# Censored Regression Models with Autoregressive Errors for Improved Time Series Estimation and Forecasting

Mikko Närhi

**School of Science**

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 23.01.2024

**Supervisor**

Dr. Jukka Kohonen

**Advisor**

Dr. Jukka Kohonen

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

**Author** Mikko Närhi

**Title** Censored Regression Models with Autoregressive Errors for Improved Time Series Estimation and Forecasting

**Degree programme** Mathematics and Operations Research

**Major** Systems and Operations Research **Code of major** SCI3055

**Supervisor** Dr. Jukka Kohonen

**Advisor** Dr. Jukka Kohonen

**Date** 23.01.2024 **Number of pages** 52 **Language** English

**Abstract**

In time series analysis, it is common to encounter censored data, where observations are only partially observed due to a detection limit. This thesis specifically addresses the case of censoring in the context of data traffic demand in radio access networks. In areas of the network with high user density and insufficient infrastructure, a discrepancy between the actual delivered data traffic and the theoretical demand arises, leading to the censoring of the theoretical data traffic demand. The primary objective of this thesis is to generate accurate estimations for the time series observations of censored data traffic demand.

We propose a Censored Linear Regression model with autoregressive errors, a novel approach designed to estimate censored time series observations. A key component of our approach is the Gaussian imputation method which is grounded on the premise that fully observed time series data can be interpreted as a realization of a multivariate normal distribution.

We apply our methodology to historical time series data from a cellular Radio Access Network comprising an entire country. The censored observations were imputed using the model developed for this thesis. We evaluate the performance of our model by forecasting future observations, both with and without the imputation process, and subsequently comparing their forecast accuracy.

The findings indicate that while our model is successful at estimating censored observations, its application in improving forecast accuracy for censored network areas is not uniformly effective. Consequently, it remains challenging to assert with certainty, the advantage of the developed method in estimating censored time series observations.

**Keywords** Censored data, Autoregressive models, Time Series Analysis, Gaussian Imputation Method, Estimation, Time Series Forecasting, Linear Regression

Aalto University
School of Science

**Aalto-yliopisto, PL 11000, 00076 AALTO**
**www.aalto.fi**
**Diplomityön tiivistelmä**

**Tekijä** Mikko Närhi

**Työn nimi** Sensuroidut regressiomallit autoregressiivisillä virheillä parannettuun aikasarjojen estimointiin ja ennustamiseen

**Koulutusohjelma** Matematiikka ja operaatiotutkimus

**Pääaine** Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3055

**Työn valvoja** FT Jukka Kohonen

**Työn ohjaaja** FT Jukka Kohonen

**Päivämäärä** 23.01.2024 **Sivumäärä** 52 **Kieli** Englanti

**Tiivistelmä**

Aikasarja-analyysissa on yleistä kohdata sensuroitua dataa, jossa havainnot ovat vain osittain havaittavissa havaitsemisrajan vuoksi. Tämä diplomityö käsittelee sensurointia radioliityntäverkon dataliikenteen kysynnän yhteydessä. Verkon alueilla, joissa on suuri käyttäjätiheys ja riittämätön infrastruktuuri, syntyy ero todellisen toimitetun dataliikenteen ja teoreettisen kysynnän välille, joka johtaa teoreettisen kysynnän sensurointiin. Tämän diplomityön ensisijainen tavoite on tuottaa tarkkoja estimointeja sensuroidun dataliikenteen kysynnän aikasarjahavainnoista.

Ehdotamme tämän ongelman ratkaisemiseksi sensuroitua lineaarista regressiomallia, jossa on autoregressiivisiä virheitä. Tämä on uusi lähestymistapa sensuroitujen aikasarjahavaintojen estimoimiseen. Keskeinen osa lähestymistapaamme on Gaussin imputointimenetelmä, joka perustuu oletukseen, että täysin havaittu aikasarjadata voidaan tulkita monimuuttujanormaalijakauman toteutumana.

Sovellamme menetelmäämme historialliseen aikasarjadataan, joka koostuu koko maan kattavasta radioliityntäverkosta. Sensuroidut havainnot imputoidaan tässä diplomityössä kehitetyllä menetelmällä. Arvioimme mallimme suorituskykyä ennustamalla aikasarjan tulevia havaintoja sekä imputointiprosessin kanssa että ilman, ja vertaamalla molempien ennustetarkkuutta.

Havainnot osoittavat, että vaikka mallimme onnistuu arvioimaan sensuroituja havaintoja, sen soveltaminen ennustetarkkuuden parantamiseen sensuroiduilla verkon alueilla ei ole aina hyödyllistä. Tämän seurauksena on haastavaa varmuudella määritellä kehitetyn menetelmän etua sensuroitujen aikasarjahavaintojen estimoinnissa.

**Avainsanat** Sensuroitu data, Autoregressiiviset mallit, Aikasarja-analyysi, Gaussin imputointimenetelmä, Estimointi, Aikasarjojen ennustaminen, Lineaarinen regressio

# Preface

I want to thank my supervisor, Jukka Kohonen, for his continued guidance and support throughout this thesis project. I would also like to thank Omnitele for their assistance and contribution of data to this thesis.

Furthermore, I wish to express my gratitude to my friends and family for their unwavering love and support. Their presence has been instrumental in my journey, providing the motivation I needed to see both my studies and this thesis through to completion. Additionally, I extend my heartfelt thanks to my family, whose consistent support has been something I could always rely on throughout my entire educational journey, from the earliest days of my education, to this significant milestone.

Otaniemi, January 31, 2024

Mikko Närhi

# Contents

# List of Abbreviations

AIC Akaike Information Criterion

AR Autoregressive

ARIMAX Autoregressive Integrated Moving Average with Explanatory Variables

BS Base Station

CLR-AR Censored Linear Regression Model with Autoregressive Errors

CQI Channel Quality Indicator

DL Downlink

KPI Key Performance Indicator

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MNO Mobile Network Operator

MSC Mobile Switching Center

MSE Mean Squared Error

NCAR National Center for Atmospheric Research

OPEX Operational Expenditure

PRB Physical Resource Block

QoS Quality of Service

RAN Radio Access Network

RMSE Root Mean Square Error

UE User Equipment

VIF Variance Inflation Factor

# 1 Introduction

With the rapid adoption of electronic devices, notably smartphones, the global demand for mobile data has surged dramatically. In 2014, the average monthly data traffic per smartphone was 1 GB, a number that had swelled to 15 GB by 2022 (Paraschiv et al., 2022; Ericsson, 2023). Such a rapid rise causes challenges that have to be managed continuously.

As the demand for data traffic rises, it becomes increasingly important for the underlying network infrastructure to evolve in tandem. In an era where users expect consistent and high-speed connections, any lapse in the Quality of Service (QoS) can be detrimental. Ericsson's Mobility Report 2022 predicts that by 2028, the monthly average data traffic per smartphone will rise to an astonishing 46 GB, nearly tripling the current rates (Ericsson, 2023). Due to this phenomenon, accurate forecasts are extremely important for Mobile Network Operators (MNOs). With accurate forecasting, MNOs can strategize their resource allocation, ensuring that the rising demand does not impede QoS. On the other hand, neglecting forecasting may lead to service degradation, negatively impacting user experiences.

Figure 1 displays the evolution of data traffic from May 2020 to June 2023. This representation, which aggregates data from a cellular network serving an entire country, provides a glimpse into the challenge at hand.

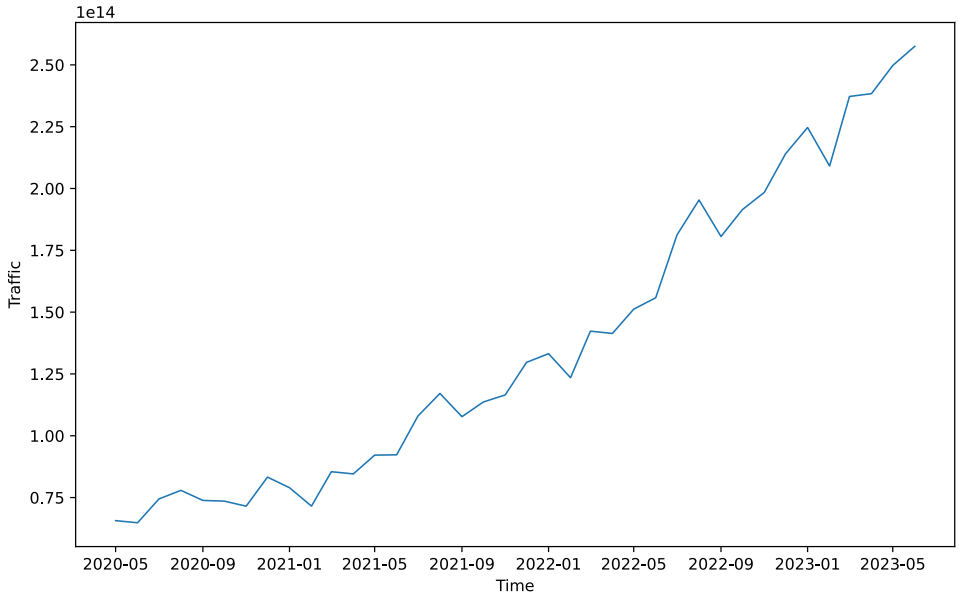


Figure 1: Month by month average Data Traffic (kbit) calculated by taking the average of every sector in the cellular network. The dataset ranges from May 2020 to June 2023, and it was provided by Omnitele, our client organization.

In this thesis, a critical distinction is drawn between two variables: data traffic and the demand for data traffic. It is essential to recognize these as separate entities, particularly when addressing the intricacies of cellular network behavior.

- Non-congested parts of the Network: Here, the data traffic and its demand are largely equal. This alignment is due to the fact that there are sufficient network resources available for users, ensuring optimal Quality of Service (QoS) and minimal buffer times. In these parts, the demand for data traffic is almost identical to what the users receive and utilize.

- Congested Parts of the Network: This is where the divergence between data traffic and its demand becomes evident. The phenomenon can be traced back to QoS degradation. As users fight for the same network resources, there is an inevitable decline in both user throughput and overall QoS. This reduction subsequently leads to diminished actual data usage. Consider, for instance, a user attempting to stream a video on YouTube. If they are consistently interrupted by buffering, it is likely that they will abandon the video before completion. The theoretical demand in this scenario would be the entirety of the video's data, but due to bad network performance, the actual data consumption falls short.

This divergence between actual data usage and its latent demand poses a significant challenge for MNOs. Their task is to pinpoint where the demand for data traffic intensifies and to quantify its increase. However, in congested parts, the demand becomes unobserved, and thus complicating the MNOs' objectives.

Figure 2 offers a visual representation of congestion's impact on network data traffic, specifically highlighting the constraints posed by limited network resources. In July 2022, new infrastructure is added to the sector which can be seen as a drastic increase in data traffic. The addition of new infrastructure causes the traffic to jump from approximately 15 billion kbit to over 30 billion kbit in just three months, nearly doubling it. Such a jump can be interpreted as the network's reversion to equilibrium, where the actual data traffic converges with its demand. With the new infrastructure in place, the network now possesses enough capacity to enable free data consumption for its entire user base.

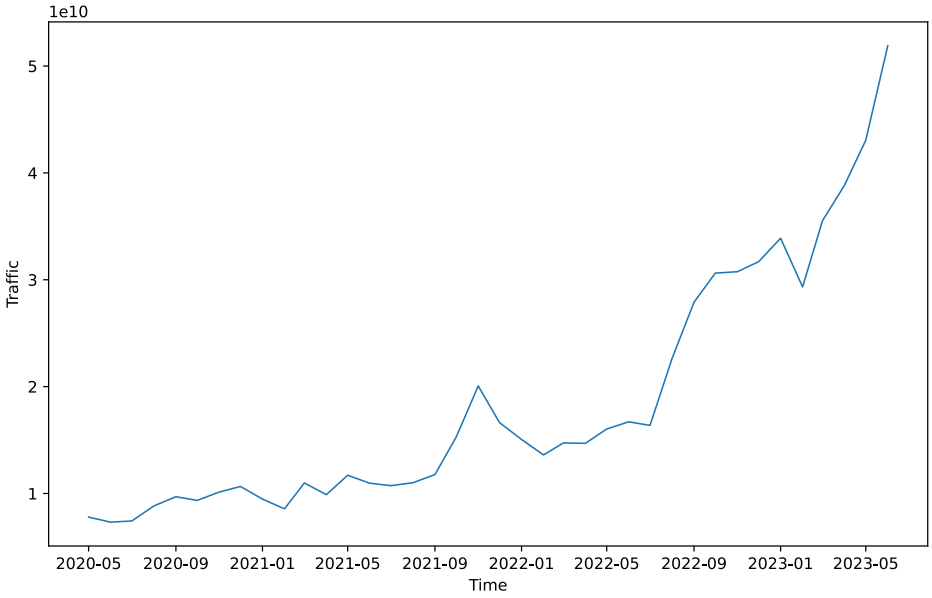


Figure 2: Month by month development of data traffic (kbit) for a sector in the network. The data is from May 2020 to June 2023.

For MNOs, accurate data traffic forecasting is a necessity that helps in making crucial decisions around network expansions. The predictive insights from such forecasts guide where and when to deploy new infrastructure, particularly in regions expected to witness substantial growth in data traffic. In scenarios where network sectors remain non-congested, forecasting future demand for data traffic is straightforward. This is because, in such cases, the recorded data traffic directly mirrors its demand. However, the challenge becomes apparent when forecasting for congested network sectors.

Suppose for the purpose of illustration that in Figure 2, we are restricted to data from May 2020 to July 2022. Traditional forecasting methodologies, like the Autoregressive (AR) models, would fail in this setting. The stagnation in observed data traffic obscures the underlying demand. Yet, the subsequent surge in data traffic after the addition of new infrastructure in July 2022 indicates that the latent demand for data traffic was steadily rising, even when the observed data did not reflect it. Understanding this unobserved demand is crucial for MNOs. If harnessed accurately, it could significantly enhance network planning, ensuring timely capacity expansions precisely where it is imperative.

## 1.1 Motivation

MNOs are always searching for the optimal allocation of resources across different locations. A cellular network business has to strategically determine the optimal site/sector/cell where to place new capacity, while minimizing the Operational

Expenditure (OPEX). Wrong or inaccurate forecasting results can lead to suboptimal decisions, which cause additional expenses and lead to QoS degradation.

This thesis aims to contribute to the network planning by estimating the demand for data traffic in congested parts of the cellular network. By estimating the demand for data traffic, our aim is to help the Mobile Network Operators gain a better understanding of the requirements for their network in the future and allow them to pre-emptively act before QoS degradation and loss in profitability.

## 1.2 Research Objectives

Understanding the future demand for data traffic for each part in the network is crucial for a MNO when they are planning a network expansion. A general trend demonstrates an ascending trajectory in data traffic demand (see Figure 1). However, right now, MNOs do not have a clear method for estimating the actual demand for data in congested parts of the network. This leads to an neglect of sectors where expansions are most needed.

The objective of this thesis centers on bridging this gap. The aim of the thesis is as follows:

- Estimation of congested data: The aim is to develop a method to estimate the latent demand for data traffic in congested sectors using their data traffic time series. The inherent nature of congestion implies that these sectors experience so-called *right censoring*. That is, while it is clear that a data point surpasses a certain threshold, the precise extent remains ambiguous. This thesis tries to answer the following question: Can we, given this censoring, estimate the unobserved data traffic demand from the observed time series?

- Model Validation: Beyond the estimation model itself, an important aspect is to quantify the estimation model's reliability. Thus, a critical objective is to assess the precision with which the model estimates the unobserved demand in congested sectors in the network. The methodology for validation is to develop a forecasting method capable of accurately predicting future data traffic, and starting the forecast from data points where new infrastructure has been added to a sector in the network. At such times, actual data traffic and its demand are expected to converge, giving us a method to gauge our model's accuracy.

By addressing these objectives, this thesis strives to offer MNOs an empirical tool, empowering them with more informed decision-making capabilities for future network expansions.

## 1.3 Structure

The remainder of the thesis is organized as follows: Section 2 goes through background information about the topic of this thesis, radio access networks, censoring and existing literature on censored time series analysis. Section 3 describes the dataset used in the thesis. Additionally, in Section 3, we go over the calculations used for defining

new variables, which are derived from the existing variables of the raw dataset. Section 4 presents the system model used for estimating the data traffic demand for congested sectors and forecasting their future values in the cellular network. The system model describes the entire system that is used to acquire raw data from the client organization, preprocessing the data for modelling to estimating the demand for data traffic and forecasting the future data traffic using a forecasting method and then evaluating the performance of the developed models. The results of the estimation and forecasting are analyzed in Section 5, while discussion about the chosen methods and results are presented in Section 6.

# 2 Background

## 2.1 Radio Access Network

In this section, we will introduce the term Radio Access Network (RAN) and explore its high-level architecture. We will also go over a few key concepts that are relevant to the topic of this thesis. RANs are a critical component of a wireless network, as they facilitate the transmission of data traffic between mobile devices and the core network infrastructure (Alsabaan et al., 2008). A typical RAN is constructed of transmission lines, antennas, Base Stations and communication control units. Our focus will be on Base Stations, more commonly referred to as cell towers, which are crucial for the cellular network's functionality.

In a cellular network, each Base Station has its own coverage area, visualized as part of a cellular grid, to which a user connects based on their current location. These Base Stations, fixed in location, form the physical infrastructure that MNOs continuously strive to expand and optimize to enhance the QoS. The wireless communication within the RAN happens between these Base Stations and User Equipment (UE), typically mobile devices. The mobility of UEs introduces complexities in maintaining stable and reliable connections, as they move within the grid-like structure formed by the Base Stations (Sirotkin, 2020).

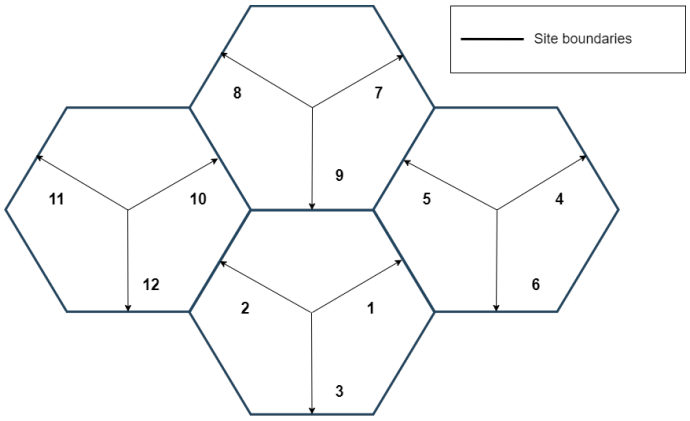


Figure 3: The cellular network layout with four sites and three sectors per site. Illustration inspired by Bin Sediq et al. (2015).

The architecture of a cellular network containing four sites and three sectors per site is shown in Figure 3. Each site contains three sectors, shown as the black arrows, and each sector contains three cells. In this thesis, we focus on estimating and forecasting the demand for data traffic at sector level. The original data used in this thesis was received at cell level which had to be aggregated to sector level. The sectors together cover an entire country as a grid, and they form the Radio Access Network.

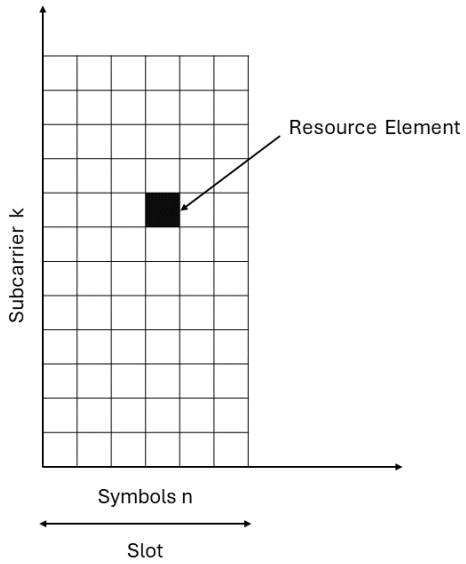


Figure 4: Resource grid of a cell. The entire $12 \times 6$ grid forms a single Physical Resource Block (PRB).

In cellular networks, the UEs are allocated to a Base Station at all times. A Physical Resource Block (PRB) visualized in Figure 4 is assigned to an user device when connected to a Base Station. It is the smallest amount of resources that can be allocated to an user (Liberg et al., 2020). In a single PRB, there are 12 subcarriers each of which defines a specific frequency range and either 6 or 7 symbols, meaning that a single PRB is a $12 \times 7$ ($12 \times 6$) grid of Resource Elements (Zöchmann et al., 2015). A single cell with 20 MHz bandwidth can schedule a maximum of 100 PRBs at once.

Congestion is an important concept in our thesis. PRB utilization, which is the ratio of used PRBs and available PRBs in the cell, tells us how congested a cell is. If the PRB utilization is high, the allocation of resources becomes unreliable and delayed, leading to QoS degradation (Hwang and Park, 2017). This is why PRB utilization is a great indicator of congestion and it is used as a main variable to determine congested periods for sectors in the dataset. High PRB utilization is ultimately solved by increasing the capacity of the cellular network.

## 2.2 Censoring

In statistics, the phenomenon where observations are only partially observable is known as *censoring*. Time series measurements can exhibit data irregularities, including censoring. In this thesis, we consider the time series data to experience right censoring. Right-censored observations are instances where the value is known to exceed a detection limit, but the exact value remains unobserved.

In the specific context of this thesis, censoring occurs in the high congestion sectors of a cellular network. While we fully observe the data traffic, the actual demand for data traffic in these congested sectors is not directly observed. This is a result of high usage combined with insufficient infrastructure, which leads to a scenario where the network's capacity to meet data traffic demand is exceeded, but the extent of this excess demand remains unknown.

An example of a censored time series from a different context is presented in Figure 5. This data has been previously analyzed in Park et al. (2007), Schumacher et al. (2017), Aydin and Yilmaz (2021) and Sousa et al. (2023) which are all relevant research papers for the topic of this thesis. The data was originally collected by the National Center for Atmospheric Research (NCAR), based on hourly observations in San Francisco and recorded during the month of March 1989. The log-transformed data is available in the package *ARCensReg* (Schumacher et al., 2016), which is implemented in R. This data is not analyzed in this thesis but rather given as an example of what censoring in a time series looks like. From the Figure, it is evident that the true ceiling height occasionally exceeds the upper detection limit, which is just shy of 5. The data of Figure 5 is also right censored, similarly to the dataset analyzed in this thesis.

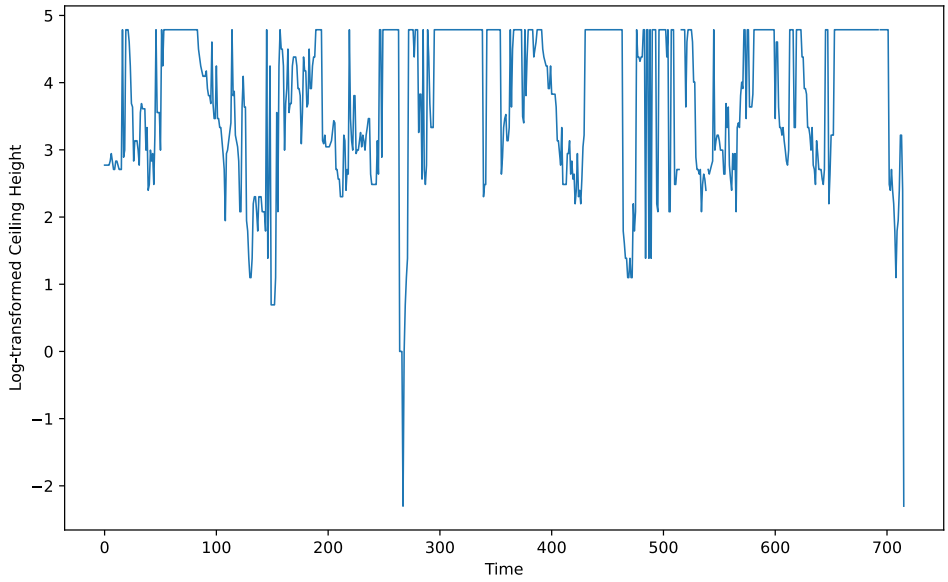


Figure 5: Censored time series data of log-transformed hourly cloud ceiling height in San Francisco during March 1989. The data was sourced from the *ARCensReg* package (Schumacher et al., 2016) in R. Plot created by the thesis author.

The censoring rate for the time series in Figure 5 is 41.62 %. In this thesis, the congested sectors of the cellular network considered to experience censoring, do not have missing observations and the censoring rate can vary significantly from approximately 10 % to 50 % by sector. An example of a censored time series in the context of this thesis is presented in Figure 2, where we consider the time series to be censored before the addition of infrastructure in July 2022.

In the time series measurements of this thesis, there does not exist an exact detection limit as the traffic for a congested sector can grow even after extreme congestion. Most relevant studies on censoring define a clear detection limit but in our thesis we define the detection limit separately for each censored observation. This process is further explained in Section 4.7. Due to lack of clear detection limit, the starting point of congestion is not trivial to define. In our thesis, we leverage expert knowledge to determine the definition for a censored observation in a congested sector.

## 2.3 Censored Time Series Analysis

There are two general approaches to overcome censoring in existing literature. The first is to discard the censored observations and continue the analysis with only the uncensored data, and the second is to treat the censored values as observed. A simple method to handle censored data is to substitute the censored values with a constant exceeding/preceding the detection limit. According to Helsel (1990), this method has been used under circumstances where the censoring rate is not large (below 20

%). However, in this thesis, this assumption is not valid as the censoring rate can be very high, even close to 50 %. Consequently, we need to estimate the parameters in the chosen estimation models based on censored data in a way that the results are more accurate than those of the simple methods mentioned above.

The first known paper in censored time series analysis is Robinson (1980), who considered imputing the censored part of the time series with the conditional expectation of the completely observed part. This method, however, may be infeasible for many consecutive censored observations. The first study to propose an estimation of censored regression model with autocorrelated errors was introduced by Zeger and Brookmeyer (1986). The study presents a full likelihood estimation and an approximation method for the parameters of the censored regression model with autocorrelated errors. The authors mentioned that a high censoring rate might not be feasible with the full likelihood. As a remedy, the authors have suggested to use a pseudolikelihood estimation. Park et al. (2007) propose an imputation method to estimate the parameters of an ARMA model from censored time series data. In the method, the fully observed time series data is regarded as coming from a multivariate normal distribution. The censored values are imputed from the conditional multivariate normal distribution given the observed data. After the imputation is complete, the time series can be analyzed using normal time series methods.

Schumacher et al. (2017) also develop a censored regression model with autoregressive errors of order $p$, similar to Zeger and Brookmeyer (1986). However, the parameters of the model are obtained as the maximum likelihood estimates from the stochastic approximation of the expectation-maximization (EM) algorithm. The results of the study are promising, even with a censoring rate of over 40 %, the proposed method is able to provide consistent results.

Aydin and Yilmaz (2021) focus on developing a nonparametric time series regression model with autoregressive error for censored observations. They divide techniques dealing with censored time series data into parametric and nonparametric methods and try to discern which will provide a better estimation for auto-correlated censored data. They found that increasing the censoring rate decreases the quality of the estimation. Also, the developed nonparametric models performed better than a Naïve $AR(1)$ model. The paper did not compare other parametric models with the developed nonparametric models.

Sousa et al. (2023) propose a model in a Bayesian framework for estimating a linear regression model with autocorrelated errors from censored observations. The proposed algorithm implements a Gibbs sampler with data augmentation where the data augmentation is completed by calculating the mean of multiple simulations which improves the accuracy of the algorithm.

Informed by the background knowledge of this chapter, our thesis adopts a Linear Regression model with Autoregressive errors, influenced by the works of Zeger and Brookmeyer (1986) and Schumacher et al. (2017), due to its ability to manage autocorrelated errors in censored datasets. Additionally, we incorporate the Gaussian imputation method, inspired by the approach of Park et al. (2007), who treat fully observed time series data as part of a multivariate normal distribution. This method

is suited to our context, given the varying censoring rates in our dataset, and the aim to provide a more nuanced and accurate estimation compared to simpler methods.

# 3 Dataset description

The considered dataset contains aggregated monthly cell-level data collected from May 2020 to June 2023 (i.e., 37 months). The dataset forms a sizeable and diverse cellular network consisting of both the more established fourth generation (4G) and the advanced fifth generation (5G) cells. Initially, these were given as two separate datasets but were later merged to form a comprehensive single dataset for this thesis.

In terms of scale, this dataset offers a panoramic view of a cellular network serving an entire country. It comprises:

- Sites: 3352 in total, which are the physical locations housing cell equipment.
- Sectors: 8896, representing the subdivision of a site, with each sector covering a specific direction or part of the site.
- Cells: 26688, the individual units responsible for wireless communication to and from devices.

Each dataset entry is structured on a cell-level and captures monthly aggregates. A good way to understand this is by considering the Traffic metric: it represents the cumulative volume of data transmitted within a specific cell over a month. To streamline the data for our modelling process, we aggregated the dataset further from the cell level to the sector level. This comprehensive dataset was provided by Omnitele, our client organization.

The original dataset contained a large number of features, 62 in total, some of which were irrelevant for the objective of this thesis. The relevant features were identified through feature correlation analysis and by leveraging expert knowledge. Further discussion on this feature selection process is detailed in Sections 4.2 and 4.3. The features chosen for analysis or used for calculating relevant Key Performance Indicators (KPIs) in this thesis are listed in the subsequent section.

- Node: Node is a radio network element of the cellular network
- Traffic: Total volume of Downlink (DL) data traffic sent in a cell, in kilobits (kbit).
- Avail_PRB: The average number of DL Physical Resource Blocks (PRB) available.
- Used_PRB: The average number of DL PRBs used.
- Avg_active_users: Average number of active users during the month in the DL in a cell.

- Avg_conn_users: Average number of connected users during the month in the DL in a cell.
- Conn_attempts: The number of connection requests.
- Time: Total duration of data transmission in a cell, with the precision of 1 ms.
- CQI_quality_(0-15): Channel Quality Indicator (CQI) indicates DL channel quality. There are separate columns in the raw data ranging from 0 to 15, and together they indicate the overall quality of the DL channel.

While the raw dataset offers valuable insights, diving deeper into the nuances of cellular networks requires the derivation of additional metrics. That is why we introduce KPIs as additional features in this thesis. The KPIs are used to shed light on the resource utilization, scheduling activity and channel quality for each sector in the cellular network. The KPI Utilization is the same as PRB utilization, discussed in Section 2.1. The KPIs have been calculated using specific equations derived from the raw dataset features. The equations are as follows

$$\text{Utilization} = \frac{\text{Used\_PRB}}{\text{Avail\_PRB}} \tag{1}$$

$$\text{Sched\_activity} = \frac{\text{Time}}{\text{\# of milliseconds in a data period}}. \tag{2}$$

Given the dataset of this thesis and columns identified with the prefix 'CQI_quality', for each value $c$ in such columns, let the weight $w_c$ be the integer value extracted from the column name. The average CQI, CQI_avg, for each row can then be represented as:

$$\text{CQI\_avg} = \frac{\sum_c cw_c}{\sum_c c}. \tag{3}$$

The descriptions for the KPIs are listed below.

- Utilization: This KPI measures the usage of Physical Resource Blocks (PRB) in the LTE system. It serves as an indicator for network resource consumption. A higher Utilization value is indicative of more extensive network resource usage.
- Sched_activity: Represents the level of operations performed by the cell scheduler. It indicates how frequently the scheduler is allocating resources to user equipment (UE) based on various criteria such as QoS requirements, channel conditions, and fairness considerations. Elevated Sched_activity values can denote a high number of UEs, fluctuating network conditions, or both.
- CQI_avg: This KPI, standing for Average Channel Quality Indicator, provides a composite measure of the channel conditions experienced by all UEs in the network. It is a weighted average across different CQI quality brackets. A higher CQI_avg value generally suggests better overall channel conditions, which can lead to higher data rates and improved user experiences.

In summary, the set of features that have been selected for the analysis includes Traffic, representing the volume of data flowing through the network, Avail_PRB and Used_PRB which indicate available and utilized Physical Resource Blocks respectively, and metrics like Avg_active_users and Avg_connected_users that provide insights into user engagement levels. Additionally, Conn_attempts sheds light on the network's demand, Utilization offers a perspective on network resource consumption, Sched_activity gives an understanding of LTE scheduler operations, and CQI_avg serves as an aggregate measure of channel conditions. As the analysis progresses, there will be an inclusion of an additional feature, Site-level traffic. The discussion and aggregation technique of this feature will be detailed in Section 4.1.

# 4 System model

This section introduces the system model developed to estimate the demand for data traffic. We go through every phase, including data preprocessing, which covers data validation, cleaning, and aggregation. Subsequently, we delve into feature selection, data transformation, data splitting (into training and test sets), feature importance calculation, performance metrics for the model, and the selected estimation and forecasting models for this thesis. A visual overview of the system model is illustrated in Figure 7.

The initial phase involves gathering raw data from multiple locations within the client organization's database. Due to the data's fragmented nature, it was essential to merge these pieces into a cohesive dataset. Upon data consolidation, rigorous validation and cleaning measures were applied to ensure data quality. The following list provides a look into the preprocessing steps:

- Data Type Consistency: Correcting data types to ensure uniformity across the dataset.
- Quality Checks: Addressing and fixing poor quality data entries. For instance, replacing the placeholder "NIL" (refer to Figure 6) with the Python-friendly "NaN" representation.
- Unit Transformation: Transforming units to be consistent across the dataset. An instance of this would be converting units in the Traffic column, which varied between gigabits (Gbit) and kilobits (kbit) across different data fragments.
- Column Name Standardization: Establishing uniform column names for 4G and 5G data. Initially, despite representing identical metrics, the 4G and 5G data had varying column names.

Figure 6 showcases an example of the data cleaning undertaken during this thesis. It is essential to highlight that this figure shows a hand-selected example, and missing values in the raw data were sparse, accounting for less than 0.1% of the dataset. Nevertheless, data preprocessing was still required to enhance data integrity.

For the purposes of this study, maintaining the time series structure was necessary, especially given that both our estimation and forecasting models leverage autoregressive errors. Thus, sectors with incomplete data entries in the raw data were removed to simplify the subsequent analysis.

| Date | gNodeB Name | Cell Name | N.PRB.DL.A | N.PRB.DL.U | N.PRB.UL.A |
|---|---|---|---|---|---|
| 2020-08 | SF1084N | 10843N | NIL | NIL | NIL |
| 2020-08 | SF5922N | 59221N | NIL | NIL | NIL |

Figure 6: An example of data that had to be cleaned. Missing values were given in the raw data as "NIL" which is not a supported format for a missing value in Python.

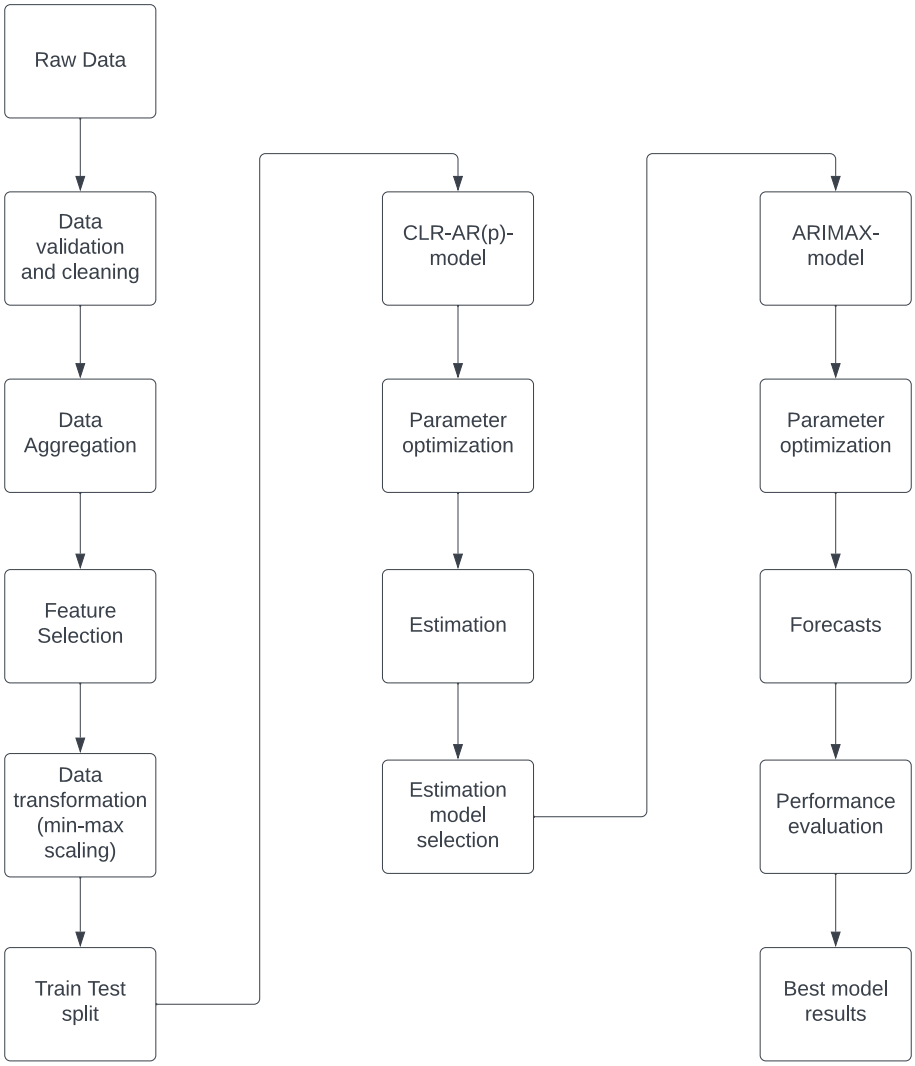


Figure 7: Visual presentation of the system model and the modelling workflow.

## 4.1 Aggregation of datasets

Refer to Section 3 for a comprehensive breakdown of the dataset used in this research. Given our goal to analyze data traffic at the sector level, the raw dataset, which is initially at the cell level, must be aggregated to a sector level dataset. Some features require diverse aggregation methodologies due to their representation as monthly aggregates in the original raw dataset. For example, features like Traffic, Time, CQI_quality_(0-15) and Conn_attempts are aggregated using the summation method. Conversely, features initially presented as averages, such as Avail_PRB, Used_PRB, Avg_active_users, Sched_activity, and Avg_conn_users, are aggregated using an averaging approach.

Let's consider the entire dataset to be represented by $N_t = \{N_{c1t}, N_{c2t}, ..., N_{cit}\}$, where $N_t$ is the complete set of network sectors at time $t$ and $N_{cit}$ is all features $c$ of a cell $i$ at time $t$. The Equation 4 is the aggregation mechanism for features cumulated using the summation method at time $t$ for each sector. We sum the values from all the cells that belong to that sector:

$$A_{ct} = \sum_{i \in I} N_{cit}, \tag{4}$$

where $A_{ct}$ is the aggregated value of feature $c$ at time $t$ for a sector. Equation 5 is used for features, where we aggregate using averages at time $t$ for each sector, by averaging values from all cells that belong to that sector:

$$A_{ct} = \frac{\sum_{i \in I} N_{cit}}{n}, \tag{5}$$

where $n$ is the number of cells in that sector.

The method outlined in Equation 5 is also employed to derive an added feature called Site-level traffic. This metric is computed by taking the aggregate Traffic across all sectors and then dividing by the sector count. By integrating the Site-level traffic, we aim to capture prevailing trends across sectors and decrease the influence of Traffic outliers.

After converting our datasets from cell-level to sector-level, the next task was to combine the 4G and 5G datasets. Preliminary data exploration indicated that the 4G sectors, on average, carried substantially more data, accounting for roughly 90% of the total volume, while 5G sectors contributed the remaining 10%. Given this disparity, a simple merging of datasets could disproportionately amplify the influence of the 5G dataset values. To address this, we adopted a data-weighted mean approach when combining the 4G and 5G datasets.

The data-weighted aggregation formula used for each time $t$ per sector, is given in Equation 6:

$$DW_t = \frac{\sum_{j \in J} w_j \cdot \text{Traffic}_j}{\sum_{j \in J} w_j}, \tag{6}$$

where $w_j$ is the weight, and $\text{Traffic}_j$ is the traffic volume from $j$. With the datasets appropriately aggregated and merged, we have prepared a comprehensive data

foundation that is beneficial for the subsequent analysis and modeling phases of this thesis.

## 4.2 Feature selection

The dataset provided by the client organization, Omnitele, contained a large number of features, 62 to be precise. While multivariate data can often be an asset, in this context, it presented a challenge. Not all of the features were relevant to the goals of our thesis. Moreover, some features were highly correlated, meaning their combined presence in the estimation and forecasting models would not add any substantial value and might even cause redundancy or multicollinearity. The concept of *multicollinearity* will be explained in detail shortly.

To streamline the dataset, we developed a feature selection process. Our approach goes as follows:

- Correlation Analysis: Before diving deep into modeling, it is crucial to understand the relationships between features. Correlation analysis offers insights into the linear relationships between variables. By analyzing these relationships, we can identify and remove variables that are highly correlated, ensuring that our model does not have redundant information. Our aim was to pick features that individually and collectively improved the forecasting model's predictive accuracy.

- Leveraging Expert Knowledge: Data-driven decisions, while powerful, can sometimes miss the nuances captured by domain expertise. Hence, we also collaborated with experts from the client organization to fine-tune our feature selection. Their insights ensured that we weren't merely relying on numbers but also on the intricate knowledge of how cellular networks operate.

Ultimately, we narrowed down our feature list to eight distinct features, in addition to the main variable Traffic. To visually capture the relationships between these features, we plotted the pairwise correlations in Figure 8.

The correlation between any two variables was computed using Pearson's correlation coefficient, denoted by $r$, as given in Equation 7. This metric provides a value between $-1$ and 1, indicating the strength and direction of the linear relationship between the two variables.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}, \tag{7}$$

where $r$ is the correlation coefficient, $n$ is the sample size, $x$ is the value of the independent variable and $y$ is the value of the dependent variable.

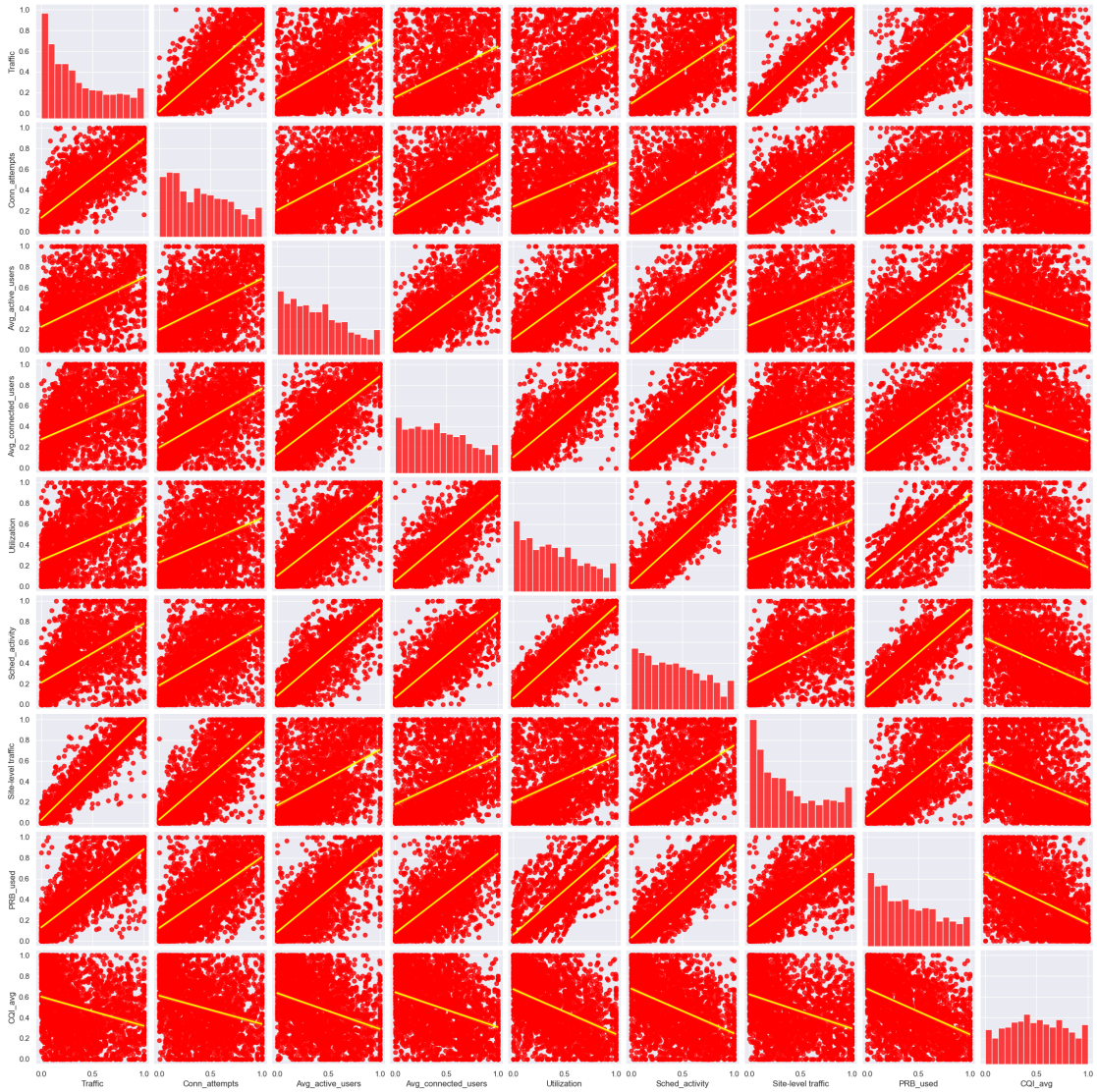


Figure 8: Pairwise correlation plot for all chosen features.

After examining the pairwise correlation plot presented in Figure 8, several patterns emerge. Predominantly, Traffic exhibits a positive correlation with nearly all features, with CQI_avg being the only exception. The correlation is particularly significant between Traffic and certain features, such as Conn_attempts, Site-level traffic, and PRB_used.

The strong correlation between Traffic and Conn_attempts suggests that as data traffic increases, there is a corresponding rise in the number of attempts to establish connections on the network. Similarly, elevated Traffic aligns well with higher Site-level traffic. This observation is intuitively understood: when a site experiences high traffic, its individual sectors are likely to be busy as well. The strong association with PRB_used indicates that as traffic volume grows, a greater number of Physical Resource Blocks are consumed to accommodate this traffic.

Further observations indicate that other metrics, including Avg_active_users,

Avg_conn_users, Utilization, and Sched_activity, also share a direct correlation with Traffic. This observation is consistent with our intuitive understanding. An increase in Traffic is typically a result of more users actively engaging with data services, establishing an elevated number of connections to the cellular network. This alignment between empirical observations and theoretical expectations underscores the robustness of the chosen features in capturing network dynamics.

While considering potential predictor variables for the estimation and forecasting models, the inclusion of multiple variables could enhance model performance. However, this introduces the challenge of multicollinearity. Multicollinearity is a phenomenon where two or more predictor variables display mutual correlation. In the realm of statistics, multicollinearity can inflate the standard errors of coefficients, making them unstable and causing certain predictor variables to appear statistically insignificant when they are, in fact, influential (McClendon, 2002; Daoud, 2017). Consequently, this leads to false conclusions on the model as the most influential predictors are prevented from being captured into the model.

While the pairwise correlation plots provide an initial glimpse into relationships among variables, it is essential to further validate the presence of multicollinearity, especially in the realm of multiple regression models. For this purpose, we employ the Variance Inflation Factor (VIF), a diagnostic tool specifically designed to identify multicollinearity among predictors in such models (Belsley et al., 1980). In essence, VIF evaluates the degree to which the variance of the estimated regression coefficient of a particular predictor is magnified due to multicollinearity. Each predictor in a regression model is assigned a VIF, encapsulating its contribution to the multicollinearity problem.

In a model with multiple predictors $X_i, i = 1, ..., k$, the VIFs can be perceived as the diagonal elements $r_{ii}$ of the inverse of the correlation matrix $R_{k\times k}$ of the predictor variables (Belsley et al., 1980). Mathematically, the VIF for predictor $i$ can be expressed as:

$$r_{ii} = \frac{1}{1 - R_i^2}, \qquad i = 1, ..., k, \tag{8}$$

where $R_i^2$ is the multiple correlation coefficient of $X_i$ regressed against the other predictors in the model.

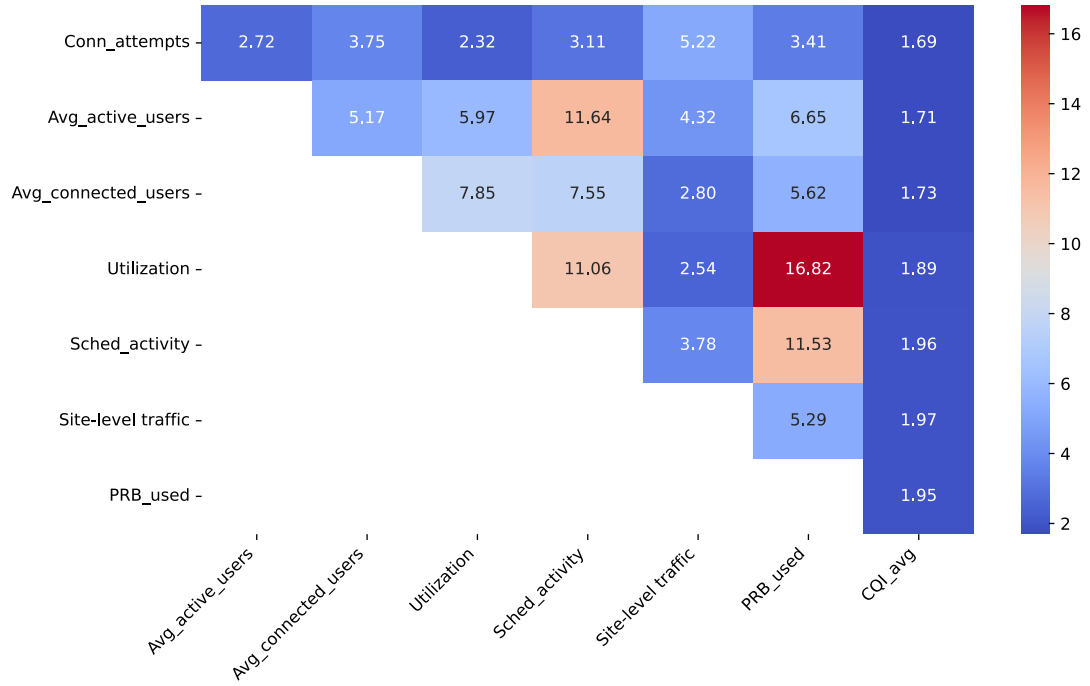


Figure 9: Pairwise VIFs for all chosen predictor variables.

The heatmap presented in Figure 9 shows the multicollinearity among our chosen predictor variables. High VIF values are cautionary signals of collinearity between predictor variables. According to Kutner et al. (2004), VIF values surpassing 10 are said to be a symptom of multicollinearity. Yet, a threshold of 5, frequently used in modern statistical practices (Sheather, 2009), suffices to flag high multicollinearity. Guided by this benchmark, we've adopted the VIF threshold of 5 in this study, removing predictor variable combinations with high multicollinearity from being considered in estimation and forecasting modelling.

In the heatmap, Utilization and PRB_used have significant multicollinearity, which is expected as Utilization is derived from PRB_used. Sched_activity also consistently exhibits high VIF values, especially when paired with Avg_active_users, Utilization, and PRB_used. Furthermore, the user counters, Avg_active_users and Avg_connected_users, display strong interdependence and also exhibit high multicollinearity with multiple features. It is apparent from these observations that there exists an intrinsic overlap in the information these features convey. To ensure the reliability of subsequent analyses, all predictor variable combinations with VIF values exceeding 5, are excluded from subsequent analysis to address potential multicollinearity concerns.

## 4.3 Data Normalization

When dealing with multivariate data in forecasting models, it's extremely important to ensure consistency across features. The dataset in this thesis consists of a large number of features, each with distinct units. For instance, while Traffic is quantified in kilobits, Utilization is listed as percentages. This difference in measurement scales poses potential issues: features with larger scales could influence model outcomes, leading to skewed results.

To fight this inconsistency and to simplify analysis and interpretation of relationships between variables, we use min-max scaling, a form of data normalization. Min-max scaling adjusts all chosen features to a consistent scale, ensuring correct representation in the model. Mathematically, min-max scaling is expressed as:

$$z_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{9}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the data, respectively This transformation scales the data to a $[0, 1]$ range, ensuring uniformity across all features.

## 4.4 Train-Test split

Following the feature selection and data normalization, the next critical step involves partitioning the data into training and test datasets. However, our approach is different from traditional methodologies, such as the typical 80 – 20 split for training and testing datasets. Our splitting criterion focuses on the specific datapoints that correspond to the addition of new infrastructure within a cellular network sector.

This criterion for data splitting is chosen due to our objective to measure the influence of data traffic demand estimations. By starting our forecasts from the time of new infrastructure addition, we can compare forecast results with and without estimations, by analyzing the differences in performance metrics. The idea is straightforward, augmenting a sector with new infrastructure increases its capacity, subsequently reducing congestion. This decrease in congestion, results in convergence of actual data traffic and its latent demand. The specific train-test splits for nine distinct sectors are presented in Table 1.

| Sector | Train | Test |
|---|---|---|
| 1 | 01/05/2020 - 01/05/2022 | 01/06/2022 - 01/06/2023 |
| 2 | 01/05/2020 - 01/05/2022 | 01/06/2022 - 01/06/2023 |
| 3 | 01/05/2020 - 01/05/2022 | 01/06/2022 - 01/06/2023 |
| 4 | 01/05/2020 - 01/07/2022 | 01/08/2022 - 01/06/2023 |
| 5 | 01/05/2020 - 01/06/2022 | 01/07/2022 - 01/06/2023 |
| 6 | 01/05/2020 - 01/05/2022 | 01/06/2022 - 01/06/2023 |
| 7 | 01/05/2020 - 01/10/2022 | 01/11/2022 - 01/06/2023 |
| 8 | 01/05/2020 - 01/08/2022 | 01/09/2022 - 01/06/2023 |
| 9 | 01/05/2020 - 01/09/2022 | 01/10/2022 - 01/06/2023 |

Table 1: Train-test splits for nine sectors in 'DD/MM/YY' format.

## 4.5 Feature Importance

In predictive modeling, understanding the relative importance of features or exogenous variables is paramount. This importance not only reveals which variables significantly influence predictions but also aids in model refinement, ensuring that the most impactful variables are included. To this end, a method centered around the Mean Absolute Error (MAE) metric was devised to quantify the significance of each exogenous variable in our model. In the calculation of feature importances, we largely follow the permutation feature importance algorithm based on Fisher et al. (2019).

The first step in this process is to determine a reference point for comparison, which we call the baseline MAE. Mathematically, the baseline MAE represents the optimal performance within the dataset and is defined as:

$$\text{MAE}_{\text{baseline}} = \min(\mathcal{M}), \tag{10}$$

where $\mathcal{M}$ represents the collection of all MAE values in the dataset. The collection of all MAE values $\mathcal{M}$, comprises from predictions made with varying combinations of exogenous variables to calculate these MAE values, allowing us to assess the variable's significance in prediction performance.

With this baseline established, the importance of each exogenous variable $v$ is computed. This entails identifying the subset $\mathcal{M}_v$ of MAE values associated with models that incorporate variable $v$. The average MAE for this subset is then calculated as:

$$\text{MAE}_{\text{avg},v} = \frac{1}{|\mathcal{M}_v|} \sum_{m \in \mathcal{M}_v} m, \tag{11}$$

where $|\mathcal{M}_v|$ denotes the count of MAE values in the subset $\mathcal{M}_v$. The feature importance, $\text{FI}_v$, is subsequently derived from the difference between this average and the baseline:

$$\text{FI}_v = \text{MAE}_{\text{avg},v} - \text{MAE}_{\text{baseline}}. \tag{12}$$

The interpretation of these importance values is straightforward. If $\text{FI}_v > 0$, it indicates that, on average, models that include variable $v$ have a performance that deviates from the best-performing model. Conversely, a value of $\text{FI}_v$ close to zero

implies that the variable's inclusion provides forecasting performance close to the optimal model.

While this method offers an intuitive way to calculate feature importance, it is anchored to the MAE metric. Hence, its insights and implications are most valuable in scenarios where MAE is chosen as the primary performance metric.

## 4.6 Model Performance Metrics

In cellular network forecasting, the data traffic demand, particularly for congested sectors, remains an unobservable variable, making its accurate estimation and forecasting crucial. Imprecise predictions can have significant consequences, highlighting the importance of determining the reliability of our models. The performance metrics selected in this thesis are not just a formality but also a direct measure of the estimation model's capability of accurately estimating the unobserved data traffic demand, thus establishing the model's credibility.

To ensure a comprehensive and robust evaluation, we have selected a set of widely accepted performance metrics. Specifically, we have adopted the Mean Squared Error (MSE) and its more interpretable variant, the Root Mean Square Error (RMSE), to measure the magnitude of the model's errors. The Mean Absolute Error (MAE) complements these by providing a direct average of absolute differences, capturing the model's error uninfluenced by extreme values. For understanding errors in relative terms across varied scales, the Mean Absolute Percentage Error (MAPE) becomes essential. These metrics are instrumental in measuring the difference between predicted outputs of the forecasting model and the actual data from our test dataset split (Nabi et al., 2023).

Concise explanations for each performance metric are listed below, and their mathematical formulas are displayed in Equations 13 – 16.
Mean Square Error (MSE): This metric represents the average squared difference between the actual and predicted values.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{13}$$

Mean Absolute Error (MAE): This metric represents the average absolute difference between the actual and predicted values.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{14}$$

Root Mean Square Error (RMSE): This metric represents the square root of the MSE, offering a measure of the average magnitude of the errors.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{15}$$

Mean Absolute Percentage Error (MAPE): This metric represents the forecast errors as a percentage, facilitating comparisons across different scales.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{16}$$

where $N$ is the number of data points, $y$ is the vector of real, measured traffic, $\hat{y}$ is the vector of predicted traffic.

It is important to note that the forecast errors are not sign independent. For a MNO, underestimating data traffic demand has more severe consequences than overestimating, given that underestimated sectors are prone to QoS degradation. Even though an asymmetric cost function is not directly incorporated into our performance metrics, its implications are examined, both visually and through discussions, in Section 5.

## 4.7 Censored Linear Regression model with autocorrelated errors of order $p$

In this thesis, we refer to the observed variable as Traffic and denote it by $y_t$. However, we operate under the assumption that Traffic is not fully observed at a given time $t$. Considering right censoring, the recorded data for the observed variable at time $t$ can be represented as $y_t = (v_t, c_t)$, where $v_t$ is the uncensored value and $c_t$ stands for the censoring indicator defined as:

$$\begin{cases} y_t \geq v_t, \text{ if } c_t = 1 \\ y_t = v_t, \text{ if } c_t = 0. \end{cases} \tag{17}$$

Given by the definition in Equation (17), whenever $c_t = 1$, it signifies that the corresponding observed variable value, $y_t$, is not fully observed. On the other hand, when $c_t = 0$, it is an indication that we have a completely observed value.

Now consider the classic linear regression model with autocorrelated errors defined as a discrete time autoregressive AR($p$) process. Let us denote this model as LR-AR($p$) model. The discrete time representation of this model for the observed variable Traffic $y_t$ at time $t$ is given by:

$$y_t = \boldsymbol{x_t}\boldsymbol{\beta} + \epsilon_t \tag{18}$$

$$\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_p\epsilon_{t-p} + \eta_t, \qquad \eta_t \sim N(0, \sigma^2), \tag{19}$$

where $y_t$ is the observed variable, $\boldsymbol{x_t} = (x_{t1}, x_{t2}, ..., x_{tk})$ is a vector of explanatory variables or features of dimension $k$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)$ is the vector of regression coefficients of dimension $k$, $\epsilon_t$ is a stationary AR($p$) process with Gaussian disturbance $\eta_t$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_p)$ is the vector of autoregressive coefficients for the AR($p$) process satisfying the stationarity conditions.

In this thesis, we encounter the challenge of censored values in our observed variable $y_t$. In the domain of time series forecasting, censoring causes challenges in model

development and the accuracy of predictions. This is because censored observations provide limited information, capturing only a partial view of the underlying dynamics. To address this challenge, we introduce the Censored Linear Regression model with autoregressive errors, providing the necessary means to tackle the challenges of censoring. Denoted as CLR-AR($p$) for simplicity and clarity, its formulation is as follows:

$$\begin{aligned} &\begin{cases} y_t & \geq v_t, \quad \text{if } c_t = 1 \\ y_t & = v_t, \quad \text{if } c_t = 0 \end{cases} \\ &y_t = \boldsymbol{x_t}\boldsymbol{\beta} + \epsilon_t \\ &\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_p\epsilon_{t-p} + \eta_t, \qquad \eta_t \sim N(0, \sigma^2), \end{aligned} \tag{20}$$

Going forward, we will denote $\mathbf{y} = (y_1, y_2, ..., y_T)$ as the observed variable, $\mathbf{v} = (v_1, v_2, ..., v_T)$ as the corresponding latent variable, $\mathbf{X}$ is the $T$ x $k$ matrix of regressors and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi})$ is the parameter vector.

### 4.7.1 Gaussian imputation method

Censored data can lead to biases and inaccuracies if not properly addressed (Schumacher et al., 2017). Recognizing this, we adopt the Gaussian imputation method to counteract the effects of censored observations and to provide a structured pathway for imputation. The Gaussian imputation method is grounded on the assumption that the observed time series data can be conceptualized as a realization of a multivariate normal distribution. The Gaussian imputation method was first introduced in Park et al. (2007), and has since become a popular strategy in addressing censored observations in time series data. The method can be represented as:

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{21}$$

with $N_n$ being an $n$-dimensional multivariate normal distribution. This distribution is characterized by a mean vector $\mathbf{X}\boldsymbol{\beta}$ and a covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2\mathbf{M}_n(\boldsymbol{\phi})$, which can be detailed as:

$$\mathbf{M}_n = \frac{1}{\boldsymbol{\sigma}^2}\begin{bmatrix} \gamma_0 & \gamma_0 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0, \end{bmatrix} \tag{22}$$

where the elements $\gamma_0, \gamma_1, \ldots, \gamma_{n-1}$ denote the autocovariances of the process. Autocovariance, being a measure of how much the current value of a series is correlated with its past values, is integral in ensuring the continuity and consistency of the imputation. These autocovariances are modeled as a component of the truncated multivariate normal distribution.

Handling complete data, the Gaussian imputation method is straightforward, as illustrated by Equation 21. However, when we consider censored data, things get

more complicated. Considering censored responses as depicted in Equation 17, we consider the truncated multivariate normal distribution

$$\mathbf{y} \sim TN_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}; A), \tag{23}$$

where $TN_n(,;A)$ denotes the truncated multivariate normal distribution constrained within interval $A$. The interval $A$ dictates the boundaries based on the censoring condition in the following way:

- For non-censored data, where $c_t = 0$, the range is unrestricted, spanning $(-\infty, \infty)$.
- For censored data, where $c_t = 1$, the range is asymmetric, spanning $(-\infty, v_t]$.

The fundamental premise behind this methodology is to acknowledge the fact that censoring imposes boundaries on our data, and a truncated multivariate normal distribution respects these boundaries, ensuring that the imputed values are statistically plausible and coherent with the nature of our data.

Next, we separate the observed data from the censored data. This separation is important to effectively impute the censored observations by using the information from the observed data. Thus, it becomes essential to partition our dataset accordingly and organize our dataset into two parts: one for observed data and one for censored data.

Let the observed part be $\mathbf{y}_O$ and censored part be $\mathbf{y}_C$ so that $\mathbf{y}$, $\mathbf{v}$, $\mathbf{X}$, $\boldsymbol{\Sigma}$ can be partitioned as:

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}_O \\ \mathbf{y}_C\end{bmatrix}, \mathbf{v} = \begin{bmatrix}\mathbf{v}_O \\ \mathbf{v}_C\end{bmatrix}, \mathbf{X} = \begin{bmatrix}\mathbf{X}_O \\ \mathbf{X}_C\end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{OO} & \boldsymbol{\Sigma}_{OC} \\ \boldsymbol{\Sigma}_{CO} & \boldsymbol{\Sigma}_{CC}\end{bmatrix}. \tag{24}$$

After partitioning the data into observed and censored parts, we can analyze the conditional distribution of $\mathbf{y}_C$ in relation to $\mathbf{y}_O$. This conditional distribution is modeled as a multivariate normal distribution characterized by parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$, as outlined by Anderson (2003). The objective is to substitute censored values in $\mathbf{y}_C$ by drawing samples from the conditional distribution derived from the observed values. Thus, we have:

$$\begin{aligned}\mathbf{y}_O &\sim N_{n_O}(\mathbf{X}_O\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{OO}}) \\ \mathbf{y}_C|\mathbf{y}_O &\sim TN_{n_C}(\boldsymbol{\nu}, \boldsymbol{\Delta}; A),\end{aligned} \tag{25}$$

where $n_O$ is the number of observed observations, $n_C$ the number of censored observations and

$$\begin{aligned}\boldsymbol{\nu} &= \mathbf{X}_C\boldsymbol{\beta} + \boldsymbol{\Sigma}_{CO}\frac{1}{\boldsymbol{\Sigma}_{OO}}(\boldsymbol{y}_O - \boldsymbol{X}_O\boldsymbol{\beta}) \\ \boldsymbol{\Delta} &= \boldsymbol{\Sigma}_{CC} - \boldsymbol{\Sigma}_{CO}\frac{1}{\boldsymbol{\Sigma}_{OO}}\boldsymbol{\Sigma}_{OC},\end{aligned} \tag{26}$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$ denote the conditional mean and covariance of the observed part of the data, meaning the parameters of a non-truncated version of a conditional multivariate normal distribution and $A$ is the censoring interval.

We describe the estimation process in detail, in Algorithm 1, to make the process easier to understand. The core principle of the imputation algorithm is to replace censored observations with estimated values using the conditional sample from the completely observed observations.

**Algorithm 1** The imputation algorithm

1: Generate the truncated multivariate normal distribution: $TN_{n_C}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}; A)$.

2: Obtain the initial values: $\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\boldsymbol{\sigma}}^{2(0)}, \widehat{\boldsymbol{\phi}}^{(0)}$:

$$\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\widehat{\boldsymbol{\sigma}}^{2(0)} = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \mathbf{x}_t^T\widehat{\boldsymbol{\beta}}^{(0)})$$

$$\widehat{\boldsymbol{\phi}}^{(0)} = pacf_p(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(0)}),$$

where $pacf_p$ is the partial autocorrelation function for lags $p = 1, \ldots, p$ calculated with the $pacf()$-function available in the statsmodels library in Python (Seabold and Perktold, 2010).

3: Calculate the conditional mean $\widehat{\boldsymbol{\nu}}^{(0)}$ and variance $\widehat{\boldsymbol{\Delta}}^{(0)}$ based on the censored part of the data as follows:

$$\widehat{\boldsymbol{\nu}}^{(0)} = \mathbf{X}_C\widehat{\boldsymbol{\beta}}^{(0)} + \widehat{\boldsymbol{\Sigma}}_{CO}^{(0)}(\widehat{\boldsymbol{\Sigma}}_{OO}^{(0)})^{-1}(\boldsymbol{y}_O - \boldsymbol{X}_O\widehat{\boldsymbol{\beta}}^{(0)})$$

$$\widehat{\boldsymbol{\Delta}}^{(0)} = \widehat{\boldsymbol{\Sigma}}_{CC}^{(0)} - \widehat{\boldsymbol{\Sigma}}_{CO}^{(0)}\widehat{\boldsymbol{\Sigma}}_{OO}^{(0)})^{-1}\widehat{\boldsymbol{\Sigma}}_{OC}^{(0)}.$$

4: Generate the sample $\mathbf{y}_C^{(1)}$ for the censored data points from the truncated multivariate normal distribution: $TN_{n_C}(\widehat{\boldsymbol{\nu}}^{(0)}, \widehat{\boldsymbol{\Delta}}^{(0)}; A)$.

5: Construct the augmented data from the observed data and the imputed sample for the censored data:

$$\mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{y}_O \\ \mathbf{y}_C^{(1)} \end{bmatrix}.$$

**Algorithm 2** The imputation algorithm (Continued)

6: Re-estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\phi}$ based on $\mathbf{y}^{(1)}$ and update the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$, as shown in Step 3. The parameters are updated with the following estimates for the $k$th iteration:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^T\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^T\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\widehat{\mathbf{y}^k}$$

$$\begin{aligned}\widehat{\boldsymbol{\sigma}^2}^{(k+1)} = \frac{1}{n}\Big(&\mathrm{tr}(\widehat{\mathbf{y}^{2(k)}}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})) - 2\widehat{\boldsymbol{\beta}}^{(k)T}\mathbf{X}^T\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\widehat{\mathbf{y}^{(k)}} \\ &+ \widehat{\boldsymbol{\beta}}^{(k)T}\mathbf{X}^T\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)}\Big)\end{aligned}$$

$$\begin{aligned}\widehat{\boldsymbol{\phi}}^{(k+1)} = &-\frac{n}{2}\log\left[(-1,(\widehat{\boldsymbol{\phi}}^{(k)})^T)D(\widehat{\mathbf{y}^k},\widehat{\boldsymbol{\beta}}^{(k)})(-1,\widehat{\boldsymbol{\phi}}^{(k)})^T\right] \\ &-\frac{1}{2}\log\left[\prod_{i=1}^{p}(1-\phi_i^{(k)^2})^{-i}\right].\end{aligned}$$

7: Repeat Steps 3 – 6 until the parameter estimates converge, meaning that $\psi < 0.0001$ which is the convergence threshold for this thesis. $\psi$ is calculated using the following convergence rule:

$$\psi = \sqrt{(\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)})^T(\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)})},$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}, \widehat{\boldsymbol{\phi}})$ and $k$ is the $k$th iteration.

Before presenting the chosen forecasting model of this thesis, it is important to note that the estimation model used for censored time series data, namely the CLR-AR($p$) model, does not impose any restrictions on the selection of prospective forecasting methods. After completing the estimation process, all suitable methods can be employed to forecast future data traffic values.

## 4.8 Forecasting model

Let $\mathbf{y}_{\text{obs}}$ be the observed variable of length $n_{\text{obs}}$ and $\mathbf{y}_{\text{pred}}$ be the prediction variable that predicts $n_{\text{pred}}$ steps into the future. By replacing the censored observations currently existing in $\mathbf{y}_{\text{obs}}$ with the estimated values $\hat{\mathbf{y}}$, we get the complete dataset $\mathbf{y}^*_{\text{obs}}$ for the observations. Before prediction, the complete dataset is structured as:

$$\widetilde{\mathbf{y}^*} = (\mathbf{y}^*_{\text{obs}}, \mathbf{y}_{\text{pred}}) \tag{27}$$

In this thesis, we have chosen the ARIMAX model as a predictor of $\mathbf{y}_{\text{pred}}$. The ARIMAX model is a composition of parts that together produces $\mathbf{y}_{\text{pred}}$ as an output. The ARIMAX model is comprised of the following parts: the autoregressive (AR) part, moving average (MA) part, integrated (I) component and the exogenous component (X). The AR component captures the inherent autocorrelation within a time series. By regressing the present observation against its historical values, each weighted by

an autoregressive coefficient, it effectively accounts for temporal dependencies. The MA component, on the other hand, focuses on the relationship between the current observation and past error terms. It quantifies the impact of previous errors on the present value through moving average coefficients, addressing short-term fluctuations and noise. Incorporating the integrated (I) component involves differencing the time series, a transformation that facilitates stationarity. This critical step renders the data suitable for modeling with both the AR and MA components, enhancing the model's ability to capture underlying patterns. The X component introduces external exogenous variables that can influence the target time series (Makridakis et al., 2008). By augmenting the model with such exogenous factors, additional information is integrated, thereby contributing to improved forecasting accuracy. This approach makes use of the complexities within time series data, the impact of external variables, and the fundamental principles of autocorrelation and stationarity.

When combined, the components discussed above create the ARIMAX model for $\hat{y}_{\text{pred}}$ given $\mathbf{y}^*_{\text{obs}}$ at time $t$, presented as:

$$\begin{aligned}\hat{y}_{\text{pred},t} = \phi_1 y_{\text{obs},(t-1)} + \ldots + \phi_p y_{\text{obs},(t-p)} \\ + \tau_1 \zeta_{t-1} + \ldots + \tau_q \zeta_{t-q} \\ + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t} + \zeta_t,\end{aligned} \tag{28}$$

where $\hat{y}_{\text{pred},t}$ is the predicted value, $\mathbf{y}_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},t})$ is the observed complete data, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ is the vector of autoregressive coefficients of order $p$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)$ is the vector of moving average coefficients of order $q$, $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ is a vector of exogenous variables of dimension $k$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is a vector of regression parameters of dimension $k$ and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_t)$ is the white noise error term.

In the realm of forecasting, the ARIMAX model emerges as a prominent choice, driven by its simplicity and comprehensibility while yielding meaningful results. The selection of the ARIMAX framework is underpinned by its balanced complexity, rendering it powerful enough to capture underlying trends and patterns in data without becoming excessively intricate. This choice supports the objective of achieving accurate forecasts while maintaining the capacity to interpret and communicate the results effectively.

The core purpose of forecasting extends beyond mere prediction. It serves as a litmus test for the accuracy of our estimation process. The forecasted values stand as a benchmark against which the estimated values are evaluated. The underpinning principle is straightforward: if the accuracy of our forecasts improves when built upon estimation-derived inputs, then the estimation itself can be deemed accurate. Implementing both estimation and forecasting serves as a dual validation mechanism, reinforcing the validity of our approach and results.

# 5 Results

In Section 5.1 the results of the estimation are examined for an example sector in the cellular network. The candidate models, CLR-AR(1), CLR-AR(2) and CLR-AR(3) are presented and their results are displayed in Table 2. The best estimation model is determined by finding the minimum Akaike Information Criterion (AIC) value, which is a statistical measure used for model selection. A lower AIC value indicates a better fit of the model to the data. Of the considered estimation models, CLR-AR(1) performed best, and thus we continue the analysis with this model in Section 5.2. In this section, the future data traffic demand is forecasted using results of the best estimation model as training data to be utilized by the forecasting model. The difference in forecasting with original data and estimated data is investigated in Section 5.2.

The data for the example sector contains monthly observations recorded from May 2020 to October 2022, consisting of 30 observations where nine observations are censored, and consequently, the censoring rate is 30 %. There were zero missing observations in the data.

## 5.1 Estimation results

The best estimation model is chosen based on the smallest AIC value. For model selection, the raw data was fitted for three CLR-AR($p$) candidate models, with $p = 1, 2$ and 3. Of the considered models, CLR-AR(1) had the smallest AIC value, and therefore was selected for further research. Figure 10 presents the original raw data and the estimated data based on the parameter estimates of a CLR-AR(1) model. This figure gives an estimation for data traffic demand for a congested sector where the data traffic demand is not fully observed due to capacity limitations and congestion.

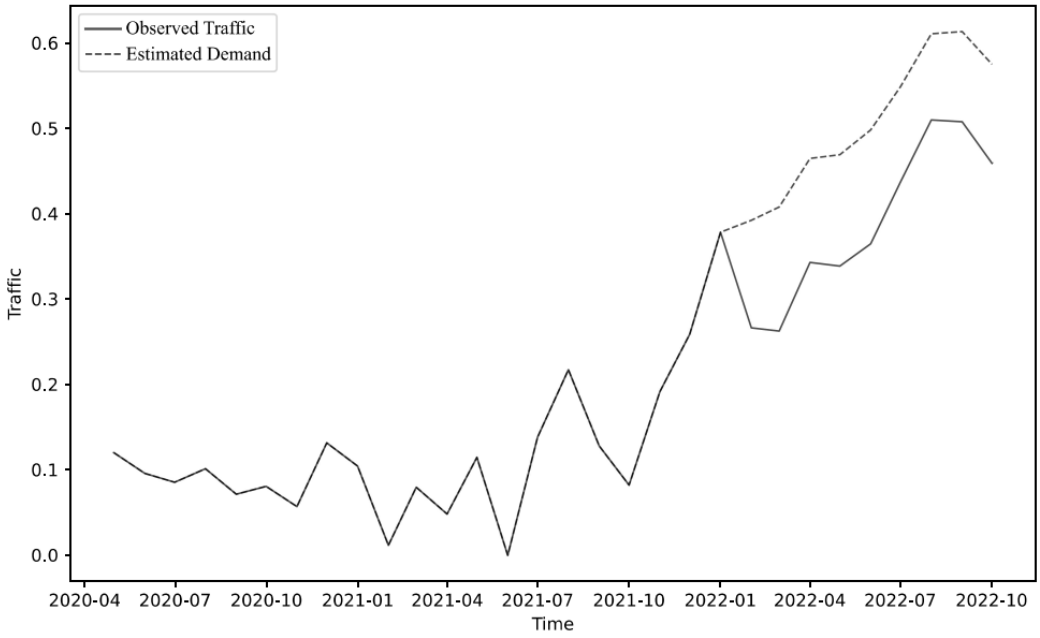


Figure 10: The time series of the original data containing censored observations and estimated data based on the CLR-AR($p$) model. The dashed line is the estimated series based on the fitted CLR-AR(1) model.

| p | AIC | $\hat{\beta}_0$ | $\hat{\sigma^2}$ | $\hat{\phi}_1$ | $\hat{\phi}_2$ | $\hat{\phi}_3$ |
|---|---|---|---|---|---|---|
| **1** | **-36.1** | **0.279** | **0.00855** | **0.904** | | |
| 2 | -35.6 | 0.290 | 0.00653 | 0.706 | 0.234 | |
| 3 | -33.7 | 0.293 | 0.00598 | 0.703 | 0.143 | 104 |

Table 2: Parameter estimates for the three candidate models. AIC is displayed as the criterion for model selection. The best model is presented in bold.

The results of the estimation are displayed in Table 2. It contains the AIC values which we use to determine the best model, and the parameter estimates for the three candidate models. The effect of the $p$-parameter in the CLR-AR($p$)-model is not drastic. The AIC values and the parameter estimates are fairly similar. The best $p$-parameter does not provide significantly better estimates.

Based on Figure 10, the estimates for the censored observations look reasonable. However, with the methods discussed in this section, we currently lack the means to measure whether the estimation of censored observations leads to improvements in forecasting future data traffic demand. Consequently, we are unable to ascertain the accuracy of our estimates for censored observations. This challenge, however, will be addressed in Section 5.2. In the upcoming section, we will employ forecasting techniques to assess the impact of our estimation approach on predicting future data traffic demand in congested sectors. We will compare forecasts made using both the original data and the estimated data derived from the fitted CLR-AR(1) model. The

goodness of fit for these forecasts will be evaluated using a test dataset that starts from a data point where new infrastructure is added to a sector.

## 5.2 Forecast performance of censored versus estimated data

In this section we present the results of the forecasting method for the same example sector as in Section 5.1. Additionally, we consider the forecasting results for the aggregate performance across 100 randomly selected sectors from the entire network, and separately evaluate the best fit for nine randomly chosen sectors within the cellular network. The Figure 11 shows the forecast for the data traffic using both the original data and the estimated data as training data. The forecasts are received from the ARIMAX model. Both forecasts use the exact same model parameters and exogenous variables, the only difference between the two being the training data for the forecasting model. The red vertical line in the Figure denotes the point where we transition from training the model using historical data on the left side to testing the forecasting accuracy on the right side.

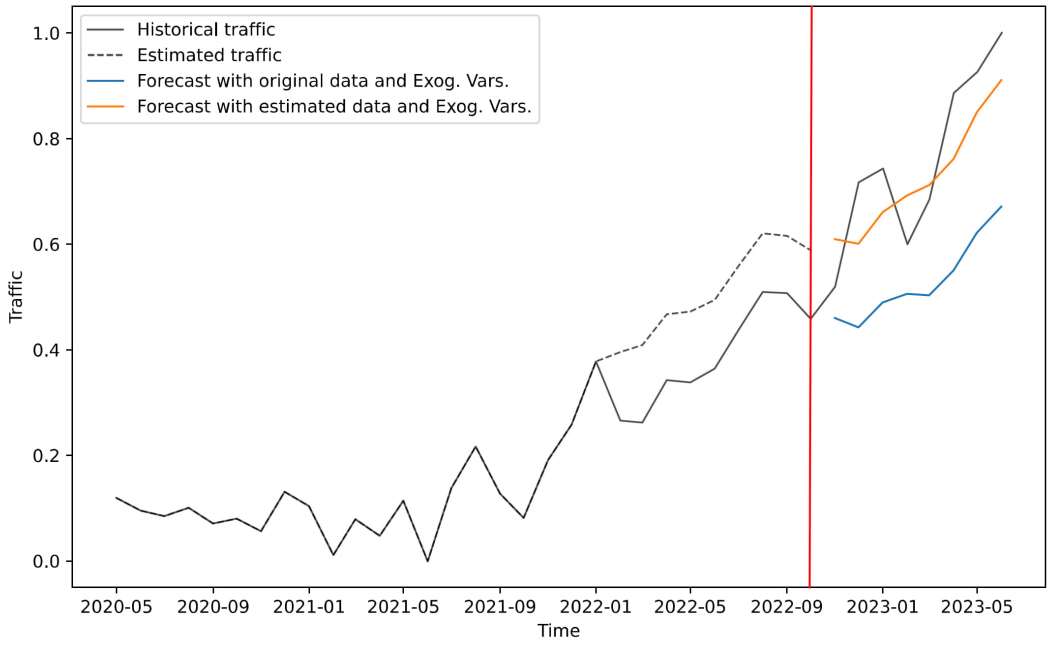


Figure 11: The forecasts using both the original and estimated data as training data.

The performance for the forecast using the estimated data, the exogenous variables Conn_attempts and Avg_connected_users and the ARIMAX model parameters $p = 1$, $d = 2$ and $q = 1$ is substantially better compared to the forecast with the original data, the same exogenous variables and model parameters. The performance metrics for the forecasts are presented in Table 3. All performance metrics for the model trained with the estimated data outperform those of the model trained with the original data. The improved performance of the forecast with the estimated data shows that the demand for data traffic in congested sectors can be estimated

accurately, and the estimation improves the predictability of future data traffic, at least for some sectors. However, as we will later discuss in this section, the effectiveness of data estimation largely varies across sectors and scenarios.

| Type | MSE | MAE | RMSE | MAPE |
|---|---|---|---|---|
| Original data | 0.0380 | 0.170 | 0.200 | 28.7 |
| Estimated data | 0.0058 | 0.072 | 0.076 | 13.5 |

Table 3: Performance metrics for the forecasts with original censored data and estimated data in Figure 11.

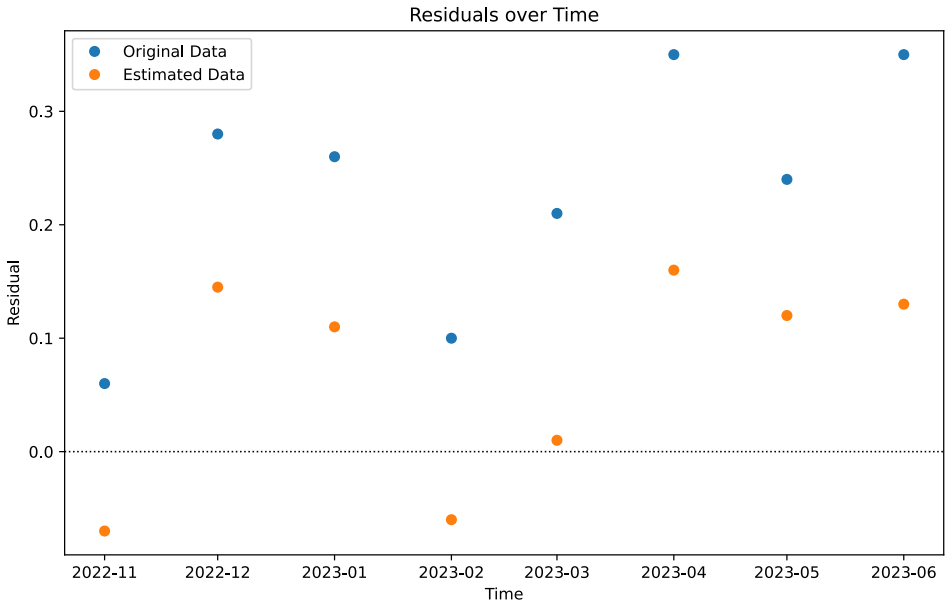


Figure 12: Scatter plot illustrating the distribution of forecast residuals derived from the original and estimated data presented in Figure 11.

In Figure 12, the residuals for the estimated data do not display a growing trend over time, suggesting no systematic increase in forecasting error. However, there is a clear and consistent pattern of positivity in the residuals for five out of the eight forecasted values, for both the original and estimated data. This shows a recurring under-prediction by our forecasting model. A plausible reason for this under-prediction is the recent infrastructure addition in October 2022. This addition has led to a significant surge in Traffic, which makes forecasting large jumps challenging, especially when the model is trained on historical data without any infrastructure additions.

It is evident from Figure 12 that the residuals for the original data are consistently larger for almost every observation, further highlighting the accuracy of our estimation model, at least for this individual sector in the network.

We have now showcased the enhanced predictability achieved through data traffic demand estimation in an example congested sector. To more conclusively address our primary research question concerning the accurate estimation of unobserved data traffic demand in congested sectors, we need to broaden our examination to include a larger number of sectors within the cellular network. Thus, we undertook a randomized selection of 100 sectors from the cellular network. While we aimed for a randomized selection of sectors to maintain objectivity in our research, we strategically imposed one condition: the chosen sectors must have experienced a capacity expansion between June 2022 and October 2022. This decision is rooted in our broader research objective to gauge the influence of data traffic demand estimations. As explained earlier in the thesis, initiating our forecasts from the point of new infrastructure addition enables us to compare forecast results with and without demand estimations. Essentially, when a sector undergoes infrastructure addition, its capacity increases, leading to a subsequent reduction in congestion. This reduced congestion causes the actual data traffic to converge with its latent demand. By setting this criterion, we could then observe how our estimations compared in this scenario.

Within this sample of 100 sectors, we assessed combinations of the following parameters:

- $p = 1, 2, 3$.
- $d = 1, 2$.
- $q = 1, 2, 3$.
- Exogenous variables: Conn_attempts, Avg_active_users, Avg_connected_users, Utilization, Sched_activity, Site-level traffic, PRB_used and CQI_avg.

Every possible combination of model parameters and exogenous variables is considered through an exhaustive grid search, excluding a few combinations based on the results of the feature correlation discussed in Section 4.2. The results for the best combination and a selected few for comparison are presented in Table 4. All values in the table are calculated using the mean for the 100 randomly chosen sectors.

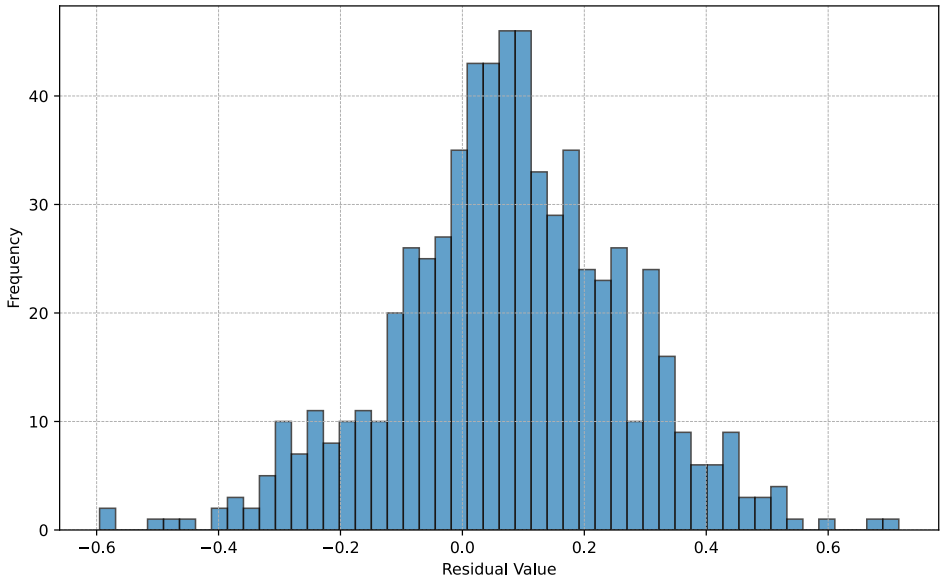


Figure 13: Histogram showing the distribution of forecast residuals for all 100 randomly chosen sectors. The residuals are derived from the best performing model in Table 4.

Figure 13 presents the histogram of forecast residuals across all 100 randomly chosen sectors. The histograms are acquired from an ARIMAX model with parameters $p = 2$, $d = 2$ and $q = 3$, and exogenous variables Site-level traffic and PRB_used which is the model with the best performance. The shape of the histogram reveals characteristics that resemble a normal distribution, a bell-shaped curve centered around a value greater than zero. This central value suggests a possible systematic bias in the model's predictions. While normality in residuals is often desirable, the non-zero centering raises concerns about the potential for consistent underpredictions by the model.

In the process of model evaluation, the choice of an appropriate performance metric is important in ensuring that the model's predictions align with the underlying patterns and nuances of the data. After analyzing the residuals of our model for an example sector in Figure 12, and given our observations from the histogram of residuals in Figure 13, it becomes clear that there might be systematic deviations in our model's predictions. These deviations can lead to potential biases if we do not select an appropriate performance metric to evaluate the goodness of fit. In this case, metrics that assume zero-centered normally distributed residuals are not a good fit. Given these insights and to address the potential biases, the Mean Absolute Error (MAE) was chosen as the primary performance metric for our models. MAE provides a straightforward interpretation by representing the average magnitude of errors between predicted and observed values, without giving undue weight to outliers (Willmott and Matsuura, 2005). Moreover, its linear penalty for errors ensures a balanced representation of the model's accuracy, making it particularly suitable

for our analysis. By leveraging MAE, we aim to minimize the average deviation of predictions from actual observations, thus ensuring robust and reliable model performance.

| Type | $p$ | $d$ | $q$ | Variables | MSE | MAE | RMSE | MAPE |
|---|---|---|---|---|---|---|---|---|
| Estimated data | 3 | 2 | 3 | Conn_attempts, Avg_connected_users, PRB_used | 0.028 | 0.129 | 0.149 | 19.8 |
| Estimated data | 1 | 2 | 1 | None | 0.067 | 0.190 | 0.218 | 24.65 |
| Estimated data | 1 | 2 | 1 | Conn_attempts, Site-level traffic | 0.030 | 0.136 | 0.155 | 20.44 |
| Original data | 1 | 2 | 1 | None | 0.081 | 0.211 | 0.249 | 29.7 |
| Original data | 2 | 2 | 3 | Site-level traffic, PRB_used | 0.024 | 0.111 | 0.130 | 21.02 |
| Original data | 1 | 2 | 2 | Avg_active_users, Site-level traffic, CQI_avg | 0.026 | 0.112 | 0.133 | 20.74 |

Table 4: Performance metrics for the optimal parameter combination and a few other parameter combinations for comparison. All forecasts are given by the ARIMAX model presented in Section 4.8.

Upon analyzing the results in Table 4, a key observation arises that contradicts the results for the example sector in Figure 11. The best forecasting results are now achieved with ARIMAX$(p, d, q)$ model that has model parameters $p = 2$, $d = 2$ and $q = 3$, and exogenous variables Site-level traffic and PRB_used, when trained on original historical data. This model achieves the best MAE of 0.111. However, when we consider models with no exogenous variables, the one trained on estimated data with parameters $p = 1$, $d = 2$ and $q = 1$ outperforms its counterpart from the original data. This suggests that estimating the training data with the CLR-AR$(p)$ model does not increase the prediction accuracy for the ARIMAX model.

This pattern emphasizes the importance of the quality of original data over extensive model adjustments. It reminds us that in predictive modeling, the quality of the input data can be just as crucial as the model's complexity. All results in Table 4 are aggregated for 100 sectors in the cellular network and they are presumed to represent the entire network.

We also see some clear patterns regarding the best exogenous variables. For models using original data, the best performing external variables are Site-level traffic, PRB_used, CQI_avg, and Avg_active_users. These keep showing up in the top-performing models, suggesting they play a key role in making accurate forecasts. On the other hand, when we look at models using estimated data, Conn_attempts, Avg_connected_users, and PRB_used stand out as the top variables. This tells us that the best variables can change depending on whether we are using original or estimated data.

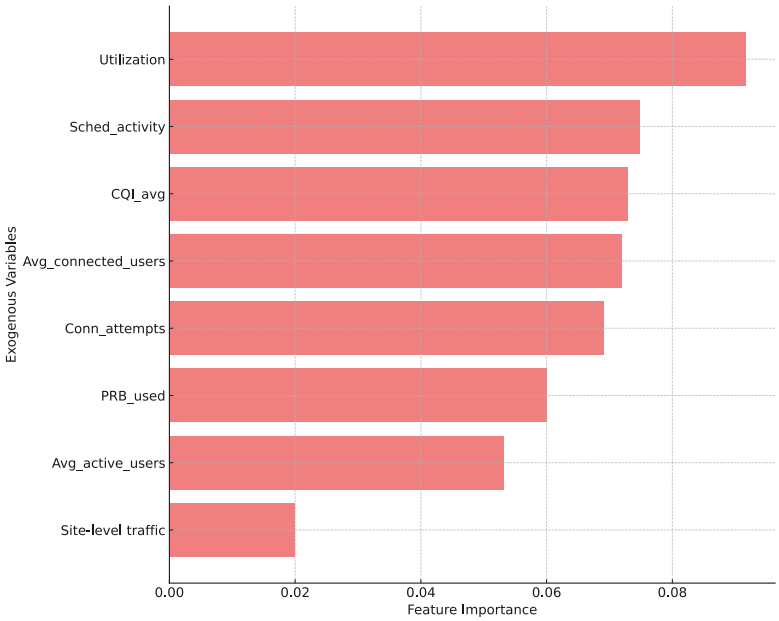


Figure 14: Feature importances for every exogenous variable and ARIMAX model trained with original data. Site-level traffic is easily the most important exogenous variable.

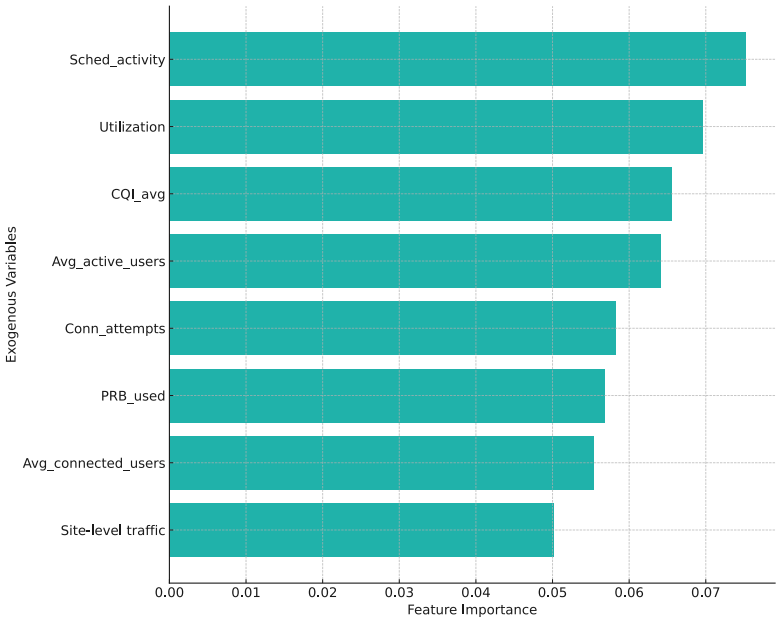


Figure 15: Feature importances for every exogenous variable and ARIMAX model trained with estimated data.

The visualized feature importances in Figures 14 and 15 for models utilizing estimated and original data, respectively, offer a deeper understanding of the significance of each exogenous variable in our predictive models. These values are calculated

using the aggregate results for every parameter combination for the 100 sectors. At a fundamental level, feature importance provides a comparative measure of how the inclusion or exclusion of a specific feature impacts the prediction accuracy. A smaller feature importance means the model's performance, with that feature, is closer to the baseline MAE, representing the best model. In essence, a feature with minimal importance contributes positively to the model's accuracy, moving it closer to the optimal performance. Conversely, a larger importance suggests that the feature introduces complexities that may challenge the model's predictive accuracy.

As previously stated, the results about the best exogenous variables for models utilizing original data are further validated. The feature Site-level traffic seems to clearly be the most important variable. Other important variables include Avg_active_users, PRB_used and Conn_attempts. As for the models utilizing estimated data, the results are also in line with Table 4, with Conn_attempts, PRB_used and Avg_connected_users remaining as important variables. Notably, also Site-level traffic maintains its influential stature across both data types, highlighting its value as an exogenous variable.

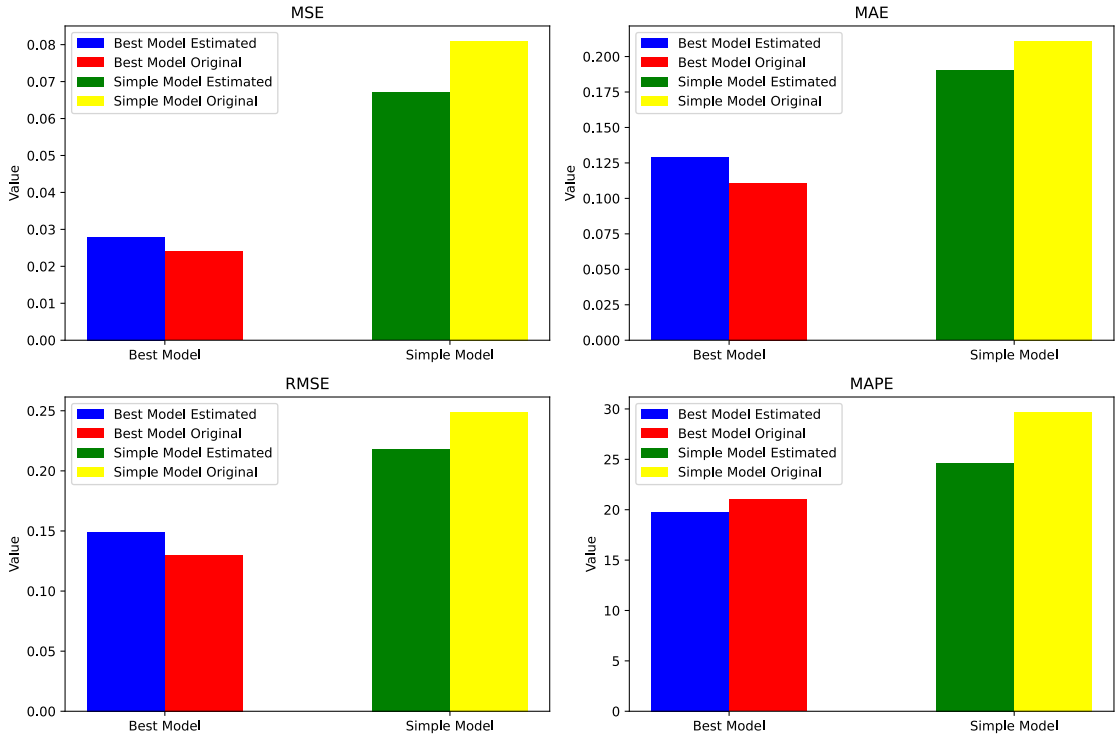


Figure 16: A comparative analysis of error metrics between the best-performing models and a simple model, ARIMAX model with $p = 1, d = 2, q = 1$ and no exogenous variables, applied to both the estimated and original data. Each subplot represents a specific performance metric's value and each bar is the performance value for a forecasting model.

Figure 16 further visualizes our observations from Table 4. It provides a visual presentation of error metrics between the best-performing models and a simple model applied across both data categories. A notable difference is clear in the error metrics between the two approaches. The best-performing models exhibit lower errors, underscoring their tailored efficiency to specific data types. Conversely, the simple model's error increases for both data categories.

In comparing the best-performing models, an intriguing observation is the MAPE metric. The MAPE for the model trained on estimated data is slightly lower than its counterpart trained on original data, indicating that while the best model for estimated data might have some absolute discrepancies, it captures the proportionality of errors more effectively. It is a noteworthy insight as it suggests that the strength of a model might not universally translate across all metrics, and certain contexts might prioritize relative errors over absolute ones.

As to the simple models, it is evident that estimated data consistently outperforms original data across all performance metrics. This indicates that even without the nuances of a tailored model, there's inherent value in the estimated data, likely due to its ability to capture underlying patterns missed in the original dataset.

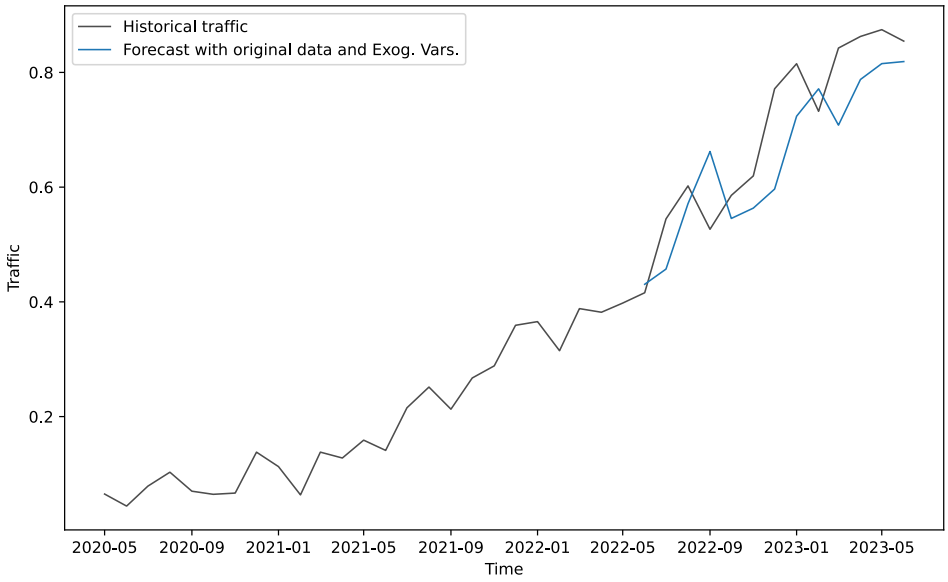


Figure 17: Comparison of historical traffic and the average of the forecast values across all 100 sectors. The forecast values are acquired from the best performing ARIMAX model in Table 4.

In Figure 17, we present a visual comparison between the historical traffic patterns and the average forecasted values for 100 sectors. The black line represents the historical traffic, while the blue line indicates the average predictions from our ARIMAX model across all 100 sectors. The forecasted values capture the overall trend and seasonality of the historical data, however they consistently fall below the

actual traffic levels.

This consistent underestimation is largely attributed to recent changes in the network. As outlined earlier, all the 100 sectors selected for this study underwent infrastructure additions between June 2022 and October 2022. It is quite trivial to understand that such infrastructure additions would lead to an increase in traffic. Forecasting such drastic shifts is challenging, especially when the model is trained on data that does not account for these recent additions.

From analyzing Table 4, it is evident that there is no one-size-fits-all approach to selecting model parameters and exogenous variables; the best combination tends to differ from one sector to another. While the metrics in Table 4 give an average performance across all sectors, many individual sectors have their unique optimal configurations that outperform this average. For example, the example sector depicted in Figure 11 achieves its best results with Conn_attempts and Avg_connected_users as exogenous variables. However, this specific combination isn't universally optimal for all sectors.

In real-world scenarios, to maximize accuracy, it is essential to tailor the parameters to each sector individually. However, this fine-tuning also comes with potential dangers, such as overfitting, as the choice of best exogenous variables, and values for $p$, $d$, and $q$ change noticeably based on the sector in question. Furthermore, the quality of the model's fit is not consistent across all sectors. Therefore, striking the right balance between customization and overfitting is crucial for achieving accurate forecasts.

| Sector | Type | Variables | $p$ | $d$ | $q$ | MSE | MAE | RMSE | MAPE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Original data | Utilization, Site-level traffic, CQI_avg | 3 | 2 | 3 | 0.0017 | 0.026 | 0.042 | 17.14 |
| 2 | Original data | Conn_attempts, Site-level traffic, PRB_used | 1 | 2 | 2 | 0.0017 | 0.033 | 0.041 | 16.5 |
| 3 | Original data | Site-level traffic, PRB_used, CQI_avg | 2 | 2 | 1 | 0.0005 | 0.016 | 0.021 | 13.6 |
| 4 | Original data | Conn_attempts, Avg_connected_users, PRB_used | 2 | 2 | 2 | 0.0025 | 0.040 | 0.050 | 8.53 |
| 5 | Original data | Conn_attempts, Sched_activity, Site-level traffic | 1 | 2 | 3 | 0.0032 | 0.044 | 0.057 | 23.6 |
| 6 | Original data | Avg_active_users, CQI_avg | 3 | 2 | 3 | 0.0017 | 0.033 | 0.041 | 15.6 |
| 7 | Original data | Conn_attempts, Avg_connected_users, Utilization | 1 | 2 | 1 | 0.0021 | 0.033 | 0.046 | 14.9 |
| 8 | Original data | Conn_attempts, Avg_connected_users, PRB_used | 1 | 2 | 3 | 0.0014 | 0.030 | 0.038 | 13.0 |
| 9 | Original data | Utilization, Site-level traffic | 1 | 2 | 3 | 0.0027 | 0.040 | 0.052 | 22.03 |

Table 5: The best unique exogenous variables and forecasting model parameters for nine individual sectors.

A key observation from Table 5 is the variability in the choice of exogenous variables across sectors. While some variables like Site-level traffic make frequent appearances, their combinations with other variables differ across sectors. For instance, Sector 1 pairs Site-level traffic with Utilization and CQI_avg, while Sector 2 integrates it with Conn_attempts and PRB_used. This suggests that while some variables remain consistently important, the context in which they yield the best results is sector-specific.

The variability is not just limited to exogenous variables. The ARIMAX model parameters ($p$, $d$, and $q$) themselves also show variability. Some sectors, such as Sector 2 and Sector 8, display a preference for a $p$ value of 1, while others, like Sector 1 and Sector 6, opt for a value of 3. This variation reflects the unique internal dynamics of each sector and how these dynamics influence the model's parameters.

There is a distinct difference in the MAPE values across sectors, indicating varying levels of forecasting accuracy. While Sector 4 achieves a low MAPE of 8.53, Sector 5 records a MAPE of 23.6. This broad spectrum of MAPE values means

that even with optimized model parameters, the relative forecasting error can vary significantly from one sector to another. The insights from Table 5 emphasize that it is vital to recognize the unique characteristics of each sector and adjust model parameters accordingly to achieve the best forecasting results. However, caution must be exercised to avoid overfitting and to ensure generalizability.

Overall, the best combination for the 100 examined sectors prominently featured the use of the original historical data in 78% of the cases when considering the optimal MAE value. This finding speaks directly to one of the core research questions posed in this thesis: does estimating data for congested sectors improve their forecasting accuracy. When we look at the numbers from Table 4, it is clear that using the original historical data in training the forecasting model often gives us better forecasting results.

Comparing the numbers, the best MAE we got with original data was 0.111, while for the estimated data, the lowest was 0.129. This shows that the original data usually has a slightly better forecasting accuracy. Although the estimation method outlined in Section 4.7 has its merits in some scenarios, the overarching finding remains clear: the original data often delivers more accurate predictions for future data traffic, at least for the methods, models and sectors used in this thesis.

# 6 Conclusions

In sectors of a Radio Access Network characterized by limited network infrastructure and a high user demand, congestion often emerges as a prevalent issue. This congestion not only degrades the QoS but also leads to reduced user throughput, both of which pose significant challenges for Mobile Network Operators. In these congested sectors, constraints on data usage result in users not fully tapping into the available data capacity. This constraint causes a distortion in the historical time series of data traffic, making it hard to paint an accurate picture of data growth, primarily due to the influence of congestion.

The central objectives of this thesis were to generate accurate estimations of data traffic demand for cellular network sectors suffering from congestion, and define their reliability. The overarching goal focused on finding the true data traffic demand in sectors where limited network capacity caused suppressed user behaviors. By addressing this challenge, this thesis aims to contribute to the enhancement of both network management and user experience, allowing for a more precise understanding of data usage trends even in congested network segments.

This thesis undertakes a systematic approach to address the research objectives by implementing a Censored Linear Regression model with Autocorrelated Errors of order $p$, denoted as CLR-AR($p$), to estimate censored observations. The focal point of this methodology lies in its ability to effectively handle censored data. This estimation process is pivotal in enabling subsequent analyses and forecasts. After estimating the censored observations, the thesis advances to the task of forecasting. Both the estimated and original censored data are employed to predict future data traffic values, allowing for a comprehensive evaluation of estimation quality.

The forecasting model chosen for this thesis is the ARIMAX model, detailed in Section 4.8. This choice was influenced by the model's balance between sophistication and simplicity. The ARIMAX model effectively captures core data trends without becoming overly intricate, ensuring predictions that are both robust and easy to understand. This is particularly crucial given that our forecasting results are intended for real-world clients, thus the need for transparency and understandability in our forecasts extends beyond mere academic preference.

In addressing our primary research question on whether censored observations of data traffic can be accurately estimated to improve forecast accuracy, it became clear that estimating the data did not universally enhance the forecast accuracy for congested sectors within the cellular network. We employed the CLR-AR($p$) model with the anticipation that these estimated observations would provide superior forecasts. However, when compared against censored observations for future traffic prediction, the original data often proved more reliable.

Our analysis showed that the original historical data generally gave more accurate results. When looking at 100 sectors from the cellular network, the original data provided the best MAE values in 78% of the cases. This means that even though the CLR-AR($p$) model works well in some situations, it often does not outperform predictions made using the original data.

Following the evaluation of the CLR-AR($p$) model, another significant finding

emerged regarding the role of exogenous variables. Their inclusion consistently improved the performance among all tested ARIMAX models. The ARIMAX model, utilizing the original data with parameters $p = 2, d = 2, q = 3$ and the inclusion of Site-level traffic and PRB_used as exogenous variables achieved the best forecasting accuracy with the following metrics: a MSE of 0.024, MAE of 0.111, RMSE of 0.130 and MAPE of 21.02.

Given the data range between 0 and 1, these results highlight the nuances and complexities involved in predicting traffic within cellular networks. While the errors are moderate, they are indicative of the challenges in capturing the intricate patterns and volatilities inherent in the data. One of the challenges was the consistent underprediction in our forecasts. However, this is not a random occurrence. As detailed in Section 1, our data partitioning strategy was deliberate, aiming to assess the accuracy of data traffic demand estimation. Consequently, every sector analyzed encountered infrastructure additions during June 2022 to October 2022. Naturally, infrastructure expansion leads to an increase in traffic, presenting unique challenges for models trained on datasets that lack the context of infrastructure additions.

In conclusion, the study has advanced our understanding of the challenges in estimating data traffic for congested cellular network sectors. By utilizing the CLR-AR($p$) model, we achieved a methodological advancement in estimating censored observations, revealing the latent traffic demand that was previously masked by network limitations. However, when we used these estimates for forecasting, we found that the original, unaltered data often gave better forecast accuracy in a significant majority of the examined sectors. This underscores an important insight: while estimation techniques are pivotal, their direct applicability in forecasting within dynamic environments like cellular networks can be complex. An additional important takeaway from our research is the undeniable value of exogenous variables. Their consistent contribution to enhancing forecast accuracy emphasizes their importance in the complex task of cellular network traffic prediction. To sum up, this research both broadens our methodological toolkit for traffic estimation and forecasting, and underscores the nuanced challenges faced when forecasting in the domain of cellular networks.

Throughout this thesis, we have explored a single method for both estimation and forecasting. Moving forward, a natural progression would involve a comprehensive comparison of diverse methodologies for estimating censored observations and forecasting future values using the estimated time series data. Implementing various imputation algorithms and comparing their influence on forecasting accuracy enables us to pinpoint the optimal estimation method for censored observation estimation. This comparative analysis not only helps us uncover the most effective approaches for estimating censored observations in congested sectors but also enhances our forecasting based on these estimations.

Furthermore, a significant improvement in forecasting accuracy could be achieved by augmenting the training data for the forecasting model with information of infrastructure additions. By introducing binary flags or using exogenous variables that encapsulate details about capacity enhancements, we could provide the model with additional context, potentially improving its predictive power.

Beyond the immediate domain, the exploration of broader applications could be an avenue for future research. Extending these proposed methods to other industries and domains could unveil novel insights from diverse datasets, thereby enhancing the generalizability of our findings. For instance, the dataset illustrated in Figure 5 could serve as an invaluable resource for achieving this objective. Also, addressing missing observations within the data could be studied in the future, further refining the methodologies' robustness and practical utility.

# References

M. Alsabaan, W. Zhuang, and P. Wang. Link layer priority techniques for real-time traffic in CDMA wireless mesh networks. In *2008 IEEE International Conference on Communications*, pages 4133–4137, 2008. doi: 10.1109/ICC.2008.776.

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471360919. URL https://books.google.fi/books?id=Cmm9QgAACAAJ.

Dursun Aydin and Ersin Yilmaz. Censored Nonparametric Time-Series Analysis with Autoregressive Error Models. *Computational Economics*, 58(2):169–202, August 2021. doi: 10.1007/s10614-020-10010-8. URL https://ideas.repec.org/a/kap/compec/v58y2021i2d10.1007_s10614-020-10010-8.html.

D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. Wiley, 1980. ISBN 9780471058564. URL https://books.google.fi/books?id=ALjuAAAAMAAJ.

Akram Bin Sediq, Rainer Schoenen, Halim Yanikomeroglu, and Gamini Senarath. Optimized distributed inter-cell interference coordination (ICIC) scheme using projected subgradient and network flow optimization. *IEEE Transactions on Communications*, 63(1):107–124, 2015. doi: 10.1109/TCOMM.2014.2367020.

Jamal I. Daoud. Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, 949(1):012009, December 2017. doi: 10.1088/1742-6596/949/1/012009. URL https://dx.doi.org/10.1088/1742-6596/949/1/012009.

Ericsson. Ericsson mobility report, 2023. URL https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2023.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20:177:1–177:81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

Dennis R. Helsel. Less than obvious – statistical treatment of data below the detection limit. *Environmental Science and Technology*, 24(12):1766–1774, December 1990. doi: 10.1021/es00082a001.

Sunghyun Hwang and Seungkeun Park. On the effects of resource usage ratio on data rate in LTE systems. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*, pages 78–80, 2017. doi: 10.23919/ICACT.2017.7890060.

M.H. Kutner, C. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. Irwin/McGraw-Hill series in operations and decision sciences. McGraw-Hill/Irwin, 2004. ISBN 9780073014661. URL https://books.google.fi/books?id=3_AlAQAAIAAJ.

Olof Liberg, Mårten Sundberg, Y.-P. Eric Wang, Johan Bergman, Joachim Sachs, and Gustav Wikström. *Chapter 7 - NB-IoT*. Academic Press, Second edition, 2020. ISBN 978-0-08-102902-2. doi: https://doi.org/10.1016/B978-0-08-102902-2.00007-8. URL https://www.sciencedirect.com/science/article/pii/B9780081029022000078.

S. Makridakis, S.C. Wheelwright, and R.J. Hyndman. *Forecasting Methods and Applications, 3rd Ed*. Wiley India Pvt. Limited, 2008. ISBN 9788126518524. URL https://books.google.fi/books?id=nxt0CgAAQBAJ.

M.K.J. McClendon. *Multiple Regression and Causal Analysis*. Waveland Press, 2002. ISBN 9781577662433. URL https://books.google.fi/books?id=kSgFAAAACAAJ.

Syed Tauhidun Nabi, Md. Rashidul Islam, Md. Golam Rabiul Alam, Mohammad Mehedi Hassan, Salman A. AlQahtani, Gianluca Aloi, and Giancarlo Fortino. Deep learning based fusion model for multivariate LTE traffic forecasting and optimized radio parameter estimation. *IEEE Access*, 11:14533–14549, 2023. doi: 10.1109/ACCESS.2023.3242861.

Dorel Mihai Paraschiv, Emilia Titan, Manea Daniela Ioana, Ionescu Crina-Dana, Mihaela Mihai, and Octavian Șerban. The change in e-commerce in the context of the coronavirus pandemic. *Management & Marketing. Challenges for the Knowledge Society*, 17:220–233, 06 2022. doi: 10.2478/mmcks-2022-0012.

Jung Wook Park, Marc G. Genton, and Sujit K. Ghosh. Censored time series analysis with autoregressive moving average models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(1):151–168, 2007. ISSN 03195724. URL http://www.jstor.org/stable/20445244.

Peter M Robinson. Estimation and forecasting for time series containing censored or missing observations. In *Time Series: Proceedings of the International Conference Held at Nottingham University, March 1979*. North-Holland, 1980. ISBN 9780444854186.

Fernanda Lang Schumacher, Christian Galarza Morales, and Victor Lachos. R package 'arcensreg': Fitting univariate censored linear regression model with autoregressive errors. 2016. doi: 10.13140/RG.2.1.3400.9209.

Fernanda Lang Schumacher, Victor Hugo Lachos, and Dipak K. Dey. Censored regression models with autoregressive errors: A likelihood-based perspective. *Canadian Journal of Statistics*, 45, 2017.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010.

S. Sheather. *A Modern Approach to Regression with R*. Springer Texts in Statistics. Springer New York, 2009. ISBN 9780387096070. URL https://books.google.fi/books?id=zS3Jiyxqr98C.

S. Sirotkin. *5G Radio Access Network Architecture: The Dark Side of 5G*. IEEE Press. Wiley, 2020. ISBN 9781119550914. URL https://books.google.fi/books?id=v28LEAAAQBAJ.

Rodney Sousa, Isabel Pereira, Maria Silva, and Brendan McCabe. Censored regression with serially correlated errors: a Bayesian approach, Jan 2023. arXiv preprint; arXiv: 2301.01852.

Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005.

Scott L. Zeger and Ron Brookmeyer. Regression analysis with censored autocorrelated data. *Journal of the American Statistical Association*, 81(395):722–729, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289003.

Erich Zöchmann, Stefan Schwarz, Stefan Pratschner, Lukas Nagel, Martin Lerch, and Markus Rupp. Exploring the physical layer frontiers of cellular uplink. *Eurasip Journal on Wireless Communications and Networking*, 2016, 2015. URL https://api.semanticscholar.org/CorpusID:2280186.