MULTIPLE TIME SERIES FORECASTING: COMPARING FACEBOOK'S PROPHET MODEL TO SARIMA

SARINA KASIEMKHAN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2068352

EXTERNAL SUPERVISOR

drs. P. Korteknie

COMMITTEE

dr. D. Hendrickson
dr. N. Venhuizen

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2023

WORDS

8798

MULTIPLE TIME SERIES FORECASTING: COMPARING FACEBOOK'S PROPHET MODEL TO SARIMA

SARINA KASIEMKHAN

Abstract

This thesis provides insights into the performance of Facebook's Prophet algorithm, SARIMA, and the moving average model for predicting parcel volume per area per day. In the available literature, these models have not been compared in this specific application before. The dataset used in this study is provided by PostNL and contains multiple time series data from January 2020 until November 2022. The comparisons conducted in this thesis reveal that Prophet generates more accurate predictions as evidenced by the error scores. Specifically, Prophet's predictions are found to be very close to the actual values, indicating that it captures seasonality well. On the other hand, SARIMA is observed to not handle multiple seasonalities well, as it keeps predicting values close to the mean. Additionally, Prophet is found to adapt well to the lockdown period, indicating that it handles sudden changes well. However, SARIMA seems to be a better model when it comes to generalizing to other areas, as Prophet fits too well on the data it trained on.

Data Source and Code

The owner of the dataset used for this thesis is PostNL. The author of this thesis does not have any legal claim to this data. Additionally, the data is not publicly available. Work on this thesis did not involve collecting data from human participants or animals. All figures and tables displayed in this thesis are produced by the author. The code written for this thesis was written by the author of the thesis and is available upon request.

1 PROBLEM STATEMENT AND RESEARCH GOAL

1.1 *Context*

Forecasting has become an integral part of every industry, especially for those dealing with seasonal items. It helps companies with managing expectations, making informed business decisions and developing new strategies. One of these companies is PostNL.

PostNL is a Dutch delivery company for mail and parcels. With over 4,800 locations and 11,000 postboxes, they are present in almost every area in the Netherlands. On average 8 million letters and 1.2 million parcels are collected every day. By predicting the volume in m^3 of parcels collected per day, per postal code (PC) area in the Netherlands, PostNL may better determine its staffing needs, optimize planning and reduce costs whilst keeping customers and retailers satisfied. A retailer refers to a drop-off point for customers, where they leave their parcel to later get picked up by a chauffeur who takes it to a depot, there it gets sorted and sent to its destination.

The objective is to investigate the performance of two time series prediction methods, Facebook's Prophet algorithm and the Seasonal Autoregressive Integrated Moving Average (SARIMA), whilst predicting parcel volume per area, per day. Volumes of the past two years for every PC and every day will be used to create the predictions. The Prophet algorithm is a flexible and intuitive method for predicting using a time series data. It is based on an additive model and is designed to handle trends, seasonality, and holidays (Taylor & Letham, 2018). SARIMA is a type of time series model that is used to capture the seasonality and temporal dependencies in data. It is based on a combination of autoregressive (AR) and moving average (MA) models, and includes a seasonal (S) component Divisekara, Jayasinghe, and Kumari (2021). The ability to capture seasonality and temporal dependencies can be useful for predicting parcel volume at PostNL, which may be affected by factors such as a lockdown.

1.2 *Thesis relevance*

Practical and societal relevance

The results of this thesis can help optimize resource allocation, predict demand, and improve efficiency. When PostNL knows how much m^3 of parcels is at a retailer they can plan accordingly and send the right number of chauffeurs to pick the parcels up. Sending too many chauffeurs results in unnecessary expenses and carbon emissions. Sending not enough chauffeurs and not picking up parcels means that retailers have less free space in

their shop. For customers it means that their parcel may not be delivered the next day as is promised by PostNL. Thus, not picking up parcels may result in a lower satisfaction from retailers and consumers. On average the daily volume is 5,445 m^3, chauffeurs can take approximately 4m^3 with them thus 1.361 pick-ups are needed per day, however, currently PostNL plans 20,087 pick-ups per day. This is excessive and can be reduced with accurate predictions,

Scientific relevance
The Prophet algorithm is relatively new and has not been as extensively reviewed compared to traditional machine learning and time series algorithms. In addition, literature where parcel volume per day or similar problems are predicted is scarce. Literature about predicting parcel volume using the Prophet and SARIMA models was not found. Investigating the performance of these models for predicting parcel volume is scientifically relevant because it may provide insights into the feasibility and effectiveness of these methods for this particular application. This research may also contribute to understanding of the strengths and limitations of these methods, as well as their potential for generalization.

1.3 *Research questions*

The performance and generalizability of Prophet and SARIMA for predicting package volume per PC area per day can be affected by a number of factors, including seasonalities and covid-19 related lockdowns. By considering these factors and evaluating the errors in the models, we may gain a better understanding of the strengths and limitations of these methods in different contexts. In particular, answering the questions outlined below can provide valuable insights into the performance and generalizability of the models.

Main research question

> *To what extent can we predict package volume per PC area per day using Facebook's Prophet algorithm?*

To answer the main research question several sub-questions are formed. We will compare the performance of Prophet, SARIMA and three baseline models. These baselines will be calculated using the moving average method on a 7-day, 30-day, and 90-day window. The models will predict the volume per day for every PC area. This leads to the first sub-question:

RQ1 *To what extent do the Prophet and SARIMA models outperform the baseline models when predicting the daily collection volume per PC area?*

In the exploratory data analysis, we observe that on weekends, parcel volume is typically lower than on other days of the week. To assess the ability of the models to capture this weekly seasonality, we will compare the predicted and actual values for volume on the weekdays. This forms the basis for the second sub-question:

RQ2 *Given the seasonalities identified in the exploratory data analysis, do the errors in the models capture these seasonalities?*

Additionally, as the dataset includes data from the period of the Covid-19 pandemic, during which there was a lockdown, we will evaluate the adaptability of the models to sudden changes. We will do this by comparing the error score during the lockdown period with the error score outside the lockdown period. This serves as the foundation for the third sub-question.

RQ3 *How big is the difference in errors in periods where there was a covid-19 related lockdown and no lockdown?*

Finally, to assess the generalizability and limitations of the models, we will evaluate their performance when trained on a single PC area and tested on the rest of the PC areas. The PC area that will be used to train on is an area that has volume values close to the mean of all PC areas. This leads us to the final sub-question

RQ4 *How well do the Prophet and SARIMA models generalize when training on only one PC area and testing on the rest of the PC areas?*

1.4 *Findings*

Both Prophet and SARIMA outperform the moving average baseline models. Generally, Prophet creates more accurate predictions, when evaluating using the RMSE. It is also better at capturing the weekly seasonality in the dataset, SARIMA predicts a value close to the mean throughout the entire period. Compared to SARIMA, Prophet is also the better prediction model when having sudden changes in your time series such as a lockdown period where all stores are closed. However, the error scores are not as good as for the period without a lockdown. Neither model generalizes well to other areas. To have accurate predictions it is better to have a separate model for every area. Nevertheless, SARIMA generalizes better than Prophet.

2 RELATED WORK

2.1 *Parcel delivery*

Accurate forecasting of parcel volume is crucial for carriers in order to optimize delivery schedules and anticipate staffing needs. Despite its importance, there is a lack of research on predicting parcel volume, particularly using the Prophet and SARIMA models. The objective of this literature review is to evaluate the performance of these models in predicting parcel volume per day per PC area, and to identify any gaps in the existing literature on this topic.

One approach to forecasting parcel volume is to examine the number of orders placed with stores that use a certain parcel carrier. Research has shown that purchase behavior may vary seasonally, with notable changes occurring at the end of the year (Zitzlsperger, Robbert, & Roth, 2009). Chen and Li (2020) found that holiday-related promotions and lower prices during this period can lead to an increase in sales volume. It is important to consider the potential impact of seasonality on forecast accuracy, as consumers with seasonal patterns in their transactions may affect the accuracy of the forecast (Zitzlsperger et al., 2009). This approach to forecasting parcel volume is based on the idea that changes in the number of orders placed with stores that use a parcel carrier may be an indicator of future parcel volume.

