A Time Series Analysis

On

Rejsekort Check-ins

Course: Predictive Analytics

CBS
COPENHAGEN
BUSINESS SCHOOL
HANDELSHØJSKOLEN

MSc. Business Administration and Data Science

Submitted By: S M Ahasanul Karim - S158793

Number of Characters (with spaces): 13,425

August 18th, 2023

Abstract

This paper explores the predicted number of rejsekort journeys in Denmark in 2023 using several time series analysis models. Although Rejsekort check-in numbers do not represent all the commuting passengers, it corresponds to the biggest percentage of them. Accurately predicting future journeys from the check-in data in different seasons and situations is an essential aspect for companies like Nordjyllands Trafikselskab, Midttrafik, Sydtrafik, Fynbus, BAT, Moviatrafik and the Danish Public Transport Operators so that they can introduce more means of transport accordingly while also determining whether the fares should be increased to achieve a certain annual profit margin. This study utilizes a comprehensive dataset of 3652 observations from all the daily rejsekort check-ins from January 2013 to December 2022 while taking into consideration data from 2021 to make the forecasts due to covid-19 structural break. To retain the forecasts several ARIMA models have been tried out along with a basic ETS model. This paper highlights the significance of Predictive Analytics as a mandatory tool for transport data prediction. Successful application of these time series models opens the door to enhancing accuracy for pricing and enabling transport companies to further investigate when they should increase their means of transport leading to better passenger experience. Further advancements in this field can contribute to the development of automated systems that assist transport authorities in effectively identifying and treating these issues.

Keywords: Predictive Analytics, Time Series Analysis, SARIMA, ETS, Rejsekort check-ins, Danish Public Transport.

Introduction

The Danish public transport system is greatly aligned by rejsekort, a unique digital electronic ticketing system card which covers all means of public transport. Rejsekort aggregates and automates ticketing for all the zones, time intervals, and discount schemes to unite all passengers and public travel operators in harmony. This system dating back to 2003 introduced by the Danish Transport Operators, now has more than 2.5 registered users who represent half of the Danish population and are spread across all around Denmark while mostly living in Copenhagen.[1] Rejsekort is jointly owned by DSB; Copenhagen Metro, Movia, Nordjyllands Trafikselskab, Midttrafik , Sydtrafik and FynBus. It was regionally implemented in 2011 and has now been implemented for journeys all across the nation. There are also many types and forms of Rejsekort like rejsekort personal, rejsekort flex, rejsekort anonymous, rejsekort business and rejsekort commuter card with different perks and benefits.[2]

It can be easily inferred that the aggregate ticket prices for the rejsekort journeys hold the largest portion of revenue for the Danish Public Transport Operators and is a prime estimator for the annual profit calculation. According to DSB passenger revenue was 1412 million DKK in 2022, which is 185 million DKK more than the previous year. Also, the recorded 163.7 million journeys were a 39 percent increase from the previous year. However, this increase could not bring about more profit. The profit for the year 2022 was only DKK 229 million in comparison to the profit of DKK 805 million in 2021. This was due to the fact of increased prices of diesel and electricity. Although many efforts to reduce consumption, DSB experienced a 344 million DKK higher energy expense than in 2021.[3] To address this issue, DOT has increased the ticket prices by an average of 4.9 percent from January 15th.[4] According to data released by the DOT, regular Rejsekort passengers will pay an average of 8.6 percent more for their trips. The price of a three-zone trip will increase by 9.3% from current rates.[5] Estimating and forecasting an annual number of journeys can be very useful for determining the fare increase as well as increasing the means of transport if necessary. The highest number of rejsekort check-ins was around 179 million in 2019 which dates back to before covid.[6] It is important to know if 2023 will surpass that. Therefore, forecasting the number of annual rejsekort journeys in 2023 has been attempted in this paper with ARIMA and ETS models.

---

1 https://www.rejsekort.dk/

2 https://www.globalrailwayreview.com/article/30425/rejsekort-unifying-different-ticketing-elements-create-common-use-system/

3 https://www.dsb.dk/globalassets/arsrapport/2022/dsb-annual-report-2022.pdf

4 https://www.thelocal.dk/20221115/danish-public-transport-to-cost-more-in-large-parts-of-country

5 https://dinoffentligetransport.dk/media/2757/takstblad-2023.pdf

6 https://passagertal.dk/

Dataset Description

The passagertal.dk website contains all means of public passenger data in Denmark including rejsekort data, S-trains, regional trains, busses and flights. Among them, the rejsekort data has the most observations useful for further analysis. However, the website is heavily protected against any link sharing/translation as every separate tab has the same link. The data has been downloaded from the left first visualisation graph in the rejsekortrejser tab. The visualization shows a bar chart which shows the total annual rejsekort journeys from 2013 to 2022. The data behind this bar chart can be found out with the "vis data" button and the yearly aggregate can be transcribed into monthly and daily estimates by clicking on the leftmost plus icons. The data can then be downloaded with the "Exporter data till excel" button in the rightmost corner.

