Predicting Patient Waiting Times at an On-Call Emergency Line at a Dutch GP Office

Using ARIMA-family Time Series Forecasting

Katarína Barteková

4914309

Supervisor: Gerard Vreeswijk
Second Reader: Palina Salanevich
External Supervisor: Daan Ooms
External Company: Esculine
Collaboration Team: Yaqin Tian
Number of words: 15 060

Submitted 5th July 2023 in Partial Fulfillment of the
Requirements for the Degree of

Master of Science in **Applied Data Science**

at Utrecht University, Graduate School of Natural Sciences

Abstract

A challenge is posed by the prediction of patient waiting times at an emergency call line operated by a GP office in the Netherlands. The office currently employs a prediction approach based on a Discrete Event Simulation model, which has not been compared to different approaches. Previous studies suggest that ARIMA-family models and LSTM might perform well in predicting patient waiting times. We select the most accurate model among ARIMA, SARIMA and SARIMAX models for this particular case and compare it to the current simulation-based model. We also compare the most accurate ARIMA-family model with LSTM in terms of their prediction accuracy of patient waiting time prediction for this particular case.

The data used in this thesis were provided by an external company and they span from January 2022 to April 2023. The time series is analyzed using an hourly frequency and the predicted outcome is the average patient waiting time per hour measured in minutes. Box-Jenkins method and external forecast validation are applied in modelling and selection of the most accurate model for forecasting. The most accurate among the ARIMA-family models is the SARIMAX model. This model uses exogenous variables: calendar variables and number of incoming calls per hour to model and predict the average patient waiting time per hour. The SARIMAX model with incoming calls is also more accurate than the current simulation model for short- and long-range predictions. However, LSTM has better prediction accuracy for this case than any ARIMA-family model. The implementation of LSTM is recommended for this case. We also provide tentative results regarding the effect of the number of staff available for patient waiting time predictions, and we suggest that it is investigated in greater detail and with more data available in future work.

Keywords: patient waiting time, time series forecasting, emergency line, ARIMA, SARIMA, SARIMAX, LSTM, calendar variables

Table of Contents

1. Introduction

This paper focuses on models used to predict patient waiting times for the specific case of a single GP practice that operates an on-call out-of-office emergency line in the Netherlands. There are two commonly-used and well-performing methods in predicting patient waiting times and patient arrivals: the linear approaches of time series forecasting such as ARIMA (autoregressive integrated moving average)-based methods and non-linear methods such as LSTM (Zhao et al., 2022). However, there is not an academic consensus about which prediction approach is the most efficient for predicting patient waiting times for callers at the emergency lines at GP offices. In this paper, we suggest testing these two different approaches for the case of a single GP practice in the Netherlands that operates the telephonic out-of-office emergency line.

Although the practice already uses a prediction system for patient waiting times provided by the client company, it may not be the most effective one. Currently, the predicted waiting times are determined by a simulation (Discrete Event Simulation) that draws on empirically-determined distributions of arrival rates and service times using calendar information. However, the current system has not been benchmarked against different models and therefore it is possible that the current system is not the most precise one. Academic literature suggests that there might be more accurate and simpler methods available (Sudarshan et al., 2021).

There are several reasons why predicting patient waiting times is important. Recent research has shown that the Emergency departments (EDs) notoriously experience long patient waiting times and overcrowding (Kuo et al., 2020). Overcrowding can happen when there is a failure to account for changing demand of patients, which in turn can lead to decreased quality of care (Bernstein, 2009; McCarthy, 2011). Moreover, patient overcrowding is one of the main causes of staff burnout in ED (Wargon et al., 2009), yet a good approximation of patient waiting time might help the provider with resource management (Kuo et al., 2020). Therefore, it is both in patients' as well as ED providers' interests to obtain a good approximation of how long patients need to wait.

Accurately predicting patient waiting times for our Dutch GP case is crucial, as it is currently experiencing overcrowding and long patient waiting times. In the Dutch context, an equivalent to an ED for milder cases is the out-of-office emergency service on-call at a general practitioner (GP) office, called "huisartsenpost" (Hanneke, 2022). GP offices in the country run emergency lines on-call, which are mostly intended for mild to high urgencies, not for life-threatening conditions. However, many patients with low urgency problems also call (Hanneke, 2022). Therefore, the line is often busy and the staff on-call experiences patient overcrowding in the form of incoming calls. Henceforth, it is important to predict patient waiting times for this case in the Netherlands, as patients experience long waiting times, including those who have a very serious problem.

We pose the main Research Question: Which one of the commonly-used methods – time series forecasting with ARIMA-family models or LSTM – predicts patient waiting times at the single GP on-call emergency line in the Netherlands with higher accuracy? In order to answer this research question, we also formulate two sub-research questions. The first sub research question is: Which time series approach among the ARIMA-family predicts hourly patient waiting times in our case of the single Dutch GP emergency on-call line with the highest accuracy? The second sub-research question is: Does the most accurate ARIMA-family model produce more accurate hourly predictions of patient waiting times than the current simulation-based model used in the client's GP emergency on-call line in the Netherlands?

The aims and scope of this thesis are closely linked to the research questions. Our goal is to compare the accuracy of commonly-used linear time series forecasting methods and LSTM in patient waiting time prediction for a single GP practice operating an emergency line in the Netherlands. We aim to test and compare which ARIMA-family approach predicts hourly patient waiting times at the Dutch GP office with the highest accuracy. We also aim to test if ARIMA-family methods improve the accuracy of waiting times prediction as compared to the accuracy of the current simulation-based model. After diligent comparison, we plan to suggest the most suitable model for further deployment for the single GP practice in our case. In order to do this, we utilize historical operational data for a single GP office in the Netherlands. Thus, the scope of the thesis is limited to the given case that is used to model and fine-tune the methodological approach, but the generalizability is limited and performance might vary from a practitioner office to office. It is also not intended to establish a general comparison of the linear ARIMA-based and non-linear LSTM approaches for patient waiting time prediction.

Focusing on the case data of a single GP office, this paper is divided into several parts. Firstly, we review the relevant literature to propose a suitable theoretical framework. Secondly, we describe a detailed method for ARIMA-based models that is applied to our case. Thirdly, we present the results of the ARIMA-based models and select the best performing one for further comparison. Fourthly, we compare the best linear model with the best non-linear model and suggest one for further deployment. Lastly, we discuss the limitations of this thesis and conclude the work.

2. Background

2.1 Case Description and the Current Model

Our case consists of a Dutch GP office that operates an emergency call line during the out-of-office hours, on which patients experience high waiting times. The line is open between 17:00:00 until 00:00:00 and 00:00:00 until 8:00:00 on the weekdays, and from 00:00:00 until 00:00:00 on the weekends and holidays. These times correspond to the hours when the physical GP office is closed. When the patients call the emergency line at the GP office, they hear a voice-recording through which they can select whether they experience a problem of high or lower urgency. The patients who experience health issues of high urgency enter a separate line to receive faster treatment. On the other hand, all the patients with lower urgencies undergo a triage that further determines the urgency of their issue. Depending on the number of callers in front of them, they are given advice right after their call is picked-up, or they automatically enter a waiting line if the operator is busy. As the number of incoming calls is oftentimes higher than the number of operators, the line oftentimes experiences overcrowding and long patient waiting times. The GP office tries to schedule more staff on the shifts for the times when the waiting times are longer. In order to do this, they need to have accurate predictions of when this happens.

Currently, the GP office uses a simulation model to determine the length of patient waiting times during the times it specifies in the model. The current model was developed by the client company collaborating on this thesis work. This model is based on empirical distributions of arriving calls that are used in a Discrete Event Simulation. It consists of three types of data – the frequency of arriving calls, length of a call and the number of staff on shift available to operate the calls. The distributions of the number of arriving calls are created based on the historical data creating scenarios based on grouping similar observations together. Similar calls in terms of their predictor conditions such as weather, day of the week and time of the day, are grouped together into buckets by averaging their values. After obtaining thousands of scenarios (buckets) with similar distributions, the average Poisson distribution of the number of arriving calls per hour is drawn for the scenario the GP office selects in the system. For example, the GP office might be interested in a scenario for the upcoming Tuesday. The Discrete Event Simulation determines whether a call occurs or not for each minute of the day. If a call occurs, the simulation consults the number of shifts available to operate the call. If the number of calls exceeds the number of operator shifts available, the calls are entered into a queue and all the proceeding calls enter the queue as well, until the distribution for the whole day is created. As the last step, the data on the length of the queue and information on patient waiting time is collected and provided to the GP office. This model is the simulation-based model that is employed for comparison of models proposed in this thesis.

2.2 Related work

2.2.1 Studies from call centers and ED setting

Our case resembles both an ED and a call center, which are common settings that have been studied by past literature. On the one hand, the case is similar to an ED due to the healthcare setting, in which patients with a serious health problem arrive. On the other hand, unlike in a traditional ED, the patients tend to suffer from mostly non-life-threatening health conditions, and they arrive by phone rather than physically, which resembles the situation of a call center. Both of these case situations were modeled in previous literature focusing on the number of arriving patients or calls rather than direct waiting time prediction. Most of the previous studies from call centers focused on developing and testing prediction models of call arrivals (Ibrahim et al., 2016; Rausch et al., 2022) and call volume (Baldon, 2019). Similarly to call centers, most previous studies from the healthcare setting predicted the number of emergency department presentations during a specific time period as emphasized in a literature review (Gul & Celik, 2018). Even though patient or call arrivals have been studied more widely, we still focus on waiting time prediction as this was requested by the client since they are developing a separate prediction tool for patient arrival predictions. We aim to provide the client with two model alternatives, one that uses the arrival number of calls and one that does not. Thus, in this sense, the task of this thesis is slightly different from most previous literature reviewed as it seeks not to predict the number of patient call arrivals, but to predict the patient waiting times, which can be closely related to patient call arrivals.

Past studies vary greatly in the frequency of observations they use and the length of study and test period of their prediction models. To illustrate, Mai and colleagues (2015) who predicted the number of emergency department presentations, focused on a monthly frequency of observed data. Their model was developed over the period of 7 years and used 5 years of data for testing the model (Mai et al., 2015). Champion and colleagues (2007) also forecasted the average ED patient presentations for each month, yet they used an hourly frequency of observations. Their prediction model was trained on data that spanned 5 years and their model was tested on the observations for the first five months of the following year (Champion et al., 2007). Similarly, other studies also used hourly frequency data to forecast ED occupancy (Cheng et al., 2021) and patient arrivals (McCarthy, 2011; Zhang et al., 2022). Other studies focused on daily frequencies of patient arrivals to ED (Abraham et al., 2009; Carvalho-Silva et al., 2018; Tuominen et al., 2022). These studies developed prediction models using training data of 2 years and more and tested their forecasts on new observations that spanned 1 year and more. The choice of the frequency of observations employed depends on the intended use of the model (Champion et al., 2007). Monthly frequencies are usually preferred for large hospitals, while hourly frequencies and forecasts are mostly aimed for shift planning and short-term predictions (Champion et al., 2007), which applies to our case. Also, the data available spans a much shorter period than in previous studies, where it was several years. In our case, the span of data available is a little over a year.

Previous studies from the health setting as well as call centers commonly used meteorological and calendar information to obtain a better estimate of patient or call arrivals (Boyle et al., 2012; Channouf et al., 2007; Jones et al., 2008; McCarthy, 2011; Zhang et al., 2022). Several past studies suggest that calendar and meteorological information improves the prediction of ED arrivals (Jones et al., 2008; McCarthy, 2011; Zhang et al., 2022). The influence of different variables on the prediction accuracy of patient arrivals was investigated by Tuominen and colleagues (2022) using high-dimensional multivariate data. This study found that calendar variables were among the dominant predictive features of forecasted patient arrivals, while meteorological data were not. Similarly, Marcilio and colleagues (2013) found that while calendar variables alone were able to detect patterns of daily variability in the number of patient arrivals to ED, weather information did not improve forecasting accuracy at all. Several previous studies also included patient information next to weather and calendar information (McCarthy, 2011; Rosychuk et al., 2015). However, this is not our case since we do not have access to the patient information nor the content of the call due to its sensitivity. As calendar data seems to improve prediction in similar settings to our case, we focus on its inclusion in our prediction model.

2.2.2 Time series

Waiting time prediction belongs to the time series analysis as it concerns sequential data. Time series is a special type of sequential data that consists of observations made at subsequent points in time for a single or more attributes (Stagge, 2020). Our case requires this type of analysis since the observations of patient waiting times are sequential and might fluctuate over time. The prediction of time series is oftentimes labelled time series forecasting (Choudhury & Urena, 2020), which has been traditionally used in various other fields such as economics and finance (Choudhury & Urena, 2020). Time series forecasting has been widely used to forecast patient arrivals (Choudhury & Urena, 2020; Wargon et al., 2009), but also can be used to forecast patient waiting times (Duarte et al., 2021; Stagge, 2020).

A widely-used theoretical framework used to model waiting times in different time series scenarios such as call centers is Queuing Theory. For a detailed explanation of Queuing Theory, see Ross (2007). The emergency line at the GP office in our case resembles a so-called queuing system. Queuing systems are characterized by customers who are waiting for a limited resource as they cannot access it simultaneously and thus a queue (or line) of the waiting customers is formed (Ross, 2007; Stagge, 2020). In our case, the customers are patients calling and the limited resource is the advice from the healthcare provider. Using the framework proposed by Queuing Theory further for our case, patient waiting time can be modelled through the demand side (number of arriving calls) and the supply side (availability of personnel on a shift at the time to serve the arriving calls). Both of these influence the overall waiting time of a calling patient. Yet, previous studies that employed Queuing Theory use the Poisson arrival process to generate call arrival forecasts, and thus focus on modelling the demand side (Rausch et al., 2022). However, as pointed out by Rausch and colleagues (2022) a key feature of call center and ED arrivals might not align with the Poisson assumption of time independence as the arrivals oftentimes exhibit strong seasonal patterns that repeat themselves in cycles of different lengths. Thus, we investigate the influence of the demand and supply side on the overall waiting time prediction using prediction algorithms that find underlying patterns, rather than Poisson arrivals from Queuing Theory. However, since there was limited data available on the supply side in our case, we focus mostly on the demand side influencing the predicted waiting time while the prediction incorporating the supply side is analyzed using less data to provide only tentative results intended for future analysis. Overall, Queuing Theory is an important framework for our modelling approach and underpins the relevance of previous research that focused on patient arrivals and staff schedules for waiting time prediction.

Previous literature identifies two groups of models used to analyze time series through finding underlying patterns in them: linear and non-linear (de O. Santos Júnior et al., 2019; Gul & Celik, 2018). Linear approaches include time series forecasting methods such as ARIMA and ARIMA-derived models (de O. Santos Júnior et al., 2019). On the other hand, non-linear approaches consist of machine learning approaches and deep-learning approaches such as neural networks (de O. Santos Júnior et al., 2019). These two approaches differ mostly in their assumptions of the underlying patterns that characterize the time series at hand. They are either assumed to be linear for linear approaches or non-linear for non-linear approaches (de O. Santos Júnior et al., 2019; Gul & Celik, 2018). While this thesis focuses on the linear approaches, the non-linear approach adopting LSTM neural network can be found in a parallel thesis work by Tian (2023)[1]. Both of these approaches have been extensively studied by previous literature and were found to be highly effective in predicting patient arrivals at ED (de O. Santos Júnior et al., 2019; Gul & Celik, 2018; Wargon et al., 2009).

Linear time series forecasting models from the ARIMA family have been widely used in previous studies in the health context. Several studies used non-seasonal ARIMA models to forecast emergency department presentations finding high accuracy of the models (Carvalho-Silva et al., 2018; Champion et al., 2007; Channouf et al., 2007). Similarly, Kim and colleagues (2019) developed an ARIMA-based forecasting model for predicting febrile ED visits the next day with high accuracy. Choudhury and Urena (2020) even found that an ARIMA model performed better than a neural network model for their case, exhibiting the highest forecasting accuracy in hourly forecasts for ED arrivals.

However, several studies found that the seasonal ARIMA model (SARIMA) was better suited for the ED patient arrivals and potentially waiting time. These studies found that ED patient arrivals exhibit seasonal patterns and thus SARIMA models should be preferred (Abraham et al., 2009; Cheng et al., 2021; Marcilio et al., 2013; Rosychuk et al., 2015). Kim and colleagues (2020) focused on comparing ARIMA and SARIMA models to predict outpatient visits. They found that SARIMA had superior prediction accuracy over all the time-spans forecasted. Similarly, Jones and colleagues (2008) and Becerra and colleagues (2020) also found that the ED demand patterns showed seasonality and that the SARIMA model exhibited good prediction performance. Moreover, Baldon (2019) found that SARIMA performed even better than neural networks at a daily frequency scale, yet neural networks outperformed SARIMA at an hourly frequency scale (Baldon, 2019). Thus, as suggested by previous literature, seasonality might be dependent on the frequency scale employed. Oftentimes, there are multiple levels of seasonality at play, such as daily and weekly seasonality for hourly data, while SARIMA models do not capture multiple seasonality well (Cheng et al., 2021).

Past studies oftentimes accounted for multiple seasonality and other important characteristics in ARIMA-based approaches using external predictors in their models. A study found that for forecasted daily emergency department arrivals, regression with ARIMA errors (ARIMAX) was the most accurate model integrating external variables such as calendar variables (Tuominen et al., 2022). Similarly, Jones and colleagues (2009) used multivariate time series approach to modeling and forecasting demand in the emergency department finding superior prediction results. These finding are supported by Aboagye-Sarfo and colleagues (2015) who compared multivariate and univariate time series approaches of forecasting emergency department demand. Moreover, Cheng and colleagues (2021) found that the inclusion of higher-level seasonality through the use of external predictors for forecasted hourly ED occupancy improved the predictive performance of their seasonal model SARIMAX.

2.3 Time Series Models: Theory and Motivation

Time series can be studied through examining its underlying system. The system underlying the time series is unknown (Wargon et al., 2009) and our goal is to estimate it as closely as possible to be able to draw predictions for future situations from it. The system is commonly considered to be a sum of

[1] We refer to the parallel thesis work by Yaqin Tian throughout this thesis in the form (Tian, 2023)

shocks that affect the time series, and the variations that are observed correspond to the responses to these shocks (Wargon et al., 2009). A time series can be decomposed in four main patterns: a trend, which is the sloping of the time series; a repetitive pattern over the time called seasonality; cycles, which are fluctuations of the data which are not a real pattern; and a random component known as noise (Balgon, 2019). To approximate the underlying system, a minimum number of observations need to be present for a sound model development. The minimum number depends on the characteristics of the time series and the model, i.e. a seasonal model might require more observations than the non-seasonal model. A commonly-used rule is that the time series should have at least 100 observations for a sound application of ARIMA-family models (Box & Tiao, 1975).

Time series related to ED-data usually estimate patient arrivals as a combination of long-term trends and repeated variations that can be related to seasonality of the data, and the effects of random events (Wargon et al., 2009). An equivalent is expected to hold true for patient waiting times that consist of the number of patient visits and the number of available shifts. Previous literature emphasizes that seasonal variations in ED-data can be driven by multiple seasonal patterns on more levels depending on the frequency of the observations (Wargon et al., 2009). In time series with hourly frequency, the seasonal patterns can be mainly driven by the cycles repeating each day (Wargon et al., 2009). The use of a seasonal model would thus be recommended. Even though several studies found that a seasonal variation of an ARIMA model (SARIMA) performed better than the non-seasonal alternative, we still need to investigate possible seasonality in our time-series as each time series can have different underlying patterns.

2.3.1 ARIMA-family models

The autoregressive integrated moving average time series model (ARIMA) was proposed by Box and Jenkins in 1970 (Daellenbach & Flood, 2002), thus this type of model and the models stemming from it are also known as Box-Jenkins methods (de O. Santos Júnior et al., 2019). This type of model is among the most widely used time series methods (Wargon et al., 2009).

The ARIMA model is characterized by three processes, with three corresponding parameters that determine the order of each process: p, d and q. The order of the autoregressive (AR) process is represented by p. This process treats each value in the time series as a linear function of p past values. This reflects the fact that the variable of interest is regressed on its own previous values (i.e. current values of waiting time depend on past values of waiting time). The integration process (I) relates to the long-term trend of the time series and assumes that the difference between two consecutive values is invariable. This assumption is tested with statistical tests for stationarity and unit root. The degree of differentiation used is represented by d. The usual differencing method consists of subtracting the past value from the current given value when developing the model. In such case, the parameter d of the process is equal to 1. The order of the moving average (MA) process is represented by q. This process assumes that each value depends not only on the error intrinsic to the value, but also on the sum of the errors affecting q past values, indicating that the regression error is a linear combination of past error terms (Vollmer et al., 2021; Wargon et al., 2009). The order of these processes is usually reported with the model in the form ARIMA(p, d, q). For a more detailed explanation see Shumway and Stoffer (2000) and chapter 7 in Wheelwright and colleagues (1998).

The ARIMA model holds several assumptions. The main assumptions are serial dependency of the observations, normality and stationarity (Box & Jenkins, 1976). Stationarity in a time series indicates that a series' statistical attributes, such as mean and variance, are constant over time (i.e., it exhibits low heteroskedasticity). It is commonly assessed by statistical tests for stationarity and graphical methods. In addition, the model assumes that the future observations are a linear function of the past p values, thus it is considered a linear modelling technique (de O. Santos Júnior et al., 2019).

We believe that the previous observations of waiting time might indeed be linearly related to future observations as was suggested by previous research (Wargon et al., 2009). It is also assumed that the model residuals follow a normal distribution with constant variance over time. These assumptions are assessed using statistical tests such as the Ljung-Box test and Goldfeld-Quandt test, and graphical methods. Furthermore, ARIMA models assume that there are no missing variables present in the time series, and that there are regular intervals between observations (Shumway & Stoffer, 2000). The model also assumes that there are no significant outliers present in the time series (Shumway & Stoffer, 2000). These assumptions are necessary to consider when developing our models.

