Exploratory Data Analysis and Forecasting of a Time-Series using a SARIMA Model

Presentation

Pedro Miguel Leite Martins up202007065
Nina Lichtenberger up202408285
Mariana de Saavedra Lourenço up201906985

Department of Computer Science
Faculty of Sciences of University of Porto (FCUP)

January 2025

Table of Contents

Data Description

- San Francisco International Airport Report on Monthly Passenger Traffic Statistics by Airline.
- Data ranges from July 1999 until August 2024.
- Includes 302 observations.
- Information includes airline, type of travel, price category code, . . .

SFO
SAN FRANCISCO
INTERNATIONAL AIRPORT

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

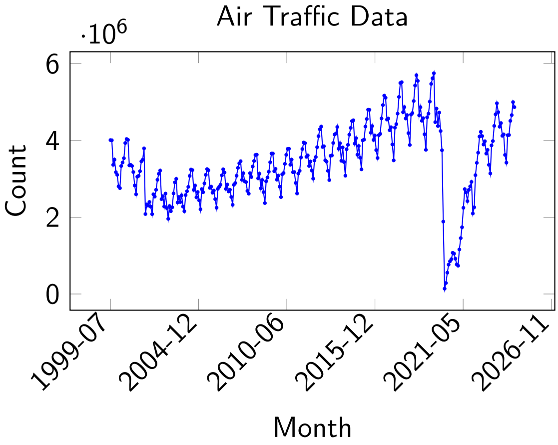
Why is this dataset so important?

Motivation:

- Improve the route lines
- Enhance flight logistics
- Support optimization of air traffic management
- Provide valuable insights for aviation analysis

EDA

The initial stage of analysis for time-series data is exploratory data analysis (EDA). We tried to focus mainly in the passenger counts over the time interval of 1999 to 2020.

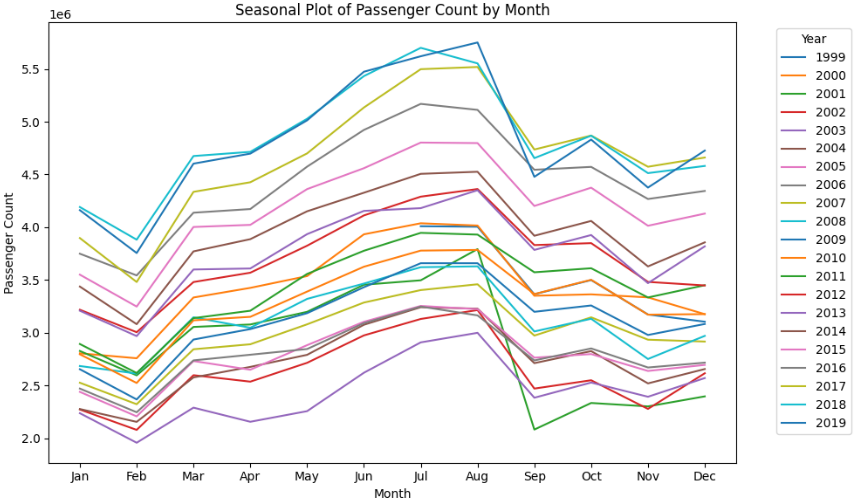


1. Seasonal pattern that has been consistent over time, along with an overall trend and some heterocedasticity.
2. Sharp decline around 2019-2020 caused by COVID-19.
3. Between 2001 and 2004, we have a decline which may have been caused by 9/11.

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

EDA

Seasonality and trend:



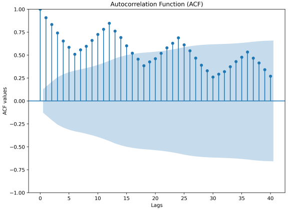
U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

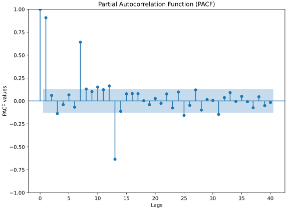
Data Transformation

Differencing:

- EDA indicated presence of seasonality and trend
- ACF and PACF also indicate non-stationarity

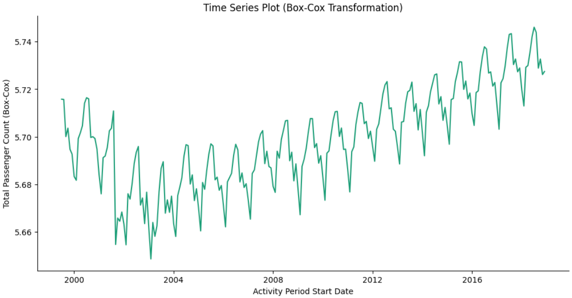
$\rightarrow$ d = D = 1 necessary





Box-Cox Transformation:

- Used to stabilize variance before model building



Approach for Finding the Best Model

Procedure for Model Selection

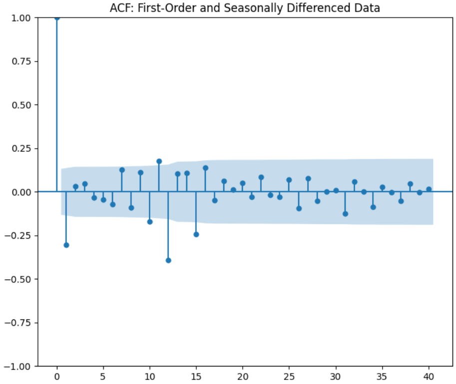
1. Analysis of ACF and PACF
2. Fit models based on the ACF and PACF
3. Inspect correlation of residuals (ACF and Ljung-Box test p-values)
4. Check parameters for significance
5. Inspect Q-Q plot to check normality of residuals

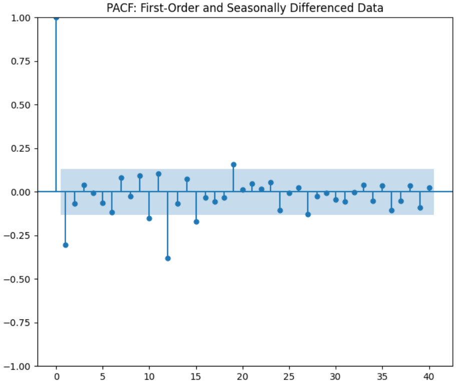
$\longrightarrow$ Refit models based on the results

$\longrightarrow$ Compare best models based on forecasting performance and Information Criteria

First Model

- ADF test confirms stationarity of differenced data
- ACF and PACF of differenced data suggest $q = p = 1$ or 0 and high P as well as Q
- Starting point: $\mathrm{SARIMA}(0,1,0)\times(2,1,2)_{12}$

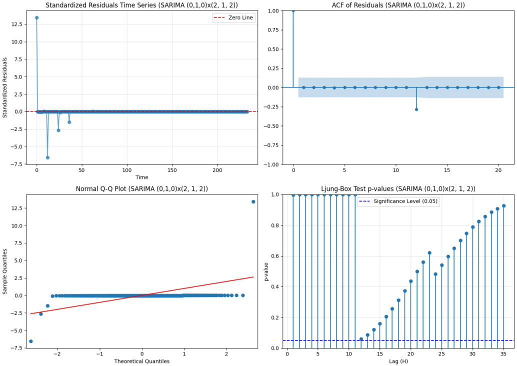




SARIMA$(0,1,0)\times(2,1,2)_{12}$

- Significant correlation at lag 12
- Insignificant seasonal MA coefficients

$\rightarrow$ Indicate necessity of higher P and lower Q



Parameter	**Coef.**	**Std Err**	z	$P>\|z\|$
ar.S.L12	-0.54	0.02	-22.37	0.00
ar.S.L24	-0.23	0.02	-9.59	0.00
ma.S.L12	0.02	0.07	0.28	0.78
ma.S.L24	-0.12	0.09	-1.28	0.20
AIC		-1689.649		
BIC		-1673.258		

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Best Models

Overview

Manually found (all coefficients significant):

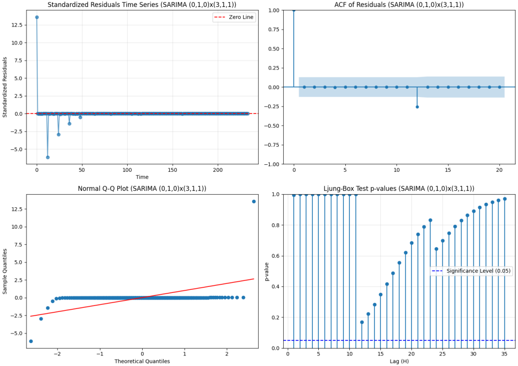
- SARIMA$(0,1,0)\times(3,1,1)_{12}$
- SARIMA$(0,1,1)\times(3,1,1)_{12}$

Automatically found (optimized AIC, insignificant MA12):