The downloaded data contains two columns with the date and passenger check-in count from January 2013 to December 2022 in 3783 rows. Apart from daily counts, it contains the monthly counts, yearly counts and a total count of 11,69,192,496 journeys. Therefore, any rows with non-date entries in the first column have been removed before analysis. After cleaning the data has 3652 rows.

Exploratory Data Analysis

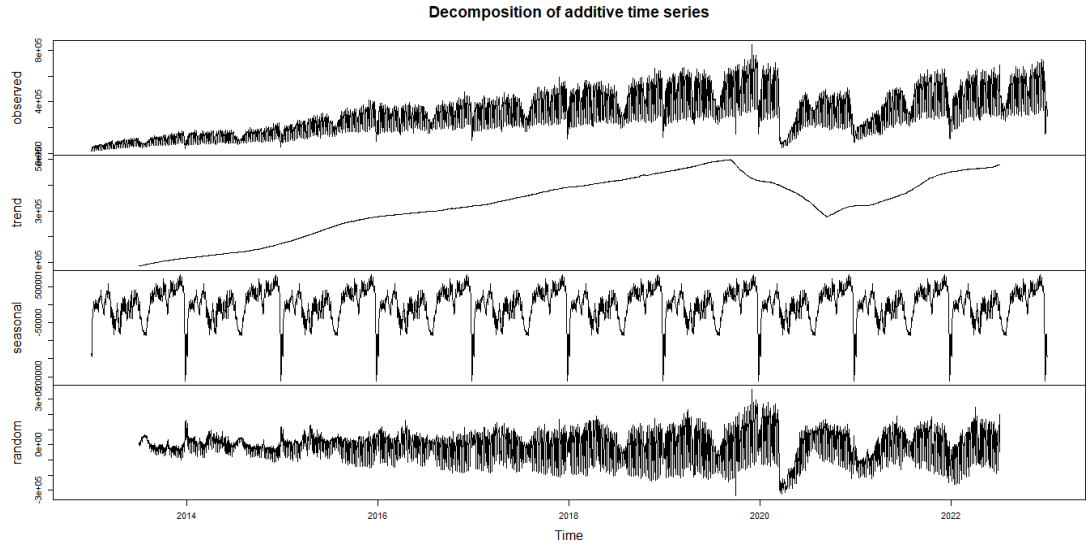

Figure 1
Additive Components of the Decomposed Rejsekort Check-in Data

From the Data the decomposition function of an upwards trend is visible except for a sudden fall during the covid period. It also displays heavy seasonality. Using the summary function in R it can be discovered that there is 320,151 average number of passengers in a rejsekort journey daily while the median value is 309,941 which indicates some level of asymmetry and a little right-skewed distribution with the presence of outliers. However, the maximum value which represents to the most number of daily check-ins is 846,026 in November 2019.

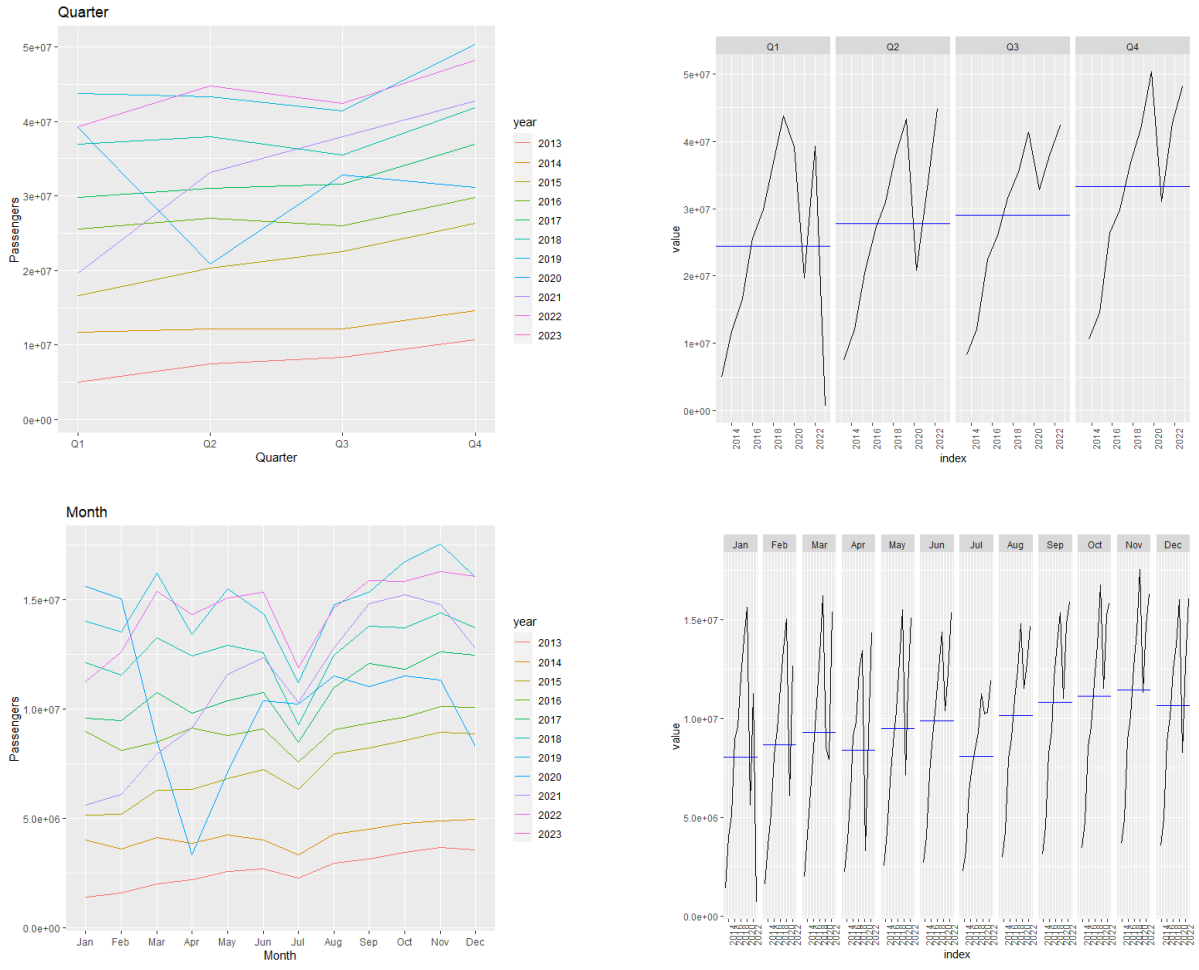

Figure 2
Quarterly and Monthly Seasonal Pattern

The 'ggseasonplot' and 'ggsubseries' shows the underlying pattern of the data for quarters and months. It is observed that except for covid years, the data tends to show a zig-zag pattern across the four quarters. However, in the first three quarters, it shows somewhat of a mixed behaviour across the years but always rises on the fourth quarter due to winter when people bike less. From the monthly seasonplot, it can be

also observed that the data shows common seasonality except the year 2020 and 2021 which were affected due to covid-19 pandemic. The monthly plot over the years shows the number of passengers rises the most in November each year because of the start of the winter and goes down the lowest in July when most people are out on summer vacation.

The subseries plots however greatly point out a structural break. Therefore a QLR test has been done on the dataset with a result of sup.F = 2088.5, p-value < 2.2e-16 which rejects the null hypothesis of the dataset with no structural breaks. Also, most of the QLR graph was above the threshold. With the breakpoints function the breakpoint was found to be on the 2627th observation which corresponds to the break date of 2020(72) which is 12th March 2020. The first Covid-19 lockdown in Denmark was introduced on 13th March 2020. There was also another covid lockdown in December 2020 when a second corona wave hit Denmark.

Thus, the data was segmented from January 2021 to make the forecast more recent and uniform. Since the data was a daily observation, the segment was enough to produce a good forecast. The growth rate of the subset data was fluctuating but had a similar seasonal pattern annually. The subset data from January 2021 were split into train and test sets with an 80-20 per cent measure. The train data with its corresponding autocorrelation function and partial auto-correlation function is plotted below:

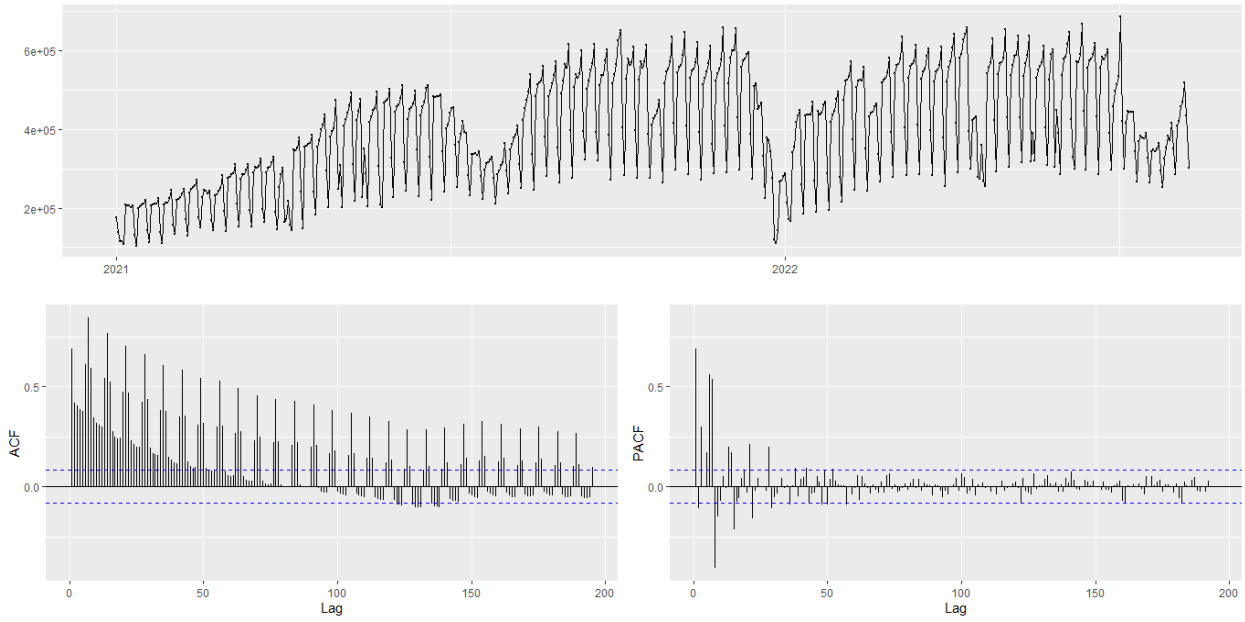

Figure 3
Time Series Plot with ACF and PACF of the Training set

It is eminent from looking at the lags that the ACF is tailing off and the PACF is cutting off at the first lag. This is highly suggestive of a “Moving Average” model. The ACF also shows a common pattern every 7 lags further pointing down to a “Seasonal Moving Average” or a Seasonal ARIMA model with parameters (p,d,0).

But before moving into that, some diagnostic measures had to be performed. Firstly, the value for BoxCox.lambda(train) was found to be 1 which indicated any further transformation like (log transformation) was unnecessary because they would produce results identical to the original data.[1]

- **Deterministic Trend Testing:** The training set has been checked with a low power adf test with lags 1. The output test statistic for the training data is -2.969 and Based on the output, the value of the test statistic is less than the critical value of -1.95 at the 5% significance level. Since the test statistic is more negative than the critical value, the null hypothesis is rejected. Which suggests that a deterministic trend is likely present in the data. Therefore we take the first derivative to remove it. After the differentiation, the data has no clear trend visible.

- **Unit Root and Stationarity Testing:** The first derivative of the training data were further checked into ADF and KPSS tests. The first ADF test was done with the trend and the t-value was found to be -27.341 which is less than the critical value -3.96 at the 1% significance level denoting a potential stationarity with the trend. Then it was checked again with ADF test with drift and again the t-value was found to be -27.3607 which is less than the critical value -3.43 at the 1% significance level denoting potential stationarity with the drift. It was further checked without any trend or drift and the t-value was found to be -27.3834 which is less than the critical value -2.58 at the 1% significance level. Thus we could not reject the null hypothesis and concluded the data to be stationary.

  Then we moved to the KPSS test with type ‘tau’ and found the value of the test statistic as 0.0301 which is less than the critical value 0.216 thus we can not reject the null H0 that data is not stationary. Also another KPSS test with type ‘mu’ found the value of the test statistic as 0.0879 which is less than the critical value of 0.739 thus we can not reject the null H0 that data is not stationary.

  From the overall analysis above we can confirm that the differentiated training data is stationary and has no unit root.[2]

- **Seasonality:** from the Correlogram and Decomposition of the differentiated data we can observe that the ACF lags kind of repeat every 7 lags and the PACF is significant at lag 7 as well.

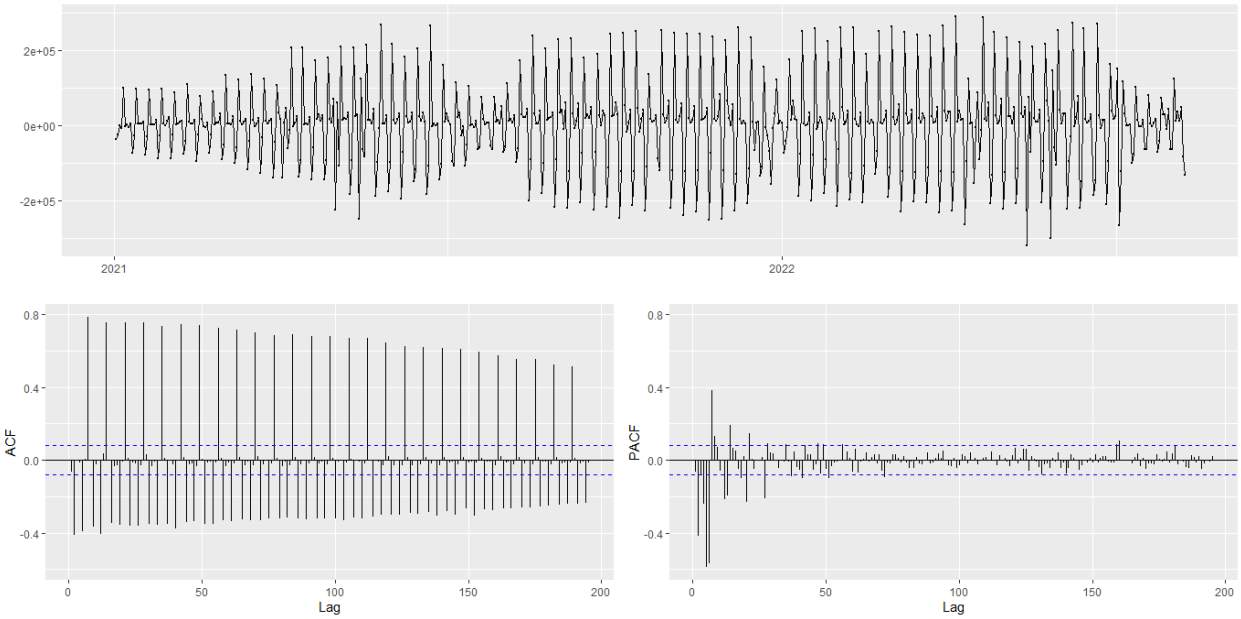

Figure 4
Time Series Plot with ACF and PACF of the Training set Derivatives

Therefore we use seasonal differencing with lag=7 to remove the seasonality from the differenced train data. After doing it the ACF and PACF look much more random and suppressed under the 95% confidence interval. The seasonality is therefore eradicated.

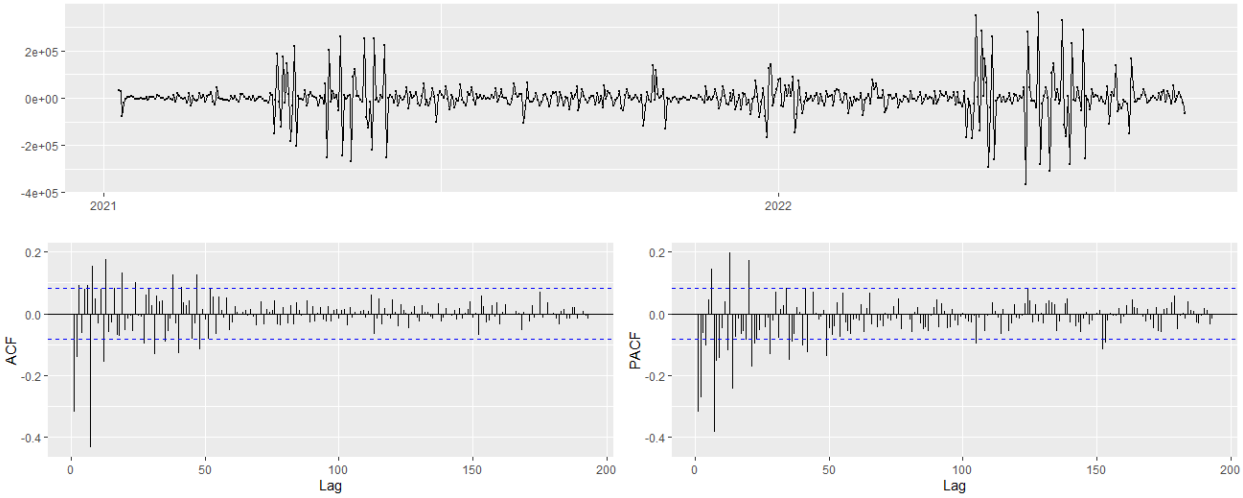

Figure 5
Time Series Plot with ACF and PACF of the Seasonal Differentiated Training set Derivatives

- **Structural Break:** Another QLR test has been done on the training dataset to check if there are any more structural breaks left. With a result of sup.F = 149.15, p-value < 2.2e-16 which rejects the null hypothesis of the dataset with no structural breaks. The IS test of breakpoints further shows that the data has another break point at the 346th observation on 12 December 2021. However, it was chosen to be ignored due to further not shortening the total observations.

Methodology & Diagnostic Tests

- **Auto-ARIMA:** To get an overall idea of the ARIMA estimates an "auto.arima" function has been tried with the train set. The result caused a non-seasonal forecast although the "seasonal" parameter had been set to "TRUE". This results in an ARIMA (2,1,3) model with AIC:14959.83 and BIC:14986.06. It describes a poor residual plot where it can be inferred that the residuals are not white noise by inspecting their autocorrelation functions.

- **Guessed-SARIMA:** From Figure4 it is difficult to determine any values for p and q, let's get back to Figure3 which displayed a clear indication of p=1 and q=0 because the ACF was exponentially decaying and the PACF was cut off at lag 1. As the training data was differentiated, we know that d=1. Again from the graph of the seasonal differences with lag seven, we can discover significant lags in lag 7 in both ACF and PACF charts. Thus the seasonal components P , D and Q for guessed SARIMA with lag 7 can be all inferred as 1. Using this parameter the forecast was much

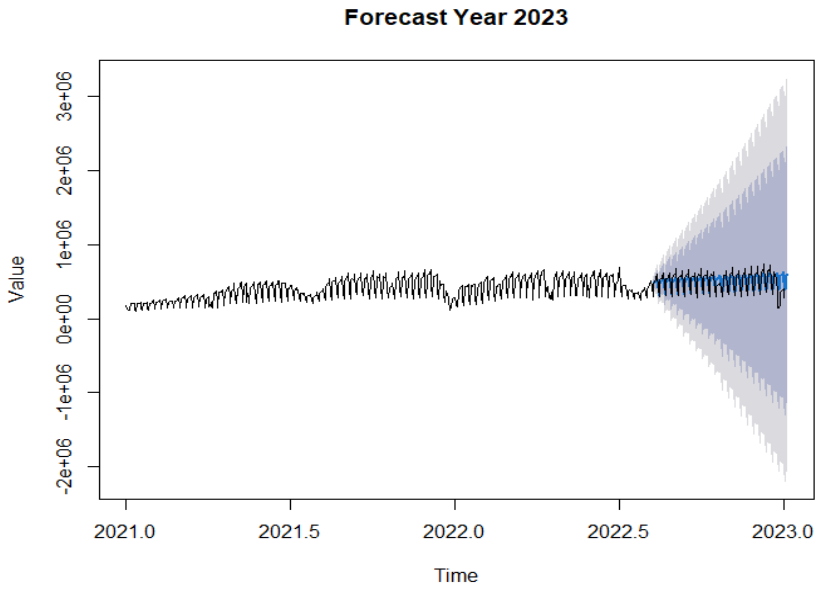

Figure 6
Forecast with ARIMA(1,1,0)(1,1,1)[7]

better and with AIC = 14249.96. Also, the residuals displayed a white noise with 95% ACF lags lying in their corresponding confidence interval. It is also more parsimonious than the auto-generated ARIMA (2,1,3). Therefore, it can be clearly seen that the guessed SARIMA model is the better model. [3]

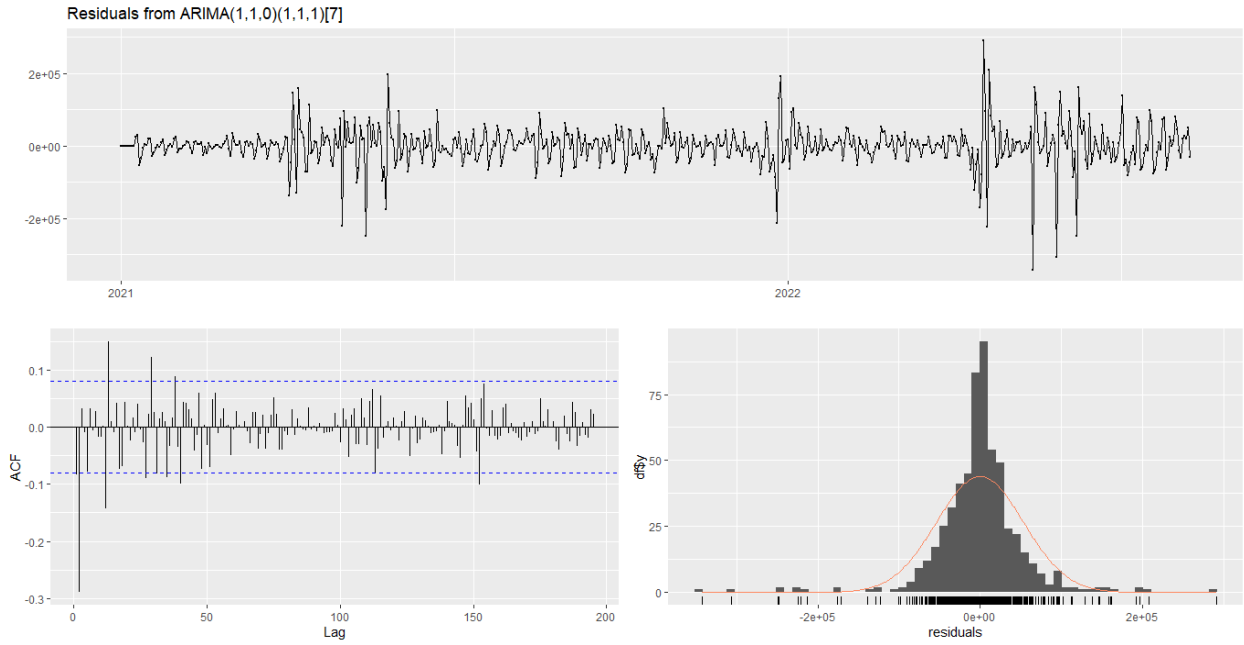

Figure 7
Residual Plots from ARIMA(1,1,0)(1,1,1)[7] Forecast

However, there are still some significant spikes that are still present in the residual plot and the low p-value (2.207e-05) in the Ljung-Box test suggests strong evidence against the null hypothesis of independence of residuals. This could be attributed to the existing structural break.

- **Exponential Smoothing Model:** In this section, an Exponential Smoothing Model has been tried to fit to the data directly. Since the data has a daily frequency and ETS can not handle data frequency of more than 24, it was automatically fit to an ETS(M,N,N) model. It is eminent that the function fails to capture the seasonality as the component was not functional but it has, however, managed to cover the variance in its 95 percent confidence interval. It has an AIC value of 17188.72 and a BIC value of 17201.84 which is greater in comparison to the seasonal ARIMA models. Also, the Ljung-Box test p-value is p-value = 2.207e-05 which is very small and denotes limited capturing of all the underlying patterns in the data.

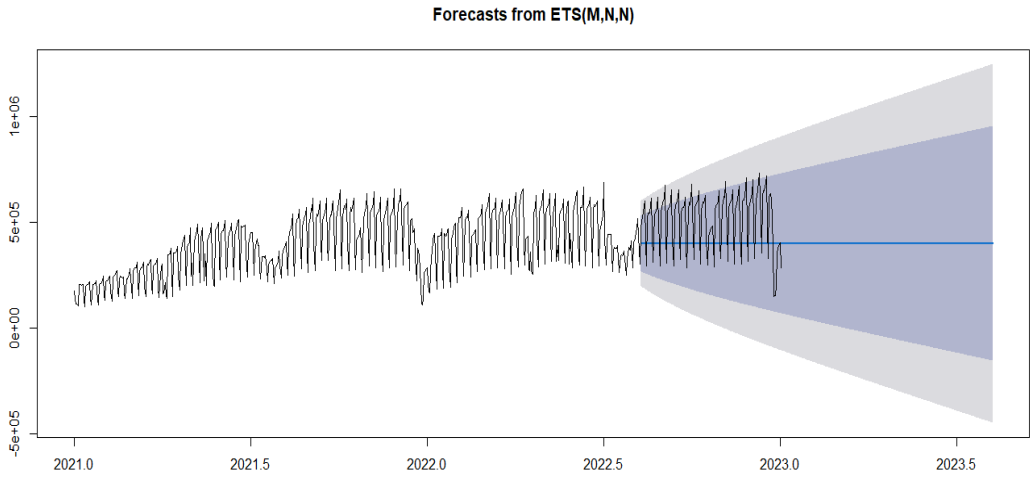

Figure 8
Forecast with ETS(M,N,N)