Another assumption of the simple ARIMA model is that the data that it models is non-seasonal (Box & Jenkins, 1976). However, empirical studies in the ED context have also found that there could be some seasonality in the patient arrivals and waiting time present. These studies suggest that the Seasonal ARIMA (SARIMA) models perform better than non-seasonal ARIMA models due to their capability to capture seasonal trends in the data (Jones et al., 2008; Kim et al., 2020).

Seasonal autoregressive integrated moving average (SARIMA) models extend on ARIMA models. They build on ARIMA models by allowing for the incorporation of a repetitive periodic pattern, such as the weekly pattern observed in daily ED patient volumes (Jones et al., 2008). Seasonal ARIMA (SARIMA) models are usually described using the notation SARIMA (p,d,q)(P,D,Q)m. The parameters p,d and q have the same meaning as in the non-seasonal model version. P and Q indicate the order of the seasonal AR, and seasonal MA terms respectively, while D indicates the order of seasonal differencing. The last parameter of the model, m, indicates the order of seasonality. In other words, it indicates after how many observations the seasonal pattern repeats itself (Hyndman & Kostenko, 2007). See Shumway and Stoffer (2000) and Wheelwright and colleagues (1998) for a detailed description of SARIMA models.

Moreover, the ARIMA and SARIMA models are suitable only for univariate time-series modelling. However, in our case, there are several external variables available that could help in the prediction accuracy of the outcome variable of the time series, patient waiting time. This proposition stems from the findings of previous literature that found that the use of predictor variables such as calendar information helped in the prediction accuracy of the model (Tuominen et al., 2022; Whitt & Zhang, 2019). Therefore, we also propose to compare the results obtained from the ARIMA model to the (S)ARIMAX model that accounts for the external predictors.

The ARIMAX model adds additional independent covariates to the ARIMA model in conjunction with the univariate historical signal. This model is also referred to as regression with ARIMA errors or ARIMAX (Tuominen et al., 2022). The model uses independent covariates as exogenous parameters when fitting the model. Therefore, as in regular regression, the covariates need to be normalized to be on a similar scale not to bias the predictions (Shumway & Stoffer, 2000). There is also a seasonal alternative of the ARIMAX model, SARIMAX, that was also found to have a superior predictive performance in the hospital context (Cheng et al., 2021). For a detailed explanation of ARIMAX and SARIMAX models see Shumway and Stoffer (2000) and Chapter 6.6 in Brockwell and Davis (2002).

2.4 Translation of the Case into a Data Science Problem

2.4.1 Forecasting and Validation

In addition to traditional statistical modelling, this study also adopts a forecasting and a validation approach. Predicting future values by a time series model and validating its predictions has been widely used in previous literature. To illustrate, the systematic review by Wargon and collegues (2009) indicates that almost all of the studies reviewed used some form of forecast validation to test the accuracy of the time series model predictions. In our case, the predictions for the average patient waiting time in minutes are generated and validated in an hourly interval. This time interval is the most suitable for the aim and practical application of this thesis (see Appendix II.I.II for more explanation). Thus, the predicted variable is also expressed per hour as the average patient waiting time per hour in minutes. We predict this variable for all patients regardless of which waiting line stream they entered as the waiting times from the two streams are interrelated. In order to compare the proposed models in terms of their accuracy, we need to investigate the forecasts of each model for three time intervals: short-range, mid-range and long-range. These ranges correspond to one day in the future, one week in the future and one month into the future. These time ranges are relevant for the client as they need to generate not only short-, but also long-range predictions for planning purposes. Obtaining predictions of patient waiting time for each model allows us to compare these models with each other.

In our case, it is necessary to investigate two types of SARIMAX models based on the inclusion of the number of incoming calls as an exogenous predictor. The number of incoming calls per hour might be an important predictor of patient waiting times based on the suggestions of their dependence by Queuing Theory. However, this variable is different than other exogenous predictors such as the calendar variables, as it is not known at the time of forecasting. Thus, it needs to be predicted as well. It was indicated by the client that the prediction of the number of incoming calls would be available in the future and it is thus outside of the scope of this research. Yet, we aim to provide two models for the client to have a practical comparison: one with the number of incoming calls each hour and one without, to see if the number of calls improves the prediction quality. This way in case the predicted incoming calls deviate from the real values too greatly, the client company would still have an indication of how well the model without calls performs.

Moreover, we provide a separate analysis for the inclusion of the number of shifts in the model. Since we do not have enough historical data available, this step acts as a preliminary exploration for future work. Although the number of shifts is available for a very short time period with few observations available, it includes important information for the client aiming to schedule the right number of shifts. The results are provided nevertheless, yet they are only tentative aiming to provide a base for future analysis of this variable.

2.4.2 Validation Metrics

There are several well-established validation metrics identified in the previous literature that are used to assess the accuracy of time series prediction models: RMSE, MAE, and MAPE (Wargon et al., 2009). RMSE represents the root mean square error. It is the root mean of the squared differences between the values predicted by the model and the actually observed values (Choudhury & Urena, 2020; Wargon et al., 2009). MAE represents the mean absolute error. The mean absolute error assigns equal weight to large and small errors, unlike RMSE that assigns more weight to large errors (Cerqueira et al., 2017). Moreover, RMSE is almost always larger than MAE, and is equal only in those cases, when all of the errors have the same magnitude (Cerqueira et al., 2017). Low values of RMSE and MAE indicate a high accuracy of the model's predictions as they represent a good match between the predicted and the actual values (Wargon et al., 2009). RMSE and MAE are directly related to the mean value of the outcome variable analyzed and the unit in which they are expressed is the unit of the predicted variable. These error metrics can be potentially less easy to interpret for models that predict variables that have different mean values (Wargon et al., 2009). However, in our case, we are using the same outcome variable in all of the models. Thus, MAE and RMSE are a fitting choice for our case.

On the other hand, the MAPE metric is not suitable for our study. MAPE represents the mean absolute percentage error (Sudarshan et al., 2021). Although MAPE offers a better comparability across models with different outcome variables since it does not depend on the mean value of the predicted variable (Wargon et al., 2009), it is not suitable for our case. In our case, there are observations when the waiting time is zero in those cases when the call line is closed. However, MAPE produces undefined or infinite error estimations in such cases as it cannot operate with zero values or values close to zero (Sudarshan et al., 2021). It can also produce misleading error estimations for different scales of time series.

To obtain these metrics, comparing the actual value with the predicted value is necessary. In order to do this, the dataset needs to be split into a training set and a test set as was done by most of the previous studies (McCarthy, 2011; Wargon et al., 2009). Several studies used the hold-out or sample-splitting method, in which a dataset is split into a train set and a test set (Boyle et al., 2012; Zhang et al., 2022). This means that part of the dataset (the test set) is not used to train the data. This allows for comparison of the predictions with actual data to obtain prediction accuracy (Boyle et al., 2012; Zhang et al., 2022). Moreover, several studies validated their model's predictions on an additional external (out-of-sample) dataset (Boyle et al., 2012). Using an external dataset of data that were not previously "seen" by the model allows us to approximate how the model would perform in the future on new observations.

3. Method

3.1 Time Series Implementation and Model Selection

We adopt a well-established statistical modelling method for ARIMA-family models, the Box-Jenkins method. This method was originally developed by the inventors of ARIMA, George Box and Gwilym Jenkins in the 1970s for the ARIMA model (Daellenbach & Flood, 2002). Through this method, we aim to find the most fitting ARIMA model order (de O. Santos Júnior et al., 2019). The Box-Jenkins method follows a systematic and iterative process that consists of five main steps: data preparation and investigation, model identification, parameter estimation, statistical assumptions checking and forecasting (Daellenbach & Flood, 2002). The Box-Jenkins method emphasizes the iterative nature of time series analysis between the steps and their inter-relatedness, which we adopt in this study. This five-step approach is an expansion on the original approach that did not include the data preparatory and the application phases (thus the process can also be found in its shorter three-step form in some literature) (Wheelwright et al., 1998). We use a smaller portion of the dataset (training set) for model identification, parameter estimation, assumptions checking and forecasting. Over time, several case-relevant ARIMA-extensions have been developed, namely the Seasonal ARIMA, ARIMAX and SARIMAX as mentioned in the Background segment. This method in full length can be used also for these extensions of the original ARIMA model for each model separately except for the preliminary data preparation phase (de O. Santos Júnior et al., 2019). The order of model development we employ is from the simplest to the most complex model, as recommended by Box and Jenkins (Wheelwright et al., 1998). Thus, the order of model development is as follows: ARIMA, SARIMA, SARIMAX without calls as an exogenous predictor and SARIMAX with calls as an exogenous predictor.

In this study, we expand on the traditional statistical Box-Jenkins modelling approach by using an approach that is conventionally used more in machine-learning. After performing the Box-Jenkins approach for each model type we select the most accurate model. Accuracy is assessed by comparing well-established validation metrics MAE and RMSE using the same test dataset for all models developed. Once the most accurate model is selected, we compare this model's accuracy of predictions with the simulation model the client currently uses on a new external validation dataset. The results obtained from the validation on the external dataset are also used for comparison with the LSTM approach (see parallel work by Tian (2023)). See Appendix II.II.I for a schematic illustration of the full method with all steps.

3.1.1 The Box-Jenkins Approach

3.1.1.1 Data preparation and Investigation

We first preprocess and transform the data into the applicable format for ARIMA models. This step involves transformations in order to stabilize the variance of the time series. Transformation of the data such as square roots or logarithms are performed in case there is a need to stabilize the variance in a series, for example if the variation changes with the level of the time series (Daellenbach & Flood, 2002). Moreover, the data is preprocessed, variables are normalized, and outliers are identified and handled. For preprocessing description, please refer to section 3.2 Preprocessing. Moreover, the dataset available is split into a training and a test set for validation using the hold-out method (described in greater detail in section 3.1.1.5 Forecasting and Model Validation). The training set consists of the whole dataset except for the last month of observations. The observations from the last month of the dataset are used as a test set in a later step (see 3.1.1.5: Forecasting and Model Validation).

Secondly, we investigate the need for data differencing and the order of differencing using graphical tools and statistical tests. Graphs such as autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are used to identify the presence of autocorrelation and seasonal patterns (Daellenbach & Flood, 2002). The ACF plot shows the correlation between the series and its lagged values, while the PACF plot represents the correlation between the series and its lagged values after removing the effects of shorter lags. Statistical tests such as the Augmented Dickey-Fuller (ADF) test for unit root, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for stationarity and the Phillips-Perron (PP) test are used to check for stationarity and unit root (Daellenbach & Flood, 2002). We also use the Canova-Hansen test to determine the stability of seasonal patterns and the possible need for seasonal differencing, if any are found (Busetti & Harvey, 2003). The results of these tests are applied later when differencing is performed in the model. Unlike the other steps in the Box-Jenkins approach, this step is performed only once.

3.1.1.2 Model Identification

We identify the model by determining the model’s order. We use Akaike’s Information Criterion (AIC) for this step. AIC is a device for estimating the predictive accuracy of models with lower AIC values indicating a better-fit model (Forster & Sober, 2011). We use AIC to help us decide between multiple model configurations of the same model type as AIC can only be used to compare models with the same outcome and predictor variables (Forster & Sober, 2011). To illustrate, AIC is used to identify parameters among ARIMA models, but cannot be used to compare the model fit with a SARIMA model because of different parameters used in this model. We use the automatic order search to find the most fitting configuration of the autoregressive, moving average and differencing order for the model based on the lowest AIC using the step-wise algorithm search. In case of seasonal models, the same process is used to also find the most fitting configurations of the seasonal autoregressive, seasonal moving average and seasonal differencing order of the model. This approach selects from among multiple differencing options to induce difference-stationarity. For determining the differencing orders that should be tested, we use the information obtained in the previous step.

3.1.1.3 Parameter estimation

After the appropriate model order is determined, the values of the model coefficients for each parameter are estimated. This step is performed algorithmically by fitting the model identified in the previous step to our time series. We estimate the coefficient values for each model order so that obtain the closest fit to our normalized data. In case of the SARIMAX model, coefficients are estimated also for each exogenous predictor. The estimation is performed using the maximum likelihood estimation (MLE).

3.1.1.4 Statistical Model Assumptions checking

After the model parameters are estimated, we assess the adequacy of the chosen model. This step involves diagnostic checking of the residuals which should be normally distributed, with zero mean and constant variance to fulfill the underlying assumptions of the ARIMA model. Alternatively, we identify any areas where the model is inadequate using graphical tools and statistical tests. We employ the Ljung-Box test to check for the independence of residuals and the Goldfeld-Quandt heteroskedasticity test to test for constant variance of the residuals (Brockwell & Davis, 2002).

3.1.1.5 Forecasting and Model Validation

Once all the previous steps are performed, we use the model to forecast future values and the forecasts are checked against actual values. The dataset has been split into a training and test set in the first part of the process (3.1.1.1 Data preparation and Investigation). We generate forecasts based on the model obtained from the previous steps using the training data. We generate predictions for three time-frames: one day in advance, one week in advance, and one month in advance. These three time-frames are selected to compare the short-, mid-, and long-term predictive power of the model. In our case, we predict the average waiting time on an hourly basis and we predict the average waiting time in minutes. The predictions are obtained in a normalized form, so we de-normalize them. For each time-frame, we compare the forecasted de-normalized values from the model with the actual values from the test set. The test set is the same for all ARIMA-family models. We evaluate the model in terms of the commonly used validation metrics MAE and RMSE (Wargon et al., 2009), which are explained in the Background section. These metrics can be interpreted in the unit of mean patient waiting time per hour in minutes. This is the most important step in our case and provides the results of model accuracy for the following step (see 3.1.2 Model Selection).

3.1.2 Model Selection based on Forecast Accuracy

Among all the models from ARIMA-family tested, we select the model with the highest accuracy. For accuracy assessment, the MAE and RMSE values obtained from the previous step are used. All of the models have been validated using the same test dataset making the MAE and RMSE values comparable among models of different types. We select the ARIMA-family model with the lowest MAE and RMSE values for the most time-horizons. We also select the most accurate model that predicts waiting time without the inclusion of the exogenous variable number of calls for practical purposes of the client. In this step, we attempt to answer the first sub-research question.

3.1.3 External validation and Final Comparison with the Currently-Used Simulation Model and LSTM

We generate and check predictions for a longer time horizon using the models we selected in the previous step. We re-train the models on all the data available to match the time-horizon of the external validation data. The selected models generate predictions for a further time horizon, which is equivalent to the range of the external validation dataset (one month: May 2023). These predictions are then validated on an external dataset that consists of actual values in this time-frame. We compare the values predicted by the most accurate model and the actual values from the external dataset using MAE and RMSE error metrics. The two models we compare our results to are the simulation model that is currently in use and LSTM. These models used the same external validation dataset to obtain their MAE and RMSE values. We compare the accuracy of all of the models using the two metrics. In this step, we answer the second sub-research question by comparing the selected model with the current model, and the main research question by comparing the selected model with LSTM.

3.2 Preprocessing

In the preprocessing phase, we perform several steps to prepare the data for analysis and forecasting. For the final external validation dataset, all of the following steps are repeated with the exception of outlier removal and missing data imputation. Details on data preprocessing can be found in Appendix II.II.II. The most important preprocessing steps are outlined below.

Firstly, we handle missing data for those variables that are used later and account for the opening times of the call line. To approximate the waiting time for missing observations, we calculate the difference between the recorded start time and answer time of the call. The difference between these two times is a good approximation of how long the person had to wait on the line stemming from the definition of these two variables (start time of the call and the time when the call was answered). However, the recorded waiting time does not include a delay caused by a tape played at the beginning of each call. To account for this delay in the imputed values, we subtract the mean delay (calculated as the difference between the differenced time and the recorded waiting time) from the differenced time. Moreover, due to the GP office's emergency call line schedule, there are periods when the line was closed, resulting in irregular intervals between observations, as some hours were not recorded due to the line being closed. Since ARIMA-based models cannot handle missing values and nor can they handle irregular time between observations (Helfenstein, 1996), we add all hours of the day to the dataset for the times the line was closed. We also assign 0 to the waiting time and number of calls incoming during these closed periods to indicate that no calls were recorded. This way, we obtain regular intervals between observations without any missing values. Additionally, we create a dummy predictor variable called "open" that takes on the value 1 when the line was open and 0 when the line was closed.

Secondly, we handle outliers as ARIMA-based models can be sensitive to outliers (Hyndman & Athanasopoulos, 2018; Ledolter, 1989). We remove all observations whose call duration was 1 second or less, as these were likely to be mistakes by the callers or system errors. Moreover, we also identify the outliers in terms of waiting time using the standard interquartile range (IQR) method (Tukey, 1977). Observations whose waiting time is more than 1.5 times the IQR below the first quartile (Q1) or above the third quartile (Q3) are considered outliers and removed from the dataset. This procedure also removes any suspicious values that could have been introduced in the imputation as these are taken care of as outliers.

Thirdly, we preprocess the data by transforming them into a more suitable format. In our case, we have time series data, thus, we transform them into a datetime format using the date and time when a call arrived as an index. Furthermore, we calculate the number of calls per hour by summing all calls whose recorded start time fell within a particular hour. Similarly, the average waiting time per hour is calculated as the mean of the waiting times (in minutes) for all calls within a particular hour. Also, we calculate the number of shifts available at the start of each hour to create the variable Number of shifts (Tian, 2023). Moreover, to ensure accurate modeling with ARIMA-based methods (Gamakumara et al., 2023), we normalize the average waiting time per hour, number of incoming calls per hour and number of shifts available using z-score normalization (Panigrahi & Behera, 2013).

Fourthly, we enrich the data with calendar information and recode all the categorical variables into dummy variables. To incorporate valuable calendar information, we enrich the data by adding calendar variables such as the day of the week, season, and holiday information based on public holidays in the Netherlands. Holiday information is synthesized into the day of the week, indicating a special value for any day that was a public holiday, regardless of which weekday it was exactly. Moreover, as these are categorical variables, we recode them into dummy variables for further analysis.

3.3 Variables and Datasets Analyzed

We use the average waiting time per hour as our outcome variable for all models tested using the same test dataset. This variable is recorded in minutes and is de-normalized after obtaining the predictions to get the actual values of average waiting time per hour. For the multivariate SARIMAX models, we use the following variables as exogenous predictors in a dummy form: open (indicating if the call line was open), day of the week (including the information on holiday), and season. We also investigate the effect of the number of incoming calls per hour, using it as an extra predictor to see if it leads to improvements in prediction accuracy in a separate SARIMAX model. However, since the number of calls might not always be available to the client at the time of forecasting, as this parameter requires being predicted as well, we also select the best model when this parameter is not included. The same dataset is used for training all models except for those using number of shifts. The test dataset for model selection is the same across all models.

Due to the short span of the historical data on number of shifts available, we provide only an exploratory analysis of the impact of the number of shifts on waiting time prediction. We train four different model combinations: one without the inclusion of the number of calls and number of shifts; one with the inclusion of number of shifts yet without the number of incoming calls; one with the inclusion of number of calls yet without the number of shifts; and one with the inclusion of the number of incoming calls and the number of shifts. We compare these models in terms of their prediction accuracy using MAE and RMSE as error metrics. For training of these models, we use the shorter dataset filtered from the point for when the number of shifts were available. The analysis of this dataset provides a tentative analysis of the impact of the number of shifts on the waiting time prediction accuracy.

4. Data Description

4.1 Data Provided

The available data includes information on the incoming calls and scheduled shifts. The data concerning calls include historical data spanning from 20 January 2022 to 17 April 2023. However, the historical data concerning scheduled shifts span 1 March 2023 to 17 April 2023, only a month and a half. This dataset is also considerably shorter. The data on shift schedules has 1 427 observations in schedules with no missing observations, while the main dataset contains 10 873 observations on an hourly basis. See Appendix II.II for a detailed description of the raw data available.

The main dataset used for our analysis includes only the data concerning calls and additional calendar information about the days when the calls occurred. The average waiting time per call is 12.73 minutes (SD: 14.9). We imputed the missing values in the variable waiting time, which had 2 786 missing values, using the method described in the previous section. Moreover, as stated in the Methods section, we removed 314 call observations that were equal to or shorter than 1 second. We also removed the outliers using the IQR method, which resulted in additional 2001 observations removed.

The shorter, filtered, dataset used for the tentative analysis of models with number of shifts includes data on calls, calendar variables and number of shifts from their overlapping period. This dataset thus includes only the most recent data spanning from 1 March to 17 April 2023, when all of the variables were available. In this dataset, the shifts last 6.6 hours on average (SD: 1.28). Similarly to the main dataset, the time series in this dataset are reported in an hourly format. See Appendix II.III for a more detailed data description.

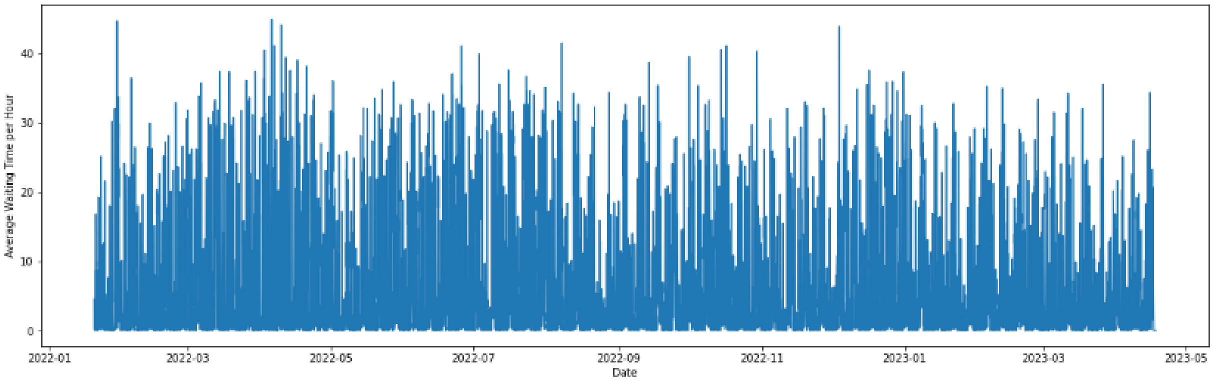


Figure 1: The average patient waiting time per hour in minutes

There is considerable variation in the observed average waiting time per hour. Figure 1 depicts the average waiting time per hour in minutes over the time. The average of the mean waiting time per hour, which we use as our outcome variable, is 5.67 minutes (SD: 8.25), showing quite a big variation. The variation becomes even more apparent in Figure 2 (left part) that clearly shows how the average waiting time changes per weekday and Figure 2 (right part) that shows differences on an hourly level. The waiting times seem to be larger during the weekend and holidays. Also, they seem to be much lower in early morning (from 1 am to 6 am).