Another way to optimize delivery schedules and anticipate collection volume is by offering consumers the option to pick up their parcel from their home. While this can improve the efficiency of the delivery process, it does not address the need for carriers to anticipate staffing needs well in advance. Morganti, Seidel, Blanquart, Dablanc, and Lenz (2014) found that consumers expect to have their parcels picked up within a few days of requesting this service. To effectively plan staffing needs, carriers must have a reliable method for predicting the volume of parcels that will be collected. This may involve using statistical models or expert judgment.

Econometric models are statistical techniques that use economic data to analyze the relationships between economic variables (such as inflation, unemployment) and parcel volume. The study by Hu and Chen (2020) found that econometric models were able to achieve higher prediction accuracy for parcel volume in the Chinese express delivery industry. Expert judgment, on the other hand, involves relying on the knowledge and experience of experts in the field (such as logistics professionals or industry analysts) to make informed forecasts. A study by Kim and Lee (2018) found that expert judgment was a useful tool for predicting parcel volume in the Korean e-commerce industry. They conducted a survey of industry

experts and used their responses to forecast parcel volume. The study found that expert judgment was able to provide more accurate forecasts than multiple linear regression. However, both econometric models and expert judgment are limited in their ability to account for unexpected events or changes in the market (Hu & Chen, 2020; Kim & Lee, 2018). This is proved by the study by Abdelkader and Aloui (2013), they found that traditional time series models were more effective than econometric models in predicting tourism demand in Tunisia. The traditional time series models outperformed the econometric models in terms of forecasting accuracy. The authors attributed the superior performance of the traditional time series models to their ability to capture the seasonal and long-term trends in the data. Based on these results traditional time series forecasting methods, may be more effective at predicting parcel volume whilst taking seasonality into account.

2.2 *Traditional time series forecasting methods*

Time series forecasting is the process of using historical data to make predictions about future events. Time series forecasting methods take into account the fact that the value of a time series variable is often influenced by its past values, as well as any seasonal or cyclical patterns that may exist. De Gooijer and Hyndman (2006) reviewed 940 papers on time series forecasting methods over the past 25 years and found that exponential smoothing, , Autoregressive Integrated Moving Aver- age (ARIMA), Multiple Linear Regression (MLR) and non-linear models such as Neural Networks were the most commonly used methods. This chapter aims to review the existing literature on time series forecasting methods such as the methods mentioned above and evaluate their possible effectiveness in predicting parcel volume per day per PC area.

The article by Sulandari, Suhartono, and Subanar (2021) about the application of exponential smoothing on a dataset with multiple time series, concludes that exponential smoothing should be used for short-term forecasting as it may not be as effective at capturing long-term trends or seasonality. This is due to how the model works, it calculates a weighted average of past values, where more recent values are given higher weights (Sulandari et al., 2021). Exponential smoothing may not be the most effective method for predicting parcel volume, as it may not be able to capture long-term trends or seasonality. However, a hybrid model appears to have performed well in a Makrikadis competition (M-Competitions). The M-Competitions are international forecasting challenges aimed at evaluating and comparing the performance of different forecasting methods to solve a specific problem. The challenges often result in applying innovative tech-

niques (Makridakis, Spiliotis, & Assimakopoulos, 2020). The winner of the M4 competition published a paper on the models that he used. A hybrid approach using exponential smoothing in combination with a Recurrent Neural Network (RNN) resulted in the lowest symmetrical mean absolute percentage error (sMAPE). Exponential smoothing was used to capture the main components of the time series while the RNN tolerated non-linear trends. A popular machine learning algorithm was not used by the winner because, according to him, they are not meant to be used on time series data sets. When one wants to use one of the popular machine learning algorithms, the data would require heavy pre-processing and they would need to apply cross-learning (Smyl, 2020).

According to the study by Athiyarath, Paul, and Krishnaswamy (2020), that compared different forecasting methods on three datasets, Recurrent Neural Networks (RNN) performed the best for making short-term predictions (1-29 time periods). ARIMA and Seasonal Autoregressive Integrated Moving Average (SARIMA) were more effective for mid- and long-term predictions (30-300 time periods and 300+ time periods). The study found that the MLR model had the highest Root Mean Squared Error (RMSE) for short-, mid-, and long-term predictions, indicating that it performed the worst compared to the other models Athiyarath et al. (2020). While ARIMA is one of the best performing methods when making mid- and long-term predictions, it does not consider seasonality (De Gooijer & Hyndman, 2006). SARIMA is a combination of the auto regression, integration, and moving average models. It is similar to ARIMA, but includes an additional component for seasonality, making it more robust for time series forecasts with seasonal influences (Divisekara et al., 2021; Hyndman & Athanasopoulos, 2018). This can be confirmed by Parviz (2020), who compared two hybrid models. The first model was a combination of SARIMA and a support vector machine (SVM), the second was a combination of SARIMA and an artificial neural network (ANN). The mean absolute error (MAE) for the SARIMA-SVM models was 18.02 while the MAE for SARIMA-ANN was 23.88. Upon further investigation, he concluded that the accurate results were due to SARIMA and its seasonality component. However, a drawback of using an ARIMA-based model is that it can become quite complicated as there is a large number of parameters involved (Parviz, 2020). In addition it requires a considerable amount of data to yield high evaluation scores. Lastly, the model's performance drops when dealing with more than one seasonality (De Gooijer & Hyndman, 2006). In terms of generalisation, ARIMA based models perform quite well in comparison to generalised regression neural network (GRNN) models as was concluded in the research by Wang et al. (2021). In this research the second wave of Covid-19 infections in the US were predicted based on data in India.

In recent time series related literature SARIMA and Prophet are often compared. Prophet proves to make more accurate predictions compared to SARIMA when trying to predict air pollution in India. The RMSE for a SARIMA model was 4.12, the RMSE for Prophet was 3.78 and when the y values from the training data were log transformed, the RMSE was 3.54 (Samal, Babu, Das, & Acharaya, 2019). Similarly, in the study by Jha and Pande (2021), who predicted supermarket sales, the Prophet model outperformed ARIMA and Holt Winter's. The RMSE scores here were 65.8, 85.7 and 151.64 respectively. Ensafi, Amin, Zhang, and Shah (2022) compared thirteen different models to predict sales of a furniture shop. The dataset had 200 units of sales. In this study, a stacked long short term memory (LSTM) was the best performing model by far, its RMSE was 128.5 while the RMSE of Prophet was 194.92. The ARIMA and SARIMA scores were 282.5 and 205.7 respectively. LSTM is a powerful forecasting method, but it can take a long time to train on large datasets (Jha & Pande, 2021).

RMSE, Mean squared error (MSE) and mean absolute error (MAE) are commonly used evaluation metrics on time series data (Athiyarath et al., 2020; De Gooijer & Hyndman, 2006; Divisekara et al., 2021). In their study, Taylor and Letham (2018) used the RMSE as well as the mean absolute percentage error (MAPE) as evaluation metrics.

In conclusion, parcel volume has been predicted using econometric models and expert judgment. However, these models are limited in their ability to account trends and seasonality. On the other hand, traditional time series methods have demonstrated their ability to capture seasonal and long-term trends in data, which may make them more effective at predicting parcel volume. Further research is needed to explore the potential of Prophet and SARIMA for predicting parcel volume per day per PC area, and to identify any potential limitations or challenges in using these models for this purpose. This research aims to fill this gap in the literature and contribute to the development of more effective approaches for predicting parcel volume.

3 METHODOLOGY AND EXPERIMENTAL SETUP

3.1 *Models*

Parcel volume will be predicted using a baseline, Prophet and SARIMA model. Before building the models it would be useful to understand them. The baseline model is the moving average. It smooths out short-term fluctuations in data by taking the average of a set of values over a specified time period (window). For this study, we will be calculating the moving

average on a 7-day, 30-day, and 90-day window. This is what PostNL is currently using.

The open-source algorithm Prophet was created in 2017 by two data scientists who work for Facebook. The paper they published to explain the model is called 'Forecasting at Scale'. Their objective was to create a time series forecasting model that is easy to use with little knowledge about time series, but with enough flexibility for a wide range of applications. The makers claim that the model can detect and handle trends and (multiple) seasonalities well (Taylor & Letham, 2018). Prophet is an additive model using three components. Equation 1 shows the components of Prophet:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

The trend function *g(t)* uses a linear model with a fourier series to capture the overall upward or downward movement in the time series. It can be used to detect whether the parcel volume is generally increasing, decreasing, or remains stable over time. The periodic changes such as weekly and yearly seasonality are calculated using a piecewise linear or logistic growth model. It is represented by *s(t)*, and helps identify which days or weeks of the year are likely to have higher or lower parcel volume. The effects of holidays which occur on potentially irregular schedules are often difficult to predict, however they are accounted for in Prophet with *h(t)*. Using this component we can model how the holidays will affect parcel volumes and determine how likely it is that the volume will change on or around certain holidays or events in the future. Any changes that the model is unable to account for are represented by the error term ε_t (Samal et al., 2019; Taylor & Letham, 2018).

The other model that will be used is SARIMA, which is based on the ARIMA model thus it would be valuable to understand this model. ARIMA combines the three components, autoregression, integration and moving average. *Autoregression* uses the lagged values of the time series as input to predict the future value. It captures the relationship between the value of a variable at a certain point in time and its past values. The lagged values used to predict parcel volume and a motivation can be found under chapter 3.4.2 (ACF & PACF plots). The *integration* component refers to the fact that the data may need to be differenced in order to make it stationary, which means that the mean and variance of the data do not change. This will be further explained in chapter 3.4.2 (stationarity). The last component, *moving average* in an ARIMA model uses the residual errors, it captures the short term fluctuations in the time series data and helps make more accurate predictions. SARIMA includes a seasonal component, whereas ARIMA does not, thus making SARIMA a more suitable model to create

predictions when dealing with a time series with a repeating pattern. SARIMA is defined as:

$$SARIMA(p,d,q)(P,D,Q)m \tag{2}$$

where the p, d and q are non seasonal components of the model and P, D, Q and m are seasonal components of the model. The p and P refer to the non seasonal and seasonal auto regressive order, the d and D is the degree of first differencing involved to make the data stationary, the q and Q are the non seasonal and seasonal moving average. The m indicates the number of observations per year and heavily influences the seasonal components of the model. SARIMA performs well when the data has a clear, repeating seasonal pattern. When a dataset has multiple seasonalities, it becomes harder to identify the appropriate D and P which makes it more challenging to accurately predict (Divisekara et al., 2021; Hyndman & Athanasopoulos, 2018).

In conclusion, Prophet and SARIMA are two different time series forecasting methods. Prophet is an additive model that uses trend, seasonality, and a specific component for holidays. SARIMA is a linear model that uses autoregression, integration and moving average, with a seasonal component. Both models can handle trend, seasonality and temporal dependencies.

3.2 *Dataset description*

PostNL provided the dataset and is not publicly accessible. The dataset contains three features, the date, volume in m^3 and postal code. It has data from January 2020 until November 2022 and approximately 463,000 PC areas, for every day and every area there is a row which holds the volume data, summing up to 47,323,116 rows. The volume is determined based on scans given in a depot. The barcodes on the parcels are scanned by a machine which has a sensor that determines the size and weight of every parcel. The PC area is assigned to a parcel based on where the barcode was first collected and got scanned.

3.3 *Pre-processing*

3.3.1 *PC areas in NL*

In this thesis we are predicting the volume per day, per PC. More specifically, we are predicting the parcel volume that will be collected from retailers per PC area. The customer drops the parcel off at a retailer, a PostNL chauffeur picks the parcel up and takes it to a depot where it gets

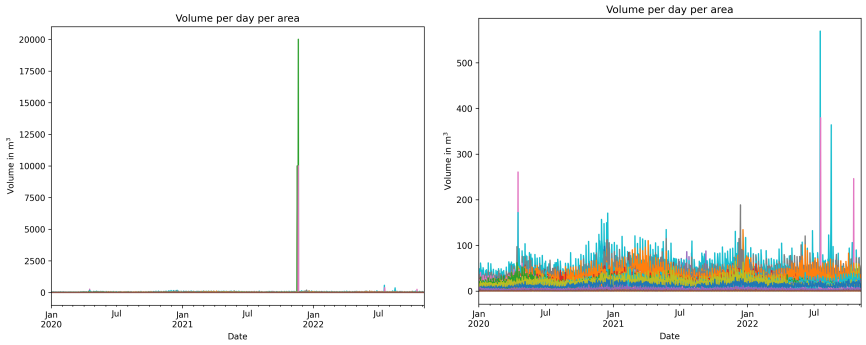
sorted and delivered to its destination. On average there are 5 retailers per PC, the volume of these retailers is summed up and put in a single row for this PC area, for this day. Furthermore, it is important to note that a postal code in the Netherlands contains six characters which is referred to as a level 6 PC, a level 3 PC refers to the first three characters. The initial dataset contains 463,000 level 6 postal codes. To reduce the number of time series in the dataset and predictions to be made, a new column with the first three characters is created, the dataset is then grouped by this new column and date, and the volume is summed up. As a result, there are 798 PC areas. Predictions will be made for those PC areas.

3.3.2 *Lockdown feature*

Analyzing how the models perform with a change in the dataset will be valuable, thus a feature explaining whether there was a Covid-19 related lockdown will be added to the dataset based on the date. Since there were different levels of a lockdown, in this thesis it refers to stores being fully closed. If there was a lockdown the value 1 is added and when there was no lockdown a 0 is added to the dataset. The details can be found in Appendix A (page 31).

3.3.3 *Outliers*

In figure 1, the volume per day per PC area is visualized. As can be seen in figure 1a there are some extreme outliers. Upon further inspection, it became evident that there was a system malfunction on this day at a certain depot. There were incorrect measurements for 4 parcels. To deal with this, in the original dataset the volume for these parcels is reduced to 0 while keeping the correct volume for the other parcels. The volume for these PC areas remains above 10 and is in line with the volumes for these PC areas before and after this day. After removing those outliers we are left with figure 1b. This plot displays the outliers which occurred because there was a big parcel volume, this occurred in five different PC areas. The plot shows that there was more than 200 m^3 collected on random days. These collection volumes are unusual and thus considered outliers and replaced with the mean volume of other PC areas for this day. After dealing with these 9 outliers, the daily collection volume per PC area appears more uniform as seen in figure 2.



(a) Original dataset with outliers where volume is bigger than 800 m^3

(b) Dataset where volume bigger than 800 m^3 is removed shows new outliers

Figure 1: Volume per day where every area is a different color displays outliers

3.4 *Exploratory data analysis*

3.4.1 *Multiple times series*

In this section the data of all PC areas is explored and analyzed. Figure 2 displays the volume per day per area in the dataset without outliers. Every color is a different PC area. As can be seen, there are a few areas who constantly have a relatively low volume while there are other areas where the volume is relatively high. What is noticeable is that the volume of all areas increases around the Black-Friday, Sinterklaas and New-Year period and decreases around the summer time meaning that all PC areas may be affected by seasonality.

In the bar plot in figure 3 the average volume per day per area is displayed. The volume is highest on workdays and is the lowest on the weekend, the volume is especially low on Sunday. On Monday the average retailer collects 7 m^3 of parcels and on Sunday it is not even 1 m^3, a possible explanation for this, is that retailers are closed on Sunday.

Figure 4 represents the distribution of the average volume per day, per PC area. As can be concluded from the figure, it is highly skewed to the right. The majority of the PC areas have a volume between 0 and 8 m^3 per day.

3.4.2 *Single time series*

The dataset used for this thesis is a multiple time series dataset, it contains the volume per PC per day. To simplify the EDA, the dataset is converted into a single time series, by grouping by date and aggregating the mean volume across PC areas. After resampling the dataset to a monthly basis, and plotting the volume (figure 5) we can see that the volume starts to

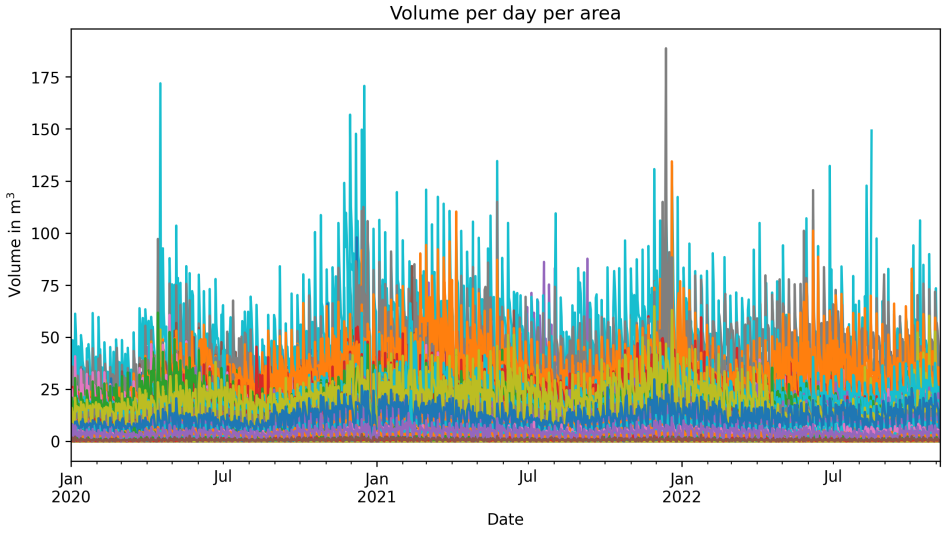


Figure 2: Daily volume where every color represents a different area

increase in September and keeps increasing until January. The volume seems to be the lowest in the summer period. The volume is in line with the findings of Zitzlsperger et al. (2009) in their research.

After resampling the data to a weekly basis and plotting the volume as can be seen in figure 6, we can see the changes in volume over time in greater detail. There is a strong decrease in volume in December 2020 -January in 2021, this might be a result of the sudden announcement from the Dutch government that all non-essential stores had to close due to the Covid-19 related restrictions. But the exact cause for this drop in volume cannot be confirmed with the available data.

Stationarity

ARIMA models assume that the time series data is stationary, which means that the mean and variance do not change over time. When the data is not stationary, it may have a trend or seasonal component which would invalidate the underlying assumptions of the model. If the data is not stationary, the model's parameters would change over time, making the predictions unreliable Hyndman and Athanasopoulos (2018). To confirm that the time series is non-stationary we can use a Unit root test such as the Augmented Dickey-Fuller (ADF) test. This is a statistical test where the null hypothesis is, that the time series is non-stationary, thus by rejecting the null hypothesis we can conclude that the time series is stationary. ADF is performed on the time series and the P- value is 0.0289, this value is smaller than 0.05 and thus we can reject the null hypothesis and assume

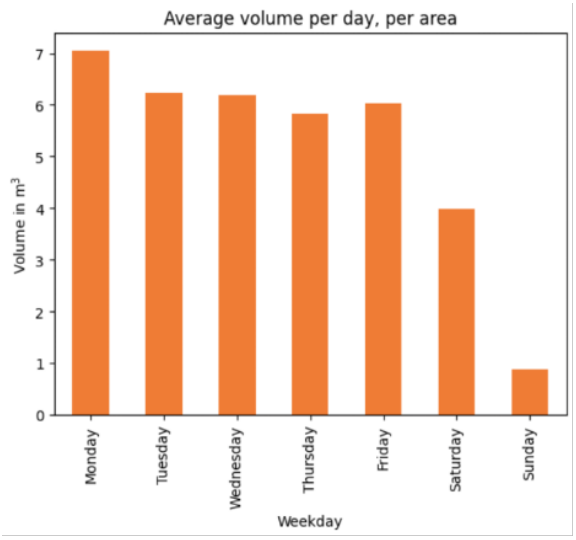


Figure 3: Average volume per weekday, the volume is lowest in the weekends and highest on Monday

that the dataset is stationary. A time series can be stationary, but still have seasonality. This means that while the mean and variance are constant over time, there are still regular and repetitive patterns in the data in a specific time period.

ACF and PACF plots

Autocorrelation is the correlation between a variable and itself at different points in time. It measures how closely related a variable is to itself over time. Partial correlation, on the other hand, is the correlation between two variables while controlling for the influence of other variables (Shumway & Stoffer, 2000). Figure 7 demonstrates the autocorrelation between the collection volume and k intervals (lags) and partial autocorrelation where the influence of other values is also accounted for. The light blue area in the plots displays the significance threshold and depicts the 95% confidence interval. Values outside this area suggest that there is likely a correlation. This is the case for the seventh, fourteenth, twenty first and twenty eighth have the highest correlation. This indicates that there is a weekly seasonality. We can also see that the more lags there are, the lower the correlation is. Thus, it is better to predict the future values with more recent values. This information aids in determining the autoregression values when building the model.

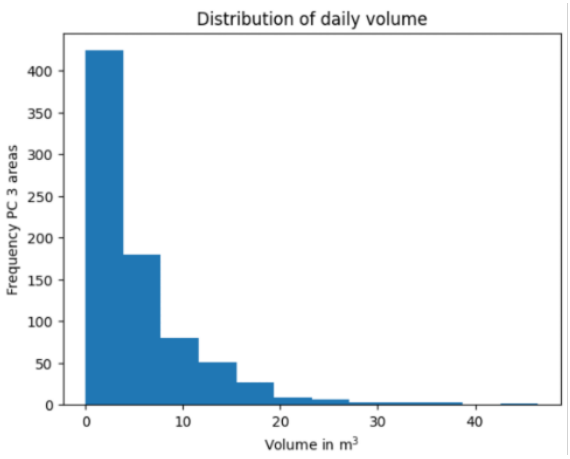


Figure 4: Right skewed distribution of volume in m^3

3.5 *Experimental procedure*

3.5.1 *Training, validation and test data*

Regression models will be built for this time series dataset using Moving average as the baseline, Prophet and SARIMA. The whole dataset will be split into a train (80%), validation (10%), and test set (10%). Since we are predicting per PC area, this results in 798 different train, validation and test sets. To plan which chauffeurs to send to which postal code area, PostNL needs to have the predictions 1 month in advance, thus the last portion of the dataset will be used for the validation and test sets. Non-essential stores were closed from 14 December 2020 until 8 February 2021 (RIVM, 2022), this period is referred to as the lockdown period in this thesis. To create a fair evaluation and to be able to analyze the errors between the lockdown and no lockdown period, the last two weeks of the lockdown period will be added to the previously made validation and test split and will be removed from the training data. The exact dates that the data was split on can be found in table 1 and is plotted in figure 8.
Figure 8 displays what the dataset looks like after the train, validation and test splits. The volume in the validation split looks similar to the training data, in the test split there is an outlier.

3.5.2 *Pipeline*

In figure 9 the pipeline used for this thesis is presented. The data will be pre-processed as was explained in chapter 3.3 In order to train the models, the data is split into train, validation and test data. The training data will be used to train the baselines, Prophet model and SARIMA. When

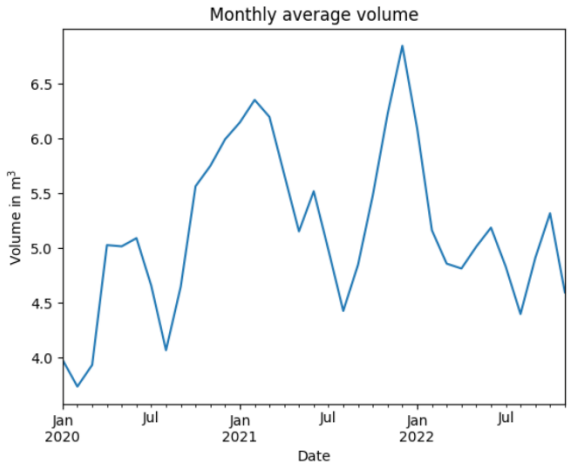


Figure 5: Monthly resampled data shows yearly seasonality

Table 1: Split dates for the dataset

Data	Date in mm/dd/yyyy format
Training data	01/01/2020 until 01/24/2021 02/08/2021 until 09/06/2022
Validation data	01/25/2021 until 01/31/2021 09/07/2022 until 10/08/2022
Test data	02/01/2021 until 02/07/2021 10/07/2022 until 11/06/2022

evaluating on the validation data, only the training data will be used to train the models, when evaluating on the test data, the training data as well as the validation data will be used to train the models. Every PC area will have its own baseline and model. These models will be optimized by performing grid search to find the best parameters and will be evaluated on the validation data. To test how well these models would perform with unseen data, they will be evaluated on the test set. With these results the first sub-question of this thesis can be answered. To answer the second sub-question about how well the models captured seasonality, the predicted and actual values of the models will be plotted and compared. To discover whether the lockdown had an effect on the predictability of volume and how well the models adapted to this period, the mean of the errors of the lockdown period and the mean error of the no lockdown period will be compared. Parallel to training the models for all areas, a model will be trained for the most average PC. The best parameters will be found using grid search and the models will be evaluated on the validation and test

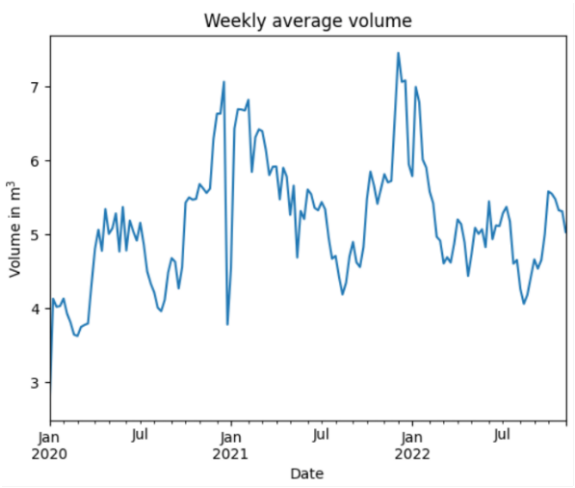
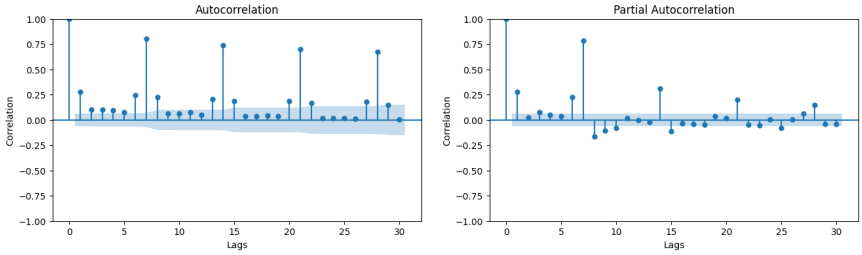


Figure 6: Weekly resampled data shows a strong decrease in volume in December 2020 -January in 2021



(a) ACF plot

(b) PACF plot

Figure 7: ACF and PACF plots used to determine SARIMA parameter values show us that the seventh lag has the highest correlation

sets of the other areas as well as on its own area. The results will tell us how well the models generalize to other areas and thus answer the fourth sub-question.

3.5.3 *Training models and grid search*

The algorithms used are Prophet and SARIMA. According to the literature Prophet is easy to use, quick and has accurate results. SARIMA was found to have relatively high evaluation scores when making mid-term predictions. A function will be created which takes the model, a list of parameters, the training and validation data as input and this function will loop over all different PC areas and a model will be fit with the best parameter combination, then a prediction will be made. The output of this function will be a list of predicted volume values for every PC area. This

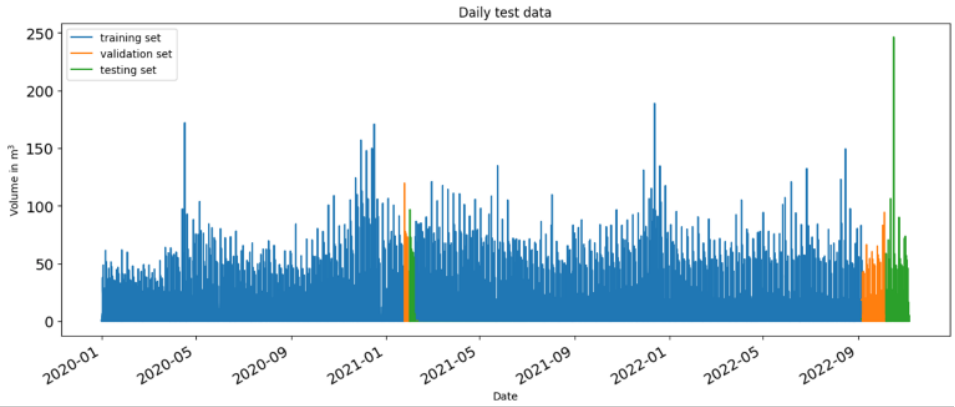


Figure 8: Training, validation and test set for all PC areas

list with predictions will then be compared to a list of actual values and the error scores will be calculated.

To find the best parameters for the models, grid search will be used. The parameters that will be experimented with for the Prophet model are listed in table 2. For every PC a model will be trained and every model will have its own set of parameters that yield the highest performance scores. A personalized dictionary of holidays will not be passed as a parameter because all the public holiday of the Netherlands are already in the default list of the model. Changepoints are the most impactful parameter, they determine the flexibility of the trend. If not done carefully it is easy to overfit. The seasonality controls the flexibility of the seasonality. The smaller the value, the smaller the magnitude of the seasonality. The values of the parameters are chosen based on the suggestions on the documentation of the library (Facebook Open Source, 2022).

The parameters that will be experimented with for the SARIMA model are listed in table 3. The p, d and q are non seasonal components of the model while P, D, Q and m are seasonal components of the model. The m indicates the number of periods in every season. As became clear from the unit root test in the EDA, the data is stationary thus differencing is unnecesary, the value for d and D is 0. From the ACF and PACF plots in figure 7 we noticed that there was a weekly trend where the seventh lag had the highest correlation, thus the m parameter will be set to 7. The other parameter values are chosen based on what was suggested in the library's documentation (Pmdarima, 2022).

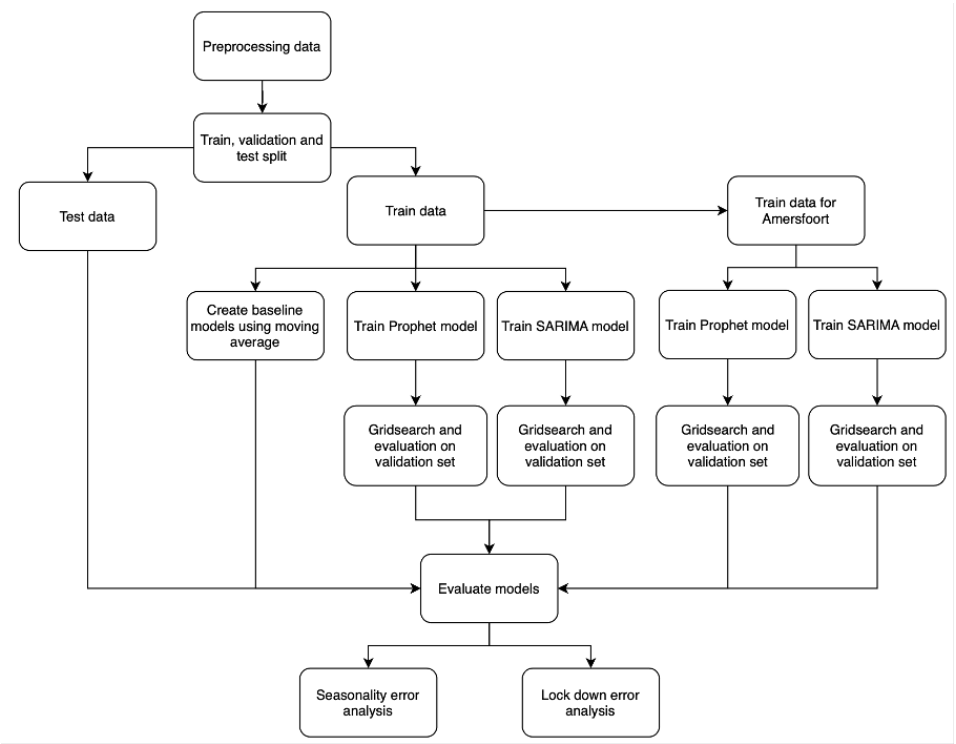


Figure 9: Data Science pipeline used for this thesis

3.5.4 *Evaluation method*

The models will be compared against each other and the baselines. The baselines will be calculated using the moving average method on a 7-day, 30-day and 90-day window. According to the literature, the most common evaluation methods for time series are MSE, RMSE and MAE. Since MSE and RMSE are quite similar, the MSE will not be used to evaluate the model scores. The primary evaluation score that will be considered is the RMSE.

As the dataset is a multiple time series and a model is trained and evaluated for every area, the best way to evaluate and compare the models is by calculating the mean of the error scores of all areas. Equation 3 shows how the RMSE of the models is calculated.

$$Mean\_RMSE = \frac{\sum RMSE_x}{N} \tag{3}$$

Table 2: Prophet parameter values used in gridsearch

Parameters for Prophet	Values
Changepoints prior scale	0.1 0.2 0.3 0.4 0.5
Seasonality prior scale	5 6 7 8 9 10

In the equation x represents the PC area. The sum of the error scores per PC area is divided by the number of areas. Equation 4 shows how the MAE is calculated. It is calculated similar to the mean RMSE score.

$$Mean\_MAE = \frac{\sum MAE_x}{N} \tag{4}$$

3.5.5 *Code implementation*

Code for this thesis is written in python using a Jupyter notebook, this is a personal preference. The notebook is written in AWS Sagemaker in a ml.m5.4xlarge instance which supports the use of large datasets and complex models. The most important libraries used are Pandas, Matplotlib, Prophet, Pmdarima and Statmodels. These libraries will aid in manipulating and plotting the data, performing exploratory data analysis, training the models and then evaluating them.

4 RESULTS

4.1 *Models*

Every PC area has its own set of parameters that result in the highest model scores. The most frequently best parameters for Prophet are 0.5 for the changepoints and 10 for the seasonality. These parameter values are equal to the default parameter values and for both parameters it means that no regularization has been applied. For SARIMA, the most common starting values for seasonal and non-seasonal moving average and auto

Table 3: SARIMA parameter values used in gridsearch

Parameters for SARIMA	Values
Start p	1 2
Start d	0
Start q	1 2
Start P	1 2
Start D	0
Start Q	1 2
m	7

regression is 2. These values indicate the number of lags in the stationary time series and in the forecasting error.

After the models were trained using these parameters, they were evaluated on the validation and test set. The corresponding RMSE and MAE scores as well as the baseline scores for the moving average are listed in table 4. Interestingly, the error scores of the baselines for the validation data are higher than for the test data. Between the scores of the moving average, using the volumes of the most recent 7 days results in slightly better scores. Both the Prophet and SARIMA outperform the baseline scores. Prophet has more accurate predictions on the test data compared to the validation data when looking at the RMSE and MAE scores. On the contrary, SARIMA has better predictions in the validation data compared to the test data. When evaluating all scores, Prophet has more accurate predictions.

The results in table 4 are the mean scores of all PC areas. For some areas the models had better predictions while for other areas the models had less accurate predictions. The highest RMSE score for a PC area using Prophet was 55.759 for the validation data and 63.676 for the test data. The lowest RMSE scores were 0.34 and 0.33 for the validation and test data. SARIMA's range of RMSE scores was smaller. The highest RMSE scores are 43.781 and 27.483 for the validation and test data. The lowest scores were 0.924 and 0.956 for the validation and test data.

Table 4: Validation and test scores of the baseline, Prophet and SARIMA models

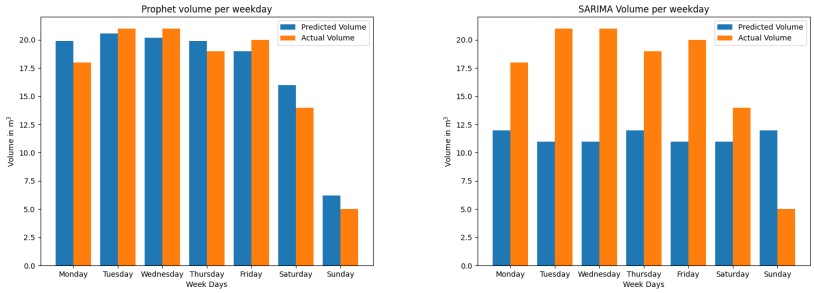
	Validation data		Test data	
	RMSE	MAE	RMSE	MAE
Moving average 7 days	6.347	3.060	5.395	3.050
Moving average 30 days	6.841	3.336	5.830	3.301
Moving average 90 days	6.980	3.443	5.969	3.415
Prophet	3.960	1.130	2.419	1.152
SARIMA	5.529	3.885	6.049	4.375

4.2 *Seasonality captured in the models*

A Prophet and SARIMA model were built per PC area, to simplify analyzing the predictions of the models, the data was grouped by date and aggregated by the mean of the predicted and actual volume. In the EDA monthly and weekly seasonality were identified. In this thesis the volume was predicted for one week in the lockdown period and the last month of the dataset. Since only the volume for one full month was predicted it is not possible to analyze whether the models captured the monthly seasonality, thus this section will be dedicated to analyzing only whether the models captured the weekly seasonality. It is important to point out that the average volume in the last month of the dataset is higher than the yearly average, due to the holiday season (Black Friday, Sinterklaas, Christmas).

As became evident in the EDA, in figure 3, on a weekly basis the volume was lowest on the weekends and higher throughout the week. In figure 10 the actual and predicted values are plotted for the Prophet and SARIMA models. It looks like Prophet did a better job at identifying on which weekdays more, or less volume would be collected compared to SARIMA. Prophet's predictions are quite close to the actual values, on Mondays and the weekends it predicts that there will be a higher volume than there actually is. SARIMA seems to predict roughly the same amount of volume for every day and on average predicts even more on Sundays than on other days. The values of the predicted volume made by SARIMA are similar to the mean values per weekday over the whole year as was plotted in the exploratory data analysis (figure 3).

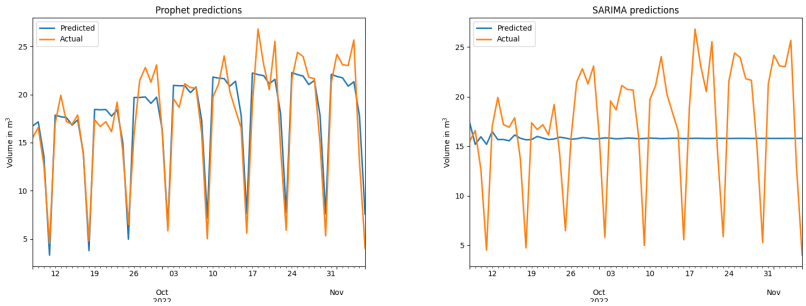
In figure 11 the predicted in contrast to the actual volumes are plotted for the validation data. The conclusions we can draw from figure 10 and 11 is similar. Prophet seems to follow the seasonality quite well while SARIMA seems to predict a value close to the mean of the actual values. There are some slight fluctuations at the beginning while towards the



(a) Prophet predicted volume per weekday (b) SARIMA predicted volume per weekday

Figure 10: Predicted and actual volumes grouped by weekday

end the predictions seem to be constant. Prophet (figure 11a) seems to predict a lower volume than the actual value for the weekends while at the end of the dataset it predicts more volume than the actual volume. On Mondays it predicts more at the beginning of the dataset while at the end the predictions for the Mondays are lower than the actual values. Overall, the line of the predicted volumes seems to follow the line of the actual volumes. When looking at figure 11b and considering that the model's predictions were grouped by date and aggregated using the mean, a question that might come to mind is whether the predictions for all PC areas are equal to the mean of the actual values, or whether half of the PC areas has a high predicted value while the other half has a low value which results in the mean score. The answer to this question is that the predictions for the analyzed PC areas have a value close to the mean volume of all PC areas.



(a) Prophet predicted vs. actual volumes (b) SARIMA predicted vs.actual volumes

Figure 11: Predicted and actual volumes from the validation data

Table 5: Model scores during the lockdown and no lockdown periods

	Validation data		Test data	
	RMSE	MAE	RMSE	MAE
Moving average 7 days errors for lockdown period	3.196	1.603	3.867	1.769
Moving average 7 days errors for no lockdown period	7.272	2.828	6.824	3.419
Moving average 30 days errors for lockdown period	3.235	1.709	3.840	1.765
Moving average 30 days errors for no lockdown period	7.517	3.789	6.4687	3.736
Moving average 90 days errors for lockdown period	3.401	1.781	3.995	1.996
Moving average 90 days errors for no lockdown period	8.054	3.920	6.586	3.247
Prophet model errors for lockdown period	2.182	1.181	2.059	1.111
Prophet model errors for no lockdown period	4.362	1.905	2.500	1.336
SARIMA model errors for lockdown period	6.052	4.125	6.624	4.660
SARIMA model errors for no lockdown period	3.212	2.824	3.501	3.114

4.3 *Lockdown analysis*

To discover whether the lockdown influenced the predictability of volume and how well the models adapted to this period, the mean of the errors of the lockdown period and the mean error of the no lockdown period will be compared using table 5. The baseline and Prophet models performed better during the lockdown period, with lower error scores compared to the no lockdown period. SARIMA, on the other hand, performed better during the no lockdown period, with a lower RMSE and MAE score compared to the lockdown period. Furthermore, it can also be seen that the baseline models performed poorer in both the lockdown and no lockdown periods compared to Prophet and SARIMA. Overall, Prophet is more robust in handling sudden changes such as the lockdown period, whereas SARIMA performs better in more stable conditions.

4.4 *Generalization of models*

To test the model's generalization capabilities, a single time series was trained with the date and volume features. The series that served as train data was PC area 381, the center of Amersfoort. It is the most average PC in the dataset. The overall mean of all PC areas was considered the true value and the volumes of all areas were compared to this true value using the RMSE. The RMSE value of PC 381 is 1.148, it is the lowest score, thus closest to the overall mean volume.

A single prediction was made with the baselines models Prophet and SARIMA, this prediction was compared with the actual values of Amersfoort and the actual values of other PC areas. The error scores are shown in table 6. Overall, we can conclude that none of the models generalize well to other PC areas. In all cases, the models have a lower error score

when the model was not generalized to other areas. When it comes to predicting for the same area, Prophet has the best score. SARIMA seems to generalize better than Prophet and the baselines, however, this might be due to constantly predicting a value close to the mean as was seen in section 4.2.

Table 6: Evaluation of how well the models generalize

	Validation data		Test data	
	RMSE	MAE	RMSE	MAE
Moving average 7 days on same PC area	4.755	3.609	6.597	5.241
Moving average 7 days on other PC areas	20.342	18.348	20.976	17.257
Moving average 30 days on same PC area	4.779	3.638	6.497	5.469
Moving average 30 days on other PC areas	21.357	18.673	21.349	17.378
Moving average 90 days on same PC area	5.019	3.528	7.022	6.211
Moving average 90 days on other PC areas	21.475	18.435	21.124	18.543
Prophet on same PC area	1.589	1.267	2.483	2.115
Prophet on other PC areas	12.357	11.746	14.710	14.127
SARIMA on same PC area	2.461	2.054	2.486	2.150
SARIMA on other PC areas	9.801	8.049	9.275	7.449

5 DISCUSSION AND CONCLUSION

5.1 *Summary and discussion of results*

This thesis aimed to determine the extent to which it is possible to correctly predict parcel volume per PC area, per day using the Prophet algorithm. This led to the following main research question:

> *To what extent can we predict package volume per PC area per day using Facebook's Prophet algorithm?*

To answer this question several sub-questions were formulated and will be answered in this section. To determine the extent to which Prophet and SARIMA outperform the baselines, the models were built and optimized using gridsearch to find the best parameters, they were then compared using the RMSE and MAE scores. The Prophet models that had the lowest error scores used parameter values 0.5 for changepoints and 10 for seasonality. These values were at the upper limit of the range tested. Therefore, it is suggested that future studies explore even higher values for these parameters. As expected based on the results of Jha and Pande (2021); Samal et al. (2019), Prophet outperforms the moving average baselines and SARIMA. The RMSE scores for Prophet were 3.960 and 2.419, whereas

the scores for SARIMA were 5.529 and 6.049 for the validation and test sets, respectively. The error scores for both Prophet and SARIMA are very similar to the scores of Samal et al. (2019) who predicted air pollution values in India. The error scores for the validation data are higher than for the test data for both models. These scores might indicate that there is a higher variation in volumes in the validation data, making it harder for the models to predict. Another reasonable explanation for the higher error scores on the validation data is that when the models were trained and evaluated on the validation data, only the training data was used. The models evaluated on the test set were trained using the training and validation data meaning that more data was available to the models to pick up on underlying trends and seasonalities. It is important to note that the results in table 4 were the mean scores of all PC areas, and for some areas the models had better predictions while for other areas the models had less accurate predictions. The range of scores for Prophet was larger than that of SARIMA, indicating that for some PC areas, SARIMA may be a better fit.

Taylor and Letham (2018) claiming that Prophet handles trends and seasonality well led to the second sub-question of this thesis. The results prove that the model indeed captures seasonality well, especially compared to SARIMA, which constantly predicts a value close to the mean volume of all PC areas. One possible reason for why SARIMA may not capture seasonality well in general, as explained by Divisekara et al. (2021), is that the seasonality in the dataset needs to be clear. When looking at figure 3, the volume on weekdays is consistently higher than on weekends, with Monday having approximately seven times more volume than Sunday. Thus, this reason may not be applicable in this thesis. Other than weekly seasonality, the dataset also contained monthly seasonality. Similar to what was found by Zitzlsperger et al. (2009) volumes are higher in the holiday season compared to the summer period. This leads to the second possible explanation. As written by De Gooijer and Hyndman (2006); Hyndman and Athanasopoulos (2018), the performance of the model drops when the dataset contains more than one seasonality. A solution to this would be to use a model which handles having multiple seasonalities well, such as Prophet. Lastly, a possible explanation to SARIMA's performance would be the parameters, especially the ones related to the moving average. The number of parameter values was limited in this thesis because it reduced the run time of the code. A suggestion for future research would be to experiment with more seasonal related parameters such as P and Q.

Compared to the actual values, Prophet seems to predict more volume in the weekends and Mondays while it predicts lower volumes on other days (figure 10a and 11a). PostNL aims for next-day delivery, so predicting

more volume and sending extra drivers to collect parcels is preferred to avoid leaving parcels at retailers' shops and delaying delivery. This also prevents dissatisfaction from retailers due to excess parcel storage. Since Prophet is the model that most closely predicts the volume and even predicts that there is more volume than there actually is on some days, this may be the better model for PostNL to use.

To answer the third sub-question, we compared how well the models adapted to the sudden changes such as a lockdown period. Overall we observed that both Prophet and SARIMA consistently outperformed the baseline models. The baseline models and Prophet perform better when there is a lockdown, whereas SARIMA performs better when the data is stable and there is no lockdown period. This conclusion is in line with what was claimed by Taylor and Letham (2018) about Prophet being able to capture changes in the data and SARIMA being able to create more accurate predicions when there is a clear pattern (Hyndman & Athanasopoulos, 2018).

The last sub-question was related to the generalization of the models. Overall, we can suggest to create separate models for every PC area to get more accurate predictions. However, as concluded by Wang et al. (2021) ARIMA based models are better to generalize to other areas. The error scores for SARIMA are the lowest, however, in this thesis this may also be due to SARIMA predicting volume values close to the average volume of all PC areas. An assumption that can be made about Prophet is that, it fits the training data too well and thus overfits when it comes to generalizing the model.

In conclusion, Prophet, SARIMA and baseline models were built to predict parcel volume per day, per area. The models were compared based on their prediction accuracy as well as how well they captured seasonality, adapted to sudden changes and how well they generalize to other areas. We have found that overall Prophet has the lowest error scores, the model captures seasonality well and performs well with sudden changes in the dataset. SARIMA predicts often predicts a value close to the mean which may be due to the multiple seasonalities in the dataset, or the parameters used to train the model. Overall it is better not to use a single model and generalize it to other areas. Prophet seems to fit very well on the data it trained on. However, SARIMA has lower error scores when generalizing and may be the better model to use in this case.

5.2 *Discussion of scientific and social impact*

This thesis made a novel contribution to the literature by comparing the performance of Prophet and SARIMA with baseline models to predict

parcel volume per day in m^3. The study found that Prophet outperforms SARIMA in predicting parcel volume, handles seasonality and sudden changes such as lockdowns better. Neither of the models should be used to generalize to other areas. By providing an analysis of these models in the context of parcel volume prediction, this research fills a gap in the literature. The findings of this thesis provide new insights into the use of these models for forecasting parcel volume and add to understanding the strengths and weaknesses of these models. These results can be used as a reference by researchers to make more informed decisions when choosing a model for their prediction problem.

This thesis presents valuable insights for the Retail department of PostNL. By comparing the predicted values, actual values and planned rides to a retailer on a daily basis, the study found that using the Prophet model for forecasting parcel volume could result in significant cost savings and efficiency gains for the company. Specifically, on the test dataset we calculated that if PostNL had used the Prophet model, they could have saved up to 591,330 rides to a retailer. This highlights the importance of accurate forecasting methods for the company, as it will help them optimize their operations and reduce unnecessary expenses.

5.3 *Limitations and future research*

This thesis used a dataset with multiple time series, and although efforts were made to make the data and models suitable for multiple time series, more accurate predictions may have been achieved by using a model specifically designed for multiple time series. Thus, future research may benefit from using deep learning methods, which are known to work better with multiple time series. Another suggestion for future research is to cluster the dataset based on PC areas. By clustering the data, it would be possible to compare which areas have similar parcel volume patterns and then use one area per cluster to make predictions for the other areas in the same cluster. This would enable the models to better generalize and make more accurate predictions. Lastly, due to the large number of predictions that needed to be made and the limited computational resources and time available it was not possible to predict parcel volume using an LSTM model. However, as was concluded by Ensafi et al. (2022) LSTM models are powerful tools and generate more accurate predictions compared to Prophet and SARIMA. Thus it would be suggested to use this model in future research.

5.4 *Conclusion*

In conclusion, this thesis has provided a detailed analysis of the performance of Prophet and SARIMA models for predicting parcel volume per day in m^3. The results showed that Prophet outperforms SARIMA in predicting parcel volume, captures seasonality and sudden changes such as lockdowns better. However, SARIMA tends to predict a value close to the mean and is better when generalizing to other PC areas. By providing a detailed analysis of these models in the context of parcel volume prediction, this research fills a gap in the literature and provides new insights into the use of these models for forecasting parcel volume.

The study also highlighted the limitations of the models and suggested future research directions. Future research should consider using models that are designed to work well with multiple time series data, or clustering the dataset based on PC areas to better generalize the predictions. Additionally, it would be beneficial for future research to consider using more advanced models such as LSTM which have been found to generate more accurate predictions compared to Prophet and SARIMA.

Overall, this thesis demonstrates the potential of the Prophet model for forecasting parcel volume and provides valuable insights for organizations in the parcel delivery industry and researchers in the field of time series forecasting.

REFERENCES

Abdelkader, Z., & Aloui, C. (2013). Forecasting tourism demand: a time series approach. *Tourism Management*, *35*, 28–38.

Athiyarath, S., Paul, M., & Krishnaswamy, S. (2020). A comparative study and analysis of time series forecasting techniques. *SN Computer Science*, *1*(3), 1–7.

Chen, C., & Li, X. (2020). The effect of online shopping festival promotion strategies on consumer participation intention. *Industrial Management & Data Systems*. doi: https://doi.org/10.1108/IMDS-11-2019-0628

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, *22*(3), 443–473.

Divisekara, R. W., Jayasinghe, G., & Kumari, K. (2021). Forecasting the red lentils commodity market price using sarima models. *SN Business & Economics*, *1*(1), 1–13. doi: https://doi.org/10.1007/s43546-020-00020-x

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – a comparative analysis. *International Journal of Information Management Data Insights*, *2*(1),

100058. Retrieved from https://www.sciencedirect.com/science/article/pii/S2667096822000027 doi: https://doi.org/10.1016/j.jjimei.2022.100058

Facebook Open Source. (2022). *Diagnostics.* Retrieved 07.10.2022, from https://facebook.github.io/prophet/docs/diagnostics.html

Hu, X., & Chen, Y. (2020). Forecasting parcel volume of express delivery industry in china: An empirical study. *Transportation Research Part E: Logistics and Transportation Review*, *139*, 101791.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Jha, B. K., & Pande, S. (2021). Time series forecasting model for supermarket sales using fb-prophet. In *2021 5th international conference on computing methodologies and communication (iccmc)* (pp. 547–554).

Kim, H., & Lee, H. (2018). Comparison of expert judgment and statistical models for forecasting package volume in the express delivery industry. *Transportation Research Part E: Logistics and Transportation Review*, *116*, 240–250. doi: https://doi.org/10.1016/j.tre.2018.06.005

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207019301128 (M4 Competition) doi: https://doi.org/10.1016/j.ijforecast.2019.04.014

Morganti, E., Seidel, S., Blanquart, C., Dablanc, L., & Lenz, B. (2014). The impact of e-commerce on final deliveries: alternative parcel delivery services in france and germany. *Transportation Research Procedia*, *4*, 178–190.

Parviz, L. (2020). Comparative evaluation of hybrid sarima and machine learning techniques based on time varying and decomposition of precipitation time series. *Journal of Agricultural Science and Technology*, *22*(2). Retrieved from http://jast.modares.ac.ir/article-23-26018-en.html

Pmdarima. (2022). *Pmdarima auto arima.* Retrieved 21.10.2022, from http://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html?highlight=auto%20arima

RIVM. (2022). *Tijdlijn van coronamaatregelen.* Retrieved from https://www.rivm.nl/gedragsonderzoek/tijdlijn-maatregelen-covid

Samal, K. K. R., Babu, K. S., Das, S. K., & Acharaya, A. (2019). Time series based air pollution forecasting using sarima and prophet model. In (p. 80–85). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3355402.3355417 doi: 10.1145/3355402.3355417

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3). Springer.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75-85. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207019301153 (M4 Competition) doi: https://doi.org/10.1016/j.ijforecast.2019.03.017

Sulandari, W., Suhartono, & Subanar. (2021). Exponential smoothing on modeling and forecasting multiple seasonal time series: An overview. *Fluctuation and Noise Letters*, *20*(04), 2130003. doi: https://doi.org/10.1142/S0219477521300032

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37-45. Retrieved from https://doi.org/10.1080/00031305.2017.138008 doi: 10.1080/00031305.2017.1380080

Wang, G., Wu, T., Wei, W., Jiang, J., An, S., Liang, B., ... Liang, H. (2021). Comparison of arima, es, grnn and arima–grnn hybrid models to forecast the second wave of covid-19 in india and the united states. *Epidemiology and Infection*, *149*, e240. doi: 10.1017/S0950268821002375

Zitzlsperger, D. F., Robbert, T., & Roth, S. (2009). Forecasting customer buying behaviour 'controlling for seasonality'. In *Proceedings of the anzmac conference.*

APPENDIX A

Below is the timeline related to Covid-19 lockdown periods. This list was used to create the lockdown feature in the dataset. Lockdown refers to stores being fully closed with no options for ordering and picking up from the store (click & collect) and no option to make an appointment to shop in the store. If there was a lockdown the value 1 is added and when there was no lockdown a 0 is added to the dataset.

- 14 Dec. 2020 non-essential stores closed
- 8 Feb. 2021 Click & collect
- 3 Mar. 2021 consumers can make an appointment to enter the store
- 28 Apr. 2021 stores are open again with no need for an appointment, opening hours until 20:00
- 26 Jun. 2021 no limitations to opening hours anymore
- 12 Nov. 2021 non-essential stores open until 20:00

- 25 Feb. 2022 normal opening hours again

RIVM (2022)