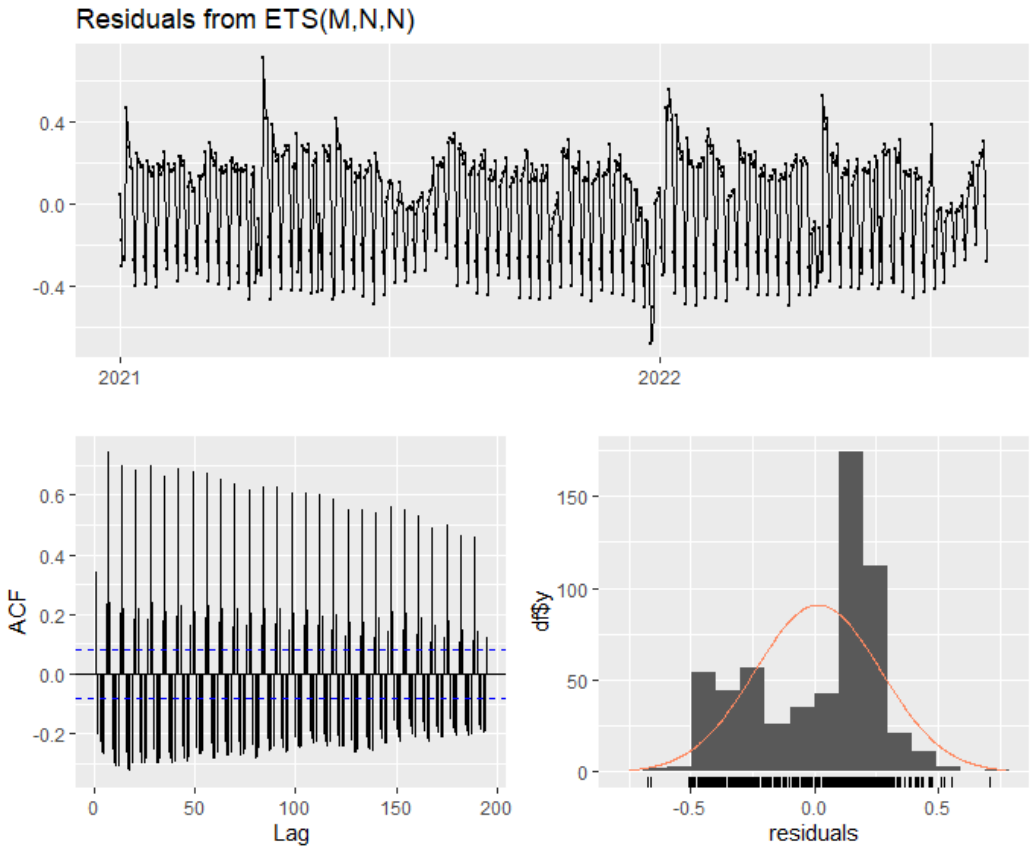

Figure 9
Residual Plot with ETS(M,N,N)

Results

Accuracy on the training set:

Model	MAE	RMSE	MASE	MAPE
ARIMA (2,1,3)	70255.14	85311.45	0.4273823	20.13001
ARIMA(1,1,0)(1,1,1)[7]	33280.78	53443.15	0.2024566	10.48547
ETS(M,N,N)	87617.97	103139.4	0.5330054	26.28352

Accuracy on the test set:

Model	MAE	RMSE	MASE	MAPE
ARIMA (2,1,3)	158003.05	178418.66	0.9611782	30.33972
ARIMA(1,1,0)(1,1,1)[7]	64034.90	84502.86	0.3895428	15.48702
ETS(M,N,N)	154101.07	173835.5	0.9374413	29.86753

From the above table we can conclude the ARIMA(1,1,0)(1,1,1)[7] has better characterized the behaviour of the rejsekort check-in data more than the other models. The model was, therefore, used to forecast and the total number of journeys predicted in 2023 amounted to **253.25 million** which is a 44 percent increase from the last year.

Limitations and Future Work

The Danish transport data from rejsekort check-in is a complicated dataset with many structural breaks. It is hard to forecast the data with many complications. It is also to be noted that, many people conduct their journey by commuter cards or passes that do not require check-in. Therefore, it is difficult to assume the total number of passenger journeys from only check-in data. The authorities could introduce more efficient ways to with the help of AI and sensors to track commuting passengers for further prediction accuracy. Also, check-in data from different busses and metro routes can be analysed and forecasted to optimize the next route for M5/M6 metro lines.

Bibliography

[1] Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252.

[2] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics, 54(1-3), 159–178. [DOI]

[3] Hyndman, Rob J and Athanasopoulos, George, Forecasting. Principles and practice. https://otexts.org/fpp3/