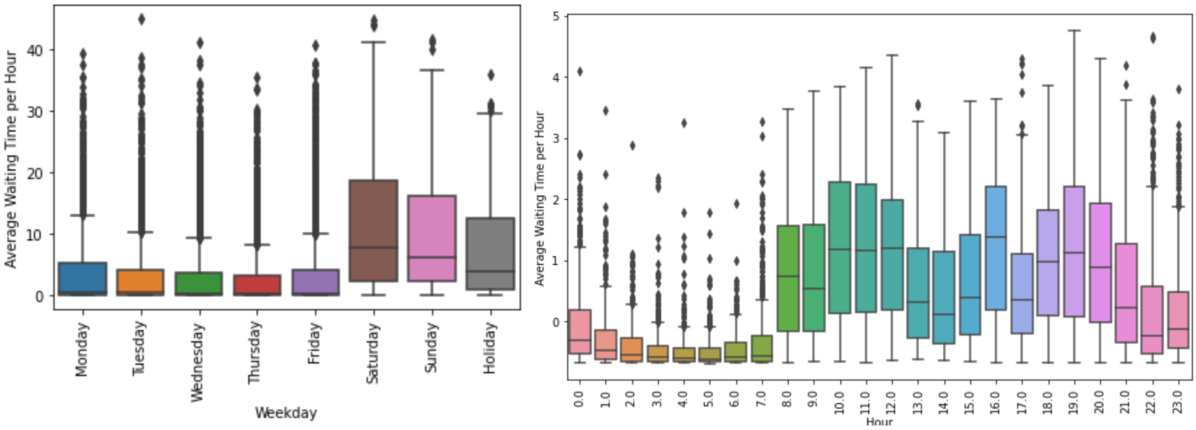


Figure 2: Left part: Average waiting time per hour on each weekday (in minutes), Right part: Average waiting time per hour each hour (in minutes)

4.2 Seasonality and Stationarity

Informed by the first step of the Box-Jenkins method, possible seasonality in the data and stationarity of the data is assessed (see 3.1.1.1 Data preparation and Investigation). ADF and PP tests indicate that the data is stationary with stable mean and variance over time. This indicates that no differencing is necessary. However, KPSS indicates that the data is currently not fully stationary and would require differencing in the model by 1. The plotted time series appears to be relatively stationary with no visible trend in the data (Figure 1). Overall, the tests suggest that the differencing order in the ARIMA-family models should be either 0 or 1.

Moreover, we suspect that there might be daily seasonality present in our data since it was recorded on an hourly basis. As visible from the Autocorrelation and Partial Autocorrelation plots (Figure 3), our data show seasonality on a daily level since a pattern repeats itself every 24 hours. Moreover, based on the Canova-Hensen test, the time series might require some seasonal differencing and that the seasonal pattern might not be fully stable in seasonal cycles of m=24. The test suggests that the seasonal differencing term in the models might be either 0 or 1. For more stationarity and seasonality assessment, see Appendix II.III.

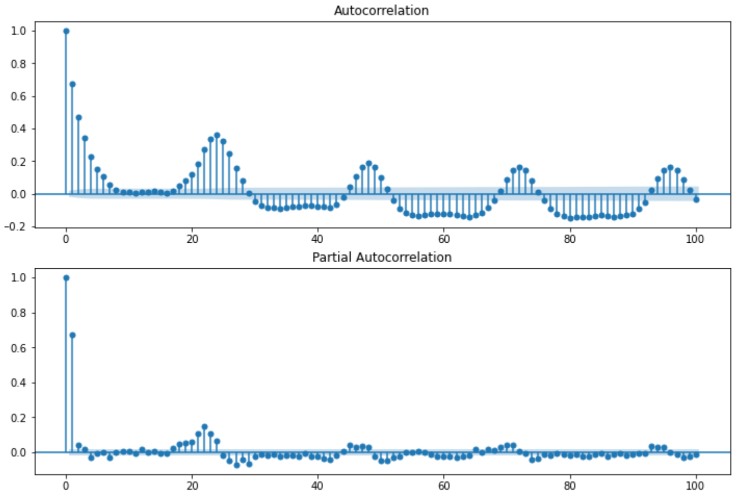


Figure 3: Top part: Autocorrelation plot, Bottom part: Partial Autocorrelation plot

5. Results

5.1 ARIMA model

This section reports the results of the non-seasonal ARIMA model. According to the suggested differencing order from the first step of the Box-Jenkins method (see 3.1.1.1 Data preparation and Investigation), the results of non-seasonal ARIMA with two options of differencing: 0 and 1 are reported for comparison. The comparison between the model with no differencing and differencing 1 is available in the Appendix III.I. The best model identified was ARIMA (1,0,1) with AIC of 23040.251 based on the second step of Box-Jenkins method (see 3.1.2 Model Identification). The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are available in Appendix III.I.I. However, based on statistical assumptions checking (see 3.1.1.4), the model with the lowest AIC met the statistical assumptions only partially. Although the residuals of the model were independent, they were not normally distributed with constant variance. More details about the assumptions assessment can be found in Appendix III.I.I.

The results from forecasting and model validation (see 3.1.1.5) are available in Table 1 and the forecasts are visualized in Figure 4. As demonstrated in Figure 4, the ARIMA model mostly predicted the average value over the month's horizon. As indicated in Table 1, MAE and RMSE were the lowest for the short-range predictions and the highest for a weekly prediction. The long-range predictions had relatively low MAE and RMSE scores.

	MAE	RMSE
1 day into future	5.14843	5.31148
1 week into future	5.21820	6.34346
1 month into future	4.92958	5.90687

Table 1: Forecast Accuracy of the ARIMA model

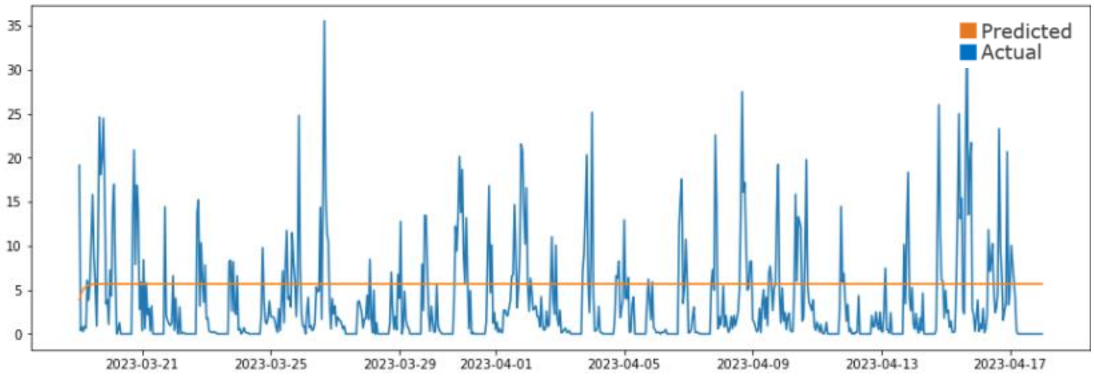


Figure 4: ARIMA model forecasts of the average patient waiting time per hour in minutes one month into the future

5.2 SARIMA model

The seasonal ARIMA model had the following results. According to the suggested differencing order from the first step of the Box-Jenkins method, the results of the two SARIMA model alternatives (normal differencing in the range 0-1, and two options of seasonal differencing: 0 and 1) are provided in Appendix III.II. The model identified in the second step based on the lowest AIC was SARIMA(0,1,2)(0,0,2) 24 with AIC of 23582.147. The estimated coefficients obtained in the third step for each model parameter are available in Appendix III.II. According to the fourth step, statistical assumptions checking, the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. More details about the assumptions assessment can be found in Appendix III.II.I.

The results from forecasting and model validation (fifth step of Box-Jenkins method) are available in Table 2 and the forecasts are visualized in Figure 5. This model had the lowest MAE and RMSE one day into the future. It had the highest errors one week into future and relatively low error scores one month into the future. As depicted in Figure 5, this model was able to capture some variation in the data rather than predicting just an average value of mean waiting time per hour.

	MAE	RMSE
1 day into future	5.33618	5.56938
1 week into future	5.46588	6.49624
1 month into future	5.10722	5.99752

Table 2: Forecast Accuracy of the SARIMA model

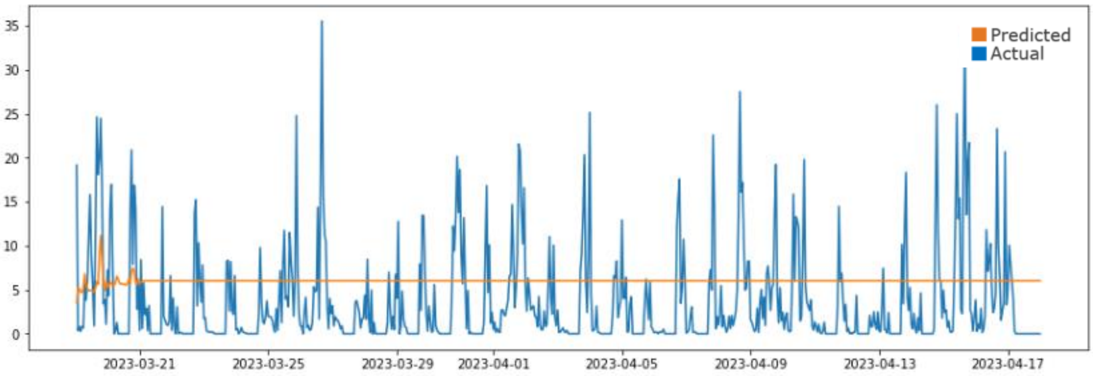


Figure 5: SARIMA model forecasts of the average patient waiting time per hour in minutes one month into the future

5.3 SARIMAX models with and without the number of calls per hour

This section reports the results of the two SARIMAX models used: a SARIMAX model without calls and a SARIMAX model with calls as a predictor. The results of the models modelled with automatically-selected normal differencing in the range of 0-1 and with the automatically-selected seasonal differencing in range of 0-1 are provided based on the information obtained from the first step of the Box-Jenkins method.

The results of the SARIMAX model without the number of calls as a predictor are presented. The SARIMAX model without calls with the lowest AIC that was identified was SARIMAX(0, 1, 5)(0, 0, 2)24 with AIC 21973.375 during the second step of Box-Jenkins approach. The coefficients estimated for each parameter for the third step are reported in Appendix III.III. Out of these coefficients, the external parameter "open" had the highest coefficient indicating the highest influence of the prediction, while the parameters indicating day of the week also had considerable influence. Based on the fourth step of Box-Jenkins method, this model partially fulfilled the statistical assumptions tested. The residuals of the model were independent, yet they did not have a constant variance. For detailed assumptions assessment, see Appendix III.III.I. The SARIMAX model without calls had the lowest prediction error of its forecasts on the short-range prediction (as seen in Table 3) (fifth step of the Box-Jenkins method). The predictions for mid-range horizons had the second lowest error scores. Yet, the long-range predictions had the highest error scores. The plotted predictions of this model can be found in Figure 6 (top part).

The results of the SARIMAX model with the number of calls as a predictor are presented. The SARIMAX model with calls that was identified in the second methodical step as having the lowest AIC was SARIMAX (1,1,1)(2, 0, 0)24 with the AIC of 21373.795. The coefficients estimated for each parameter in the third step are reported in Appendix III.III.II. Out of these coefficients, the external parameters number of calls per hour and "open" had the highest coefficients indicating the highest influence of the prediction, while the calendar parameters were not significant. This model partially fulfilled the statistical assumptions tested as informed by the fourth step of Box-Jenkins method. The residuals of the model were independent, yet they did not have a constant variance. For detailed assumptions assessment, see Appendix III.III.II. This model had the lowest error scores on the short-range prediction (as seen in Table 3) (fifth step of the Box-Jenkins method). The error scores were higher on mid-range predictions and they were the highest for long-range predictions. As seen in Table 3, this model had lower error scores for each forecast horizon than the SARIMAX model without calls. As visible in Figure 6 (bottom part), the plotted predictions of this model followed the actual values of the time series more closely than for the model without calls.

	SARIMAX model without the number of calls			SARIMAX model with the number of calls	
	MAE	RMSE		MAE	RMSE
1 day into future	2.95589	3.46144		2.38199	2.99895
1 week into future	3.85744	5.37179		3.25313	4.90553
1 month into future	4.13237	5.37641		3.56596	4.98470

Table 3: Forecast Accuracy of the SARIMAX model without the inclusion of the number of calls, and the SARIMAX model with the inclusion of the number of calls as an exogenous predictor

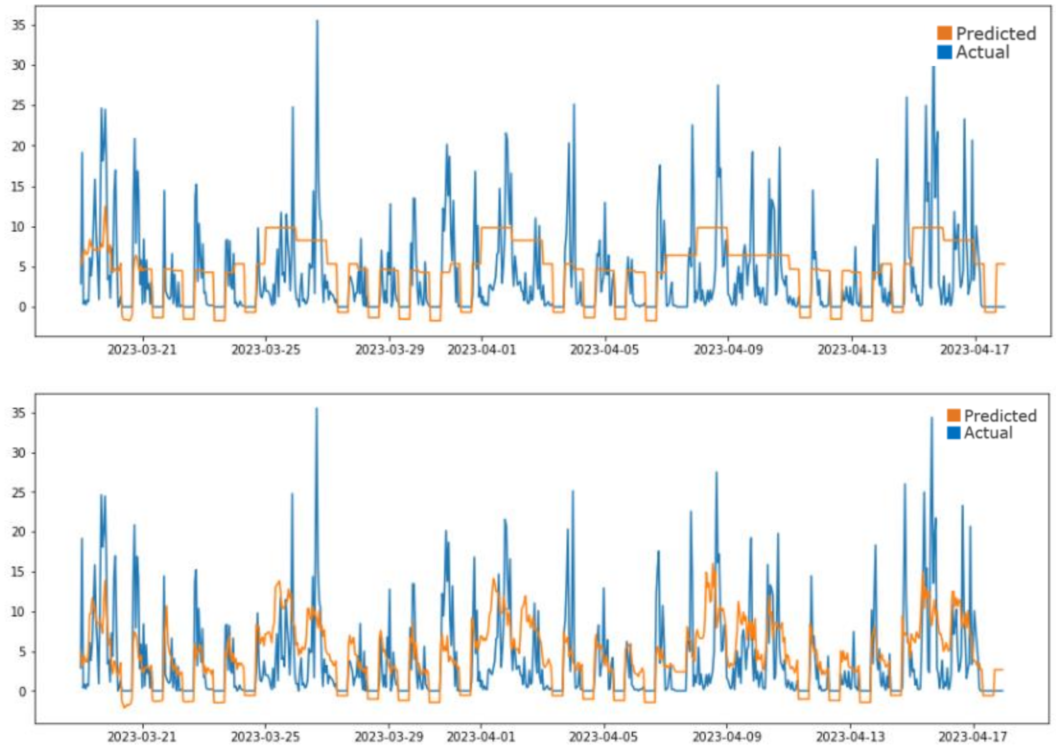


Figure 6: Top part: no calls used as predictor, Bottom part: calls included as a predictor, 30 day prediction

5.4 Tentative Analysis of the SARIMAX Model on a Shorter Dataset

The results of the tentative analysis using the shorter dataset filtered by the availability of the number of shifts are presented. The results of the individual models are available in Appendix III.IV. The model with the highest accuracy was the SARIMAX model without shifts and with the inclusion of the number of incoming calls. However, when the number of calls were not included as an exogenous predictor, the SARIMAX model with the number of shifts and without calls had higher accuracy than the SARIMAX model without both shifts and calls. The most accurate model, the SARIMAX model with the number of incoming calls, had the following form: SARIMAX(0,1,0)(0,0,0)24. It did not identify any seasonal autoregressive, differencing nor moving-average processes. The estimated parameters and statistical assumptions testing are available in Appendix III.IV.III. As can be seen in the Appendix, this model did not fulfill the statistical assumptions well. In fact, it was significantly worse than for all the other ARIMA-family models. When validated, its prediction accuracy was generally lower than most of the other models tested on the same test dataset. This model had the highest accuracy of predictions for the short- and long-range predictions. For more information, see Appendix III.IV.III.

	MAE	RMSE
1 day into future	5.05879	5.49077
1 week into future	5.09789	6.41966
1 month into future	4.97602	5.97360

Table 4: Results of the Most Accurate Model from Tentative Analysis: SARIMAX model with calls trained on a shorter dataset

5.5 Selection of the Best Model

The model with the lowest error scores and the highest accuracy was the SARIMAX model with calls. On the same test set, this model had the lowest error scores on each time horizon for both error metrics compared to the other models (see Tables 1-4). This model performed the best in the short-range predictions and its performance decreased over time. The errors however remained quite similar, indicating a relatively stable quality of predictions over time. From the visual representation of the predictions and the actual values in the test set (Figure 6), it can be seen that this model had more nuanced predictions than the model without calls. Its forecasts also followed the actual time series more closely than those of any other model analyzed (see Figure 6 compared to Figure 4 and 5).

The second best performing model was the SARIMAX model without calls. This model also had relatively low error scores for all time horizons. Overall, its error scores for both error metrics indicated better prediction accuracy than for any other ARIMA-family model except for the SARIMAX model that included the number of calls per hour (see Tables 1-4). When compared to the plotted forecast of other models, the SARIMAX model without calls had predictions that followed the actual values the second most closely after the best performing model (see Figure 6 compared to Figure 4 and 5). See Appendix III.V for a graphical comparison of model accuracies of all ARIMA-family models tested.

5.6 External Validation and Comparison with the Current Model and LSTM

The results of validation of the SARIMAX model with calls as the best model and the SARIMAX model with calls as the best model (in case number of calls are not available) on an external dataset are reported. The same error metrics were used for the three previously employed time-horizons. The results of these models are reported in Table 5. The predictions and the actual values of the external validation dataset are visualized in Figure 7. Table 5 shows that the performance for the selected model without calls on all time horizons was better than on training data.

	SARIMAX model without the number of calls			SARIMAX model with the number of calls	
	MAE	RMSE		MAE	RMSE
1 day into future	2.95054	3.17029		1.59699	1.80255
1 week into future	3.65417	4.76957		3.22853	4.50148
1 month into future	3.95944	5.36151		3.44011	5.08273

Table 5: Performance of the best-performing time series models

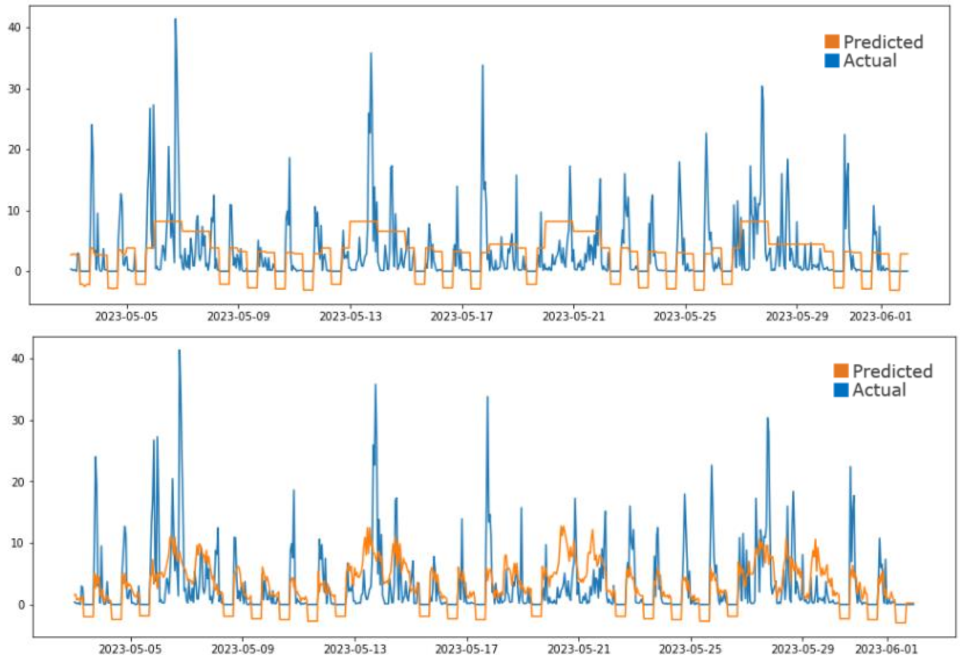


Figure 7: Top: SARIMAX without calls, Bottom: SARIMAX with calls, 30 day prediction on external validation data for best performing time series models

Moreover, the results of the current simulation-based model are reported. The current simulation model had the accuracy as depicted in Table 6. This model had the lowest error scores on the short-term predictions. The error scores increased on longer ranges for both weekly and monthly predictions. The RMSE was especially high for long-range predictions.

	MAE	RMSE
1 day into future	2.42390	4.39920
1 week into future	2.60046	4.60218
1 month into future	4.98939	9.73468

Table 6: Forecast Accuracy of the current model

The comparison between the two best SARIMAX models and the current simulation model showed mostly higher accuracy of the SARIMAX models. The SARIMAX model with calls, the best ARIMA-family model, had lower error scores on both RMSE and MAE metrics for short-range and long-range predictions than the current model (see Table 5 and Table 6). For mid-range predictions, the MAE of the current simulation model was lower (see Table 5 and Table 6), while the RMSE of the SARIMAX model with calls was lower. The SARIMAX model without calls, the second best ARIMA-family model, had lower RMSE and higher MAE on short-range prediction than the current simulation model (see Table 5 and Table 6). On mid-range predictions, the current model had lower error scores. The SARIMAX model without calls also had lower error scores for both MAE and RMSE for long-range predictions than the current model.

Lastly, the comparison between LSTM and the best SARIMAX model showed that LSTM had better accuracy. LSTM model with calls was the most accurate LSTM model in parallel work (see Tian, 2023). The results of the LSTM model with the highest accuracy are reported in Table 7. The LSTM model had lower error scores for both error metrics than both SARIMAX models selected on all time horizons (see Table 7 and Table 5).

	MAE	RMSE
1 day into future	1.17055	1.59194
1 week into future	2.28213	3.730
1 month into future	2.15150	3.68457

Table 7: Forecast Accuracy of the most accurate LSTM model: LSTM with calls (adopted from parallel work by Tian, 2023)

6. Discussion

6.1 Comparison to previous research

In our study, we developed the ARIMA-family models from the simplest non-seasonal univariate ARIMA model, through a seasonal univariate SARIMA model, to a complex seasonal ARIMA model with exogenous predictors. We discuss our results in light of previous literature, the current simulation model, and the LSTM model.

6.1.1 Seasonality

Consistent with some previous research, our data showed clear seasonality on a daily level. Seasonality in our time series was caused by the fact that the frequency of the observations was recorded in hourly intervals that were repeated each 24 hours. Although in our case, the seasonal ARIMA ended up having very comparable predictive performance to the non-seasonal model, the statistical tests as well as graphical inspection suggested a strong seasonal pattern. Moreover, the comparison of the graphs plotting the forecasts of the non-seasonal model and the seasonal model forecasts showed more nuance and closer fit to the actual values in the predictions of the seasonal model. Our findings are supported by some previous literature that also found the presence of seasonality in data (Channouf et al., 2007; Cheng et al., 2021; Kim et al., 2020). However, previous studies used different time frequencies and seasonality levels of time series such as hourly (Channouf et al., 2007; Cheng et al., 2021), daily (Champion et al., 2007; Kim et al., 2020) and weekly frequencies (Channouf et al., 2007) to study ED time series data. Accordingly, not all of the studies found seasonality in the ED data they analyzed (Channouf et al., 2007). The most similar study to our case in terms of observations frequency was the study by Cheng and colleagues (2021). This study focused on hourly forecasts of ED occupancy and similarly to our study results, it found a 24-hour seasonality. This study also incorporated higher-level seasonality cycles such as day of the week as external regressors just as we did in our SARIMAX models. However, in our case, the emergency call line was open only in certain times of the day during the weekdays, potentially adding an extra weekly seasonal pattern, which is not the case for a standard ED. Arguably, the variable “open” and day of the week could have captured some of the seasonal information on such weekly level, as they helped the SARIMAX model to distinguish between different days of the week. Thus, the weekly seasonal pattern was most likely captured in the external predictors to some extent.

6.1.2 Forecast Accuracy and Its Development Over Time for ARIMA-family models

For all ARIMA-family models, accuracy was assessed through the use of MAE and RMSE. MAE was lower than RMSE for all of the models developed. This was due to RMSE being more sensitive to extreme values (Chicco et al., 2021). Moreover, MAE was more important for assessing the magnitude of errors, while RMSE was good for assessing the direction of the errors (Chicco et al., 2021). In our case, MAE and RMSE tended to be quite similar which indicates that the errors made by the model were not too large in their magnitude and that the predictions were generally oriented in the right direction. The average of the mean waiting time per hour which we used as our outcome variable was 5.673 minutes

(SD: 8.248), while our MAE and RMSE for the two most accurate models were relatively low compared to the mean value. This indicates that overall, the selected SARIMAX models performed quite well. However, the best model SARIMAX with calls (as seen from the graph) underestimated significant fluctuations in waiting time and in other cases, it overestimated low waiting times. Overall, the model tended to underestimate more than overestimate, which was also pointed out by Aboagye-Sarfo and colleagues (2015) in the context of forecasting ED demand with ARIMA models.

The accuracy of different ARIMA-family models varied depending on the time range of their predictions. While the prediction accuracy of our two best models (SARIMAX without calls and with calls) steadily decreased over time, this was not the case for the less accurate models, ARIMA and SARIMA. Both of these models had higher prediction accuracy for long-range predictions than mid-range ones. The relatively low prediction errors for the long-range predictions in case of ARIMA and SARIMA were most likely due to the averaging effect. In other words, the increasing number of observations considered in the long-range horizon compared to the mid-range horizon caused the errors to become relatively lower although the predictions remained the same. This was visible in the plotted forecasts of these two models that deviated majorly from actual values without capturing any significant trend and only predicting the average value. On the other hand, a decrease in prediction accuracy over all time horizons would have been expected.

This further demonstrates the reason why the SARIMAX models were superior to the other models. The decrease in prediction accuracy was to be expected as the model predicts patterns in the future about which it has less information and more uncertainty regarding the behavior of the time series over time. Generally, the uncertainty increases with time. Previous studies also found a steady decrease in prediction accuracy over time (Aboagye-Sarfo et al., 2015; Boyle et al., 2012; Tuominen et al., 2022). In line with these studies, the SARIMAX models also showed a gradual decline in forecast accuracy, although the performance was relatively stable with the MAE and RMSE that did not change more than two minutes from one time-range to another.

In summary, as an answer to the first sub-research question, we found that the SARIMAX model that included the number of incoming calls per hour was the most accurate among the ARIMA-family models for our case of a single GP office that operates an emergency call line in the Netherlands.

6.1.3 External predictors

There was a stark improvement of prediction accuracy when the external predictors were introduced in the model. This demonstrates that the inclusion of external variables was indeed valuable for time-series modelling in our case. Similarly to previous research (Jones et al., 2008; McCarthy, 2011; Tuominen et al., 2022; Zhang et al., 2022), we found that the inclusion of calendar variables, namely the weekday indication, improved waiting time prediction. However, the distinction is that previous research used calendar variables mostly to predict the number of patient arrivals, which would be equivalent to the number of calls per hour in our case. Moreover, with the exception of McCarthy and colleagues (2011) who predicted hourly arrivals and Cheng and colleagues (2021) who predicted hourly ED occupancy, the calendar variables have been used mostly to improve daily predictions. Tuominen and colleagues (2022) stress the inclusion of local holidays and calendar variables as they were identified as one of the most dominant predictive features. In our case, holidays were not a significant predictor which was likely due to the short span of data, which was just over one year. Yet, the variable “open” acted similarly to a calendar variable in our case. It indicated when the emergency line was open and when closed, a value dependent on time. It was one of the most significant predictors in our SARIMAX models.

Moreover, the inclusion of the number of incoming calls in our SARIMAX model majorly improved its forecast accuracy. The inclusion of the number of calls added nuance to the predicted waiting times as was visible in the plotted predictions. This is in line with previous literature (Jones et al., 2009) that suggested that hourly waiting time and demand is driven by the exogenous patient arrival process – the count of patient arrivals equivalent to our calls. Duarte and colleagues (2021) also examined the waiting time in ED and the number of attendances together, yet they did not determine their relationship. Most importantly, the strong relationship between the number of incoming calls per hour and the predicted waiting time was suggested by the Queuing Theory. However, once the number of calls was included, most of the calendar variables lost their significance in the model, with the exception of the variable "open". This would suggest that the number of incoming calls could be collinear with the calendar variables to some extent, which needs to be investigated in future studies. However, the number of incoming calls is a predictor that is not usually known at the time of forecasting. It is necessary to be forecasted as well, and in case it is not available, calendar variables contribute to model's predictions.

6.2 Integration in the Current System and Comparison with the Current Simulation Model

The two best SARIMAX models showed relatively better performance than the current model. Furthermore, it is possible that the actual prediction accuracy once the models are deployed could be even higher since the model would not be used to generate predictions for the time when the line is closed. In its original form, both SARIMAX models predict negative patient waiting times for the times the emergency call line was closed. Consequently, the negative predictions increase the error metrics RMSE and MAE, as the error was larger than it would have been if the models were employed in practice, and they would not predict waiting times for the periods when the call line was closed. Overall, the two SARIMAX models had a more stable performance over time than the current model. The SARIMAX models were especially more accurate for long-range predictions. When the current model had RMSE of over 9 minutes, both of the selected SARIMAX models had RMSE of around 5 minutes. Such high RMSE for the current model indicates that quite a few large prediction errors occurred, which was not the case for the SARIMAX models. This also indicates that the SARIMAX models were better at predicting the direction of the errors than the current model. This is likely due to the fact that the current model is derived only from a distribution of calls rather than from a learnt-pattern as our models.

Moreover, even the second most accurate SARIMAX model without the number of calls performs relatively well compared to the current simulation model. Namely, it performed better than the current model on a short-range and long-range prediction. Its prediction and also the prediction of the model with calls, were worse in a mid-range than the current model. Moreover, as seen in the plotted predictions, the SARIMAX models were not fully able to capture large fluctuations in the average waiting time. As suggested by the literature (Aboagye-Sarfo et al., 2015), ARIMA-family models work well on general patterns, but sometimes might not be able to capture large variations from the normal pattern. On the other hand, this also makes them less prone to overfitting as some more complex techniques such as LSTM (Ko et al., 2021).

In case the client decides to implement the ARIMA-family models, we suggest that ideally the SARIMAX model with the number of incoming calls is implemented and that the number of future incoming calls is predicted as well. If this is not possible, the equivalent SARIMAX model without the number of calls also performs relatively well. The suggested models could be used mostly to improve the predicted waiting time one day in advance and one month in advance. However, for predictions one week in advance, these models had worse performance than the current model, which could still be considered, for example in a combined prediction. Caution should also be taken when considering large fluctuations as the SARIMAX models cannot capture them well. This is important also from an ethical perspective when the GP office might expect much lower waiting time during a time of a large fluctuation and the possibility of underestimation needs to be communicated. The forecasts with the SARIMAX models also require specifications for each hour. These specifications are also the adjustable parameters the client can input into the system, such as the calendar variables, number of shifts available and expected number of calls.

Moreover, it is possible to simulate the predicted waiting times based on the selected SARIMAX model. These simulations provide the estimated average predicted waiting time per hour. They can be generated multiple times and the predictions can be averaged. In case they provide values lower than 0, which occurs when the line is closed, these can be either not used or fixed to be 0.

In order to incorporate new observations, the same SARIMAX model can be used. It can be re-trained or a new model configuration can be obtained. The running time for model identification based on the lowest AIC takes a significant portion of time, thus it should only be done at night. We recommend the re-estimation of SARIMAX parameters each week or month depending on how much change occurs in the character of the time series. Re-training the model, which corresponds to parameter estimation can be performed more often, however the running time can also be approximated to take several minutes. However, both parameter re-identification and their re-estimation would be necessary in the future to make sure that the model that describes the time series character is still fitting and relevant. This would be especially advisable if a clear trend starts to occur. However, these suggestions are yet to be explored and evaluated.

Several precautions need to be taken when implementing the SARIMAX models. These models do not provide a prediction for each call, but rather an average prediction for all the calls that arrive in a specific hour. Moreover, the predictions are based on and are tested on only one source of data, thus there is limited generalizability of the best models to other GP practices. However, the provided method can be applied to each GP office individually when the data are provided to improve operations of the emergency line. More detailed suggestions for implementation are provided to the client in the documentation.

Answering our second sub-research question, we found that the best ARIMA-family model, SARIMAX with calls, had higher accuracy than the current simulation model used for our case of a single GP office operating an emergency call line in the Netherlands for short- and long-range forecasts of the average patient waiting time in minutes.

6.3 Comparison to LSTM

We recommend LSTM for future implementation as the most accurate model. The linear time-series forecasting models selected for comparison with LSTM were the SARIMAX model without the inclusion of calls and with the inclusion of calls. LSTM with calls had better performance on all time horizons. Moreover, the LSTM model was trained on less data and was thus able to include the variable number of shifts. On the other hand, ARIMA-family models were not able to achieve such high prediction accuracy on a much larger training dataset. These results are in line with some previous studies that suggested that for some cases LSTM had better performance than ARIMA-family models (Siami-Namini & Tavakoli, 2018; Zhang et al., 2022). Siami-Namini and Tavakoli (2018) emphasized that LSTM has better capabilities to capture potential non-linear trends in time series, which might also be the case for our study. This would be supported by the findings from checking the model assumptions, in which none of the ARIMA-family models fully fulfilled the model assumptions. However, such high accuracy as is the case for LSTM can oftentimes indicate that the model might overfit future data (Ko et al., 2021) while ARIMA-family models do so less often.

Answering our main research question, for our case of a single GP office that operates an emergency call line in the Netherlands, LSTM had better performance of the average patient waiting time per hour than ARIMA-family models.

6.4 Discussion of the Tentative Analysis

Initially, we intended to test the inclusion of the number of shifts as an external predictor. The need to incorporate the number of the limited supply of available staff was also stressed by Kim and colleagues (2020) and Tuominen and colleagues (2022) and is represented by the supply side in the Queuing Theory. However, due to the lack of data containing this variable, we were able to perform only a tentative analysis of this variable's effect. The effect of the number of shifts on the average patient waiting time prediction found in the tentative analysis might be scalable on larger datasets, however there was not enough data to establish this assumption. In fact, this model had poor prediction accuracy compared to the other ARIMA-family models. The poor accuracy results of the model trained on less data and its incapability to identify any seasonal components in the data indicate that a larger dataset is necessary for ARIMA-family models to learn a seasonal pattern on a higher level accurately. When the number of calls were included in the SARIMAX model, the inclusion of the number of shifts did not improve the prediction further. However, when the number of calls was not included, the number of shifts for each hour improved the prediction accuracy. Therefore, it is possible that the number of incoming calls could capture most of the trend otherwise captured by the number of shifts. On the other hand, when the number of calls would not be available in future scenarios, the number of shifts could help increase the forecast accuracy of the SARIMAX model. However, more data would be needed to draw valid conclusions.

6.5 Limitations and Future Work

Although diligent analysis was performed, there are a couple of limitations of this research. Firstly, the suggested models were validated using the external dataset spanning only one month into the future. It is necessary to test the model with future data also for different seasons and months to establish its performance across various conditions. Secondly, we modelled all patient waiting times together although there were two different streams of calls based on urgency. This choice was undertaken as the calls from these two streams might influence each other and are similar in principle. Yet, this assumption needs to be tested in the future. Moreover, we imputed the missing values of the waiting time for individual calls by using the mean calculated delay. Even though we handled outliers, which likely handled also imputed values that would be too different, future research should validate this. Two parallel analyses should be conducted to compare the results of a model trained on data using imputed values and a model discarding these values.

Furthermore, although it was intended originally, we were not able to test the effect of the number of shifts on the predicted waiting time in a desired manner. The dataset containing the number of shifts consisted of only around a thousand observations, which was very few to find meaningful patterns. We provided suggestions for how the number of shifts could influence patient waiting time, yet these suggestions need to be tested further. Future research should investigate the inclusion of the predictor number of shifts on larger amount of data.

Moreover, there is a practical limitation for future deployment regarding the running time of the best SARIMAX model. Even though three different libraries were tried in order to minimize the running time, the identification of the model parameters took several hours. Moreover, the training time of the proposed model on all of the data available was also significant, taking around 20 minutes. For this reason, also an attempted rolling prediction had to be abandoned. Yet, we suggest that future research adopts rolling prediction with the potential of improving the forecast accuracy. Thus, the selection should be performed only once over a longer time period, when the character of the time series is expected to change. To test this assumption, future research should investigate the optimal time for parameter re-identification and re-estimation.

We also suggest that future research investigates the effect of predicted values of incoming calls rather than true values with a sensitivity analysis. We suggest that different predictions obtained for the number of arriving calls should be performed to investigate the positive effect of the number of calls on the accuracy of the waiting time prediction. It is possible that if the predictions of the incoming number of calls per hour deviate from the actual values too much that the SARIMAX model without the number of incoming calls would have better performance.

7. Conclusion

In this thesis, we aimed to test and compare the prediction accuracy of ARIMA-family models with LSTM for patient waiting times at a single Dutch GP office and compare them to the current model of the client. This thesis focused on the case of a single GP office that operates an emergency call line in the Netherlands. Currently, the office predicts patient waiting times on their on-call line using a Discrete Simulation model that is based on variable distributions of past calls and staff schedule. Although plenty of studies focused on prediction of patient arrivals at the emergency departments in hospitals, there has been limited research that investigated patient waiting time at emergency departments. Moreover, to our knowledge, no previous research has aimed at predicting patient waiting times on an on-call emergency line in the Netherlands based on historical data. We adopted one of the most commonly-used time-series forecasting group of models: the ARIMA family. Using the Box-Jenkins approach and prediction validation, we iteratively performed tests of ARIMA, SARIMA and SARIMAX models and selected the most accurate model using the prediction errors metrics MAE and RMSE. Our models were trained and tested on an hourly frequency of data trying to predict the average waiting time in minutes for each hour on the horizons of one day, one week and one month in advance.

In our case, the SARIMAX model using the number of incoming calls per hour was selected as the best-performing ARIMA-family model with the lowest prediction errors for all time horizons. When this model was validated on an external dataset and compared to the model that is in current use, it outperformed the old model on all time horizons except for one week in advance. Yet, we acknowledge that the number of incoming calls per hour is a predictor that needs to be forecasted as well. For this reason, we also tested the second best model – SARIMAX without calls on the external dataset. This model had a bit higher prediction errors, yet it performed better than the current model a month in advance, and comparably well one day in advance. We also attempted to test the effect of the number of shifts on the prediction of the waiting time. Yet, the dataset available for this variable was spanning a very short time and included only around one thousand observations which allowed us to perform only a tentative analysis that might serve as a basis for future research.

Considering the analysis, we recommend the implementation of the LSTM model with calls over ARIMA-family models for patient waiting time prediction at this GP office. The LSTM model had higher prediction accuracy than the most accurate ARIMA-family model for all time horizons. Moreover, LSTM required less training data to achieve such high accuracy and was able to utilize the number of shifts available each hour, unlike the ARIMA-family models. Overall, LSTM had better performance than the ARIMA-family models for this particular case.

Bibliography

Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F. M., Preen, D. B., Stewart, L. M., & Fatovich, D. M. (2015). A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. *Journal of Biomedical Informatics*, *57*, 62–73. https://doi.org/10.1016/j.jbi.2015.06.022

Abraham, G., Byrnes, G. B., & Bain, C. A. (2009). Short-term forecasting of emergency inpatient flow. *IEEE Transactions on Information Technology in Biomedicine*, *13*(3), 380–388. https://doi.org/10.1109/TITB.2009.2014565

Baldon, N. (2019). *Time series Forecast of Call volume in Call Centre using Statistical and Machine Learning Methods* [MSc Thesis]. KTH ROYAL INSTITUTE OF TECHNOLOGY.

Becerra, M., Jerez, A., Aballay, B., Garcés, H. O., & Fuentes, A. (2020). Forecasting emergency admissions due to respiratory diseases in high variability scenarios using time series: A case study in Chile. *Science of The Total Environment*, *706*, 134978. https://doi.org/10.1016/j.scitotenv.2019.134978

Bernstein, S. L. (2009). The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, *16*(1), 1. https://doi.org/10.1111/j.1553-2712.2008.00295.x

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day. https://books.google.nl/books?id=1WVHAAAAMAAJ

Box, G. E. P., & Tiao, G. C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, *70*(349), 70–79. JSTOR. https://doi.org/10.2307/2285379

Boyle, J., Jessup, M., Crilly, J., Green, D., Lind, J., Wallis, M., Miller, P., & Fitzgerald, G. (2012). Predicting emergency department admissions. *Emergency Medicine Journal*, *29*(5), 358–365. https://doi.org/10.1136/emj.2010.103531

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Busetti, F., & Harvey, A. (2003). Seasonality Tests. *Journal of Business & Economic Statistics*, *21*(3), 420–436. https://doi.org/10.1198/073500103288619061

Carvalho-Silva, M., Monteiro, M. T. T., de Sá-Soares, F., & Dória-Nóbrega, S. (2018). Assessment of forecasting models for patients arrival at emergency department. *Operations Research for Health Care*, *18*, 112–118. https://doi.org/10.1016/j.orhc.2017.05.001

Champion, R., Kinsman, L. D., Lee, G. A., Masman, K. A., May, E. A., Mills, T. M., Taylor, M. D., Thomas, P. R., & Williams, R. J. (2007). Forecasting emergency department presentations. *Australian Health Review*, *31*(1), 83–90. https://doi.org/10.1071/ah070083

Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, *10*(1), 25–45. https://doi.org/10.1007/s10729-006-9006-3

Cheng, Q., Argon, N. T., Evans, C. S., Liu, Y., Platts-Mills, T. F., & Ziya, S. (2021). Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, *48*, 177–182. https://doi.org/10.1016/j.ajem.2021.04.075

Cerqueira, V., Torgo, L., Smailović, J., & Mozetič, I. (2017). A Comparative Study of Performance Estimation Methods for Time Series Forecasting. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 529–538. https://doi.org/10.1109/DSAA.2017.7

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/peerj-cs.623

Choudhury, A., & Urena, E. (2020). Forecasting hourly emergency department arrival using time series analysis. *British Journal of Healthcare Management*, *26*(1), 34–43. https://doi.org/org/10.12968/bjhc.2019.0067

Daellenbach, H. G., & Flood, R. L. (2002). *The Informed Student Guide to Management Science*. https://doi.org/10.5860/choice.40-1926 de O. Santos Júnior, D. S., de Oliveira, J. F. L., & de Mattos Neto, P. S. G. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, *175*, 72–86. https://doi.org/10.1016/j.knosys.2019.03.011

Duarte, D., Walshaw, C., & Ramesh, N. (2021). A Comparison of Time-Series Predictions for Healthcare Emergency Department Indicators and the Impact of COVID-19. *Applied Sciences*, *11*(8). https://doi.org/10.3390/app11083561

Forster, M., & Sober, E. (2011). Aic Scores as Evidence: A Bayesian Interpretation. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics* (Vol. 7, pp. 535–549). North-Holland. https://doi.org/10.1016/B978-0-444-51862-0.50016-2

Gamakumara, P., Santos-Fernández, E., Talagala, P., Hyndman, R., Mengersen, K., & Leigh, C. (2023). *Conditional normalization in time series analysis*. https://doi.org/10.48550/arXiv.2305.12651

Gul, M., & Celik, E. (2018). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, *9*, 263–284. https://doi.org/10.1080/20476965.2018.1547348

Hanneke, P. (2021). *Model based decision-making on the possibilities to reduce overcrowding at out-of-hours general practitioner departments* [MSc Thesis]. TU Delft.

Helfenstein, U. (1996). Box-Jenkins modelling in medical research. *Statistical Methods in Medical Research*, *5*(1), 3–22. https://doi.org/10.1177/096228029600500102

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts: Melbourne, Australia. https://otexts.com/fpp2/

Hyndman, R., & Kostenko, A. (2007). Minimum Sample Size Requirements for Seasonal Forecasting Models. *Foresight: The International Journal of Applied Forecasting*, *6*, 12–15.

Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, *32*(3), 865–874. https://doi.org/10.1016/j.ijforecast.2015.11.012

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics*, *42*(1), 123–139.

Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., & Snow, G. L. (2008). Forecasting Daily Patient Volumes in the Emergency Department. *Academic Emergency Medicine*, *15*(2), 159–170. https://doi.org/10.1111/j.1553-2712.2007.00032.x

Kim, K.-R., Park, J.-E., & Jang, I.-T. (2020). Outpatient forecasting model in spine hospital using ARIMA and SARIMA methods. *Journal of Hospital Management and Health Policy; Vol 4 (September 2020): Journal of Hospital Management and Health Policy*. https://doi.org/10.21037/jhmhp-20-29

Kim, T. H., Hong, K. J., Shin, S. D., Park, G. J., Kim, S., & Hong, N. (2019). Forecasting respiratory infectious outbreaks using ED-based syndromic surveillance for febrile ED visits in a Metropolitan City. *The American Journal of Emergency Medicine*, *37*(2), 183–188. https://doi.org/10.1016/j.ajem.2018.05.007

Ko, M.-S., Lee, K., Kim, J.-K., Hong, C. W., Dong, Z. Y., & Hur, K. (2021). Deep Concatenated Residual Network With Bidirectional LSTM for One-Hour-Ahead Wind Power Forecasting. *IEEE Transactions on Sustainable Energy*, *12*, 1321–1335.

Kuo, Y.-H., Chan, N. B., Leung, J. M. Y., Meng, H., So, A. M.-C., Tsoi, K. K. F., & Graham, C. A. (2020). An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department. *International Journal of Medical Informatics*, *139*, 104143. https://doi.org/10.1016/j.ijmedinf.2020.104143

Ledolter, J. (1989). The effect of additive outliers on the forecasts from ARIMA models. *International Journal of Forecasting*, *5*(2), 231–240. https://doi.org/10.1016/0169-2070(89)90090-3

Mai, Q., Aboagye-Sarfo, P., Sanfilippo, F. M., Preen, D. B., & Fatovich, D. M. (2015). Predicting the number of emergency department presentations in Western Australia: A population-based time series analysis. *Emergency Medicine Australasia*, *27*(1), 16–21. https://doi.org/10.1111/1742-6723.12344

Marcilio, I., Hajat, S., & Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic Emergency Medicine*, *20*(8), 769–777. https://doi.org/10.1111/acem.12182

McCarthy, M. L. (2011). Overcrowding in emergency departments and adverse outcomes: Death and admission rates are higher when length of stay is longer. *BMJ: British Medical Journal*, *342*(7809), 1220–1221. JSTOR. https://doi.org/10.1136/bmj.d2830

Panigrahi, S. S., & Behera, H. S. (2013). *Effect of Normalization Techniques on Univariate Time Series Forecasting using Evolutionary Higher Order Neural Network*.

Rausch, T. M., Albrecht, T., & Baier, D. (2022). Beyond the beaten paths of forecasting call center arrivals: On the use of dynamic harmonic regression with predictor variables. *Journal of Business Economics*, *92*(4), 675–706. https://doi.org/10.1007/s11573-021-01075-4

Ross, S. M. (2007). Introduction to Probability Models (9th ed.). Academic Press.

Rosychuk, R. J., Youngson, E., & Rowe, B. H. (2015). Presentations to Alberta Emergency Departments for Asthma: A Time Series Analysis. *Academic Emergency Medicine*, *22*(8), 942–949. https://doi.org/10.1111/acem.12725

Siami-Namini, S. & Tavakoli, N. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Stagge, A. (2021). *A time series forecasting approach for queue wait-time prediction* [MSc Thesis]. KTH ROYAL INSTITUTE OF TECHNOLOGY.

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications*. Springer; WorldCat.org.

Sudarshan, V. K., Brabrand, M., Range, T. M., & Wiil, U. K. (2021). Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study. *Computers in Biology and Medicine*, *135*, 104541. https://doi.org/10.1016/j.compbiomed.2021.104541

Tian, Y. (2023). *GP Post Call Wait Times Predictions Using LSTM* [MSc Data Science Thesis]. Utrecht University.

Tukey, J. W. (1977). *Exploratory data analysis* (2nd ed., Vol. 2). Reading, MA.

Tuominen, J., Lomio, F., Oksala, N., Palomäki, A., Peltonen, J., Huttunen, H., & Roine, A. (2022). Forecasting daily emergency department arrivals using high-dimensional multivariate data: A feature selection approach. *BMC Medical Informatics and Decision Making*, *22*(1), 134. https://doi.org/10.1186/s12911-022-01878-7

Vollmer, M. A. C., Glampson, B., Mellan, T., Mishra, S., Mercuri, L., Costello, C., Klaber, R., Cooke, G., Flaxman, S., & Bhatt, S. (2021). A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*, *21*(1), 9. https://doi.org/10.1186/s12873-020-00395-y

Wargon, M., Guidet, B., Hoang, T. D., & Hejblum, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal*, *26*(6), 395. https://doi.org/10.1136/emj.2008.062380

Wheelwright, S., Makridakis, S., & Hyndman, R. J. (1998). *Forecasting: Methods and applications*. John Wiley & Sons.

Whitt, W., & Zhang, X. (2019). Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care*, *21*, 1–18. https://doi.org/10.1016/j.orhc.2019.01.002

Zhang, Y., Zhang, J., Tao, M., Shu, J., & Zhu, D. (2022). Forecasting patient arrivals at emergency department using calendar and meteorological information. *Applied Intelligence*, *52*(10), 11232–11243. https://doi.org/10.1007/s10489-021-03085-9

Zhao, X., Lai, J. W., Ho, A. F. W., Liu, N., Ong, M. E. H., & Cheong, K. H. (2022). Predicting hospital emergency department visits with deep learning approaches. *Biocybernetics and Biomedical Engineering*, *42*(3), 1051–1065. https://doi.org/10.1016/j.bbe.2022.07.008

Appendix I: Authors contribution

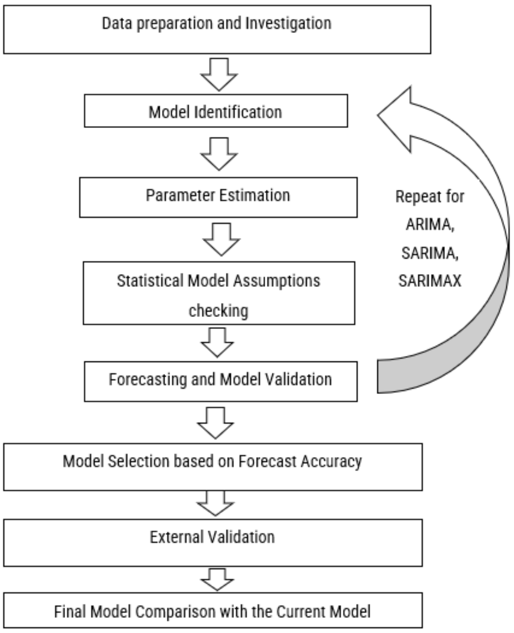
Data provider: Esculine (external company), Data preprocessing (training and test set, external dataset for comparison): Katarína Barteková, Feature Engineering: feature Number of shifts: Yaqin Tian, all other features (time index, average waiting time per hour, season, week day, holiday, hour, minute, open/not open, number of incoming calls): Katarína Barteková, LSTM model development: Yaqin Tian, ARIMA-family models: Katarína Barteková

We refer to the work by Yaqin Tian throughout this thesis in the form (Tian, 2023)

Appendix II: Method, Preprocessing and Data Description

II.I Method

II.I.I Schematic Illustration of the Method


Data preparation and Investigation
Model Identification
Parameter Estimation
Statistical Model Assumptions checking
Forecasting and Model Validation
Repeat for ARIMA, SARIMA, SARIMAX
Model Selection based on Forecast Accuracy
External Validation
Final Model Comparison with the Current Model

II.I.II Hourly Frequency Selection

We selected the hourly frequency as the interval at which the time series is studied in this thesis. The two other possible time frequencies for our time series analysis were on a minute- and day-basis. We attempted also several predictions on both a minute- and day-basis. However, these turned out to be either too granular and time-intensive in the case of minute-frequencies or too broad to capture any significant trend in case of daily-frequencies. Moreover, the training time and convergence of the models posed a major problem for even the simplest minute-frequency models as the models were not able to converge. On the other hand, for daily frequency, the dataset consisted of only 453 days and the model would not have enough data to find meaningful patterns. We selected the hourly time interval as it gives us the most granularity while still being able to capture the trends in the data compared to the possible alternatives of predictions per minute or per day.

II.I.III Packages used and their Scalability

The python packages Pmdarima, Statsforecast and Statsmodel were used in the technical part of this thesis. Originally the Pmdarima package was employed for step 3.1.1.2 model identification and step 3.1.1.3 parameter estimation to select the best model configuration. This package was originally chosen since it offers the capability to not only identify the most-fitting parameters of ARIMA-family models using step-wise selection that minimizes AIC, but to also estimate the parameter coefficients and make forecasts using the best model selected. However, the library was not scalable on such large dataset on more complex models (SARIMA and SARIMAX models) and several memory errors occurred after hours of running. The documentation for this library is available at: http://alkaline-ml.com/pmdarima/

Thus, the Statsforecast python library was used for step 3.1.1.2 model identification. We used the AutoARIMA function from the package to perform the parameter search and find the most fitting parameters and differencing order for the model based on AIC using the step-wise algorithm search. The stepwise algorithm finding the best model with the lowest AIC was essentially the same to Pmdarima library, yet it worked at a higher pace and used less memory. The library was used in its default setting using the AutoARIMA function in a manner corresponding to the type of model being fit (i.e. no exogenous predictors nor any seasonality were used for a plain ARIMA model). The documentation for this library is available at: https://nixtla.github.io/statsforecast/src/core/models.html

The parameters were estimated using the Statsmodels package (see 3.1.1.3 parameter estimation). The best model configurations providing the closest fit (obtained from the AutoARIMA function of Statsforecast) were used for training a model using the SARIMAX function from Statsmodels package in Python. This package allowed also for an assessment of the fit of the model to the time series. The package applied the Maximum likelihood estimation to estimate the coefficients as closely as possible. The reason for selecting Stasmodels over Statsforecast for this step was the fact that Statsforecast library does not provide such a detailed description of the coefficient values of the model parameters as Statsmodels library. Statsmodels library also provided an easier way to determine the fit of the model using statistical tests and graphical tools to plot the residuals. Statsmodels setting was used in default with the hourly frequency and using the parameters determined by Statsforecast library. The documentation for this library is available at: https://www.statsmodels.org/stable/index.html

II.II Raw Data Description and Preprocessing

II.II.I Raw Data Description

The raw data was provided in two datasets with the following features provided in a “str” form:

- Shift dataset
    - Code of the shift
    - Start shift
    - End shift
    - Shift type
    - Function of the one taking the shift (coordinator or triagist)
- Calls dataset
    - Line (emergency, normal)
    - Start call
    - End call
    - Answer time
    - Waiting time
    - Call time (length)
    - Urgency (highest 1-5 lowest) – determined by the triagist/operator

The Shift dataset included information about shifts of both historic and future character. However, the details about future shifts could not be accurate since the shift schedules change on a regular basis. Yet, the data on shift schedules on historical dates are accurate since they depict a real, unchanged schedule of the shifts.

II.II.II Preprocessing

We recoded the variables regarding time and date information into a datetime format. We set the datetime as an index. However, for ARIMA-family models to function properly we needed equal intervals between observations. In order to do this, we added the frequency of observations for each hour based on start of the call if the call started was whenever in that hour. For those hours, when no calls came, we added 0s as both the number of incoming calls and the average waiting time. The variable number of calls depicted the count of all calls that appeared in a specific hour and the variable mean waiting time depicted the mean waiting time for all calls that came in that hour in minutes. Yet, the equaling of time intervals also created missing variables on a daily level. These were imputed using the daily index since the daily variables applied to the whole day, not just specific hour. They were thus the same for all hours of the day. We also created the variable indicating the number of shifts by matching the number of shifts available for each call.

Waiting time for each call was provided in the dataset. However, there was a parallel software system that recorded how long the patient had to wait on the line recorded in the variable waiting time. Nevertheless, the variable waiting time varies from the differenced time between when the patient started the call and when it was answered. The differenced time (mean 842 sec) was longer than the record waiting time (mean 754 sec), even though we would expect them to be the same. After consulting this observation with the provider of the data, they expressed that the waiting time recorded by the system stored in the waiting time variable was recorded after a tape was played at the beginning of each call, while the time of the start of the call is recorded before the tape is played. Thus, we accounted for this delay in imputation of waiting time. We also performed the Little's MCAR test to make sure that we do not add any bias into the data.

All of the categorical variables were recoded into a dummy form using one-hot-encoding. This meant that the variables: open, season, weekday (including holiday) were recoded into a dummy form. For weekday the coding was the following: Day-1 indicated a public holiday, Day0 indicated Monday, Day1 indicated Tuesday, Day2 indicated Wednesday, Day3 indicated Thursday, Day4 indicated Friday, Day5 indicated Saturday, Day6 indicated Sunday. For season the coding was the following: Season1 indicated winter, Season2 indicated spring, Season3 indicated summer, Season4 indicated autumn.

The numerical variables number of shifts, number of calls and mean waiting time per hour were normalized using z-score normalization for model training. Moreover, outliers in mean waiting time per hour were handled using the IQR method.

The two originally provided dataframes were merged into one dataframe based on their index which was expressed in hours in a datetime form.

The train-test splitting was thus done on the intervals 24 (for 24 hours, which is one day), 24*7 (for a week), and 24*30 (for a month).

For the external validation dataset, all of the preprocessing steps were repeated with the exception of missing values imputation, outlier handling and dataset splitting.

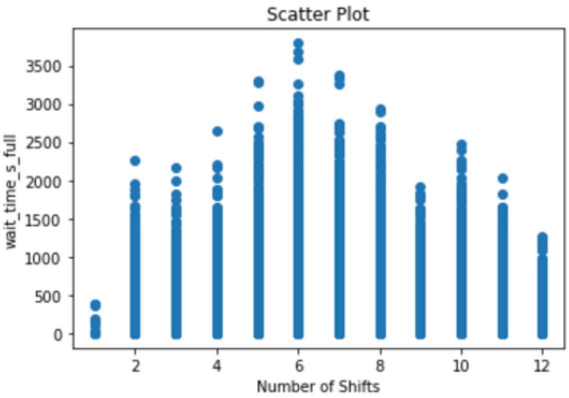
II.III Data description

We provide a more detailed data description in this section. The data on shift schedules had 1 427 observations in schedules with no missing observations. The variable number of shifts was created for each call by matching the time of the incoming call to which shifts are concurrent to the time when the call occurred and they are summed to obtain the overall number of shifts for each call arriving in a particular hour. The number of observations with the number of shifts was thus 10153, and they spanned from 1 March 2023 to 17 April 2023.

The second data set available consisted of call data. There were 9 2318 individual calls. The only variables with missing values were the waiting time variable (2786 missing values), which was also our outcome variable, the length of the call (52 missing values), and the urgency of the call determined by the triagist once the call was picked up (26998 missing values). Since neither the length of the call nor the urgency of the call were used as the predictors for our models, there was no need to impute or discard these observations. Overall, there were 314 calls equal to or shorter than 1 second and no calls longer than 2 hours (the longest call took 1,19 hours). We also removed the outliers using the IQR method, which resulted in additional 2001 observations removed.

Overall, we merged the two datasets based on the hourly index. The final dataset contained 10 873 observations on an hourly basis in the time period 20th January 2022 00:00:00 until (and not including) 18th April 2023 00:00:00 (last observation was 17th April 2023 23:00:00, this however captured the whole duration of the 23rd hour).

As visible in Appendix II: Figure 1, the number of shifts available to serve a particular call is related to the waiting time of the call with a possible trend of few shifts being assigned to times with lower waiting times and more shifts being assigned to times with possibly higher waiting times. We see the highest waiting time for the middle values of shifts available.



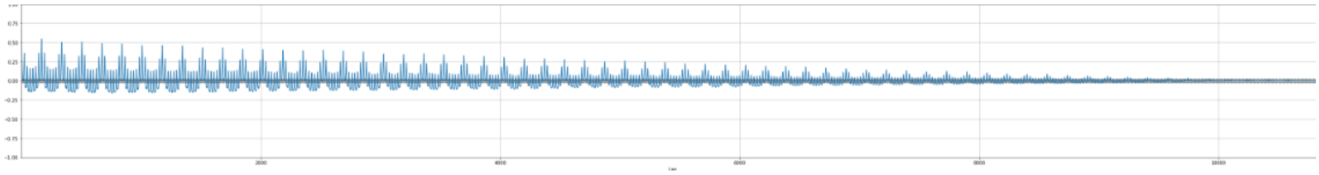
Appendix II, Figure 1: The relationship of the number of shifts and waiting time

II.III.I Stationarity testing

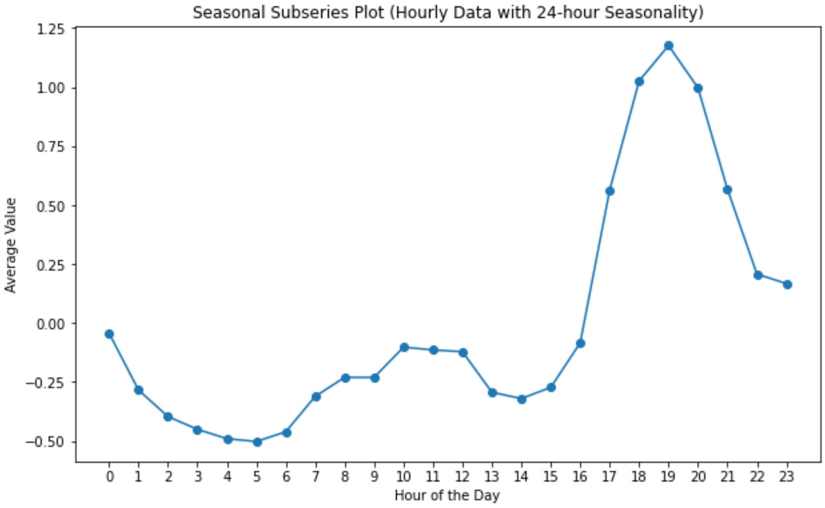
We tested the stationarity of our time series using the following statistical tests: ADF, KPSS, and the PP test. We also used the “ndiffs” function provided in the Pmdarima package with setting seasonality to 24. Ndiffs indicated no need to difference the time series as it was showing no variation in its mean and variance over time. The differencing order determined by the ADF test was 0. The differencing order determined by the KPSS test was 1. The differencing order determined by the PP test was 0. Thus, the data might become difference-stationary after differencing by 1, but it is also possible that the models would have a better fit without applying differencing. The estimated seasonal differencing term, should be either 0 (based on nsdiffs) or 1 based on The Canova-Hansen test for seasonal differences.

II.III.II Seasonality

Seasonality is clearly visible in the Autocorrelation plot that was plotted with a high number of lags so that any trend would be visible (Appendix II: Figure 2). Moreover, seasonality on a daily basis is visible also in the seasonal subseries plot that indicates the average value of the average patient waiting time per hour for different hour of the day (Appendix II: Figure 3).



Appendix II, Figure 2: Autocorrelation plot for more lags



Appendix II, Figure 3: Seasonal Subseries Plot of Waiting Time per Hour

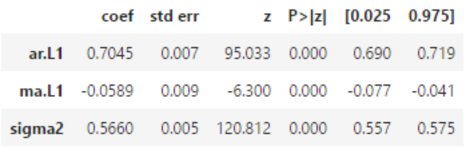
Appendix III: Additional Results to Models (Model Identification, Parameter estimation and Statistical Assumptions Checking) and Model Selection

III.I ARIMA models

We provide the detailed results of the non-seasonal ARIMA models using differencing 0 and differencing 1, as was suggested by the statistical tests conducted. The model with the lower AIC was used based on the second step of the Box-Jenkins method (see 3.1.1.2 Model Identification). It was the model with differencing 0.

III.I.I Model with Differencing 0

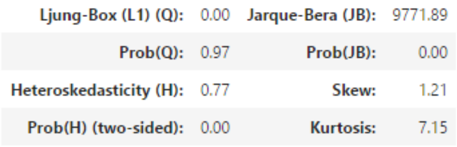
The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are available in Appendix III: Table 1. The results indicate how much each model parameter contributed to the model fit.



	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7045	0.007	95.033	0.000	0.690	0.719
ma.L1	-0.0589	0.009	-6.300	0.000	-0.077	-0.041
sigma2	0.5660	0.005	120.812	0.000	0.557	0.575

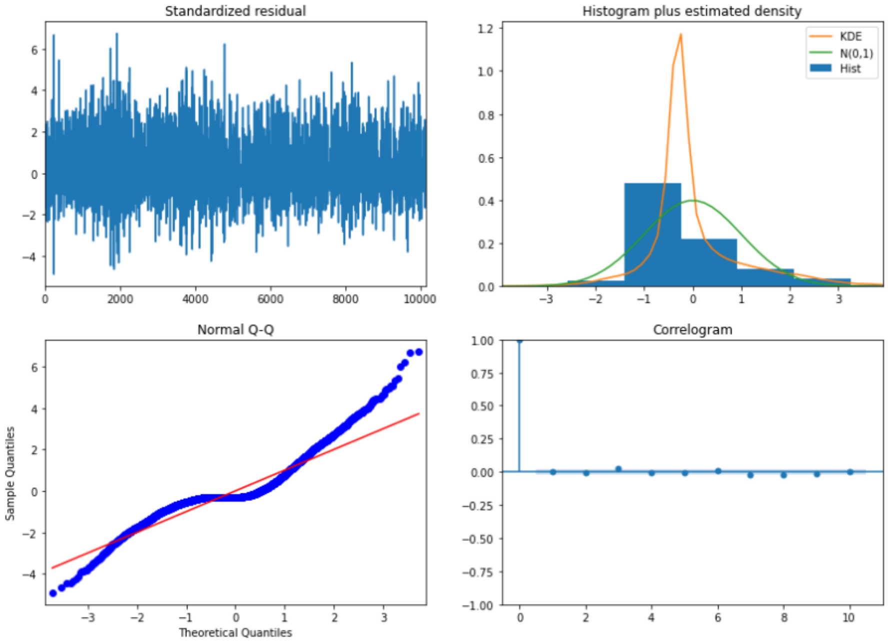
Appendix III: Table 1: Estimated Parameters

The results from statistical assumptions checking (see part 3.1.1.4) are reported. As can be seen in the statistical tests results in Appendix III: Table 2 and graphs in Appendix III: Figure 1, the model does not have a perfect fit. According to the model summary in Appendix III: Table 2, the model meets the condition of independence in the residuals (no correlation) because the p-value of the Ljung-Box test (Prob(Q)) is greater than 0.05. This means that the residuals for this model are independent, as assumed in the model. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05. Thus, there is strong evidence of heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. Furthermore, there is some skewness in the residuals, which suggests that the distribution of the errors is not fully symmetric as would be desired. Skewness is also visible in the histogram in Appendix III: Figure 1 and in the Q-Q plot in Appendix III: Figure 1, where the residuals do not align with the straight line. The AIC is 23040.



Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	9771.89
Prob(Q):	0.97	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.77	**Skew:**	1.21
Prob(H) (two-sided):	0.00	**Kurtosis:**	7.15

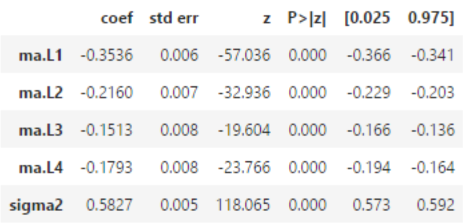
Appendix III: Table 2: Statistical Assumptions Checking



Appendix III: Figure 1: Statistical Assumptions Checking

III.I.II Model with Differencing 1

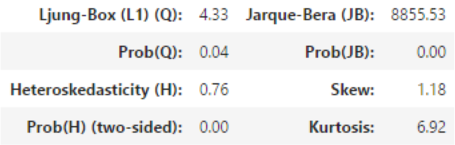
The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are available in Appendix III: Table 3. The results indicate how much each model parameter contributed to the model fit.



	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.3536	0.006	-57.036	0.000	-0.366	-0.341
ma.L2	-0.2160	0.007	-32.936	0.000	-0.229	-0.203
ma.L3	-0.1513	0.008	-19.604	0.000	-0.166	-0.136
ma.L4	-0.1793	0.008	-23.766	0.000	-0.194	-0.164
sigma2	0.5827	0.005	118.065	0.000	0.573	0.592

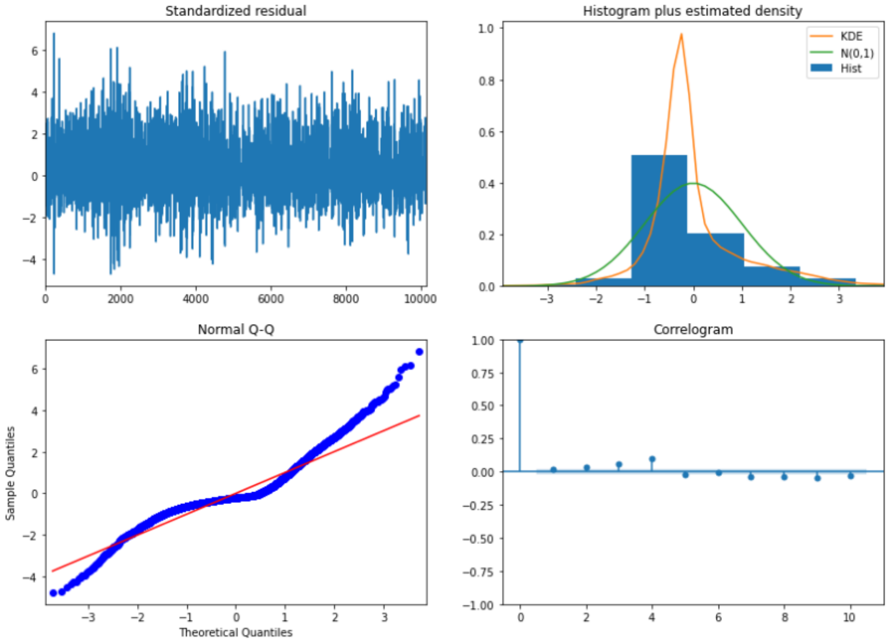
Appendix III: Table 3: Estimated Parameters

The results from statistical assumptions checking (see part 3.1.1.4) are reported. As can be seen in the statistical tests results in Appendix III: Table 4 and graphs in Appendix III: Figure 2, the model does not have a perfect fit either. According to the model summary in Appendix III: Table 4, the model does not meet the condition of independence in the residuals because the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. This means that the residuals for this model are not independent, as assumed in the model. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05. Thus, there is strong evidence of heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. Furthermore, there is some skewness in the model residuals, which suggests that the distribution of the errors is not fully symmetric as would be desired. The skewness is very similar to the model without any differencing (Appendix III: Figure 1). Skewness is also visible in the histogram in Appendix III: Figure 2and in the Q-Q plot in Appendix III: Figure 2, where the residuals do not align with the straight line. The AIC is 23339, which is higher than the AIC from the model with no differencing showing worse fit to our time series.



Ljung-Box (L1) (Q):	4.33	**Jarque-Bera (JB):**	8855.53
Prob(Q):	0.04	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.76	**Skew:**	1.18
Prob(H) (two-sided):	0.00	**Kurtosis:**	6.92

Appendix III: Table 4: Statistical Assumptions Checking



Appendix III: Figure 2: Statistical Assumptions Checking

Overall, since the AIC is lower for the model without differencing and the model meets the condition of independence better than the model with differencing of 1, we selected the model with no differencing for our main analysis of ARIMA models.

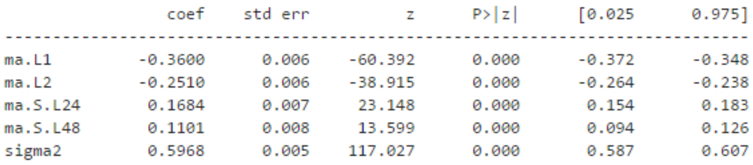
None of the model fits were perfect. Thus, we also tried the log-transformation of our time series data before fitting the models to account for the heteroskedasticity of the model residuals. However, the results were comparable without any major improvement. Another possible data transformation was Box-Cox transformation, which was however not possible with our data due to waiting time observations taking the value 0 during all the times the emergency line was closed.

III.II SARIMA models different differencing versions comparison

We set the differencing (d) to be maximally 1 in the following models. Thus, the differencing values tested when selecting the model with the lowest AIC were 0 and 1. In order to test whether seasonal differencing should be 0 or 1, we run one model with seasonal differencing (D) D = 0, and one with D = 1. The model with the lower AIC was used based on the second step of the Box-Jenkins method (see 3.1.1.2 Model Identification). It was the model with no seasonal differencing.

III.II.I SARIMA with Regular differencing max 1, seasonal differencing 0

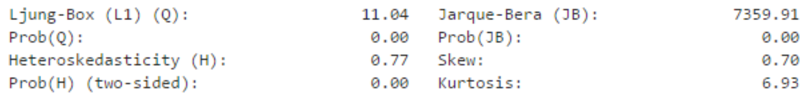
The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are available in Appendix III: Table 5. The results indicate how much each model parameter contributed to the model fit.



```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.3600      0.006    -60.392      0.000      -0.372      -0.348
ma.L2         -0.2510      0.006    -38.915      0.000      -0.264      -0.238
ma.S.L24       0.1684      0.007     23.148      0.000       0.154       0.183
ma.S.L48       0.1101      0.008     13.599      0.000       0.094       0.126
sigma2         0.5968      0.005    117.027      0.000       0.587       0.607
```

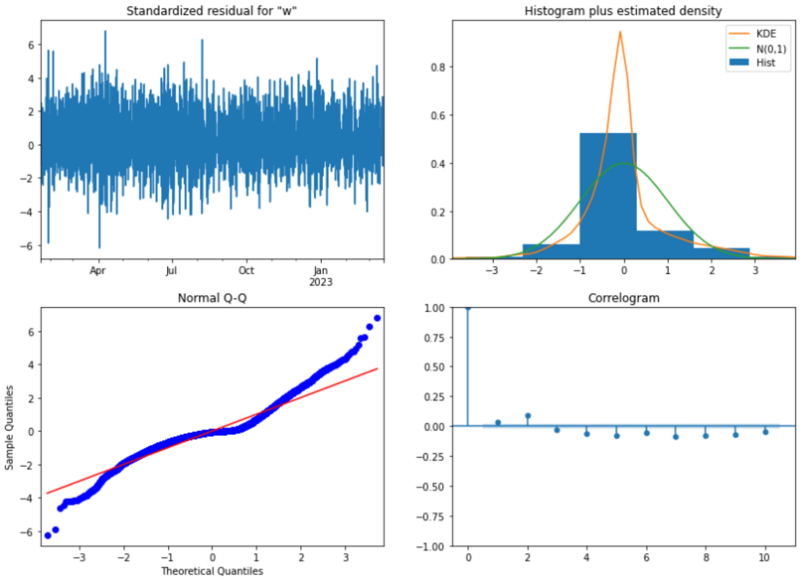
Appendix III: Table 5: Parameter Estimation

The results from statistical assumptions checking (see section 3.1.1.4) are reported. The best model with no seasonal differencing is the one where regular differencing is employed (D=0, but d was selected as 1 using stepwise algorithm). As can be seen in the statistical tests results in Appendix III: Table 6 and graphs in Appendix III: Figure 3, this model does not show a perfect fit. According to the model summary in Appendix III: Table 6, the model does not meet the condition of independence in the residuals because the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05. Thus, there is strong evidence of heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. Furthermore, there is some skewness in the residuals, which suggests that the distribution of the errors is not fully symmetric as would be assumed. However, it should be noted that the skewness is much lower than in the previous (non-seasonal ARIMA) models. The skewness of this model is also visible in the histogram in Appendix III: Figure 3 and in the Q-Q plot in Appendix III: Figure 3, where the residuals do not align with the straight line. The AIC is 23582.



```
Ljung-Box (L1) (Q):                 11.04   Jarque-Bera (JB):              7359.91
Prob(Q):                             0.00   Prob(JB):                         0.00
Heteroskedasticity (H):              0.77   Skew:                             0.70
Prob(H) (two-sided):                 0.00   Kurtosis:                         6.93
```

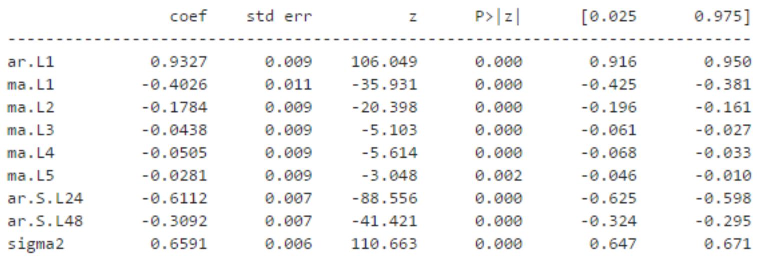
Appendix III: Table 6: Statistical Assumptions Checking



Appendix III: Figure 3: Statistical Assumptions Checking

III.II.II SARIMA with Regular differencing max 1, seasonal differencing 1

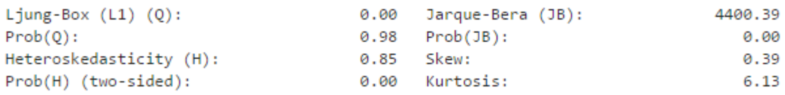
The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are available in Appendix III: Table 7. The results indicate how much each model parameter contributed to the model fit.



```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9327      0.009    106.049      0.000       0.916       0.950
ma.L1         -0.4026      0.011    -35.931      0.000      -0.425      -0.381
ma.L2         -0.1784      0.009    -20.398      0.000      -0.196      -0.161
ma.L3         -0.0438      0.009     -5.103      0.000      -0.061      -0.027
ma.L4         -0.0505      0.009     -5.614      0.000      -0.068      -0.033
ma.L5         -0.0281      0.009     -3.048      0.002      -0.046      -0.010
ar.S.L24      -0.6112      0.007    -88.556      0.000      -0.625      -0.598
ar.S.L48      -0.3092      0.007    -41.421      0.000      -0.324      -0.295
sigma2         0.6591      0.006    110.663      0.000       0.647       0.671
```

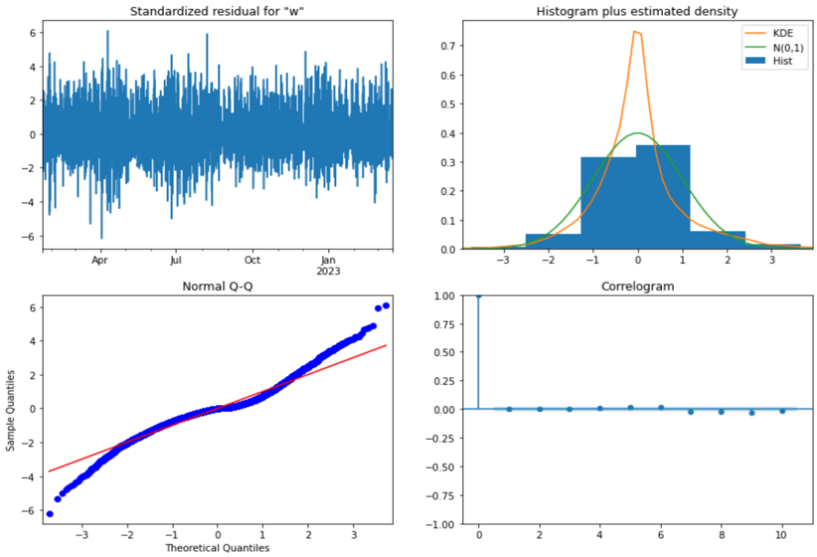
Appendix III: Table 7: Estimated Parameters

The results from statistical assumptions checking (see section 3.1.1.4) are reported. The best model with seasonal differencing 1 is the one where no regular differencing is employed (D=1, d was selected as 0 using stepwise algorithm). As can be seen in the statistical tests results in Appendix III: Table 8 and graphs in Appendix III: Figure 4, this model does not have a perfect fit either. According to the model summary in Appendix III: Table 8, the model meets the condition of independence in the residuals (no correlation) because the p-value of the Ljung-Box test (Prob(Q)) is higher than 0.05, which is an improvement compared to the previous model. This means that the residuals for this model are independent, as assumed in the model. However, similarly to the previous model, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05, indicating a non-constant variance of the residuals. Furthermore, there is some skewness in the residuals, which suggests that the distribution of the errors is not fully symmetric as would be desired. However, the skewness is lower than in the previous seasonal model. The skewness of this model is also visible in the histogram in Appendix III: Figure 4 and in the Q-Q plot in Appendix III: Figure 4, where the residuals do not align with the straight line. The AIC is 24551. Although this model has smaller skewness of the residuals and meets the condition of independence better, the AIC indicates a worse fit to the time series than in the previous seasonal model. Therefore, we select the seasonal model with no seasonal differencing for our main analysis of SARIMA models.



```
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              4400.39
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               0.85   Skew:                             0.39
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.13
```

Appendix III: Table 8: Statistical Assumptions Checking



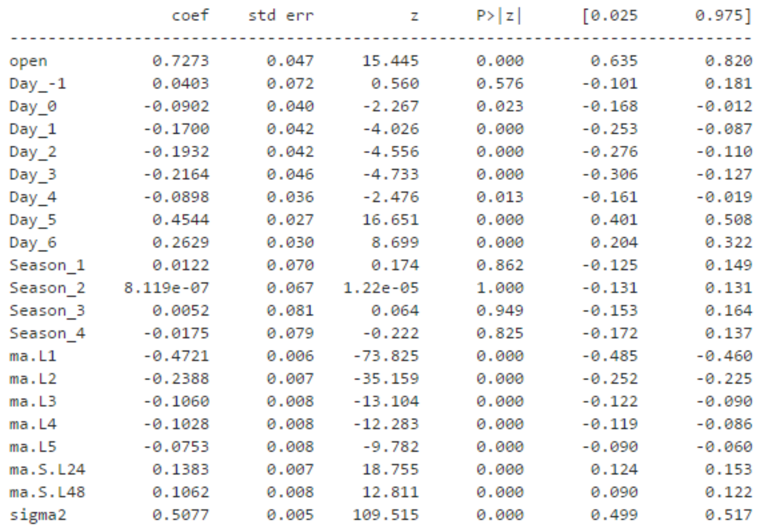
Appendix III: Figure 4: Statistical Assumptions Checking

III.III SARIMAX models comparison and coefficients

In order to determine the most fitting SARIMAX model, we test the SARIMA model with added exogenous predictors. From previous steps, we adopt no seasonal differencing, yet regular differencing is set to be maximally 1, and to be determined by the stepwise algorithmic search for the model with lowest AIC.

III.III.I SARIMAX without calls

The coefficients of the SARIMAX model without calls were obtained in the third step (see 3.1.1.3 Parameter estimation). From the coefficient values in Appendix III: Table 9, we can assess the role of individual predictors. The variable open that indicates if the emergency line has the highest coefficient value and is a significant predictor. This suggests that this variable is the most important predictor for the length of waiting time. We also find that the variables for seasons and the variable indicating a holiday day (Day -1) are not significant. This indicates that they do not contribute to the prediction of the waiting time much. Yet, this is likely due to the short span of data of over just one year. Therefore, the model might not have been able to catch any seasonal patterns on a yearly level. It is possible, if the data spanned a longer time period, that these predictors would be significant as well.

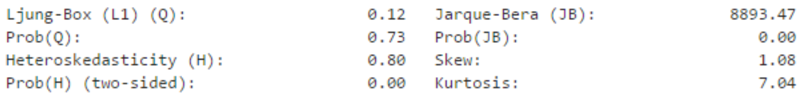


```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
open           0.7273      0.047     15.445      0.000       0.635       0.820
Day_-1         0.0403      0.072      0.560      0.576      -0.101       0.181
Day_0         -0.0902      0.040     -2.267      0.023      -0.168      -0.012
Day_1         -0.1700      0.042     -4.026      0.000      -0.253      -0.087
Day_2         -0.1932      0.042     -4.556      0.000      -0.276      -0.110
Day_3         -0.2164      0.046     -4.733      0.000      -0.306      -0.127
Day_4         -0.0898      0.036     -2.476      0.013      -0.161      -0.019
Day_5          0.4544      0.027     16.651      0.000       0.401       0.508
Day_6          0.2629      0.030      8.699      0.000       0.204       0.322
Season_1       0.0122      0.070      0.174      0.862      -0.125       0.149
Season_2    8.119e-07      0.067   1.22e-05      1.000      -0.131       0.131
Season_3       0.0052      0.081      0.064      0.949      -0.153       0.164
Season_4      -0.0175      0.079     -0.222      0.825      -0.172       0.137
ma.L1         -0.4721      0.006    -73.825      0.000      -0.485      -0.460
ma.L2         -0.2388      0.007    -35.159      0.000      -0.252      -0.225
ma.L3         -0.1060      0.008    -13.104      0.000      -0.122      -0.090
ma.L4         -0.1028      0.008    -12.283      0.000      -0.119      -0.086
ma.L5         -0.0753      0.008     -9.782      0.000      -0.090      -0.060
ma.S.L24       0.1383      0.007     18.755      0.000       0.124       0.153
ma.S.L48       0.1062      0.008     12.811      0.000       0.090       0.122
sigma2         0.5077      0.005    109.515      0.000       0.499       0.517
```

Appendix III: Table 9: Estimated Parameters

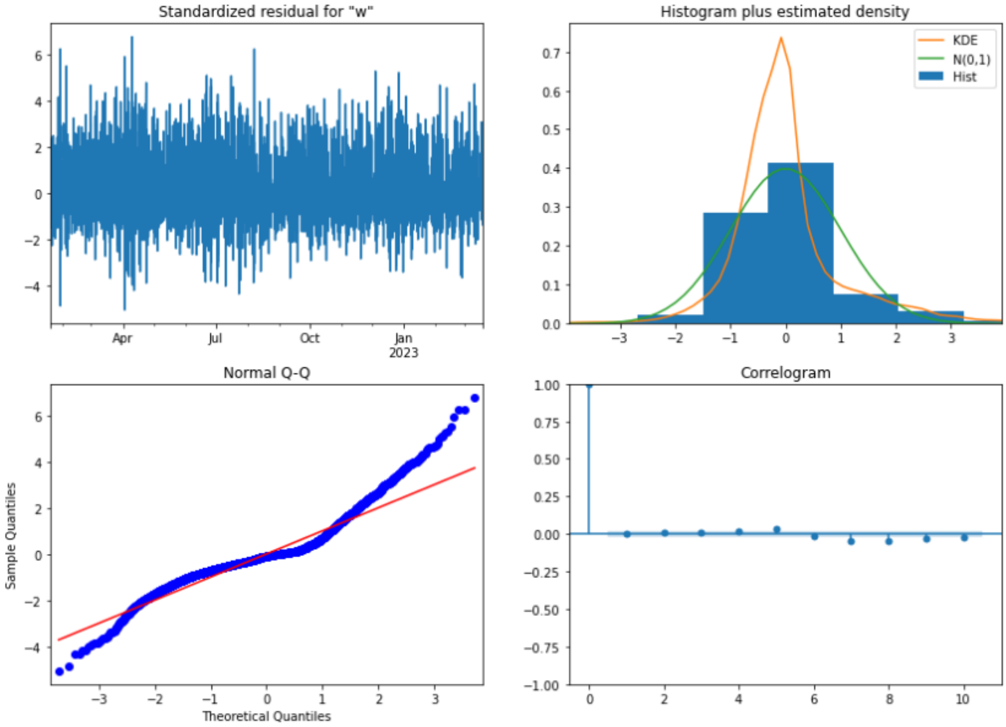
According to statistical assumptions checking (see section 3.1.1.4), the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. As can be seen in the statistical tests results in Appendix III: Table 10 and graphs in Appendix III: Figure 5 this model also does not show a perfect fit. According to the model summary Appendix III: Table 10, the model meets the condition of independence in the residuals because the p-value of the Ljung-Box test (Prob(Q)) is higher than 0.05. However, similarly to the previous models, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05, indicating a non-constant variance of the residuals. Furthermore, there is some skewness in the residuals, which suggests that the distribution of the errors is not fully symmetric as would be desired. The skewness of this model is also visible in the histogram in Appendix III: Figure 5and in the Q-Q plot in Appendix III: Figure 5, where the residuals do not align with the straight line.

However, as stated in the Results, this model has a much better performance on validation data than the previous models. Thus, the non-fulfillment of some assumptions for ARIMA-family models is secondary as in this applied problem, the good performance on a validation dataset is more important.



```
Ljung-Box (L1) (Q):                   0.12   Jarque-Bera (JB):              8893.47
Prob(Q):                              0.73   Prob(JB):                         0.00
Heteroskedasticity (H):               0.80   Skew:                             1.08
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.04
```

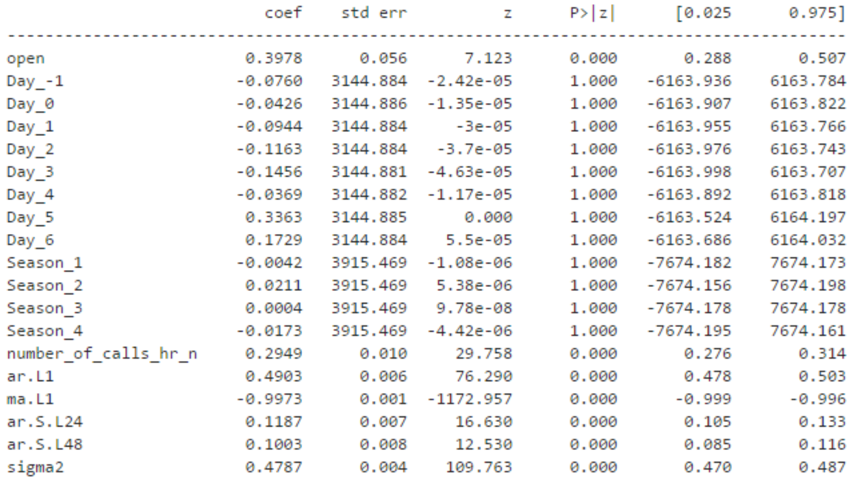
Appendix III: Table 10: Statistical Assumptions Checking



Appendix III: Figure 5: Statistical Assumptions Checking

III.III.II SARIMAX with calls

The coefficients of the SARIMAX model without calls were obtained in the third step (see 3.1.1.3 Parameter estimation). From the coefficient values in Appendix III: Table 11, we can assess the role of individual predictors. There is a crucial difference in significance of the predictors compared to the SARIMAX model without calls. When the number of incoming calls per hour is used as a predictor, all the other predictors become not significant except for the variable open. This is likely due to the variable number of calls capturing the trends that are otherwise indicated by the calendar variables. Thus, if the number of calls per hour is known or predicted with sufficient precision, the need to use calendar variables is expected to be low.

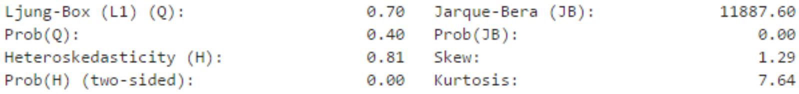


```
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
open                    0.3978      0.056      7.123      0.000       0.288       0.507
Day_-1                 -0.0760   3144.884  -2.42e-05      1.000   -6163.936    6163.784
Day_0                  -0.0426   3144.886  -1.35e-05      1.000   -6163.907    6163.822
Day_1                  -0.0944   3144.884      -3e-05   1.000   -6163.955    6163.766
Day_2                  -0.1163   3144.884    -3.7e-05   1.000   -6163.976    6163.743
Day_3                  -0.1456   3144.881  -4.63e-05      1.000   -6163.998    6163.707
Day_4                  -0.0369   3144.882  -1.17e-05      1.000   -6163.892    6163.818
Day_5                   0.3363   3144.885      0.000      1.000   -6163.524    6164.197
Day_6                   0.1729   3144.884     5.5e-05     1.000   -6163.686    6164.032
Season_1               -0.0042   3915.469  -1.08e-06      1.000   -7674.182    7674.173
Season_2                0.0211   3915.469   5.38e-06      1.000   -7674.156    7674.198
Season_3                0.0004   3915.469   9.78e-08      1.000   -7674.178    7674.178
Season_4               -0.0173   3915.469  -4.42e-06      1.000   -7674.195    7674.161
number_of_calls_hr_n    0.2949      0.010     29.758      0.000       0.276       0.314
ar.L1                   0.4903      0.006     76.290      0.000       0.478       0.503
ma.L1                  -0.9973      0.001  -1172.957      0.000      -0.999      -0.996
ar.S.L24                0.1187      0.007     16.630      0.000       0.105       0.133
ar.S.L48                0.1003      0.008     12.530      0.000       0.085       0.116
sigma2                  0.4787      0.004    109.763      0.000       0.470       0.487
```

Appendix III: Table 11: Parameter estimation

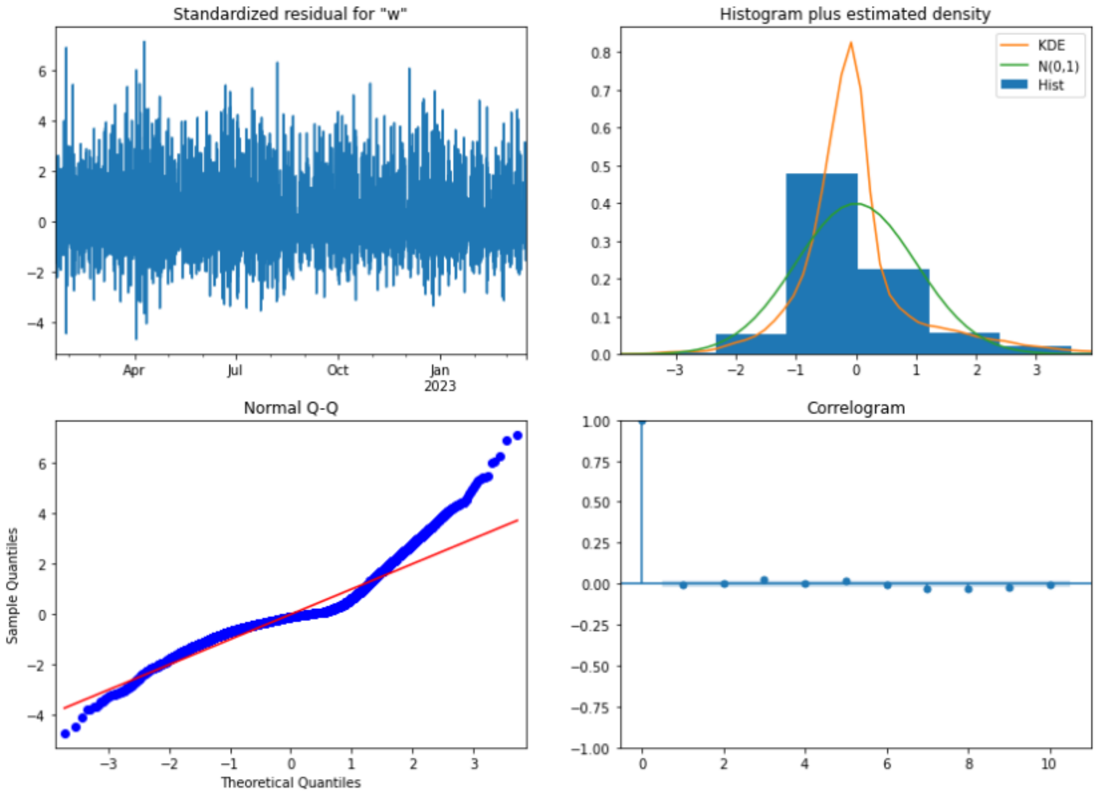
According to statistical assumptions checking (see section 3.1.1.4), the model using exogenous predictors including the number of incoming calls per hour to determine the average waiting time per hour does not show a perfect fit either. As can be seen in the statistical tests results in Appendix III: Table 12 and graphs in Appendix III: Figure 6, this model also does not have a perfect fit. According to the model summary in Appendix III: Table 12, the model meets the condition of independence in the residuals (no correlation) as the p-value of the Ljung-Box test (Prob(Q)) is higher than 0.05. This means that the residuals for this model are independent and meet the model's assumption. However, similarly to the previous models, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05, showing a non-constant variance of the residuals. Furthermore, there is some skewness in the residuals, which suggests that the distribution of the errors is not fully symmetric. The skewness of this model is also visible in the histogram in Appendix III: Figure 6 and in the Q-Q plot in Appendix III: Figure 6, where the residuals do not align with the straight line.

However, as stated in the Results, this model has the best performance on validation data than all of the previous models. Thus, the non-fulfillment of some assumptions for ARIMA-family models is secondary as in this applied problem, the good performance on a validation dataset is more important. We also select this model as the best performing model out of all the models tested.



```
Ljung-Box (L1) (Q):                   0.70   Jarque-Bera (JB):             11887.60
Prob(Q):                              0.40   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                             1.29
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.64
```

Appendix III: Table 12: Statistical Assumptions Checking



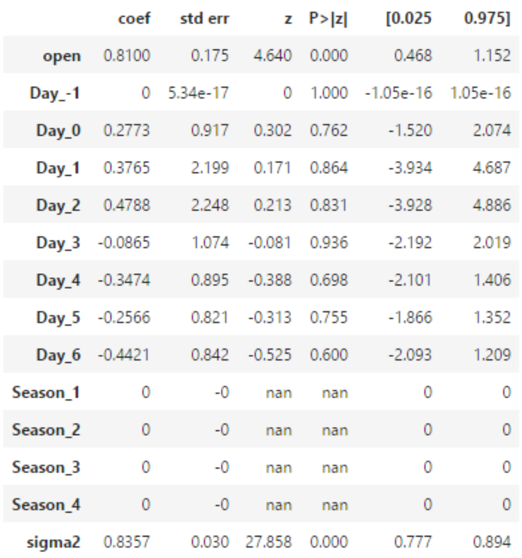
Appendix III: Figure 6: Statistical Assumptions Checking

III.IV Tentative Analysis on Filtered Data – Number of shifts

This analysis was performed on filtered data spanning from 1 March 2023 until 17 April 2023. The dataset used is equivalent to the training dataset used for all the other models with the added variable number of shifts for each hour. The following SARIMAX models include the following variables as predictors in a dummy form: open (indicating if the call line was open), day of the week (including the information on holiday) and season. The inclusion of the predictors number of calls and number of shifts is specified per model.

III.IV.I SARIMAX without calls and number of shifts

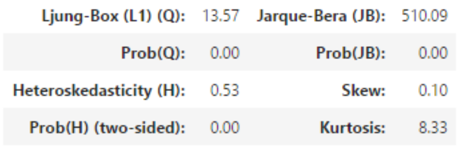
The SARIMAX model that did not include the number of shifts and the number of calls as predictors had the following results reported in this section. According to the suggested differencing order from the first step of the Box-Jenkins method (see 3.1.1.1 Data preparation and Investigation), SARIMAX was modelled with automatically-selected normal differencing in the range 0-1, and two options of seasonal differencing: 0 and 1. The model identified based on the lowest AIC was SARIMAX(0,1,0)(0,0,0)24 with AIC of 1173.74 (see 3.1.1.2 Model Identification). The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are reported in Appendix III: Table 13. The coefficients for the variables indicating season and Day-1 (corresponding to public holiday) were estimated as 0. This is due to the data spanning a very short period of time, in which the data only for one season (spring) were available and likely there were very few public holidays, deeming these variables were not used to fit the model. Similarly, the other variables that indicated weekdays were all not significant. This is also likely due to the model having too few observations to be able to learn a pattern based on these variables.



	coef	std err	z	P>|z|	[0.025	0.975]
open	0.8100	0.175	4.640	0.000	0.468	1.152
Day_-1	0	5.34e-17	0	1.000	-1.05e-16	1.05e-16
Day_0	0.2773	0.917	0.302	0.762	-1.520	2.074
Day_1	0.3765	2.199	0.171	0.864	-3.934	4.687
Day_2	0.4788	2.248	0.213	0.831	-3.928	4.886
Day_3	-0.0865	1.074	-0.081	0.936	-2.192	2.019
Day_4	-0.3474	0.895	-0.388	0.698	-2.101	1.406
Day_5	-0.2566	0.821	-0.313	0.755	-1.866	1.352
Day_6	-0.4421	0.842	-0.525	0.600	-2.093	1.209
Season_1	0	-0	nan	nan	0	0
Season_2	0	-0	nan	nan	0	0
Season_3	0	-0	nan	nan	0	0
Season_4	0	-0	nan	nan	0	0
sigma2	0.8357	0.030	27.858	0.000	0.777	0.894

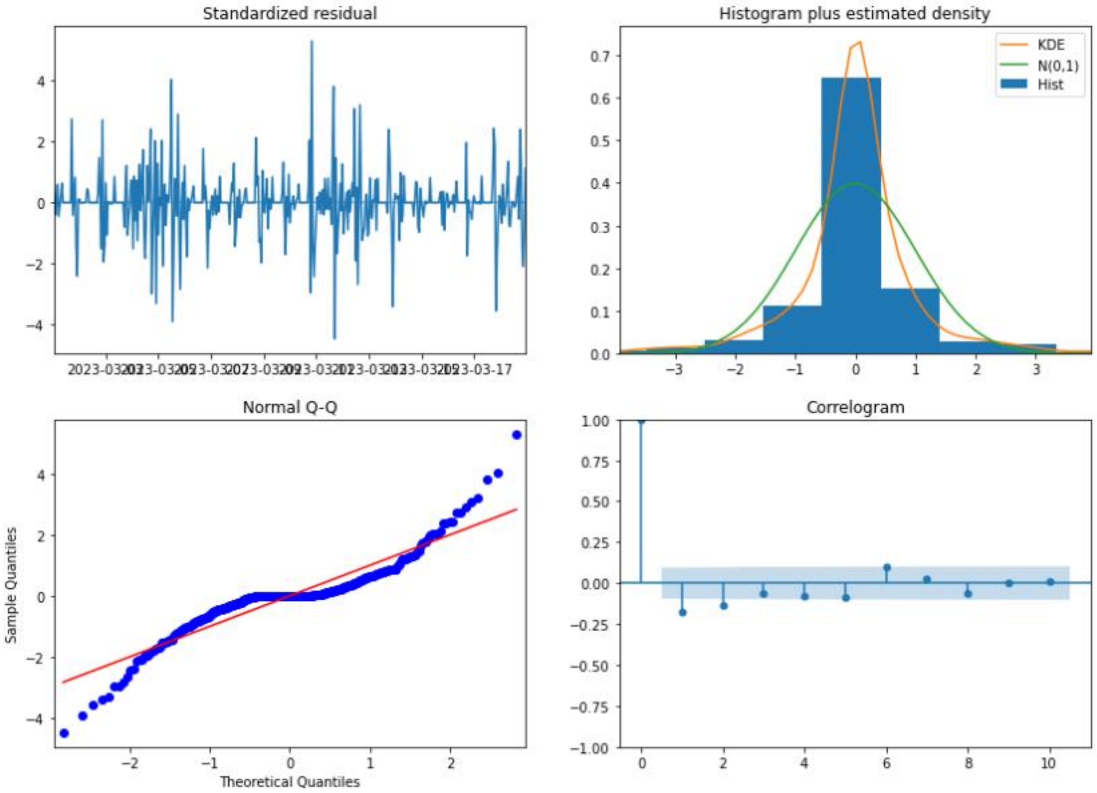
Appendix III: Table 13: Parameter Estimation

According to statistical assumptions checking (see part 3.1.1.4), the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. As can be seen in the statistical tests results in Appendix III: Table 14 and graphs in Appendix III: Figure 7, this model does not indicate a perfect fit. According to the model summary in Appendix III: Table 14, the model does not meet the condition of independence in the residuals (no correlation) because the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. The skewness of the residuals of this model is mostly visible in the Q-Q plot in Appendix III: Figure 7, where the residuals do not align with the straight line.



Ljung-Box (L1) (Q):	13.57	**Jarque-Bera (JB):**	510.09
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.53	**Skew:**	0.10
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.33

Appendix III: Table 14: Statistical Assumptions Checking



Appendix III: Figure 7: Statistical Assumptions Checking

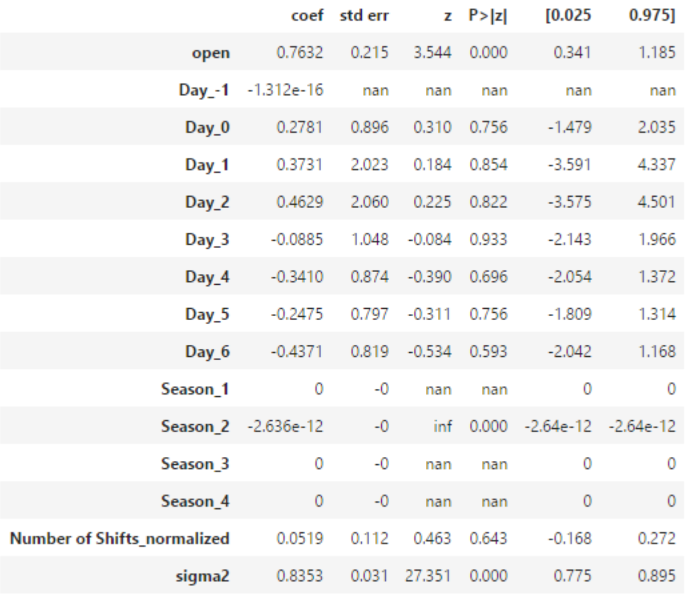
The results from forecasting and model validation (3.1.1.5) are available in Appendix III: Table 15. This model had the lowest MAE and RMSE on the long-range predictions and the highest error rates one day into the future.

	MAE	RMSE
1 day into future	6.79699	7.39479
1 week into future	6.21272	7.35057
1 month into future	5.84263	6.76676

Appendix III: Table 15: Forecast Accuracy of the SARIMAX model without callas and without the number of shifts

III.IV.II SARIMAX without calls, with number of shifts

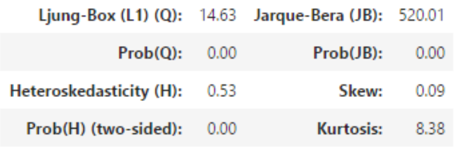
The SARIMAX model that did not include the number of calls, but it included the number of shifts available as predictors had the following results reported in this section. According to the suggested differencing order from the first step of the Box-Jenkins method (see 3.1.1.1 Data preparation and Investigation), SARIMAX was modelled with automatically-selected normal differencing in the range 0-1, and two options of seasonal differencing: 0 and 1. The model identified based on the lowest AIC was SARIMAX(0,1,0)(0,0,0)24 with AIC of 1175.57 (see 3.1.1.2 Model Identification). The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are reported in Appendix III: Table 16. The coefficients for the variables indicating season and Day-1 (corresponding to public holiday) were not estimated (except for season 2, which corresponds to spring). This is due to the data spanning a very short period of time, in which the data only for one season (spring) were available and likely there were very few public holidays, deeming these variables were not used to fit the model. Similarly, the other variables that indicated weekdays were all not significant. Moreover, also the variable number of shifts was not significant for estimating the model. This is likely due to the model having too few observations to be able to learn a pattern based on these variables.



	coef	std err	z	P>|z|	[0.025	0.975]
open	0.7632	0.215	3.544	0.000	0.341	1.185
Day_-1	-1.312e-16	nan	nan	nan	nan	nan
Day_0	0.2781	0.896	0.310	0.756	-1.479	2.035
Day_1	0.3731	2.023	0.184	0.854	-3.591	4.337
Day_2	0.4629	2.060	0.225	0.822	-3.575	4.501
Day_3	-0.0885	1.048	-0.084	0.933	-2.143	1.966
Day_4	-0.3410	0.874	-0.390	0.696	-2.054	1.372
Day_5	-0.2475	0.797	-0.311	0.756	-1.809	1.314
Day_6	-0.4371	0.819	-0.534	0.593	-2.042	1.168
Season_1	0	-0	nan	nan	0	0
Season_2	-2.636e-12	-0	inf	0.000	-2.64e-12	-2.64e-12
Season_3	0	-0	nan	nan	0	0
Season_4	0	-0	nan	nan	0	0
Number of Shifts_normalized	0.0519	0.112	0.463	0.643	-0.168	0.272
sigma2	0.8353	0.031	27.351	0.000	0.775	0.895

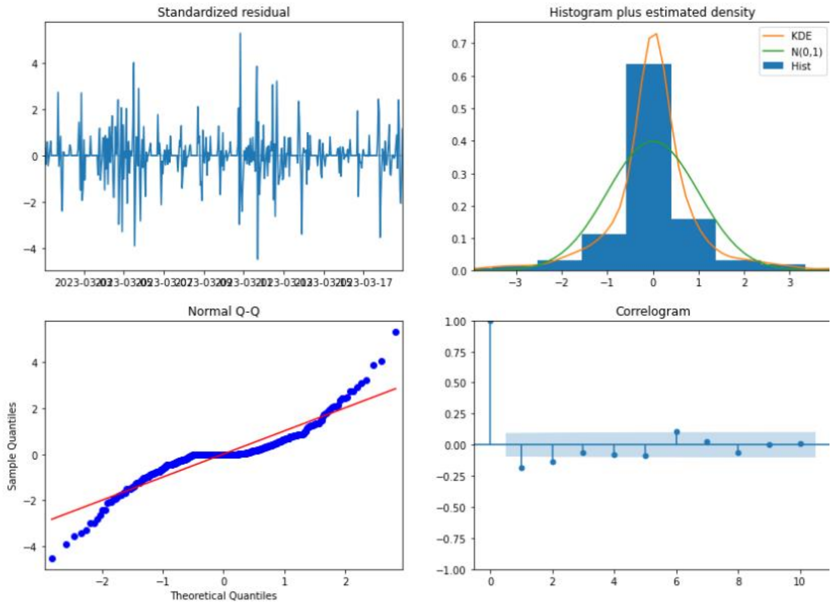
Appendix III: Table 16: Parameter Estimation

According to statistical assumptions checking (see part 3.1.1.4), the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. As can be seen in the statistical tests results in Appendix III: Table 17 and graphs in Appendix III: Figure 8, this model does not have a perfect fit. According to the model summary in Appendix III: Table 17, the model does not meet the condition of independence in the residuals (no correlation) since the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. Furthermore, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05. Thus, there is strong evidence of heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. The skewness of the residuals of this model is mostly visible in the Q-Q plot in Appendix III: Figure 8, where the residuals do not align with the straight line.



Ljung-Box (L1) (Q):	14.63	Jarque-Bera (JB):	520.01
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.53	Skew:	0.09
Prob(H) (two-sided):	0.00	Kurtosis:	8.38

Appendix III: Table 17: Statistical Assumptions Checking



Appendix III: Figure 8: Statistical Assumptions Checking

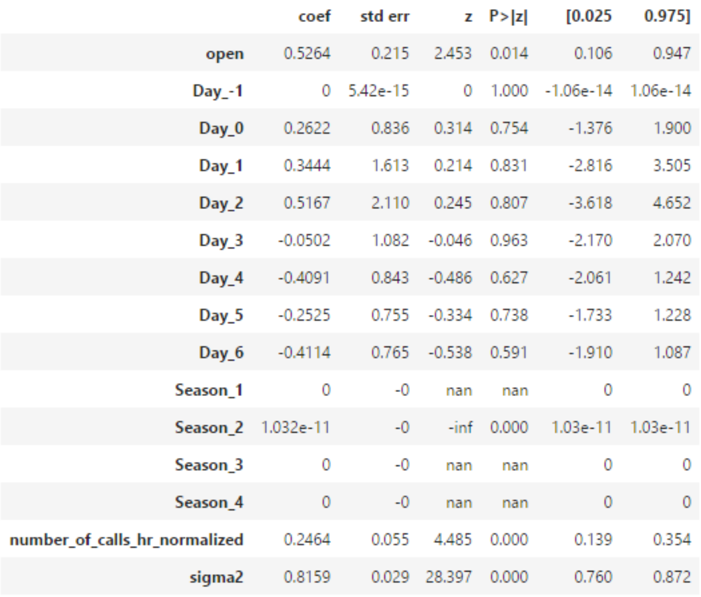
The results from forecasting and model validation (see 3.1.1.5) are available in Appendix III: Table 18. This model had the lowest MAE and RMSE on the long-range predictions. The highest error rate of this model in terms of MAE was one day into the future while in terms of RMSE it was one week into the future. However, the RMSE values were quite similar for the short- and mid-range predictions.

	MAE	RMSE
1 day into future	6.54315	7.11199
1 week into future	6.06342	7.20180
1 month into future	5.73979	6.64711

Appendix III: Table 18: Forecast Accuracy for the SARIMAX model without calls, with the number of shifts

III.IV.III SARIMAX with calls without number of shifts

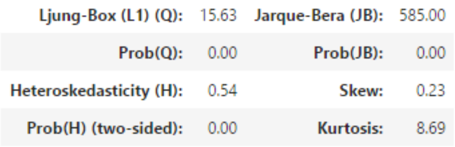
The SARIMAX model that did not include the number of shifts, but it did include the number of calls as predictors had the following results reported in this section. According to the suggested differencing order from the first step of the Box-Jenkins method (see 3.1.1.1 Data preparation and Investigation), SARIMAX was modelled with automatically-selected normal differencing in the range 0-1, and two options of seasonal differencing: 0 and 1. The model identified based on the lowest AIC was SARIMAX(0,1,0)(0,0,0)24 with AIC of 1165.42 (see 3.1.1.2 Model Identification). The estimated coefficients obtained in the third step (3.1.1.3 Parameter estimation) for each model parameter are reported in Appendix III: Table 19. The coefficients for the variables indicating season were not estimated (except for season 2, which corresponds to spring). This is due to the data spanning a very short period of time, in which the data only for one season (spring) were available and likely there were very few public holidays, deeming these variables were not used to fit the model. Even though Season2 corresponding to spring and Day-1 corresponding to public holiday were estimated and were significant based on their p-value, their coefficients were estimated to be 0. The other variables that indicated weekdays were all not significant. However, the variable number of calls was significant for estimating the model and had a relatively large coefficient. This indicates that the number of calls and the variable open were significant in fitting the model and also for future forecasting.



	coef	std err	z	P>|z|	[0.025	0.975]
open	0.5264	0.215	2.453	0.014	0.106	0.947
Day_-1	0	5.42e-15	0	1.000	-1.06e-14	1.06e-14
Day_0	0.2622	0.836	0.314	0.754	-1.376	1.900
Day_1	0.3444	1.613	0.214	0.831	-2.816	3.505
Day_2	0.5167	2.110	0.245	0.807	-3.618	4.652
Day_3	-0.0502	1.082	-0.046	0.963	-2.170	2.070
Day_4	-0.4091	0.843	-0.486	0.627	-2.061	1.242
Day_5	-0.2525	0.755	-0.334	0.738	-1.733	1.228
Day_6	-0.4114	0.765	-0.538	0.591	-1.910	1.087
Season_1	0	-0	nan	nan	0	0
Season_2	1.032e-11	-0	-inf	0.000	1.03e-11	1.03e-11
Season_3	0	-0	nan	nan	0	0
Season_4	0	-0	nan	nan	0	0
number_of_calls_hr_normalized	0.2464	0.055	4.485	0.000	0.139	0.354
sigma2	0.8159	0.029	28.397	0.000	0.760	0.872

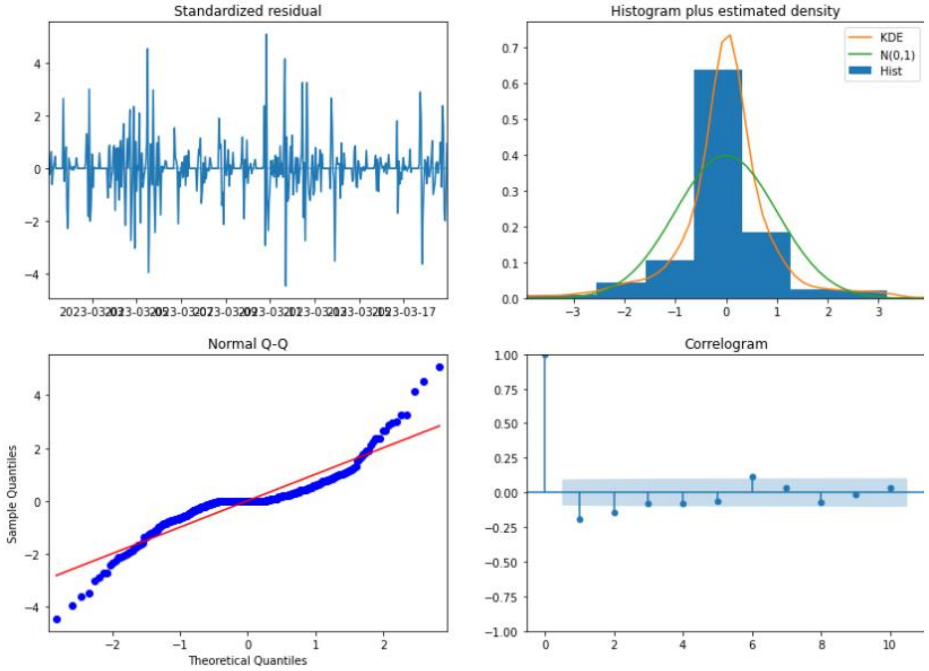
Appendix III: Table 19: Parameter Estimation

According to statistical assumptions checking (see part 3.1.1.4), the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. As can be seen in the statistical tests results in Appendix III: Table 20 and graphs in Appendix III: Figure 9, this model does not have a perfect fit. According to the model summary in Appendix III: Table 20, the model does not meet the condition of independence in the residuals because the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. This means that the residuals for this model are not independent, as assumed in the model. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates a heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. The skewness of the residuals of this model is mostly visible in the Q-Q plot and the histogram in Appendix III: Figure 9, where the residuals do not align with the straight line.



Ljung-Box (L1) (Q):	15.63	Jarque-Bera (JB):	585.00
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.54	Skew:	0.23
Prob(H) (two-sided):	0.00	Kurtosis:	8.69

Appendix III: Table 20: Statistical Assumptions Checking



Appendix III: Figure 9: Statistical Assumptions Checking

The results from forecasting and model validation (see part 3.1.1.5) are available in Appendix III: Table 21. This model had the lowest MAE on the long-range predictions. However, the RMSE was the lowest for short-range predictions. It is important to note that all the MAE values are very similar. This indicates that the model had a very stable performance.

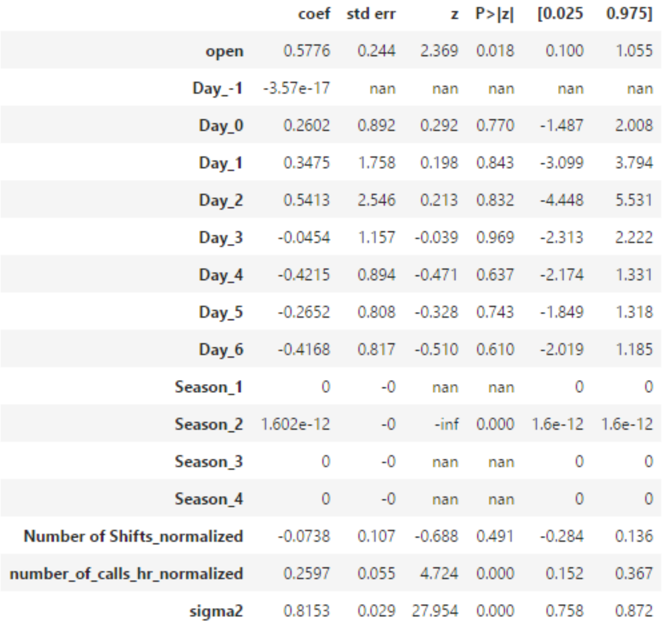
	MAE	RMSE
1 day into future	5.0587	5.49077
1 week into future	5.0978	6.41966
1 month into future	4.97602	5.97360

Appendix III: Table 21: Forecast Accuracy of the SARIMAX model with calls and without the number of shifts

III.IV.IV SARIMAX with calls and with number of shifts

The SARIMAX model that included both the number of shifts and the number of calls as predictors had the following results reported in this section. According to the suggested differencing order from the first step of the Box-Jenkins method (see 3.1.1.1 Data preparation and Investigation), SARIMAX was modelled with automatically-selected normal differencing in the range 0-1, and two options of seasonal differencing: 0 and 1. The model identified based on the lowest AIC was SARIMAX(0,1,0)(0,0,0)24 with AIC of 1167.10 (see 3.1.1.2 Model Identification).

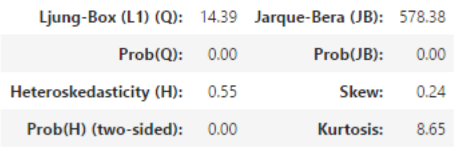
The estimated coefficients obtained in the third step (see 3.1.1.3 Parameter estimation) for each model parameter are reported in Appendix III: Table 22. The coefficients for the variables indicating season and Day-1 (indicating public holiday) were not estimated except for season 2, which corresponds to spring. However, the coefficient for this variable was very small indicating that it contributed very little to the model fit. This is due to the data spanning a very short period of time, in which the data only for one season (spring) were available and likely there were very few public holidays. This likely deemed these variables not useful to fit the model. The other variables that indicated weekdays were all not significant. Similarly, the variable number of shifts was not significant either based on its p-value. However, the variable number of calls was significant for estimating the model and had a relatively large coefficient. This indicates that the number of calls and the variable open were significant in fitting the model and also for future forecasting.



	coef	std err	z	P>|z|	[0.025	0.975]
open	0.5776	0.244	2.369	0.018	0.100	1.055
Day_-1	-3.57e-17	nan	nan	nan	nan	nan
Day_0	0.2602	0.892	0.292	0.770	-1.487	2.008
Day_1	0.3475	1.758	0.198	0.843	-3.099	3.794
Day_2	0.5413	2.546	0.213	0.832	-4.448	5.531
Day_3	-0.0454	1.157	-0.039	0.969	-2.313	2.222
Day_4	-0.4215	0.894	-0.471	0.637	-2.174	1.331
Day_5	-0.2652	0.808	-0.328	0.743	-1.849	1.318
Day_6	-0.4168	0.817	-0.510	0.610	-2.019	1.185
Season_1	0	-0	nan	nan	0	0
Season_2	1.602e-12	-0	-inf	0.000	1.6e-12	1.6e-12
Season_3	0	-0	nan	nan	0	0
Season_4	0	-0	nan	nan	0	0
Number of Shifts_normalized	-0.0738	0.107	-0.688	0.491	-0.284	0.136
number_of_calls_hr_normalized	0.2597	0.055	4.724	0.000	0.152	0.367
sigma2	0.8153	0.029	27.954	0.000	0.758	0.872

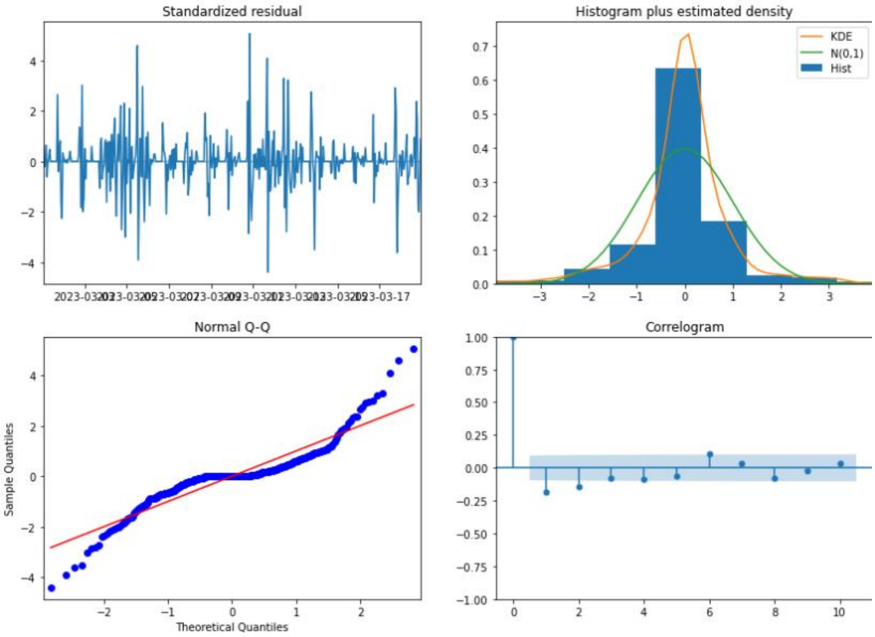
Appendix III: Table 22: Parameter Estimation

According to statistical assumptions checking (see 3.1.1.4), the model with the lowest AIC did not fully fulfill the statistical assumptions as its residuals were not independent with normal distribution. As can be seen in the statistical tests results in Appendix III: Table 23 and graphs in Appendix III: Figure 10, this model does not have a perfect fit. According to the model summary in Appendix III: Table 23, the model does not meet the condition of independence in the residuals because the p-value of the Ljung-Box test (Prob(Q)) is smaller than 0.05. This means that the residuals for this model are not independent, as assumed in the model. Moreover, the two-sided heteroskedasticity test (Prob(H)) indicates a p-value smaller than 0.05. Thus, there is strong evidence of heteroskedasticity in the residuals of our model, meaning that they do not have constant variance, as we would assume. The skewness of the residuals of this model is mostly visible in the Q-Q plot in Appendix III: Figure 10, where the residuals do not align with the straight line.



Ljung-Box (L1) (Q):	14.39	Jarque-Bera (JB):	578.38
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.55	Skew:	0.24
Prob(H) (two-sided):	0.00	Kurtosis:	8.65

Appendix III: Table 23: Statistical Assumptions Checking



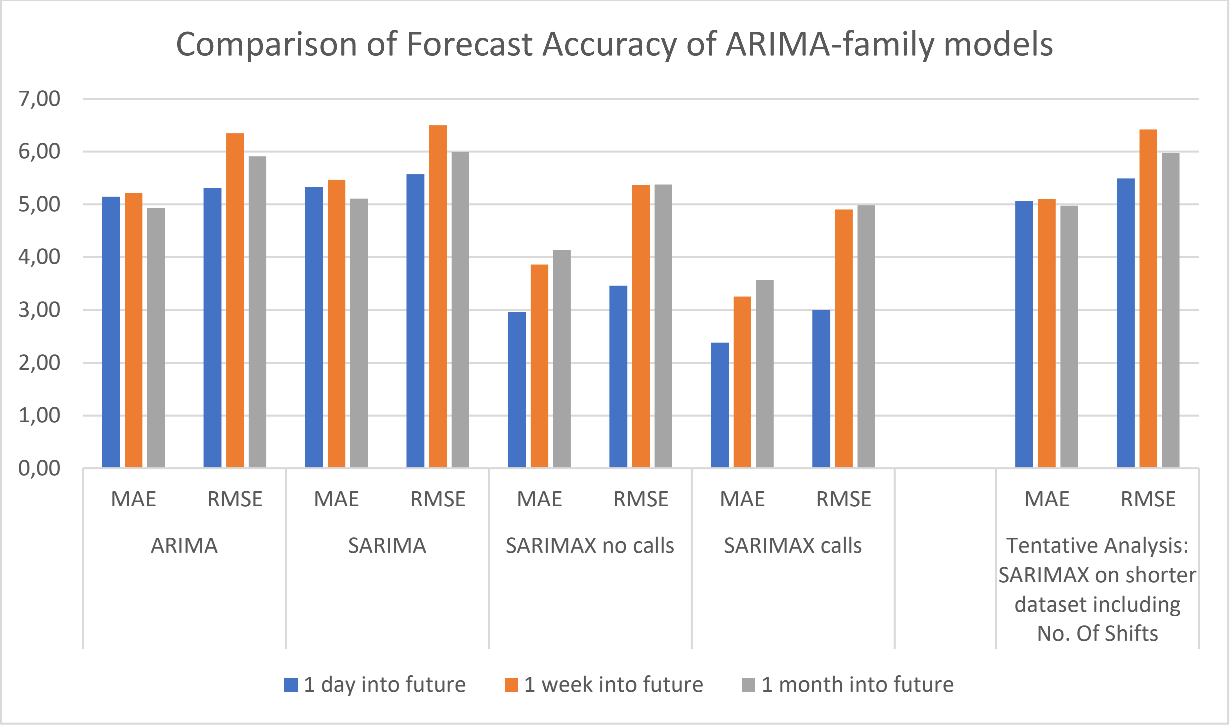
Appendix III: Figure 10: Statistical Assumptions Checking

The results of the SARIMAX model predictions are reported in Appendix III: Table 24. These results were obtained performing the fifth step of the Box-Jenkins method (see section 3.1.1.5). The forecasts of this model are quite stable over time. It has the lowest MAE on the long-range prediction, while RMSE is the lowest for short-range prediction.

	MAE	RMSE
1 day into future	5.32085	5.78049
1 week into future	5.24611	6.55977
1 month into future	5.06494	6.07399

Appendix III: Table 24: Forecast Accuracy of the SARIMAX model with calls and with the number of shifts

III.V Model Accuracy Comparison



Appendix III: Figure 11: Forecast Accuracy Comparison Among the ARIMA-family Models (tested on the same test dataset)