- SARIMA$(0,1,1)\times(2,0,2)_{12}$

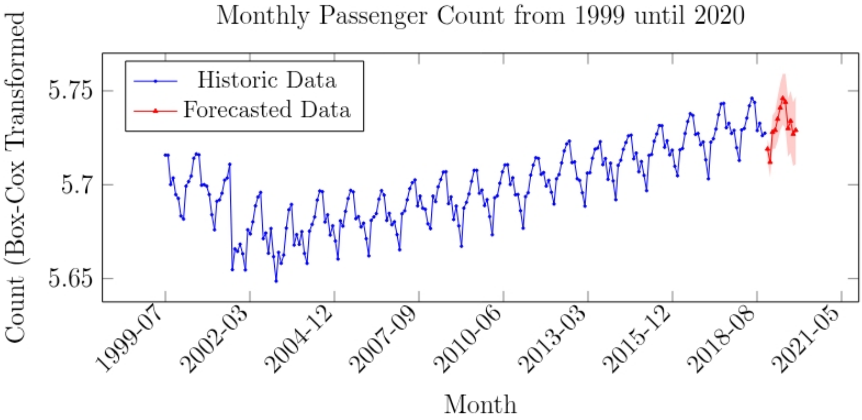
$\rightarrow$ Ljung-Box Test p-values > 0.05, BUT:
$\rightarrow$ ACF shows small correlation of residuals at lag 12
$\rightarrow$ Residuals are non-normally distributed

Exemplary Diagnostics for SARIMA$(0,1,0)\times(3,1,1)_{12}$



Forecasting

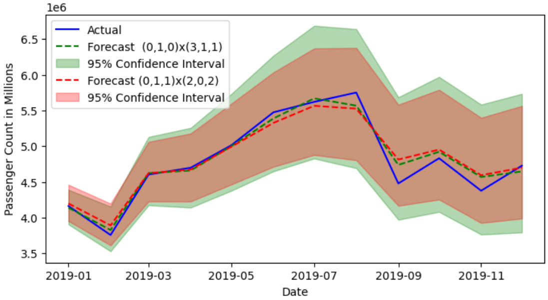
Monthly Passenger Count including forecasted values for 2019 with the 95 % confidence interval with SARIMA(0,1,0)$\times$(3,1,1)$_{12}$

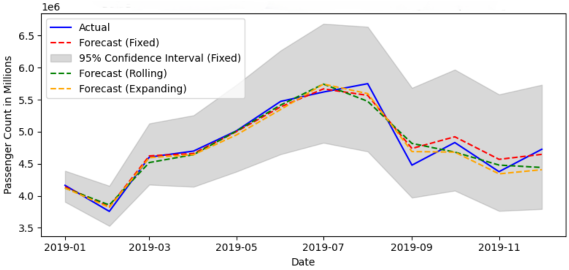


U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Forecasting

- Forecast is very accurate at the beginning and becomes less accurate towards the end.
- For the last four months, models overestimate the number of passengers. While the summer peak is forecasted to be a month too early.
- The values fall within the 95% confidence intervals for both models.
- It should also be noted that the confidence interval of the SARIMA(0,1,0)x(3,1,1) is bigger.





Forecasting

- The fixed forecasts are very similar performance wise
- But SARIMA(0,1,0)x(3,1,1)$_{12}$ has less parameters and lower RMSE
- The Information Criteria was not a valuable comparison metric for the forecasting quality in our case.

Forecast Performance of the different SARIMA Models (Fixed and Rolling Scheme computed for (0,1,0)x(3,1,1)$_{12}$)

Metric	(0,1,0)×(3,1,1)	(0,1,1)×(3,1,1)	(0,1,0)×(2,0,2)	Rolling Scheme	Recursive Scheme
Absolute Errors					
MAE	92,015	97,772	113,629	134,820	112,652
RMSE	121,600	130,875	150,522	169,719	141,099
Relative Errors					
MAE	1.921	2.041	2.372	2.814	2.352
RMSE	2.538	2.732	3.142	3.543	2.945
MAPE	1.952	2.074	2.410	2.819	2.322
Indices					
Theil's U	0.277	0.298	0.343	0.386	0.321

Conclusion

- Data could be forecasted well because of stable trend and seasonality
- SARIMA(0,1,0)x(3,1,1) with fixed scheme provided the best forecasts
- All suitable models had a relatively high accuracy under normal conditions: low relative errors and actual values within 95% confidence interval
- Limitations arise in extraordinary situations such as COVID

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO