\}}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x_1 \in C_j} d(x_i, c_j)$$

where $(C_1, \ldots, C_k)$ stands for k non-overlapping clusters, c_j represents cluster Cj, and d is a distance function.

Algorithm 4.1: K-Means Clustering

Input: $X = \{x_1, x_2, ..., x_N\}$, number of clusters k
Output: $\{c_1, c_2, ..., c_k\}$ (cluster centroids)

Initialize $p = 0$ and select k initial centroids $\{c_1^{(0)}, ..., c_k^{(0)}\}$ randomly
repeat
  Assign each x_i to the nearest centroid
  Increment p
  for $j = 1$ *to* k **do**
    Update $c_j^{(p)}$ using $c_j = \frac{1}{N_j} \sum x_i \in C_j x_i$
    where $N_j = |cluster_j|$ stands for the number of all data points in cluster j
  end
until *centroids do not change*

The K-Means clustering process itself is described in Algorithm 4.1. To initiate it, it is necessary to determine a number of clusters k. The algorithm starts with a preliminary

selection of cluster centroids $\{c_1, c_2, c_3, \ldots, c_k\}$ which are either randomly picked or chosen based on a specific heuristic method. This clustering technique employs an iterative improvement method in which, during each iteration, every data point is associated with the closest cluster centroid. Subsequently, the centroids get recalculated. The center of each cluster is calculated as the mean of all of the data points are part of that cluster:

$$c_j = \frac{1}{N_j} \sum_{x_i \in cluster_j} x_i$$

where $N_j = |cluster_j|$ means the number of all data points in cluster j.

This iterative refinement process in k means clustering continues until the cluster centroids no longer move. The computational complexity for each iteration over N data points is $O(k \times N)$. This contributes to the algorithm's widespread use, since the linear complexity assists in making it more computationally appealing.

When computing the similarity between data points, it is important to consider the distance metrics used. In this thesis, the Euclidean distance metric and Dynamic Time Warping have been used. K-Means more commonly employs the Euclidean distance metric, which measures the direct, straight distance between points. This approach takes spikes in the data into account. However, it might not always be ideal since it can, in turn, be too sensitive to shifts in a time series. In such cases, Dynamic Time Warping might offer a better alternative by aligning points based on the minimal cumulative distance between them. This approach can also be used to capture similarities and time-dependent patterns. [74]

In the course of this work, the initial approach was to utilize the K-Means clustering algorithm to identify centroids that represent the mean trend representative for each cluster. This method was applied to logically group cells based on their load patterns.

This decision was made due to the fact that the algorithm is known to process large datasets efficiently, which is an advantage for managing the large amounts of data from the cell networks. Its linear complexity ensures that even as the dataset grows, the computational demand remains manageable, allowing for a fast identification of patterns within the cell load data. Furthermore, the simplicity and speed of convergence of the K-Means algorithm make it particularly suitable for the iterative nature of the analysis, wherein multiple runs are necessary to refine and optimize the setup of cell towers. By clustering similar load patterns, K-Means facilitates the prediction of cell load distribution, enabling a more strategic allocation of resources and an optimization of cell tower placement to accommodate varying demands efficiently.

For each cell in the dataset, the trend over time was created by grouping the observations by the intervals hour, day of week, week, and hour by weekday with the most meaningful and observable trends resulting from hourly-based values. For each of the four main features (present users, connected users, active users, and PRB utilization), the K Means algorithm was performed for two up to seven clusters. Each of those iterations

was performed utilizing a standard scaler with the `TimeSeriesKMeans` function from the `tslearn.clustering` module. The results were evaluated using the previously described distance metrics: Euclidean distances as well as Dynamic Time Warping.

Another reason to utilize K-Means clustering was its wide use in similar scenarios, its straightforward approach, and its computational lightness. However, K-Means' sensitivity to the magnitude of observations prompted the need for an additional alternative method. Biclustering emerged as a viable solution, as correlation focuses on identifying patterns over time while disregarding the magnitude of each observation, therefore offering a more suitable approach for the specific use case of the thesis.

Despite this contributing to the initial approach, the K-Means clustering algorithm was ultimately not used in the further course of this thesis due to its sensitivity to the magnitude of observations. Specifically, K-Means would assign cells with identical usage patterns but different magnitudes to different clusters. For example, a cell in a shopping mall with high utilization would be clustered separately from a cell in another mall with lower utilization, even though their usage patterns were the same apart from the magnitude. This negatively influenced the relevant location information that could be sourced from the clusters – since the original goal was to differentiate between different types of locations, such as shopping malls and suburban housing areas, based on usage patterns rather than the magnitude of utilization.

Moreover, the need to account for location-specific patterns meant that distinguishing cells based solely on usage magnitude did not align with the objective of the study. The defined goal was to capture the intrinsic usage patterns of various locations, which required a method that could group cells with similar patterns regardless of their usage intensity. Therefore, the limitations of K-Means in handling the magnitude of observations necessitated exploring alternative clustering techniques. This led to the altered approach of using only Biclustering and Spectral Biclustering.

The shortcomings of K-Means become evident when taking a closer look at Figure 4.1 which displays the cluster centers arising from the K-Means algorithm. Assigning specific locations to the clusters established this way does not offer enough variation to qualify for distinguishment location-specific usage patterns.

Biclustering, which shall be discussed further in the following section, can be described as an altered approach to clustering, by simultaneously grouping rows and columns, which is different from the focus of K-Means. It features a more simplistic and heuristic method compared to the previously described one. It is particularly useful in two-dimensional data matrices where both dimensions have meaningful cluster structures that need to be identified together. In the next section, biclustering will be examined further.

The explicit desire to focus on the patterns of the cell utilization instead of the movements from one data point in the trend to the following resulted in the choice of a different algorithm. Correlation was deemed the proper choice for evaluating the grouped observations (trends), as it incorporates the patterns over time instead of the magnitude of each change.

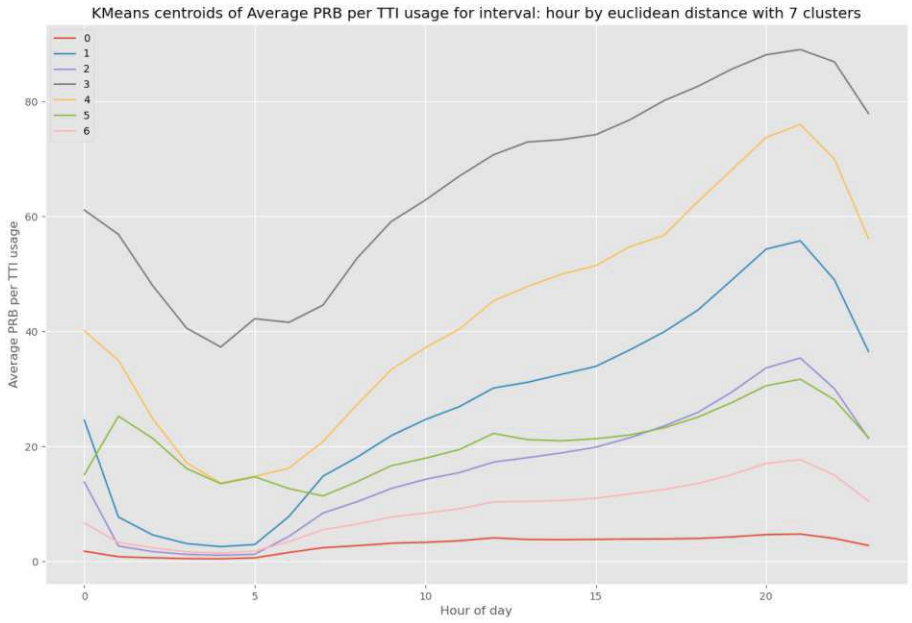


Figure 4.1: Shortcomings of the K-Means algorithm for this specific use case

This change eliminated the issue of cells with the same patterns, but different magnitudes being placed in different buckets during clustering and thus more accurately captures the idea of clustering to identify the area by looking at specific patterns.

4.2 Biclustering

The increasing use of data mining in solving real-world problems has highlighted the fact that much of the real data is and can typically be organized in a matrix format, with a correlation matrix between every cell and every other cell. A key challenge is to cluster each entity effectively. Traditional clustering methods often fall short because they do not leverage the inherent relationship between the entities. Biclustering – also called co-clustering, block clustering or cross-clustering, among others – is a technique that clusters the cells within a correlation matrix. [62] Therefore, these so-called biclusters do consist of a subset of data instances and a subset ofclosely related attributes. [16, 54] Biclustering utilizes the relationships between entities to uncover data structure more effectively. [69] Biclustering algorithms can vary significantly based on the criteria for forming biclusters. Common criteria include, as defined by Pedregosa et al. [62]:

- Identifying constant values, rows, or columns

- Detecting exceptionally high or low values
- Finding submatrices with only minimal variance
- Analyzing correlation among rows or columns

The structure of biclusters, such as block diagonal or checkerboard patterns, depends on how rows and columns are grouped. In a block diagonal structure, if every row and column is part of one unique bicluster, then reordering them can display the biclusters along the matrix's diagonal. The representation of the interspaces suggests correlations between the clusters.

4.3 Spectral Biclustering

As has been described previously, identifying effective clusters has attracted significant attention in the research and application of machine learning and pattern recognition. Spectral clustering is a technique that leverages the eigenvalues and eigenvectors of graph-related matrices to identify groups of nodes within a network. This approach was initially proposed by Donath and Hoffman as well as Fiedler in 1973. [18, 26] They suggested utilizing eigenvalues and -vectors from graph Laplacian matrices to identify clusters in undirected graphs. Spectral clustering has emerged as a popular clustering algorithm in recent years, mostly due to its simplicity and efficiency with which it can be executed. Furthermore, it is known to often outperform more conventional clustering methods like the K-Means algorithm. [77] Spectral clustering can therefore be described as a technique used to partition a dataset into distinct groups or clusters, where the partitioning is based on the spectrum (eigenvalues) of a similarity matrix derived from the data.

The foundational elements of spectral clustering are graph Laplacian matrices, which are central to spectral graph theory. [77] Unnormalized spectral clustering, a specific variant of this technique, leverages the graph Laplacian matrix without normalizing it.

The unnormalized graph Laplacian matrix is defined as $L = D - W$, with D being the diagonal matrix and W standing for the weighted matrix. [79]

The algorithm, also illustrated in Algorithm 4.2 [77], proceeds by computing the eigendecomposition of the Laplacian matrix L. For unnormalized spectral clustering, the focus is on the first k smallest eigenvalues and their corresponding eigenvectors, where k is the number of clusters to be identified. The next step involves forming a matrix $S \in R_n \times k$ from the selected eigenvectors, applying the clustering directly to the rows of S using a standard algorithm. This process effectively partitions the graph into k clusters. [79] Using Laplacian Eigenvectors for clustering leverages their ability to capture the graph's connectivity. Eigenvectors associated with smaller eigenvalues could indicate connected components or tightly knit groups, therefore identifying groups of vertices (data points) that are more densely connected to each other than to the rest of the graph.

Algorithm 4.2: Unnormalized Spectral Clustering, based on [77]

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, k clusters
Output: Clusters $A_1, A_2, \ldots, A_k$

```
Construct a similarity graph G using S and determine the weighted adjacency
  matrix W
Compute the unnormalized Laplacian L = D − W where D is the diagonal degree
  matrix
Calculate the first k eigenvectors u_1, ..., u_k of L and form matrix U ∈ R^{n×k} by
  stacking them as columns; for i = 1 to n do
    Assign to y_i ∈ R^k the vector corresponding to the i-th row of U
end
Apply K-Means to points y_1, ..., y_n to identify clusters C_1, ..., C_k
for j = 1 to k do
    Define cluster A_j as {i | y_i ∈ C_j}
end
```

4.3.1 Spectral biclustering Algorithm

Spectral biclustering is designed to simultaneously cluster the rows (samples) and columns (features) of a matrix, uncovering patterns not just within samples but also among subsets, thereby aiding in the discovery of localized structures within the data. This technique is particularly suited for datasets where the sequence or configuration of features remains constant, such as in time series data. Spectral biclustering considers the dual relationship between two types of entities (for example, documents and words, users and items). The goal is to maximize interaction within and minimize it across co-clusters, thereby revealing hidden structures by evaluating both dimensions of interest. [62]

The Spectral biclustering algorithm assumes a checkerboard structure within the input data matrix. To approximate the original matrix with a checkerboard pattern, the algorithm initially normalizes the input matrix through one of three methods: row and column normalization, bistochastization (repeated normalization to equalize row and column sums), or log normalization (adjusting based on the logarithm of the data matrix and its row, column, and overall means).[49, 62]

In the course of this thesis, unnormalized spectral biclustering was employed on previously aggregated temporal data rather than the raw data set. This approach facilitated the identification of trends, upon which the unnormalized clustering was applied to discern patterns effectively.

Algorithm 4.3: Spectral Biclustering

Input: Data matrix $A \in \mathbb{R}^{m \times n}$, number of row patterns r, number of column patterns c

Output: Matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times c}$

1 Standardize A for zero mean and unit variance across rows and columns
2 Construct row and column similarity matrices S_{row} and S_{col}
3 Compute the unnormalized Laplacians $L_{\text{row}} = D_{\text{row}} - S_{\text{row}}$ and $L_{\text{col}} = D_{\text{col}} - S_{\text{col}}$
4 Calculate the first r eigenvectors $u_1, \ldots, u_r$ of L_{row} and the first c eigenvectors $v_1, \ldots, v_c$ of L_{col}
5 Form the matrices U and V by placing eigenvectors $u_1, \ldots, u_r$ and $v_1, \ldots, v_c$ as columns respectively

4.4 Evaluating Biclustering Results

After applying the Spectral Clustering algorithm to the data for each of the four chosen features (present UEs, connected users, active users, PRB utilization) and all cells for a number of clusters in the range of two to seven clusters (theoretically allowing for a separate trend for each weekday, as trends on a larger scale are repetitive), the optimal number of clusters was evaluated with metrics described in the following. For this specific problem statement and the following goals of the Institute of Telecommunications to identify predefined areas based on the utilization of a given cell, other considerations have to be kept in mind, as well (usually increasing the number of desired clusters, as a binary distinction would not allow specific identification of all predefined areas (some examples can be, as mentioned above, shopping malls, rural housing area, industrial districts, nightlife areas, among others.).

For evaluating the quality of the biclustering results, three different metrics have been used for evaluation: the Davies-Bouldin score, the Silhouette score and the Calinski-Harabasz score. The selection of the three scores was based on their suitability for unsupervised learning, given the absence of actual labels. This approach ensures an appropriate evaluation framework for biclustering analysis in scenarios where prior label information is not available.

4.4.1 Davies-Bouldin Score

The Davies-Bouldin score is a function to evaluate the clustering performance by measuring the maximum ratio of distances withing its cluster to distances between clusters. [15] A lower score signifies the minimum possible similarity for the given clusters, therefore indicating a better clustering quality. [30, 82] The Davies-Bouldin index offers simplicity in computation compared to a Silhouette score (see below). [62]

The index is characterized as the average similarity between cluster C_i for $i = 1, \ldots, k$ and its most similar cluster C_j. According to Pedregosa et al. [62], this similarity is defined as measure R_{ij}:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

where s_i and s_j represent the average distance between each point within the cluster and it's centroid d_{ij} is the distance between the centroids of the two clusters

The Davies-Bouldin-Index is then defined as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

4.4.2 Silhouette Score

The Silhouette Score evaluates the similarity of a single object to its cluster in comparison to other clusters, with values ranging from -1 to 1. A higher score signifies a stronger match with the object's own cluster, therefore indicating better-defined clusters, while a negative score implies that the samples may have been placed in their current cluster incorrectly. [65] The Silhouette Coefficient is calculated for each individual sample in relation to other data points and consists of two scores – a (average distance between the sample and the other points in the same group) and b (average distance between the sample and the other points in the next nearest cluster) [62, 68]:

$$s = \frac{b - a}{max(a, b)}$$

For biclustering, the Silhouette Score is particularly useful because it assesses both the cohesion within clusters (the closeness of members within a cluster) and the separation between clusters (how distinct one cluster is from others).

4.4.3 Calinski-Harabasz Score

The Calinski-Harabasz score, also known as the Variance Ratio Criterion, is another measure to evaluate a model, wherein a higher score signifies better-defined clusters, measured by their denseness and separation. [9] It is defined as the ratio of the sum of between-cluster dispersion to the sum of within-cluster dispersion for all clusters. **??**

If dataset E of size n_E is clustered into k clusters, the score s is defined as the ratio of the mean dispersion between clusters to the mean dispersion within clusters. This is then given as follows [66]:

$$s = (\frac{tr(B_k)}{tr(W_k)}) \times (\frac{n_E - k)}{k \check{} 1})$$

The trace of the dispersion matrix between clusters $tr(B_k)$ is being defined as:

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

The trace of the dispersion matrix within clusters $tr(W_k)$ is being defined as:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

with C_q representing the set of points and n_q being the number of points in cluster q and also c_q being the center of cluster q and c_E being the center E. [66]

4.5 Results from the dataset

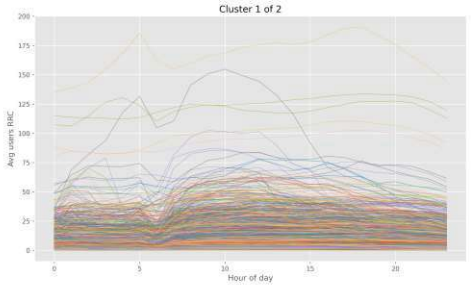
Aside from the previously described metrics, the clustering quality can be reviewed using the correlation values (or a plot thereof), which indicates the discriminatory power of the currently chosen number of clusters. Additionally, the cells within each cluster have been analyzed graphically with specific interest in a separation that allows room for future applications identifying predefined areas and plan ahead for large social events such as sport events, concerts, demonstrations and such. While the following analysis is aimed at the number of user equipment in downlink. While a similar trend holds true for other variables when increasing the number of clusters, the specific peaks during the time of day vary, but will not be elaborated on in the course of this thesis.

The cluster analysis was carried out using the features described above. The results are described and presented below using two variables: Based on the variable connected user equipment, which has shown the clearest selectivity between the Clusters and therefore serves for better illustration, and on the variable user equipment in the downlink, which was prioritized in the further course due to its practicability. This was done on the assumption that the actual use of the cellular network could provide more information than just the number of connected users.

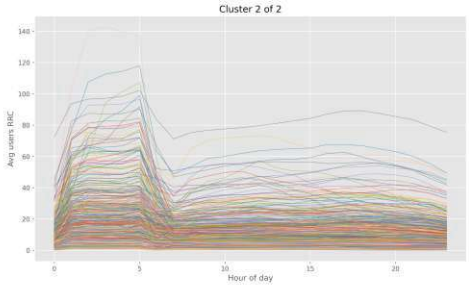
4.5.1 Two Clusters

As an example for the evaluation of the clustering quality of the users in the downlink (which imposes the direct interest of the MNO as those users are the most impactful on the network utilization and he is bound to fulfill their contracts) is analyzed. Starting with the smallest amount of two bins by the hourly trend, the correlation matrix (and its plot) suggest that for each of the ≈ 1500 cells a clean split, as two yellow cubes in figure 4.2c, seem visible. This is caused by the algorithm that shifts the cells that correlate with each other next to each other and those with weak correlation farther apart. The intra-cluster correlations are presented with low numbers resulting in blue corners in the correlation plot with the chosen color scheme.

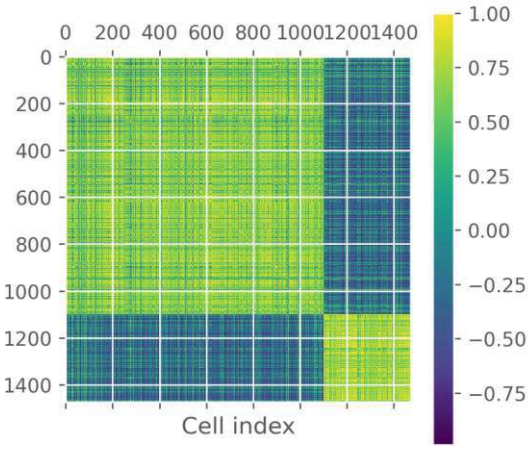
When looking at a line plot for each cell in that cluster, it becomes evident that some differing progressions over time are present for some cells. While for the first cluster, shown in Figure 4.2a, shows some cells that have a peak in the second half of the night, some display amplitudes in the first half of the day (insinuating, for example office complexes.). Something similar is true for the second cluster, as pictured in Figure 4.2b, that also has some cells with shorter peak times in the second half of the night as well as some smaller, but still distinguishable peaks between the morning and noon.



(a) Cluster 1 of 2 for connected UEs



(b) Cluster 2 of 2 for connected UEs



(c) Correlation plot for two clusters

This phenomenon clearly demonstrates the need for evaluations aside from pure numerical metrics and the power of having a look at specific data points and cells.

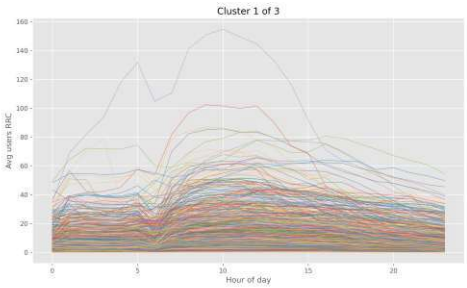
4.5.2 Three Clusters

The previously described phenomenon of some cells displaying spikes whereas others do not is dampened by introducing a third cluster. While the first cluster, shown in Figure 4.3a, still shows some noisy patterns in the second half of the night, most of the

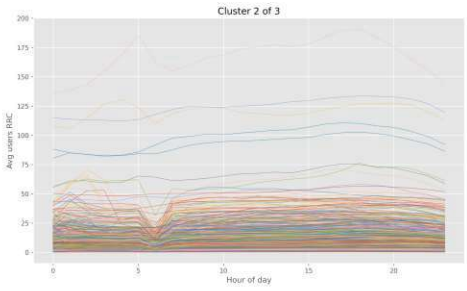
pattern-outliers have been moved to a separate cluster. The second cluster, Figure 4.3b, also looks cleaner as there are fewer cells with spikes and most of the connected user lines over time lie almost parallel – with some exceptions still.

The third, newly introduced cluster, visible in Figure 4.3c, by increasing the number of buckets, has inherited most of the cells that introduced noise to the previous distinction as well as most cells with constantly low utilization, seemingly collecting all remaining cells that did not fit the patterns of the other clusters.

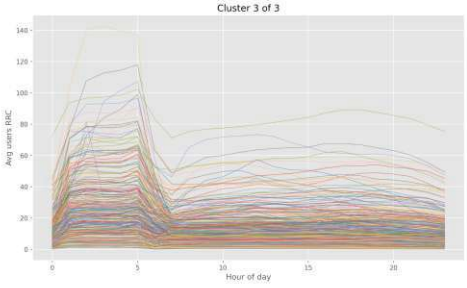
This is also reflected in the correlation plot, shown in Figure 4.3d: Cluster one and cluster three both show little correlation with the second cluster, whereas the third cluster and the first are undeniably correlated. This is explained by most cells not fitting the patterns in other clusters being shifted to the third – which happened fewer for the second cluster.



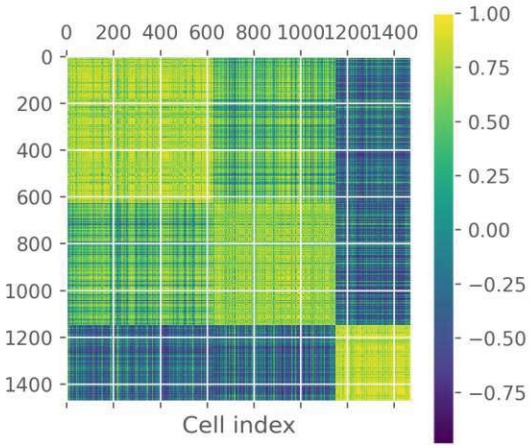
(a) Cluster 1 of 3 for connected UEs



(b) Cluster 2 of 3 for connected UEs



(c) Newly introduced, third cluster



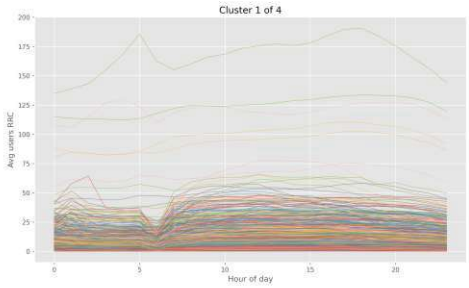
(d) Correlation plot for three clusters

4.5.3 Four Clusters

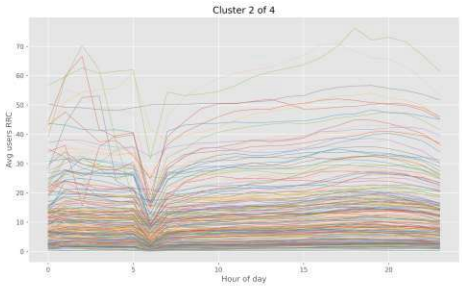
While keeping the overall pattern of distinction the same after introducing a fourth cluster, the new fourth bucket (Figure 4.4b) contains the outliers that were previously in the second cluster, and contains cells with high utilization in the especially early and late hours of the day (night heavy utilization). Other cells from the second cluster according to previous clustering separations were also moved to the third cluster, as shown in Figure 4.4c.

As those cells have previously been part of the second cluster, the correlation to the second cluster (Figure 4.4b) seems logical. Other clusters show a low intra-cluster correlation, as previously evident when looking at the correlation plot illustrated in Figure 4.4e.

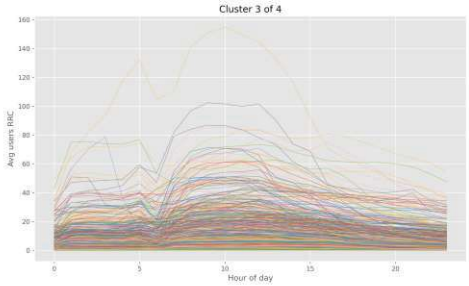
While the reasons for introducing a fourth cluster can still be justified, the benefit, uniqueness and distinguishability from the patterns of other cells in their respective previous clustering assignment does decrease in magnitde and definetly does not warrant the addition of another cluster. However, for the sake of completeness and the fact that other variables' patterns look slightly different, the addition of a fifth cluster or rather a neglection thereof will be considered in the following for completeness's sake.



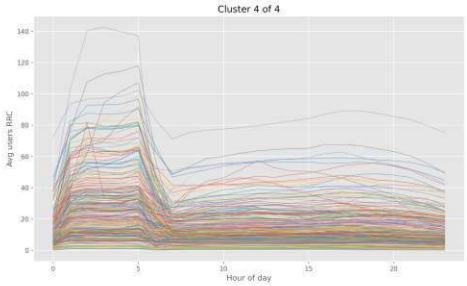
(a) Cluster 1 of 4 for connected UEs



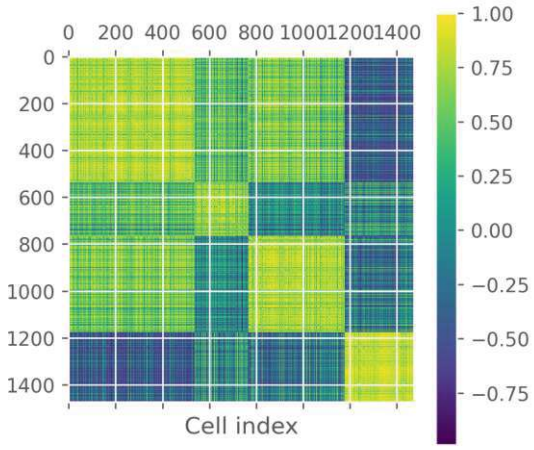
(b) Cluster 2 of 4 for connected UEs



(c) Cluster 3 of 4 for connected UEs



(d) Newly introduced, fourth cluster

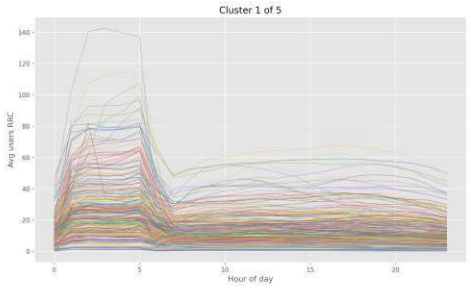


(e) Correlation plot for four clusters

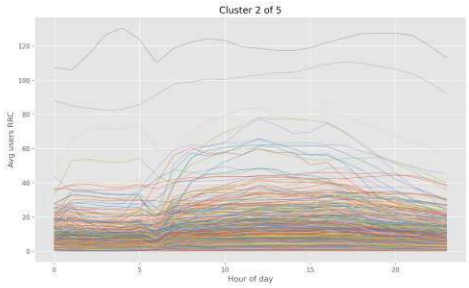
4.5.4 Five Clusters

With the introduction of an additional, fifth cluster, not many gains can be observed by looking at the line plots for each specific cluster (see Figure 4.5e). They contain specific flat lines, with the single exception of a cell that has a single amplitude in the morning

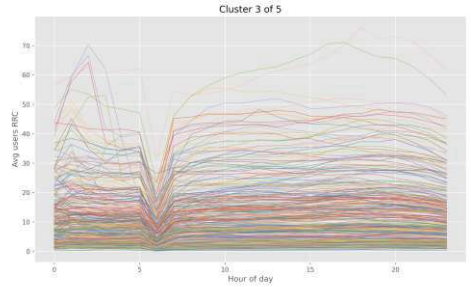
and two in the afternoon.



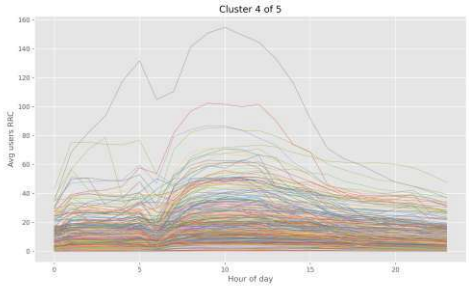
(a) Cluster 1 of 5 for connected UEs



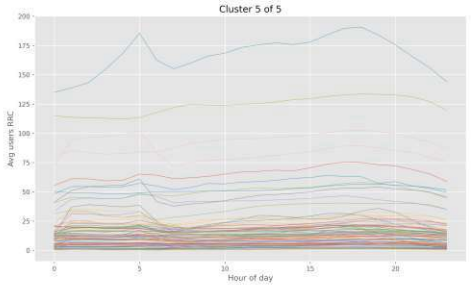
(b) Cluster 2 of 5 for connected UEs



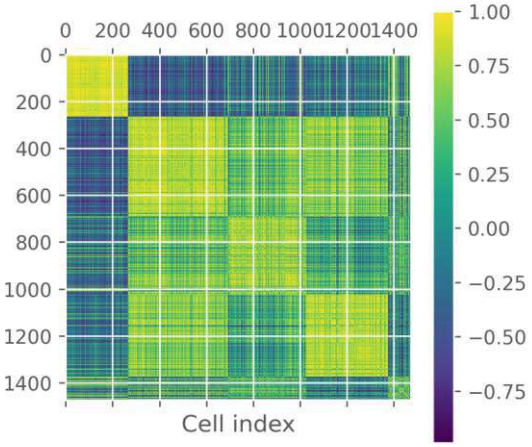
(c) Cluster 3 of 5 for connected UEs



(d) Cluster 4 of 5 for connected UEs



(e) Newly introduced, fifth cluster



(f) Correlation plot for five clusters

In summary, the introduction of a fifth cluster does neither make sense numerically nor with special consideration of geographic distinction, as it also highly correlates with three out of four other clusters, as is evident in 4.5f.

4.5.5 Identifying the optimal number of clusters

With the knowledge of all the previously described observations, it makes sense to choose five clusters for the upcoming analysis. However, the specific metrics for the Davies-Bouldin Score, the Silhouette Score as well as the Calinski-Harabasz Score have not been taken directly into consideration. Figure 4.6 displays the metrics over time when increasing the number of clusters, with metric scales visible per score on the y-axis. Purely orienting oneself to the metrics from the data, one would clearly decide for a clustering into two clusters. However, considering the importance of offering enough clusters for geographical distinction as well, and the low magnitude of change in the metrics, the choice of four clusters makes more sense. The Davies-Bouldin Score, where a lower score indicates better results, as well as the Calinski-Harabasz Score and the Silhouette Score, for both of which higher scores mean better results, do not suffer from increasing the number of clusters from three to four.

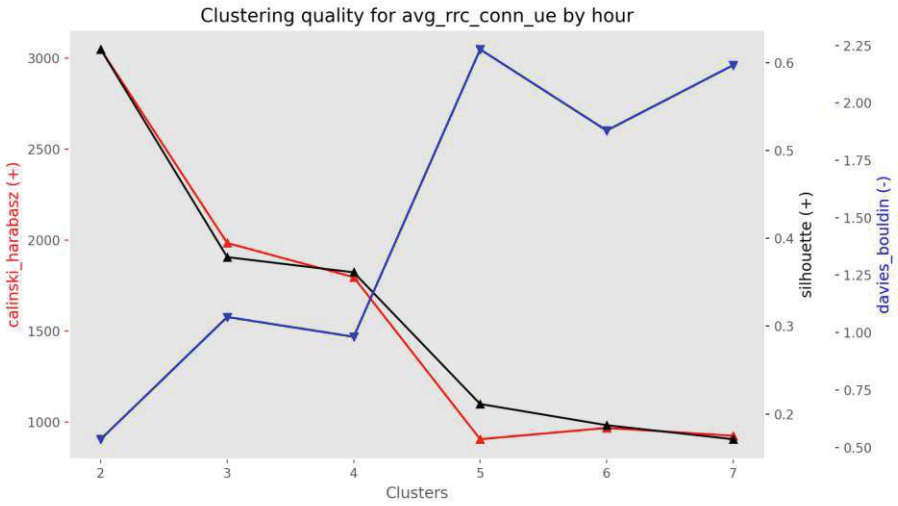


Figure 4.6: Clustering quality: Overview of the metrics' development for each number of clusters

Following the logic described above, that the fifth cluster does not offer sufficient uniqueness to warrant its introduction, and paired with the numerical backing of this decision leads to the decision that the number of clusters can be automatized. However, the analytical approach to the individual patterns within the clusters should better be done manually.

Even though the metrics vary for the other features, the same statement holds true when drawing conclusions from the evaluation metrics (see 4.6). While the measured scores are optimal for two clusters, the solutions with three clusters decrease the numeric quality of the result, but the need for more clusters warrants the introduction of even another cluster, as the metric solution only decreases slightly for a major improvement of the

possibilities in geographic tagging of areas.

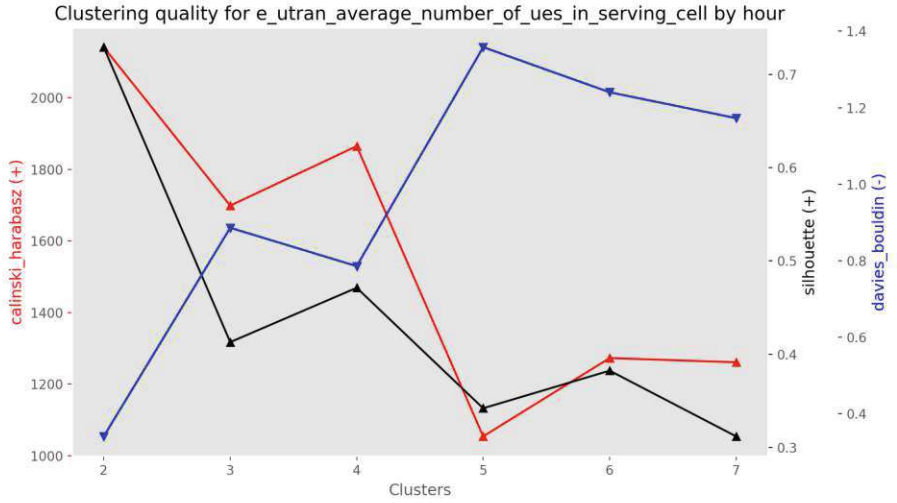


Figure 4.7: Present UEs

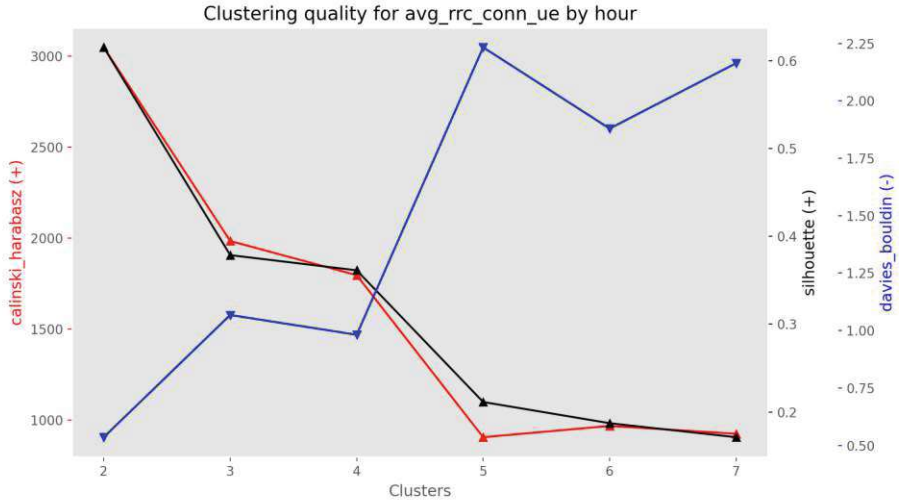


Figure 4.8: Connected UEs

This chain of logic halts at the introduction of a fifth cluster. Scores alongside the line plots for each cell and the corresponding correlation plots suggest a strong correlation with buckets already present before and the lack of uniqueness of the new clusters.

This confirms the assumption that four clusters were the optimal solution for the case at hand.

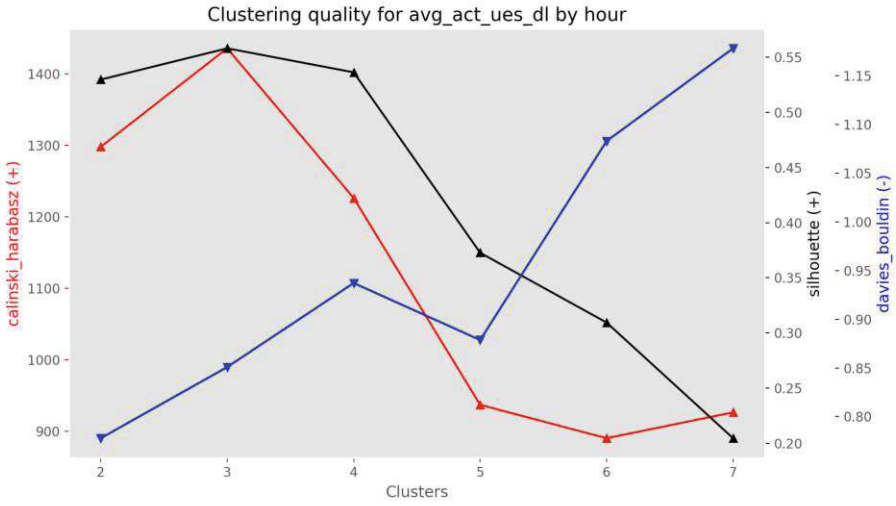


Figure 4.9: UEs in downlink

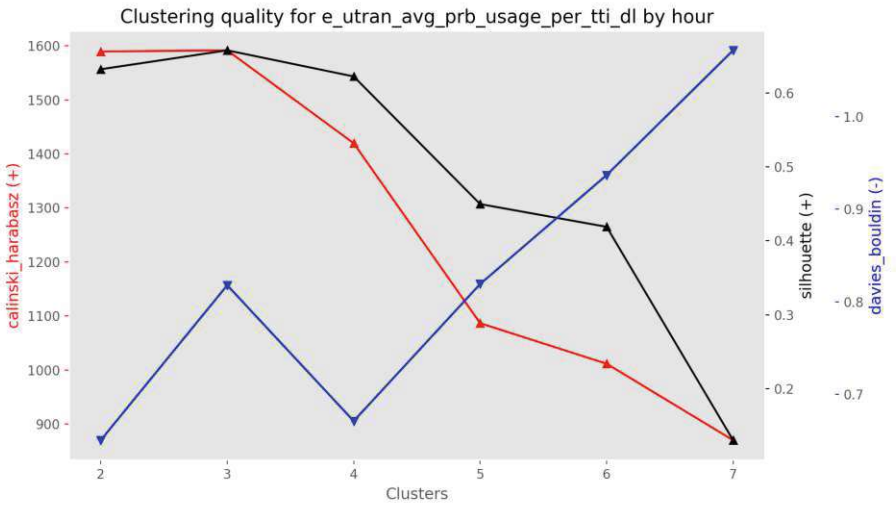


Figure 4.10: PRB utilization

Additionally, he cell towers are distributed as follows:

Figure 4.11 shows a plot of the distribution of the cell towers per frequency band as in Figure 3.2. Each cell tower is represented by a pie chart that shows how many frequencies are serviced by that specific tower, to once again gain a better understanding of the distribution and pattern of the towers in focus of this thesis.

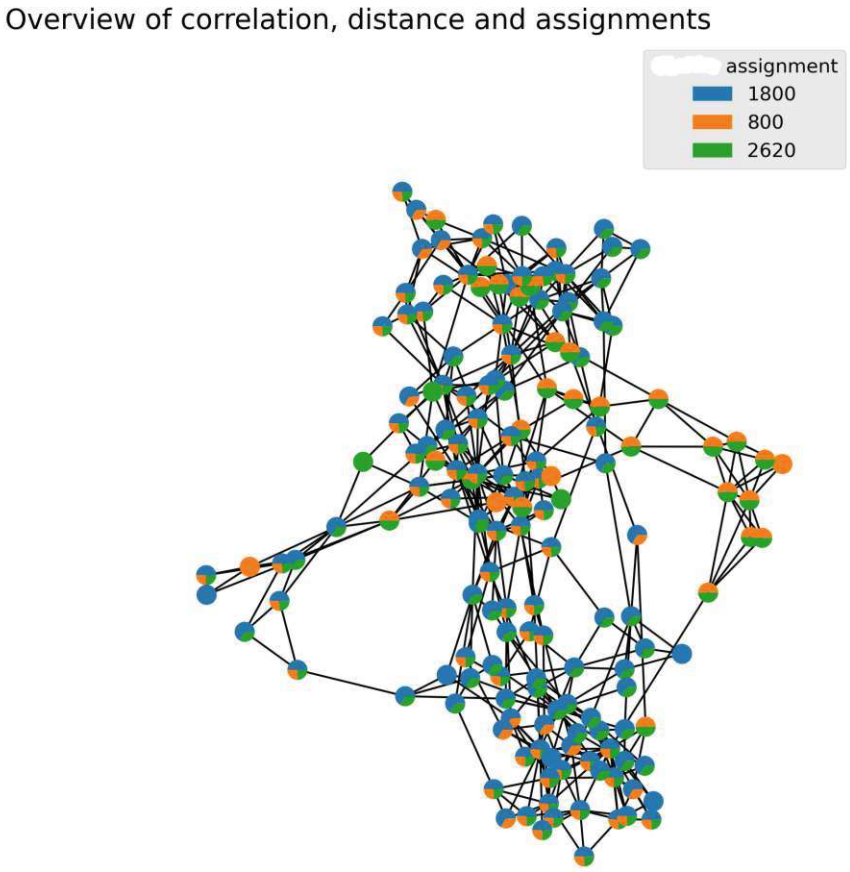


Figure 4.11: Pie charts per location for the assignments to frequency bands

CHAPTER 5

Timeseries prediction

Forecasting is a critical component in the management of telecommunications networks, enabling operators to predict future demands, optimize resource allocation, and ensure robust network performance.

Time series forecasting models consider various patterns in the data to make predictions – for example trends, seasonality and noise. Trend refers to the long-term evolution of the data, seasonality indicates regular patterns within specific time frames (such as daily, weekly, or annually) and noise shows random changes in the data that cannot be directly attributed to the aforementioned patterns.

The theoretical considerations set the stage for the application of three forecasting methods in the experiment stage of this thesis, each chosen for the use case at hand. The subsequent sections will provide an overview of the chosen models – namely, ARIMA, Holt-Winters and Prophet from the Facebook Research Team.

5.1 ARIMA

ARIMA (Autoregressive Integrated Moving Average)is a popular statistical approach for time series forecasting. The model is favored for time series without a clear seasonal pattern, which is relevant for the data set used in this thesis. It performs well in short-term prediction, but shows a decline in performance when data is missing or there are strong seasonal characteristics. [8]

It consists of three components [25]:

Autoregression (AR) This part assesses the relationship between the current observation and past observations at different time intervals. An AR(p) model is defined by the parameter p, representing the number of past errors to use within the model.

Differencing (I) This element (d) renders the data to stabilize its statistical properties by effectively eliminating trends and other seasonal variations. This includes the order of differing d values, showing how often the time series was differentiated before the actual modeling.

Moving Average (MA) This component models the error term based on previous error terms. The MA(q) model represents the order q which in turn indicates the number of past errors (lags) within the model. Modeling the correlation between the current value and the previous errors is crucial in the model.

Summarized, the ARIMA (p,d,q) model is characterized by the three parameters mentioned above - p (number of autoregressive terms), d (number of non-seasonal differences required to achieve stationarity), and q (number of lagged forecast errors).

Due to their versatility and ease of application, the ARIMA models prove to be widely successful in various disciplines, including economics, energy, transportation and finance. But despite that, they come with some limitations. ARIMA models assume that the time series is linear and requires the time series to be stationary. Additionally, choosing the parameters (p,d,q) is, to some extent, subject to individual interpretation, relying on a subjective choice of criteria. [25]

5.2 Holt-Winters

The Holt-Winters method, also known as Triple Exponential Smoothing method, is a commonly used technique for time series forecasting, particularly valued for its ability to handle seasonal patterns such as daily, weekly, or yearly cycles. It comprises three key components: level, trend, and seasonality. The level component shows the smoothed value of the time series, the trend component reflects the overall direction of the data, and the seasonality component represents periodic fluctuations.[13]

This method is suitable for time series data with either or both trend and seasonal components, making it a versatile choice for various applications. Holt-Winters is employed across multiple domains, including tourism, energy, and climate studies, where it aids in capacity planning, infrastructure development, and resource management. [53]

Despite this widespread applicability, the Holt-Winters method has some limitations. Similar to ARIMA models, it assumes linearity and requires the time series to be stationary. Additionally, it lacks precise criteria for parameter selection, which can affect forecasting accuracy. Nonetheless, its simplicity and ease of use contribute to its enduring popularity in time series forecasting. [53]

5.3 Prophet

More traditional methods like ARIMA and Holt-Winters often demand significant expertise in their field of application and some manual work. In contrast to that, Prophet

– a forecasting tool developed by Sean Taylor and Benjamin Letham of the Facebook Research Team – offers a more straightforward, automated solution. [75] It employs a decomposable time series model to forecast trends, seasonality, and holiday effects in data sets. It combines a GAM (generalized additive model) with Bayesian parameter estimation and is designed to be user-friendly, resilient to outliers and missing data and scalable in its approach. Prophet is most often used when forecasting not only yearly seasonal patterns, but also daily or weekly ones. [60]

The Prophet model has found applications across various fields, including finance, retail, healthcare, and energy. However, challenges and limitations exist, which need consideration for its optimal usage. Prophet assumes data generated by decomposable models with additive or multiplicative seasonality. It still needs users to fine-tune parameters for optimal performance and challenges may arise with irregular or high-frequency data that might require additional processing beforehand. [45]

Despite these challenges, Prophet has become a widely-used model, mainly due to its applicability for a wide range of forecasting tasks. The implementation of Prophet in this thesis will focus on both univariate and multivariate forecasting models (implemented via multiple regressors). The univariate model will predict basic load patterns, while the multivariate model will incorporate additional inputs, for example from other clusters. Although Prophet does not natively support multivariate time series forecasting, additional regressors can be incorporated to add multiple features from the dataset and using the `add_regressor()` method, allowing it to account for multiple factors influencing the primary variable. This approach then results in an ensemble that is further used for the prediction process.

5.3.1 Univariate prediction methods

In the first deployment of the Facebook model, hereinafter referred to as "Prophet Univariate," was employed to forecast the number of user equipment in the downlink. This univariate analysis focused solely on the target variable, *avg_act_ues_dl* (Average Active Users in Downlink), without incorporating any additional features or external influences. By utilizing historical data of *avg_act_ues_dl*, the model aimed to predict future values based on the inherent patterns and seasonalities present in the data. This method provides a baseline understanding of the model's performance when restricted to a single-variable input, allowing for an evaluation of its capacity to capture the underlying dynamics of user activity within the network.

5.3.2 Multivariate prediction methods

The second approach, namely "Prophet Multivariate Correlations," is a multivariate model by integrating the trend of the same feature from the five strongest correlation partners as additional features. These partners, represented by the same feature from other cells within the network, were selected based on their high correlation with the target variable. The inclusion of these correlated features aimed to enhance the accuracy

of predictions by showing interdependencies between different cells. This analysis did not consider the exact geographical positions of the correlated cells but focused on their statistical relationships with the target feature *avg_act_ues_dl*. The goal was to determine whether the inclusion of correlated data from other cells could improve the model's ability to forecast connected user entities more accurately. This model orients itself on the strongest correlation partners according to the clustering outcomes described in Chapter 4.5.

The third variant, hereinafter mentioned as "Prophet Multivariate Features," incorporated three additional features from the same cell as the target variable one. These features, detailed in Chapter 3.1 were selected based on their relevance and potential impact on the target variable, *avg_act_ues_dl*. By introducing these additional regressors, the model aimed to capture a broader spectrum of factors influencing user activity at the specific location of each cell. This multivariate approach was chosen to determine if a more comprehensive representation of the cell's and user's characteristics could enhance the forecasting performance.

5.4 Hyperparameter Tuning

For this thesis, a subset of the available cells was chosen for hyperparameter optimization. For that purpose, one cell of each of the 130 eNodeBs in the 1800 MHz (because they are by far the most common and thus most representative for this data set) frequency were chosen for hyperparameter optimization via grid search. Despite its widespread applicability, the Holt-Winters method has some limitations. Similar to ARIMA models, it assumes linearity and requires the time series to be stationary. Additionally, it lacks precise criteria for parameter selection, which can affect forecasting accuracy. Nonetheless, its simplicity and ease of use contribute to its enduring popularity in time series forecasting. [62]

For the Prophet model, the values [0.001, 0.01, 0.1, 0.5] for local holiday weights and [0.001, 0.01, 1.0, 10.0] for seasonality weights. The results of this analysis showed that the best RMSE values were achieved for $\approx 46\%$ of the cases with values of 0.01 for both variables for the univariate and multivariate models. Thus, the models were run on all cells with those hyperparameters and hyperparameter optimization was not performed on all cells in the interest of runtime optimization.

Chart 5.1 for the ARIMA model displays the distribution of the most frequently used hyperparameters for the ARIMA(p,d,q) model. The combinations of p, d, and q parameters were varied extensively to find the optimal settings. The most frequent combination was (p=4, d=0, q=4), which accounted for 21.7% of the instances, which indicates a seasonal and lagged component in the data. The second most frequent combination was (p=3, d=0, q=4), making up $17,1\%$ of the instances, followed by (p=4, d=0, q=3) and (p=3, d=0, q=3) with $13,2\%$ and $8,5\%$, respectively. This distribution again shows a strong emphasis on lag and seasonal parameters. For the further course of the forecasting, the most frequent parameter combination was used.

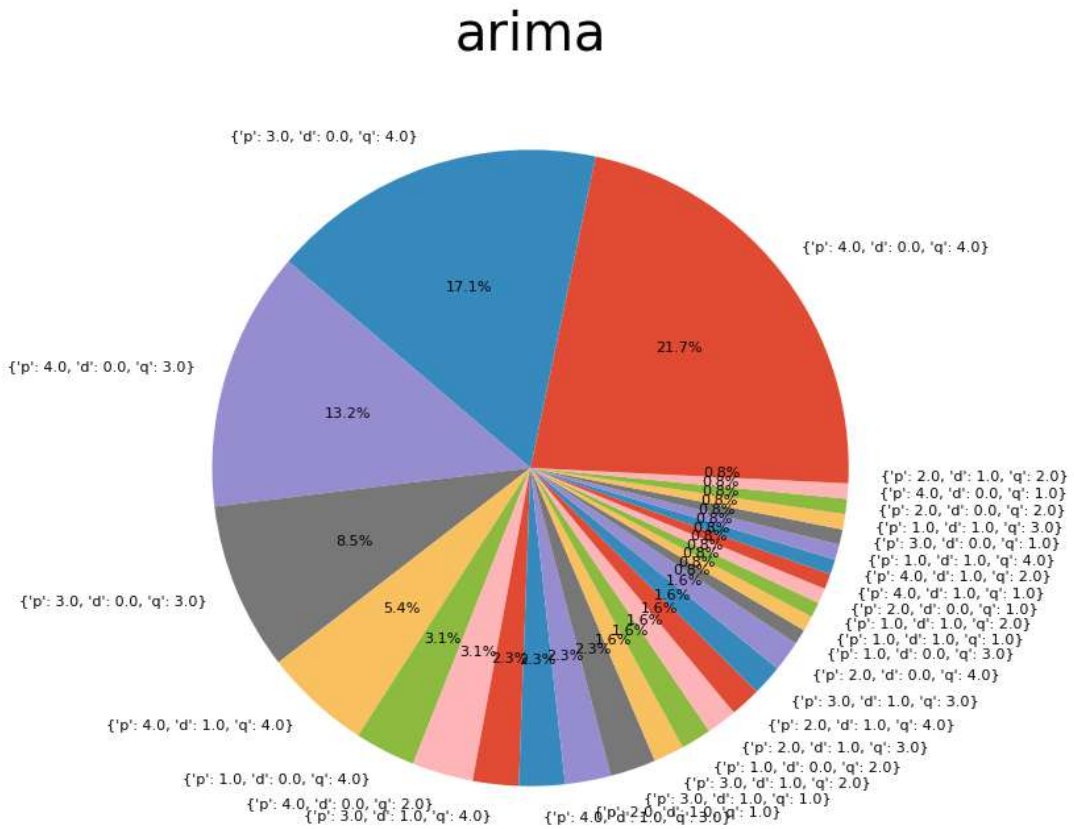


Figure 5.1: Histogram of the hyperparameters of the ARIMA model

For the Holt-Winters model, the hyperparameter tuning results shows a single hyperparameter combination: (Seasonal Periods: 12, Trend: 'add', Seasonal: 'add'), which was used 100% of the time. This shows a fixed monthly seasonality with the additive trend and seasonality components.

The Chart 5.2 for the univariate Prophet model reflects a similar diversity in hyperparameter combinations as seen with the multivariate models. The most frequent combination, (holidays_prior_scale: 0,5, seasonality_prior_scale: 0,01), was used 14,7% of the time, followed closely by other combinations with varying prior scales. This distribution highlights the importance of tuning both holiday and seasonality prior scales to accurately capture the temporal patterns in the univariate data.

The Chart 5.3 for the Prophet model with multivariate correlations shows a diverse distribution of hyperparameter combinations. The most frequent combination, (seasonality_prior_scale: 0,01, holidays_prior_scale: 0,01), accounted for 27,1% of the instances, indicating a preference for lower seasonality and holiday prior scales. Other notable

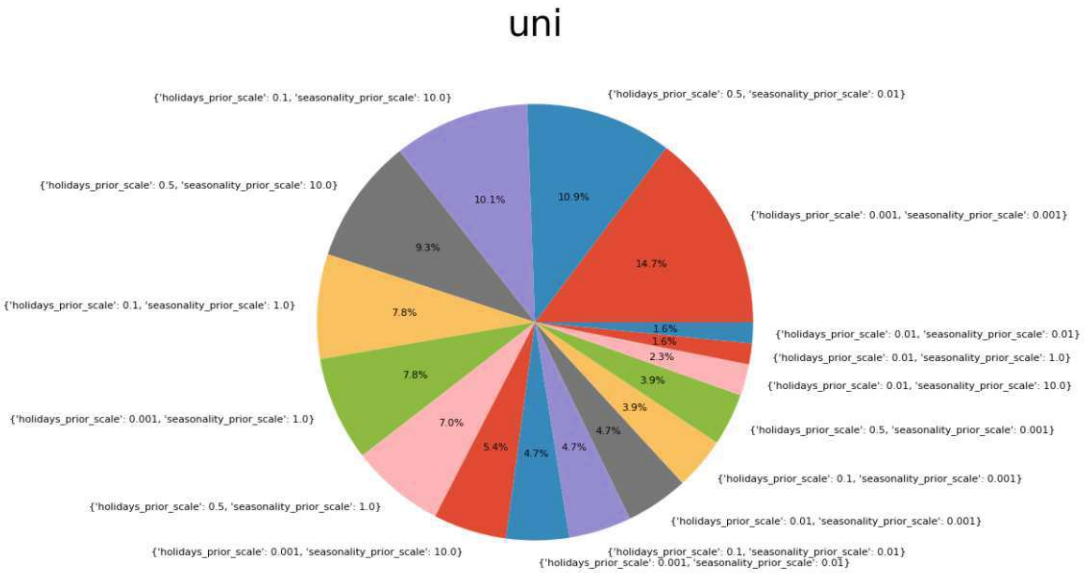


Figure 5.2: Histogram of the hyperparameters of the Univariate Prophet model

combinations included (seasonality_prior_scale: 0,01, holidays_prior_scale: 5,0) and (seasonality_prior_scale: 0,01, holidays_prior_scale: 0,1), showing varying focus on holiday effects.

The Chart 5.4 for the Prophet model with additional multivariate features also shows a diverse range of hyperparameter combinations. The most frequent combination, (holidays_prior_scale: 0,01, seasonality_prior_scale: 0,001), was used in 17,1% of the instances, followed by combinations with varying prior scales for holidays and seasonality. This variability again indicates that different features required different levels of emphasis on holiday and seasonal effects, suggesting that the model's performance depended significantly on the specific features included.

5.5 Evaluation of Results

For the time series forecasting, the data outlined in Section 3.1 was utilized with five fold cross validation. The prediction horizon for the resulting models was set to seven days, as the uncertainty grows too large after that point for most cells. In this thesis, the mean average error (MAE) and the root mean-squared error (RMSE) are used as key metrics for evaluating the model performance. These metrics will be briefly described in the following sections. MAE is an important metric for evaluating model performance because it measures forecast accuracy by averaging the absolute errors. [62]

A lower MAE means the model is performing better. Since MAE uses absolute values, it treats all errors equally, making it easy to understand and resilient to the impact of

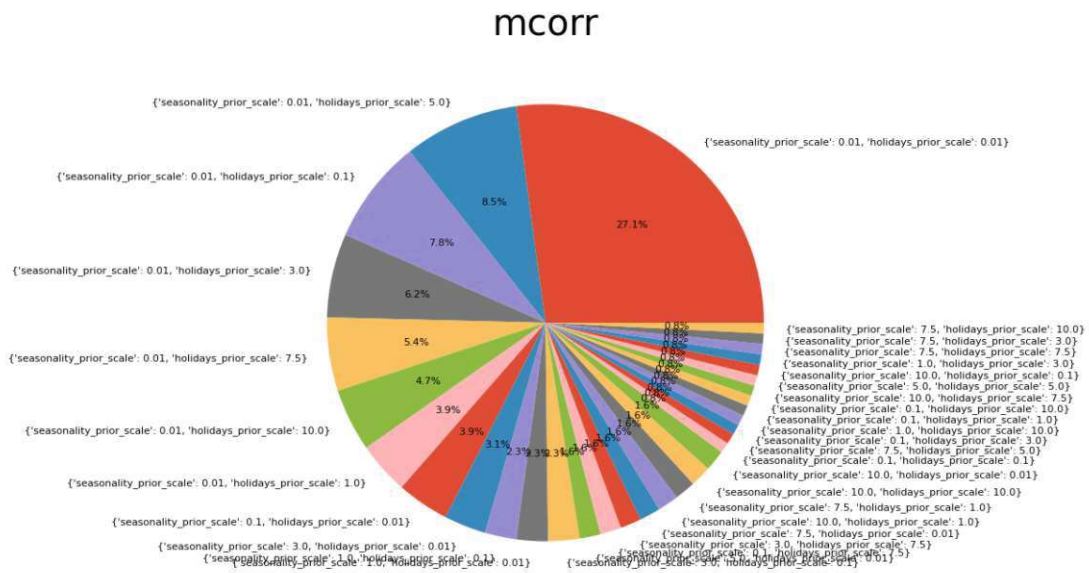


Figure 5.3: Histogram of the hyperparameters of the Multivariate Correlations Prophet Model

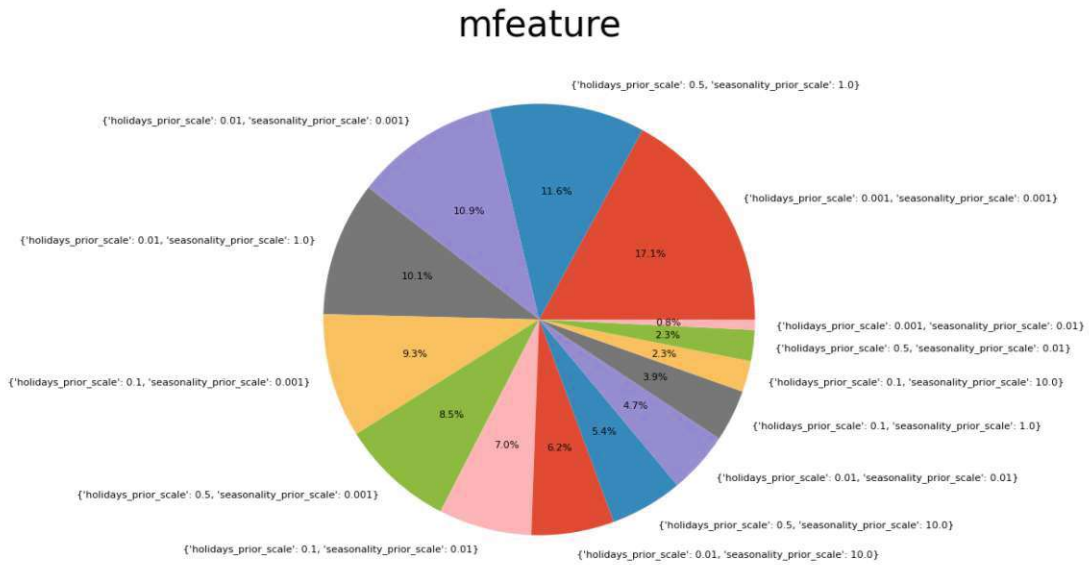


Figure 5.4: Histogram of the hyperparameters of the Multivariate Features Prophet model

outliers, making it a reliable measure of average error size. However, it does not consider whether the errors are positive or negative, which means it doesn't show if the model

consistently overestimates or underestimates the values. [11, 32] RMSE is the square root of the average of the squared differences between the forecasted values and the actual values. It penalizes larger errors more than smaller ones due to the squaring of the error terms. [62]

Similar to MAE, RMSE is also easy to interpret because it keeps the same units as the original data. However, RMSE is more sensitive to outliers, which can significantly impact the results. This sensitivity is important to consider, especially given the specific structure and data availability in this thesis. [11, 32]

The choice to use both metrics is based on their aforementioned advantages and the insights they can provide. Analyzing both MAE and RMSE offers insights into the nature of the errors, indicating whether the model's predictions are consistently off by a small margin (high MAE) or have occasional large deviations (high RMSE). Based on the research reviewed in preparation for the empirical work in this thesis, MAE and RMSE are commonly used among researchers for evaluating models in cell load prediction. [3]

Therefore, this project uses MAE and RMSE to analyze the performance of the time-series model. In the following, the results obtained from the three aforementioned forecasting techniques shall be discussed.

Modell	**RMSE**	**MAE**
ARIMA	0.24	0.17
Holt-Winters	0.36	0.29
Prophet Univariate	0.33	0.24
Prophet Multivariate Correlations	0.22	0.16
Prophet Multivariate Features	12.38	7.49

Table 5.1: Error Metrics of the models used (per cell, averaged over all cells provided in the dataset)

RMSE Results

The RMSE values for our forecasting techniques can be described as followed:

ARIMA The ARIMA model achieved an RMSE of 0.24. This indicates a relatively low level of prediction error, suggesting that ARIMA is quite effective in capturing the patterns in the cell load data.

Holt-Winters The Holt-Winters model has an RMSE of 0.36, which is higher than that of ARIMA. This indicates that while Holt-Winters can model seasonality, it might

not be as precise as ARIMA for this specific dataset and that the seasonality does not carry as much weight as anticipated. This might be improved when looking at data from an entire year thus incorporating annual trends. Another possible explanation could be the model's sensitivity to the chosen smoothing parameters.

Prophet Univariate The univariate Prophet model produced an RMSE of 0.33. This indicates moderate effectiveness, suggesting that while Prophet can capture some of the underlying patterns in the data, it might not be as finely tuned to this particular dataset as ARIMA.

Prophet Multivariate Correlation The Prophet model with multivariate correlation achieved the lowest RMSE of 0.22. This suggests that incorporating multivariate correlations significantly improves forecasting accuracy.

Prophet with Multivariate Features The Prophet model with additional multivariate features resulted in a very high RMSE of 12.38. This unusually high value indicates that the model might have overfitted the data or that the additional features introduced significant noise rather than useful information.

MAE Results

The MAE values for the forecasting techniques can be describes as followed:

ARIMA The ARIMA model yielded an MAE of 0.17. This low value suggests that ARIMA provides highly accurate predictions with minor deviations from the actual values.

Holt-Winters With an MAE of 0.29, the Holt-Winters model shows higher average error compared to ARIMA, indicating less precise predictions. Similar to the RMSE, this might indicate that due to its sensitivity to seasonality assumptions it may not align perfectly with the cell load data characteristics.

Prophet Univariate The univariate Prophet model has an MAE of 0.24. This value is lower than that of Holt-Winters but higher than ARIMA, showing moderate prediction accuracy. The MAE indicates that while Prophet captures some of the data's trends, it does not perform as consistently as ARIMA, potentially due to the lack of additional contextual features.

Prophet Multivariate Correlation The Prophet model incorporating multivariate correlation features achieved the lowest MAE of 0.16. This demonstrates the highest accuracy among all models tested, with the smallest average prediction error. The inclusion of correlated features helps the model to make more informed predictions, reducing the average error and improving reliability.

Prophet with Multivariate Features The MAE for this model is 7.49, which is significantly higher than all other models. This aligns with the RMSE result below, indicating poor performance and suggesting that the model may have struggled with the additional features.

The comparative analysis of RMSE and MAE values allows for the conclusion that the multivariate Prophet model with correlation is best suited for scenarios like the case presented in this thesis, where relevant multivariate correlations can be leveraged. This model did, averaged over all of the cells in the data set, achieve the lowest RMSE and MAE values, indicating its capacity for the most accurate and reliable predictions. Therefore, it might be assumed that is more suitable than the other models for situations where additional data sources provide meaningful context to cell load predictions. It also strenghtens the assumption that the clustering can not only be utilized for geographic tagging, but also to improve prediction quality and thus the service reliability provided by MNOs.

The ARIMA model also proved to be a robust option, especially for univariate time series forecasting when the primary focus is on capturing trends and seasonality within the cell load data. In such cases, the univariate Prophet model might, based on the data presented in this thesis, also prove to be an effective choice.

The Holt-Winters model should be considered when seasonality is a significant factor but high precision is not as much of a critical factor. Its higher RMSE and MAE values, compared to ARIMA and Prophet, suggest less precise forecasts.

Interestingly, the Prophet model with multivariate features should be approached with caution. Its very high RMSE and MAE values suggest issues such as overfitting or the inclusion of noisy features or ones with less predictive power, leading to poor performance. This model should only be considered if there is compelling evidence that the additional features significantly enhance prediction accuracy. This might be possible with different, more accurate data sets that offer a broader selection of input variables that offer more impact and more explanatory power.

In summary, the multivariate Prophet model focusing on correlation and ARIMA models emerge as the top performers for cell load forecasting, offering high accuracy and reliability. The decision between these models should be based on the availability and relevance of multivariate data.

5.6 Examplary showcasing of the results

To give a better understanding of the data at hand and the prediction results, one cell is used as an example to illustrate the entire forecasting process showcasing results of all models used. The training data covers a period of five months, showing significant variability in cell load with a rapid drop during Week 15. The test data follows the

training period and is used to validate the model's forecasting accuracy by comparing the predicted values against actual observed data.

5.6.1 ARIMA Results

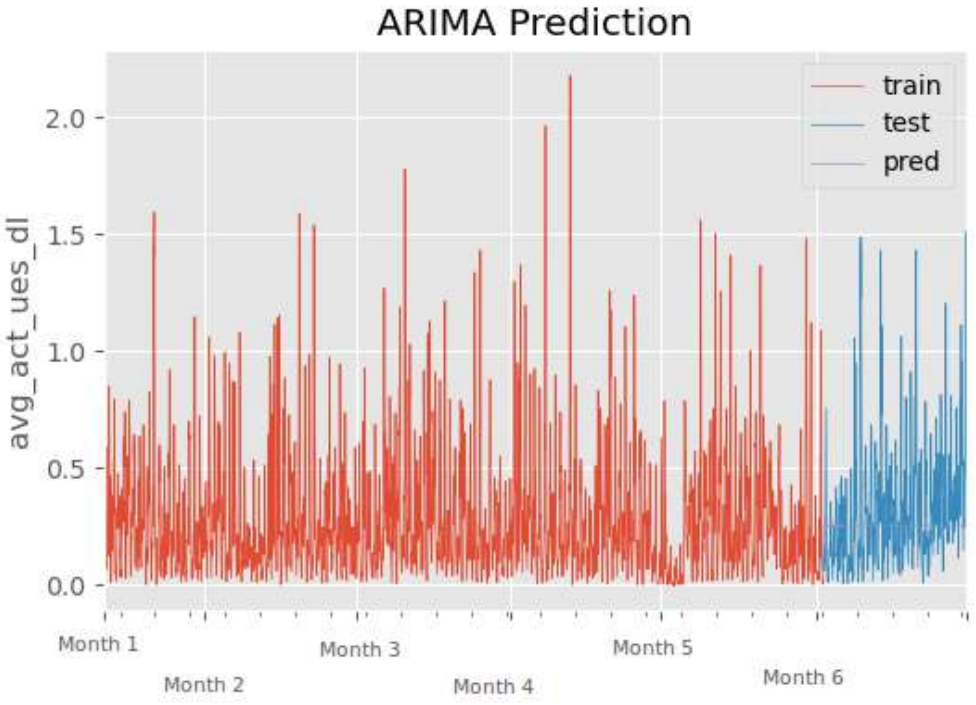


Figure 5.5: Results of Implementing the ARIMA Model

Figure 5.5 presents the ARIMA model's forecasting results for a specific cell (O533sL12). The graph is divided into three sections: training data (in red), testing data (in blue), and predictions (in purple). The predicted values align closely with the actual values in the testing period, which indicates that ARIMA is capturing the cell load pattern for this cell effectively. Given the overall prediction performance of ARIMA with an RMSE of 0.24 and an MAE of 0.17, the example of this cell too is showing that ARIMA handles the general trend of the data well. All of this should be taken into consideration with caution, as the model picked up on the recurring patterns along the lines of the rising trend well, which is also reflected in the RMSE values, but does not offer meaningful insights how the true utilization might proceed.

5.6.2 Holt-Winters Results

Figure 5.6 displays the forecasting results using the Holt-Winters exponential smoothing model for the cell. This graph is segmented into the training period (in brown), the testing period (in yellow), and the predicted values (in green). The model seems to

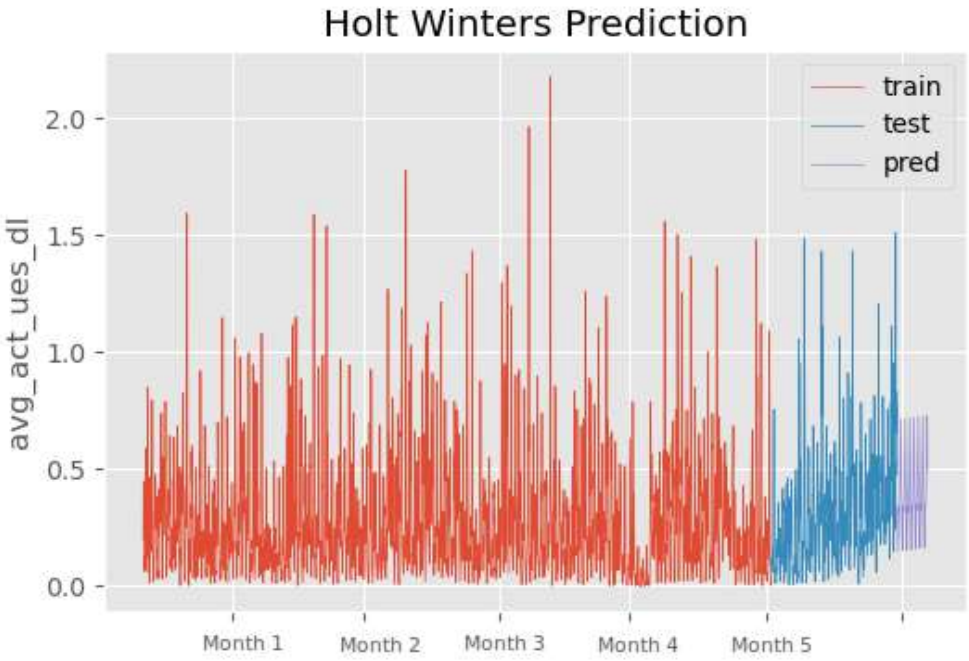


Figure 5.6: Results of Implementing the Holt-Winters Model

generally capture the overall trend, as seen in the training data, well. However, the divergence between the prediction and the actual test data suggests that while the general pattern is matched, the model does not show irregular variations in the cell load – which might be attributed to the smoothing parameters applied. Overall, the Holt-Winters model demonstrated reasonable accuracy, as implied by its RMSE and MAE scores of 0.36 and 0.29, respectively. While this may provide a solid baseline, most of the other models applied showed enhanced performance, as detailed in other sections of this thesis.

5.6.3 Prophet Univariate Results

Figure 5.7 displays the univariate Prophet model's forecasting performance. The black dots represent the individual data points, with the red vertical lines showing distinct trend change points. The plot illustrates the entire training process, as per the cross validation every data point was in the test set at one point, which is illustrated by all actual data being predicted at some point.

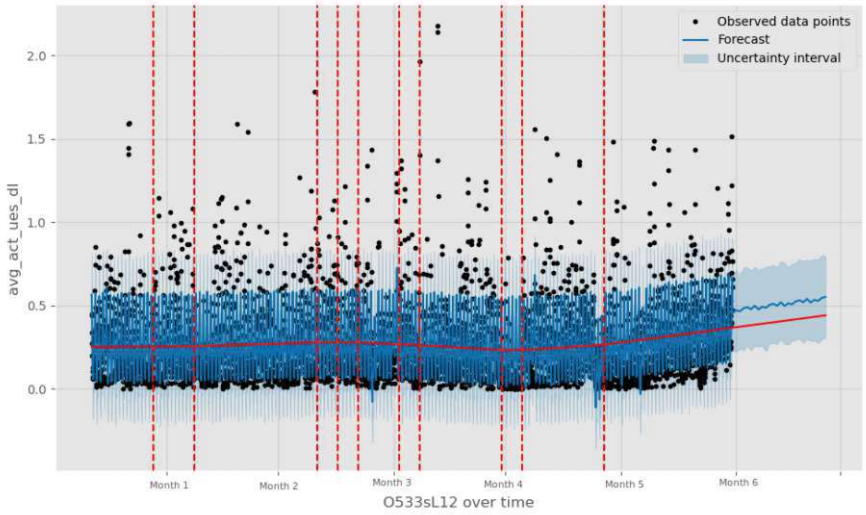


Figure 5.7: Results of Implementing the Prophet Univariate Model

By default, Prophet specifies numerous potential changepoints and applies as few of them as possible, similar to L1 regularization, to limit the actual changes used. The number of potential changepoints can be modified by the user.

The blue line with the shaded area represents the forecast and it's uncertainties, matching the overall trend but with less accuracy than the multivariate correlation model. However, the univariate model has an advantage in that is is less susceptible to overfitting, as can be seen in the application of the multivariate feature model.

The Plot 5.8 presents the RMSE of the univariate Propohet model across a forecasting horizon of one month. As can be seen the RMSE is relatively low and stable, slightly higher but similar to the multivariate correlation model detailed below.

Plot 5.9 displays an array of components and its effect on the cell load captured within the data set. While these particular manifestation of trend effects was observed within a single cell, this cell is well representative of the entire dataset. Some cells also display higher loads during the weekend, however those patterns are rather an exception than the rule within this specific collection of cells. The trend component shows a steady but rather remarkable increase in user activity over time. The next one captures the effect of specific holidays on the oerall trend, indicating a marginal, maybe even negligible influence of events like All Saints Day (November 1st) and Christmas Eve (December 24th). The weekly component shows the influence of recurring patterns across different week days – indicating a small uprise of cell load activity on the weekend. However, what might be more noteworthy is the daily component, indicating user acitvity fluctuations within a single day – with the highest user activity in the evening. This underlines the fact that the area in which the cell is situated comprises of more housing and less industrial areas.

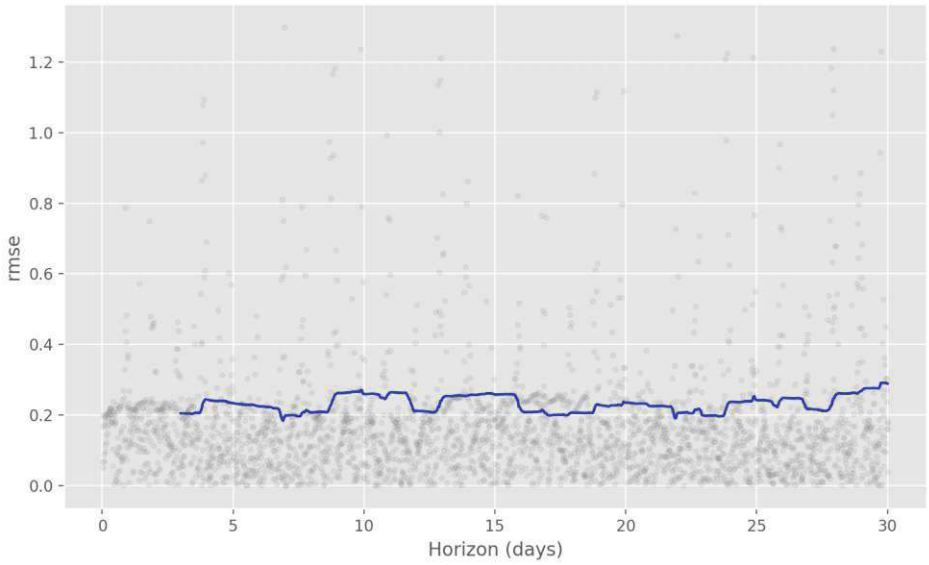


Figure 5.8: RMSE of the Prophet Univariate Model

5.6.4 Prophet Multivariate Correlation Results

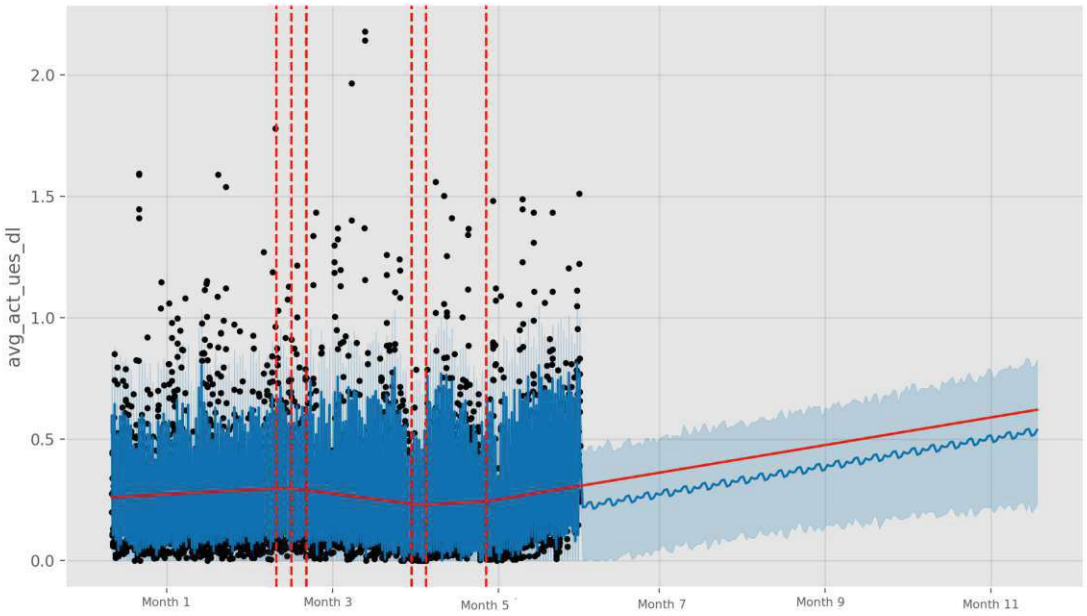


Figure 5.10: Results of Implementing the Prophet Multivariate Correlation Model

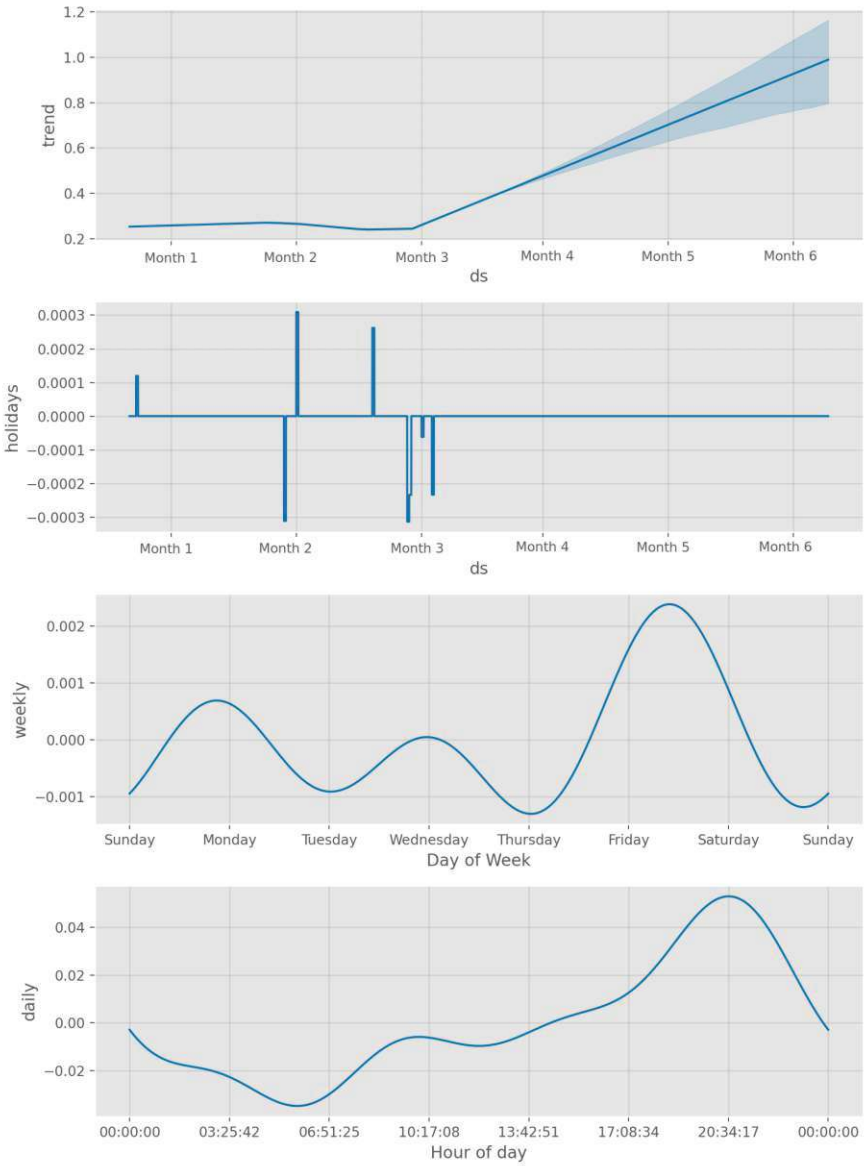


Figure 5.9: Influencing factors in implementing the Prophet Univariate Model

Figure 5.10 shows the forecasting result of the model that proved to have the best overall performance with an RMSE of 0.22 and an MAE of 0.16. The graphic consists of the same representation methods as explained above for the Univariate Prophet model and shows a slightly less steep rise in cell load over time, while indicating slightly more variation in the uncertainty intervals. The model overall captures the trend and variability in the data well.

Similar to the illustation of the RMSE of the unvariate Prophet model, figure 5.11 too

shows a relatively low and stable RMSE over the forecasting horizon of one month, indicating a slightly better forecasting performance for the cell than the aforementioned model. This, again, suggests that the model maintains accuracy over the forecasting period and that the inclusion of features from correlated features enhances the prediction accuracy.

When looking into several components influencing the time series, the first four components given in graph 5.12 present themselves similar to the one from the Univariate Prophet Model above. However, the component on extra regressors show that the influence of additional correlated features from other cells influenced the data most directly, showing more detailed variations in the data which leads to improved forecasting accuracy. This highlights the model's ability to incorporate cell correlations to gain more effective results.

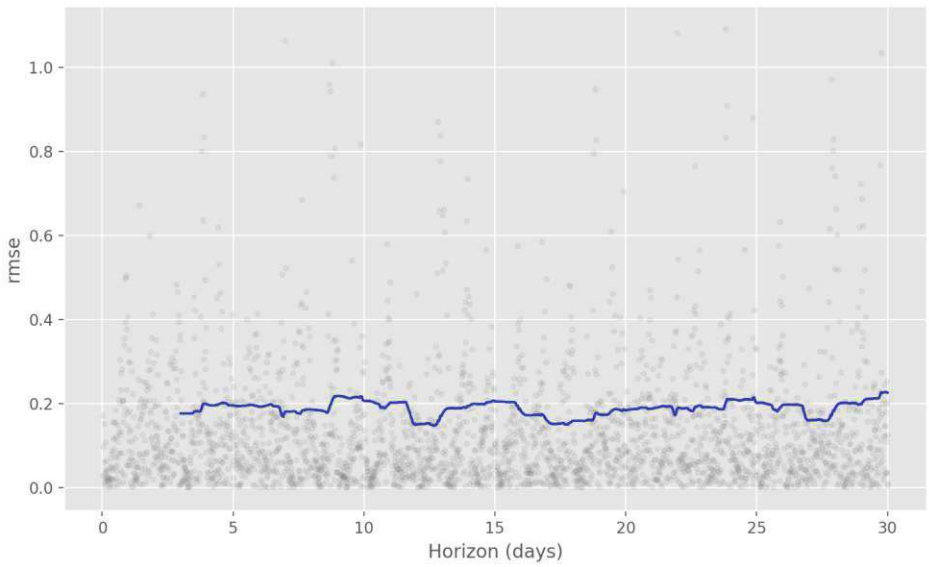


Figure 5.11: RMSE of the Prophet Multivariate Correlation Model

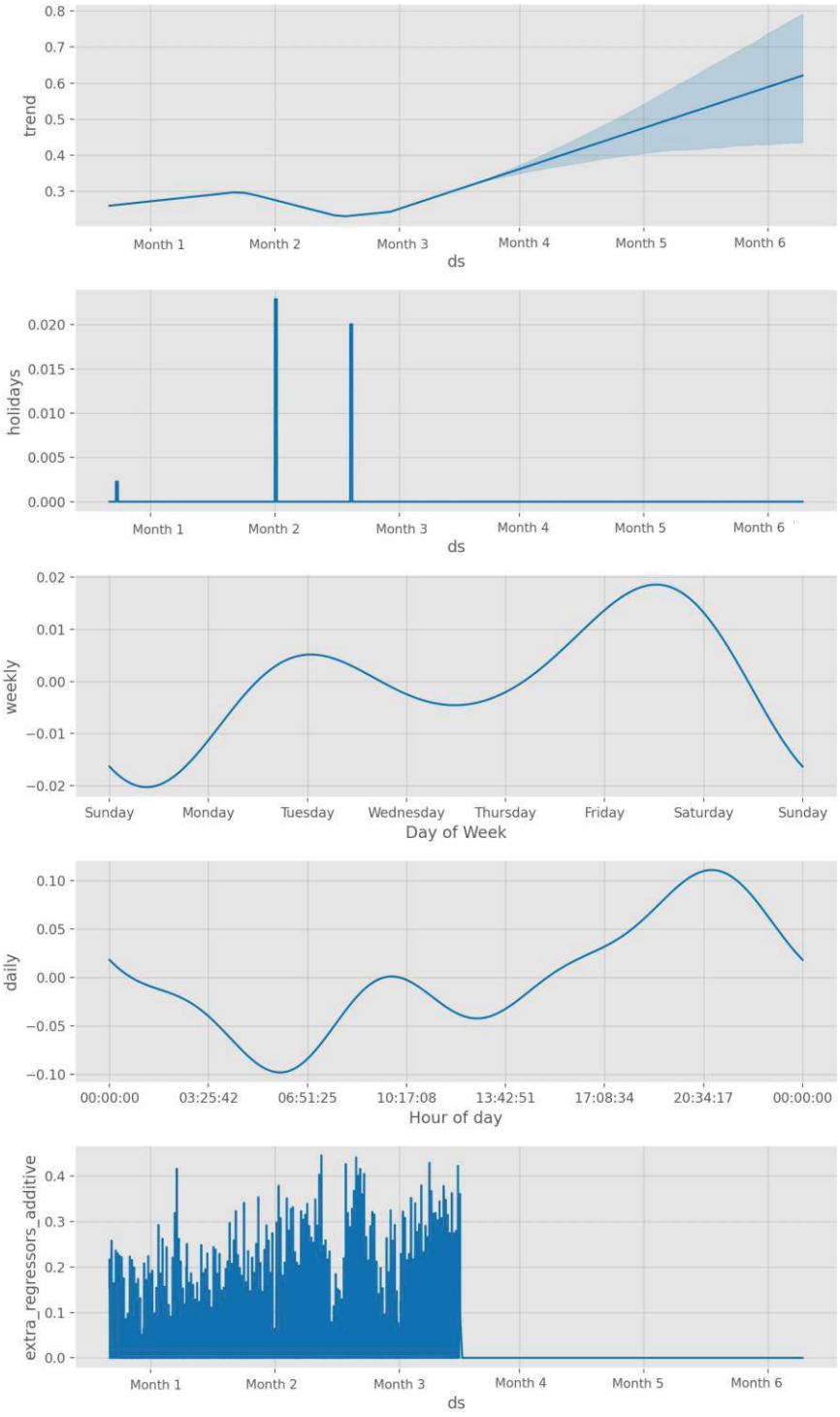


Figure 5.12: Influencing factors in implementing the Prophet Multivariate Correlation Model

5.6.5 Prophet Multivariate Feature Results

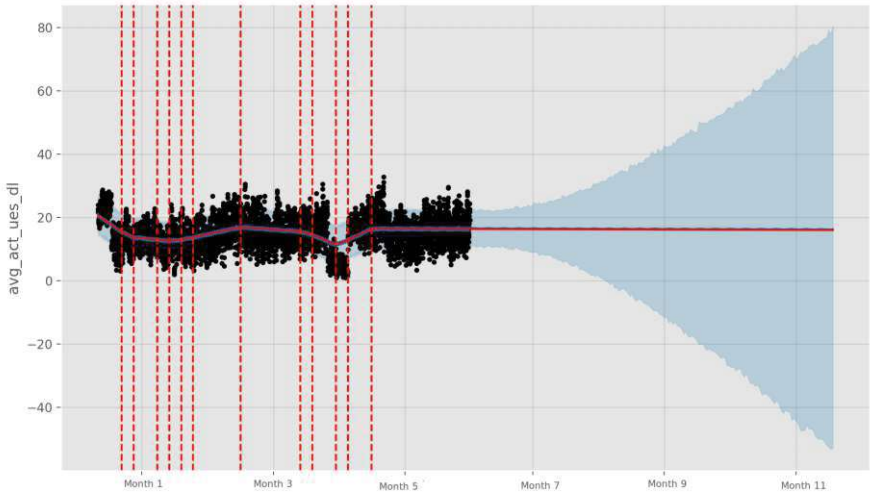


Figure 5.13: Results of Implementing the Prophet Multivariate Feature Model

Figure 5.13 shows the output of the Propeht model with additional multivariate features from the same cell. This is by far the least successful model, with an RMSE of 12.38 and an MAE of 7.49. The format is the same as with the two previous Prophet models, however the forecast shows a significant diversion in it's predictions, with a high uncertainty interval. This might indicate overfitting or noise from the additional features and shows that the predictions can not confidently be used. This suggests that, for further use of this model, a different appproach in the feature selection is needed.

The RMSE shown in figure 5.14 is comparatively much higher and fluctuates significantly more compared to the correlation model, indicating that it is also less stable.

To summarize the effects of different components noted with the Prophet Models it can be stated, that the effects of the components shown by Prophet do, for the most part, not significantly influence the cell load, however some indications can be made.

The overall trend component is essential for understanding the general direction in which the cell load is moving, which shows to be a steady increase in user activity. This could be attributed to a wide array of factors, including network expansion, increased need for communication and entertainment services from users and a general growth of market with increasing use of smartphones. In the multivariate correlation model, the trend also captures the influence of additional regressors, indicating that external factors, such as the activity in correlated cells, contribute to the overall increase.

The holiday effects component captures the changes in user activity due to holidays. The fluctuations that can be seen on those days might be attributed to increased activity due to using mobile devices more for messaging and other communication services, while the decreases might indicate that users might use their mobile devices less due to social gatherings or travel on certain holidays. The weekly seasonality also captures recurring

patterns, showing higher activity on weekends than on weekdays. This could show users having more free time to use their mobile devices.

In turn, daily seasonality captures the fluctuations in user activity within a single day, showing that the peak is typically in the evening, with most users probably being off work, and reduced activity in the late night and very early morning.

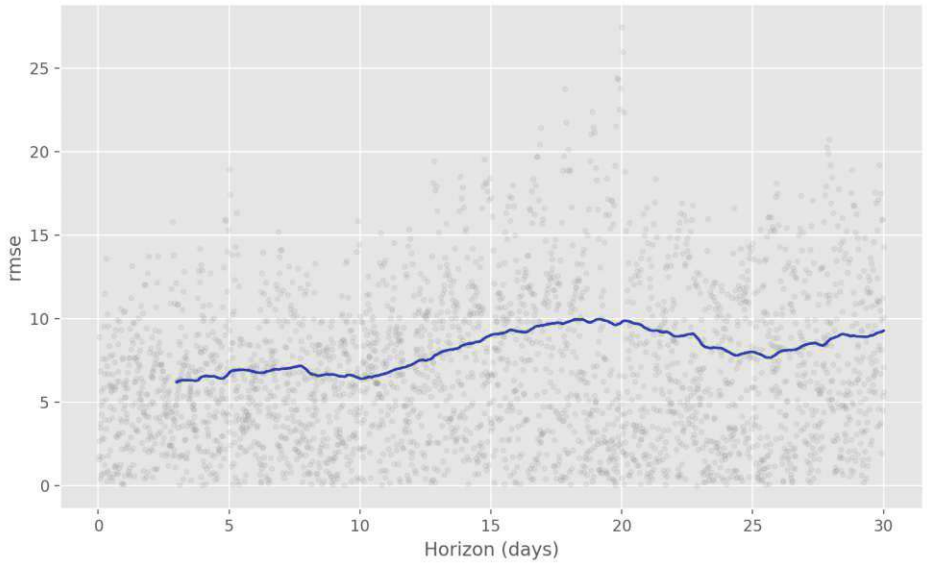


Figure 5.14: RMSE of the Prophet Multivariate Feature Model

CHAPTER 6

Conclusion

6.1 Summary and discussion

The main motivation when approaching this thesis was to assist in addressing the rising demands on cellular networks for higher data rates and therefore an increasing need for energy efficiency, driven by accurate cell load predictions. To adress this need, an approach combining different clustering techniques with traditional and new forecasting models has been chosen.

The different clustering techniques, specifically biclustering and spectral biclustering, were employed to assist in understanding the effectiveness of time-series forecasting models for cell load prediction in cellular networks. Traditionally, K-Means clustering has been the go-to method, often followed by the use of classic forecasting models like ARIMA and Holt-Winters. However, this research aimed to also introduce another approach by integrating the Facebook Prophet model too. This approach also gave the best overall results.

The use of biclustering and spectral biclustering allowed for a more nuanced segmentation of the data. As described in Chapter 4, compared to K-Means clustering, which partitions the data into distinct groups based on overall similarity, biclustering in turn clusters all cells by their overall correlation to identify patterns. Spectral biclustering was then used to further leverage the properties of the data matrix, which assisted in identifying complex patterns that might be missed by conventional clustering techniques thus recovering patterns and not scaling differences in the data set.

The integration of these clustering techniques proved to be successful with the forecasting models, especially the Prophet model, in achieving a suitable forecasting accuracy.

It might be argued that this combination was effective for a variety of possible reasons. First of all, the usage of biclustering and spectral biclustering assisted in identifying

homogenous subsets of cells that exhibited similar load patterns. This has allowed the Prophet models to be trained on more relevant data clusters, assisting in capturing trends that are closer to actual user activity. By focusing on these subsets within the data set, the model could learn and predict user activites within the whole cellular network more effectively.

Furthermore, the identified clusters could be incorporated into the multivariate correlations Prophet model effectively, resulting in the best forecasting results. The inclusion of the five strongest correlation partners, derived from other cells within the same cluster, significantly contributed forecasting accuracy by providing additional context. This highlights the effective use of correalted features from biclustering results which helped in reducing the average prediction errors.

Overall, when comparing the models to each other, the results indicate better results with the proposed approach over the more traditional methods. While ARIMA and Holt-Winters are reliable methods and have therefore been the main application methods in the field, the Prophet models showed similar or better performances. While the difference may not be a stark one, it may serve as a good indication that Prophet, especially the Prophet Multivariate Correlations Model, can be a promising application in research for the forecasting model paired with the clustering techniques applied.

In section 1.4 of this thesis, several research questions were proposed. These shall be answered in the next few sections.

How can forecasting models be improved by considering similarities and differences in the usage patterns of cellular networks within the same cluster?

In this thesis, biclustering and spectral biclustering were employed to cluster cells based on their usage patterns, which allowed for a more nuanced understanding of the data. This clustering revealed structures within the dataset, such as temporal patterns and correlations between different cells, that are crucial for accurate forecasting.

One of the main advantages of using clustered data is the reduction of noise that distracts from the underlying pattern in the forecasting models. When similar usage patterns are grouped together, the forecasting models can better identify and learn from the relevant trends without being distracted by outliers or irregular patterns from dissimilar data points. This is particularly beneficial for models like the Facebook Prophet, which relies on seasonal and trend components. By training the Prophet model on clustered data, the model showed to be more effective in capturing trends specific to each cluster, leading to more precise and reliable forecasts.

Moreover, considering differences within clusters can also refine the forecasting process. Understanding the variability within a cluster can provide insights into the factors that drive these differences. For instance, some cells might experience peak usage during specific times due to local events or demographic factors.

However, it is important to note that the prediction results mainly focus on the daily trend and are currently not well suited for long-term predictions. As can be seen in the previous section, the plots become very noisy and exhibit high uncertainty for future predictions. This limitation suggests that while the method is effective for short-term forecasting and capturing daily patterns, additional strategies or adapted data sets may be needed to enhance its performance for long-term predictions.

In summary, integrating similarities and differences in usage patterns within clusters improves forecasting models by enabling them to focus on relevant trends, account for unique local factors, and ultimately produce more accurate and reliable predictions. This method not only advances the state of predictive modeling but also contributes to more efficient and effective network management strategies.

How can clustering techniques effectively group usage patterns of individual cell towers based on their usage characteristics?

While techniques like K-Means clustering are generally useful in this context, this thesis explores new approaches that have proven to be more viable. Methods such as biclustering and spectral biclustering were methods that were explored within this thesis output clusters where each group contains cell towers with similar usage patterns, thus making it easier to analyze and predict future behaviors.

Biclustering, for instance, allows for the identification of subgroups within the data that exhibit similar behavior across specific time intervals. This is especially useful for cellular network data, where user activity can vary significantly across different times of the day, days of the week, or even seasons. By applying biclustering, it is possible to detect these subgroups and analyze their specific characteristics. Spectral biclustering goes a step further by leveraging the full data spectrum to show patterns that may not be apparent otherwise. This helps in identifying relationships between different cell towers and time points.

By grouping cell towers with similar usage characteristics, the applied clustering techniques created a more manageable and interpretable dataset. Starting from this, it may enable MNOs to implement more efficient resource management strategies. For example, they can allocate bandwidth and infrastructure support based on the specific needs of each cluster, thereby optimizing network performance and improving user satisfaction.

How can multivariate methods improve the forecast of usage patterns by capturing the trends of multiple endpoints compared to traditional univariate time series models?

Multivariate methods can significantly enhance the forecasting of usage patterns by capturing the interactions and trends of multiple endpoints simultaneously.

In this thesis, the multivariate approach was exemplified by the use of the Facebook Prophet model with additional regressors. By incorporating correlated features from other cell towers within the same cluster, the multivariate Prophet model was able to provide a more comprehensive and accurate forecast when compared to traditional methods like ARIMA. These additional regressors provided valuable contextual information that the univariate models could not capture. For instance, the inclusion of data from neighboring cells or other relevant metrics helped in understanding how local events or network-wide changes impact individual cell load. This led to improvements in forecasting accuracy, as evidenced by the lower RMSE and MAE values compared to univariate models.

One advantage that should be noted is the ability to integrate different sources into multivariate models which assists in capturing underlying patterns. For example, in cellular networks, user activity at one cell tower might be influenced by activity at nearby towers or irregular external factors such as weather conditions or special events. By incorporating these multiple dimensions into the forecasting model, multivariate methods can better reflect the real-world complexity of the network. Moreover, multivariate models are better equipped to handle seasonal and trend components that vary across different variables.

In summary, multivariate methods enhance the forecasting of usage patterns by leveraging the interactions and trends across multiple endpoints. This approach results in more accurate and reliable forecasts, better reflecting the complexity of real-world networks and enabling more effective resource management and strategic planning.

6.2 Limitations of applicability

Despite the promising results, several limitations need to be acknowledged. Arrays lend themselves to being used in loops very well.

First of all, the study was conducted on a dataset that may not have been comprehensive enough for this particular use case, given the time frame in which the data points were collected as well as the geographical distribution of the eNodeBs. Further studies might aim on including a wider variety of data to, for instance, ensure to be able to analyze annual patterns that might contribute to individual cellular network usage, such as summer holidays. Additionally, expanding the dataset to cover different geographical regions could provide further insight into the use of the methods used, for instance by focusing on another similarly structured region. This could also mean the use of synthetic data generation techniques to simlulate various scenarios.

Furthermore, as explained before, the results of the approach to incorporate additional features in the Prophet Forecasting suggests that further research could also focus on feature analysis and selection processes. This includes identifying and evaluating the most relevant features (via PCA for example) that contribute to forecasting accuracy and understanding how these features interact with each other, for example via Automated Feature Selection or by choosing a model inspection technique such as a permutation feature importance analysis.

Building up on that, another thing to consider is that the parameter choices, as explained above in Chapter 3.1, influence the clustering and forecasting models. Future research could also investigate the use of different parameters and it's influence on model performance. This could involve a systematic approach trying out and comparing various parameter settings or employing automated methods to identify optimal configurations. Additionally, understanding the impact of these parameters in different contexts and datasets could help in developing more robust and adaptable models, for example by employing various Automated Machine Learning Techniques as reference.

Another important point in this thesis was that the forecasting models required significant computational resources that required a high degree of attentiveness to and a high flexibility in the evaluation steps. This could result in an issue in near real time application scenarios and asks for optimized algorithms to improve efficiency. Focusing on Model optimization and implementing parallel processing techniques could significantly improve the processing abilities.

6.3 Outlook and future work

The integration of biclustering and spectral biclustering with the three forecasting models described in the course of this thesis showed new implications for further research.

First, the scalability of this dataset might be explored further – as described above, the empirical work conducted in this thesis was focused on a specific dataset as presented in Section 3.1. The techniques applied could also be tested on different geographical regions and over longer time spans, to assess the effectiveness of the approach in varying scenarios. In addition to this, datasets such as the CORINE land cover could be used to show the most common classes within each identified cluster. [23] As discussed, preliminary data used for this thesis suggests that certain clusters, such as industrial areas or living areas, might be more prevalent. This might again offer additional insights ito land use patterns and their impact on cell load.

A limitation as explained above was the limited focus on selecting features for the clustering and forecasting steps. An approach to include correlating features from other cells as well as additional features from the same cell was tested, but additional research could build up on that by using more advances feature selection models to optimize inputs for the models.

Furthermore, the combination of the used clustering techniques with the Prophet model showed the most promising results and might be subject of further scientific exploration. Integrating other machine learning or deep learning models with clustering methods could provide a more comprehensive approach to forecasting for MNOs. For instance, implementing hybrid models, such as combining LSTM networks with biclustering, could capture additional dependencies and patterns within the data.

6.4 Conclusion

This thesis explored the integration of advanced clustering techniques, specifically biclustering and spectral biclustering, with various time series forecasting models to improve cell load prediction in cellular networks.

The primary goal was to contribute to the understanding of cell utilization patterns for network optimization. Given the increasing demands on cellular networks for higher data rates, energy efficiency, accurate prediction of user behaviour is critical for the flexible orchestration of resources. Traditionally, models like ARIMA and Holt-Winters, often paired with K-Means clustering, have been in use for this purpose.

This thesis aimed to introduce a novel approach by incorporating the Facebook Prophet model, both in univariate and multivariate forms, alongside these clustering techniques. The results demonstrated that the combination of biclustering with the Prophet model, particularly the multivariate correlations variant, enhanced forecasting accuracy compared to the traditional methods.

As mentioned above, several limitations were noted, including the need for further exploration of feature selection and parameter optimization. Another outcome of this thesis was the underlined importance of selecting relevant features and adjusting model parameters to improve forecasting performance.

Overall, the findings suggest that the Prophet model, especially when combined with biclustering techniques, offers a potentially superior alternative for forecasting in cellular networks. This approach might assist in reducing costs and energy consumption and improving network quality for consumers by enabling more efficient resource allocation. The insights gained from this research contribute to the broader field of predictive modeling using cellular network data, offering potential applications in areas such as cell capacity planning and location-based network expansion, thereby enhancing the overall quality of service for end-users.

Overview of Generative AI Tools Used

The author used ChatGPT as an assistance for code debugging, translation of phrases as well as syntactic uncertainties regarding LaTeX. No parts of the thesis or the assistive code have been created by ChatGPT or any other generative AI tool.

List of Figures

List of Tables

List of Algorithms

Glossary

AR Augmented Reality.

ARIMA AutoRegressive Integrated Moving Average.

CDMA Code Division Multiple Access.

E-UTRAN Evolved UMTS Terrestrial Radio Access Network.

EDGE Enhanced Data rates for GSM Evolution.

eNodeB E-UTRAN Node B.

EPSG European Petroleum Survey Group.

ETSI European Telecommunications Standards Institute.

GSM Global System for Mobile Communications.

IoT Internet of Things.

KPI Key Performance Indicator.

LSTM Long Short-Term Memory.

LTE Long-Term Evolution.

MAE Mean Absolute Error.

MIMO Multiple Input Multiple Output.

MLP Multilayer Perceptron.

MME Mobility Management Entity.

MNO Mobile Network Operator.

OFDMA Orthogonal Frequency Division Multiple Access.

PCI Physical Cell Identity.

PRB Physical Resource Block.

RAN Radio Access Network.

RMSE Root Mean Square Error.

RRC Radio Resource Control.

RSRP Reference Signal Received Power.

SGW Serving Gateway.

SMS Short Message Service.

TDMA Time Division Multiple Access.

TP Transmission Point.

TTI Transmission Time Interval.

UARFCN UTRA Absolute Radio Frequency Channel Number.

UE User Equipment.

UMTS Universal Mobile Telecommunications System.

UTRA UMTS Terrestrial Radio Access.

W-CDMA Wideband Code Division Multiple Access.

WGS World Geodetic System.

Bibliography

[1] European Petroleum Survey Group (EPSG). *EPSG:31287 – MGI / Austria Lambert*. https://epsg.io/31287. Accessed: 2024-05-01. 2007.

[2] European Telecommunications Standards Institute (ETSI). *Research and Standards*. `https://www.etsi.org/research`. Last accessed July 2024.

[3] Abdulaziz Almalaq and George Edwards. "A review of deep learning methods applied on load forecasting". In: *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2017, pp. 511–516.

[4] Geoffrey H Ball, David J Hall, et al. *ISODATA, a novel method of data analysis and pattern classification*. Vol. 699616. Stanford research institute Menlo Park, CA, 1965.

[5] Ulrich Barth. "3gpp long-term evolution/system architecture evolution overview". In: *Alcatel White Paper* (2006), pp. 18–25.

[6] João A Bastos. "Forecasting the capacity of mobile networks". In: *Telecommunication Systems* 72.2 (2019), pp. 231–242.

[7] Anders Bohlin, Harald Gruber, and Pantelis Koutroumpis. "Diffusion of new technology generations in mobile communications". In: *Information economics and policy* 22.1 (2010), pp. 51–60.

[8] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[9] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[10] Rafael Saraiva Campos. "Evolution of positioning techniques in cellular networks, from 2G to 4G". In: *Wireless Communications and Mobile Computing* 2017.1 (2017), p. 2315036.

[11] Tianfeng Chai, Roland R Draxler, et al. "Root mean square error (RMSE) or mean absolute error (MAE)". In: *Geoscientific model development discussions* 7.1 (2014), pp. 1525–1534.

[12] C Chandra, T Jeanes, and WH Leung. "Determination of optimal handover boundaries in a cellular network based on traffic distribution analysis of mobile measurement reports". In: *1997 IEEE 47th Vehicular Technology Conference. Technology in Motion*. Vol. 1. IEEE. 1997, pp. 305–309.

[13] Chris Chatfield. "The Holt-winters forecasting procedure". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 27.3 (1978), pp. 264–279.

[14] Pablo Cortés and Jesús Muñuzuri. "A brief review of the state of the art in Operational Research in Telecommunications". In: *OR Insight* 21 (2008), pp. 25–34.

[15] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[16] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. "Information-theoretic co-clustering". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 89–98.

[17] Konstantinos Dimou et al. "Handover within 3GPP LTE: Design principles and performance". In: *2009 IEEE 70th Vehicular Technology Conference Fall*. IEEE. 2009, pp. 1–5.

[18] William E Donath and Alan J Hoffman. "Lower bounds for the partitioning of graphs". In: *IBM Journal of Research and Development* 17.5 (1973), pp. 420–425.

[19] Salijona Dyrmishi and Amnir Hadachi. "Mobile positioning and trajectory reconstruction based on mobile phone network data: A tentative using particle filter". In: *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE. 2021, pp. 1–7.

[20] Ayman Elnashar. "Coverage and capacity planning of 4G networks". In: *Design, Deployment and Performance of 4G-LTE Networks: Practical Approach, A* (2014), pp. 349–444.

[21] Ericsson. *Ericsson Mobility Report November 2021*. 2021. URL: www.ericsson.com/en/press-releases/2021/11/ericsson-mobility-report-mobile-data-traffic-increased-almost-300-fold-over-10-years.

[22] Ericsson. *Ericsson Mobility Report November 2023*. 2023. URL: https://www.ericsson.com/4ae12c/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-november-2023.pdf.

[23] European Environment Agency (EEA). *CORINE Land Cover*. https://land.copernicus.eu/en/products/corine-land-cover. Accessed: 2024-07-31.

[24] Muhammad Faizan et al. "Applications of clustering techniques in data mining: a comparative study". In: *International Journal of Advanced Computer Science and Applications* 11.12 (2020).

[25] Jamal Fattah et al. "Forecasting of demand using ARIMA model". In: *International Journal of Engineering Business Management* 10 (2018), p. 1847979018808673.

[26] Miroslav Fiedler. “Algebraic connectivity of graphs”. In: *Czechoslovak mathematical journal* 23.2 (1973), pp. 298–305.

[27] Chris Fraley and Adrian E Raftery. “How many clusters? Which clustering method? Answers via model-based cluster analysis”. In: *The computer journal* 41.8 (1998), pp. 578–588.

[28] Xiaohu Ge et al. “User mobility evaluation for 5G small cell networks based on individual mobility model”. In: *IEEE Journal on Selected Areas in Communications* 34.3 (2016), pp. 528–541.

[29] Hana Gebrie, Hasan Farooq, and Ali Imran. “What machine learning predictor performs best for mobility prediction in cellular networks?” In: *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE. 2019, pp. 1–6.

[30] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. “On clustering validation techniques”. In: *Journal of intelligent information systems* 17 (2001), pp. 107–145.

[31] Andrea Hess, Ian Marsh, and Daniel Gillblad. “Exploring communication and mobility behavior of 3G network users and its temporal consistency”. In: *2015 IEEE international conference on communications (ICC)*. IEEE. 2015, pp. 5916–5921.

[32] Timothy O Hodson. “Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not”. In: *Geoscientific Model Development Discussions* 2022 (2022), pp. 1–10.

[33] Tongyi Huang et al. “A survey on green 6G network: Architecture and technologies”. In: *IEEE access* 7 (2019), pp. 175758–175768.

[34] European Telecommunications Standards Institute. *2nd Generation (GERAN)*. 2024. URL: www.etsi.org/technologies/mobile/2g.

[35] European Telecommunications Standards Institute. *3rd Generation (UMTS)*. 2024. URL: www.etsi.org/technologies/mobile/3g.

[36] European Telecommunications Standards Institute. *4th Generation (LTE)*. 2024. URL: www.etsi.org/technologies/mobile/4g.

[37] European Telecommunications Standards Institute. *5G*. 2024. URL: www.etsi.org/technologies/mobile/5g.

[38] European Telecommunications Standards Institute. *5G; NR; Layer 2 measurements (3GPP TS 38.314 version 17.1.0 Release 17)*. Tech. rep. TS 138 314 V17.1.0 (2022-08). 3GPP, 2022.

[39] European Telecommunications Standards Institute. *Digital cellular telecommunications system (Phase 2+) (GSM); Mobile radio interface layer 3 specification; Radio Resource Control (RRC) protocol; Iu mode*. Tech. rep. TS 144 118 V15.0.0 (2018-07). 3GPP, 2018.

[40] European Telecommunications Standards Institute. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRAN); Physical channels and modulation (3GPP TS 36.211 version 16.6.0 Release 16)*. Tech. rep. TS 136 211 V16.6.0 (2021-08). 3GPP, 2006.

[41] European Telecommunications Standards Institute. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRAN); Services provided by the physical layer (3GPP TS 36.302 version 8.0.0 Release 8)*. Tech. rep. TS 136 302 V8.0.0 (2008-11). 3GPP, 2008.

[42] European Telecommunications Standards Institute. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRAN); X2 general aspects and principles (3GPP TS 36.420 version 12.0.0 Release 12)*. Tech. rep. TS 136 420 V12.0.0 (2014-09). 3GPP, 2014.

[43] European Telecommunications Standards Institute. *LTE; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced)*. Tech. rep. TS 136 913 V17.0.0 (2022-05). 3GPP, 2022.

[44] Anil K Jain. "Data clustering: 50 years beyond K-means". In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.

[45] Garima Jain and Rajeev Ranjan Prasad. "Machine learning, Prophet and XGBoost algorithm: analysis of traffic forecasting in telecom networks with time series data". In: *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE. 2020, pp. 893–897.

[46] Andreas Janecek et al. "Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, pp. 361–370.

[47] K Kameshwaran and K Malarvizhi. "Survey on clustering techniques in data mining". In: *International Journal of Computer Science and Information Technologies* 5.2 (2014), pp. 2272–2276.

[48] Georgios Karagiannis et al. "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions". In: *IEEE communications surveys & tutorials* 13.4 (2011), pp. 584–616.

[49] Yuval Kluger et al. "Spectral biclustering of microarray data: coclustering genes and conditions". In: *Genome research* 13.4 (2003), pp. 703–716.

[50] Reginald Tomas Yu-Lee. *Essentials of capacity management*. John Wiley & Sons, 2002.

[51] William Lehr, Fabian Queder, and Justus Haucap. "5G: A new future for Mobile Network Operators, or not?" In: *Telecommunications Policy* 45.3 (2021), p. 102086.

[52] Miriam Leopoldseder et al. "Benchmarking lightweight user mobility predictors on operational wlan data". In: *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE. 2019, pp. 1–5.

[53] Susana Lima, A Manuela Gonçalves, and Marco Costa. “Time series forecasting using Holt-Winters exponential smoothing: An application to economic data”. In: *AIP conference proceedings*. Vol. 2186. 1. AIP Publishing. 2019.

[54] Sara C Madeira and Arlindo L Oliveira. “A linear time biclustering algorithm for time series gene expression data”. In: *International workshop on algorithms in bioinformatics*. Springer. 2005, pp. 39–52.

[55] Atif Mahmood et al. “Capacity and frequency optimization of wireless backhaul network using traffic forecasting”. In: *IEEE Access* 8 (2020), pp. 23264–23276.

[56] Guowang Miao et al. *Fundamentals of mobile data networks*. Cambridge University Press, 2016.

[57] Ajay R Mishra. *Fundamentals of network planning and optimisation 2G/3G/4G: evolution to 5G*. John Wiley & Sons, 2018.

[58] Nikhil Nayak et al. “5G Traffic Prediction with Time Series Analysis”. In: *arXiv preprint arXiv:2110.03781* (2021).

[59] Kaveh Pahlavan and Allen H Levesque. *Wireless information networks*. John Wiley & Sons, 2005.

[60] A Parizad and CJ Hatziadoniu. “Using prophet algorithm for pattern recognition and short term forecasting of load demand based on seasonality and exogenous features”. In: *2020 52nd North American Power Symposium (NAPS)*. IEEE. 2021, pp. 1–6.

[61] Utpal Paul et al. “Understanding traffic dynamics in cellular data networks”. In: *2011 Proceedings IEEE INFOCOM*. IEEE. 2011, pp. 882–890.

[62] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[63] Jordi Pérez-Romero et al. “Artificial intelligence-based 5G network capacity planning and operation”. In: *2015 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE. 2015, pp. 246–250.

[64] Andras Racz, Andras Temesvary, and Norbert Reider. “Handover performance in 3GPP long term evolution (LTE) systems”. In: *2007 16th IST Mobile and Wireless Communications Summit*. IEEE. 2007, pp. 1–5.

[65] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[66] Sina Sabzekar, Mohammad Reza Valipour Malakshah, and Zahra Amini. “Unsupervised Learning for Topological Classification of Transportation Networks”. In: *arXiv preprint arXiv:2311.13887* (2023).

[67] Guto Leoni Santos et al. “When 5G meets deep learning: a systematic review”. In: *Algorithms* 13.9 (2020), p. 208.

[68] Ketan Rajshekhar Shahapure and Charles Nicholas. “Cluster quality analysis using silhouette score”. In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748.

[69] Hanhuai Shan and Arindam Banerjee. “Bayesian co-clustering”. In: *2008 Eighth IEEE international conference on data mining*. IEEE. 2008, pp. 530–539.

[70] Chaoming Song et al. “Limits of predictability in human mobility”. In: *Science* 327.5968 (2010), pp. 1018–1021.

[71] Gordon L Stüber and Gordon L Steuber. *Principles of mobile communication*. Vol. 2. Springer, 2001.

[72] Diane Tang and Mary Baker. “Analysis of a metropolitan-area wireless network”. In: *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*. 1999, pp. 13–23.

[73] Azar Taufique et al. “Planning wireless cellular networks of future: Outlook, challenges and opportunities”. In: *IEEE Access* 5 (2017), pp. 4821–4845.

[74] Romain Tavenard. *An introduction to Dynamic Time Warping*. `https://rtavenar.github.io/blog/dtw.html`. 2021.

[75] Sean J Taylor and Benjamin Letham. “Forecasting at scale”. In: *The American Statistician* 72.1 (2018), pp. 37–45.

[76] Igor Tomic, Eoin Bleakley, and Predrag Ivanis. “Predictive capacity planning for mobile networks—ML supported prediction of network performance and user experience evolution”. In: *Electronics* 11.4 (2022), p. 626.

[77] Ulrike Von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and computing* 17 (2007), pp. 395–416.

[78] Marin Vukovic, Goran Vujnovic, and Darko Grubisic. “Adaptive user movement prediction for advanced location-aware services”. In: *SoftCOM 2009-17th International Conference on Software, Telecommunications & Computer Networks*. IEEE. 2009, pp. 343–347.

[79] Chunpai Wang. *Spectral Clustering*. Chunpai’s Blog. Accessed: January 2024. Mar. 2016. URL: `https://chunpai.github.io`.

[80] Jiang Wang et al. “From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases”. In: *IEEE access* 6 (2017), pp. 1718–1729.

[81] Yiqiao Wei and Seung-Hoon Hwang. “Optimization of Cell Size in Ultra-Dense Networks with Multiattribute User Types and Different Frequency Bands”. In: *Wirel. Commun. Mob. Comput.* 2018 (2018), 8319749:1–8319749:10. URL: `https://api.semanticscholar.org/CorpusID:53291466`.

[82] Junwei Xiao, Jianfeng Lu, and Xiangyu Li. “Davies Bouldin Index based hierarchical initialization K-means”. In: *Intelligent Data Analysis* 21.6 (2017), pp. 1327–1338.

[83] Chaoyun Zhang, Paul Patras, and Hamed Haddadi. “Deep learning in mobile and wireless networking: A survey”. In: *IEEE Communications surveys & tutorials* 21.3 (2019), pp. 2224–2287.

[84] Hongtao Zhang and Lingcheng Dai. “Mobility prediction: A survey on state-of-the-art schemes and future applications”. In: *IEEE access* 7 (2018), pp. 802–822.

[85] Yiqing Zhou et al. “An overview on intercell interference management in mobile cellular networks: From 2G to 5G”. In: *2014 IEEE International Conference on Communication Systems*. IEEE. 2014, pp. 217–221.

[86] Jim Zyren and Wes McCoy. “Overview of the 3GPP long term evolution physical layer”. In: *Freescale Semiconductor, Inc., white paper* 7 (2007), pp. 2–22.