Executive Summary

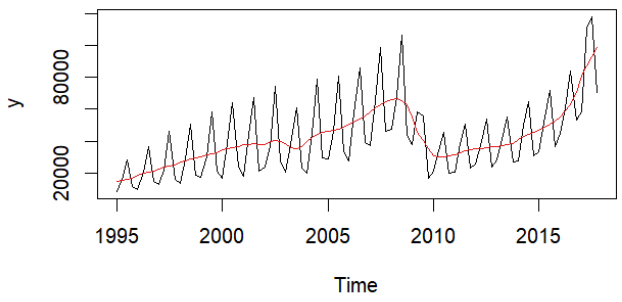
The project for forecasting had the origin country as Mexico and the destination as Canada, where the purpose was to forecast tourism flows among the two countries, and where two datasets were provided for the task. The aim was also to implement forecasts for the year 2019, where the data from 1995 to the end of 2017 was used for training the tourism data, whereas the data from the start to the end of 2018 was used for validation in order to assess the generalization, while the data from the start to the end of 2019 was used for the test set after using the validation data. This approach was done in order of ensuring robustness in the forecasts, and also avoiding overfitting on the historical data. The data for tourism flows that was extracted was Mexico to Canada, and the models that were implemented were based on a time series with a frequency of 4 because of focusing quarterly. Usually the GDP (Gross Domestic Product) is calculated annually such as in the second dataset, but in order to make it equal to fit in with the first dataset that includes tourism flows quarterly, the macroeconomic data was transformed where every value was repeated through four quarters where the feature GDP was used because of it being the most suitable and because it is the measure that is usually and mostly used. There were no missing values found within the tourism data and the data for GDP, which facilitated the process of forecasting. The following models that were implemented for assessing the robustness were such as Seasonal Naive as a benchmark, ETS and ARIMA for trend and seasonality, and regression model with lags that included both of the tourism flows and the GDP data.

Seasonal naive was chosen and did serve as a benchmark for baseline with the assumption that the year of 2019 might repeat the seasonal pattern of the previous year. ETS was chosen based on its ability of capturing the trends, the levels and also the structure that is seasonal. ARIMA was chosen based on its dependence that is seasonal in the data. Regression with lags was chosen with addition to the GDP data where 1 to 5 quarters was used for lagged tourism, and the GDP for both of the countries as variables for explanatory purposes. The section of variables was performed through stepwise regression in order to only get the predictors that are significant, retained. The forecasts were accessed using error metrics such as MAE, RMSE, and MAPE. The reason why these error metrics was chosen for evaluation was because of that MAE is simple to interpret when it comes to units such as the count of tourists, RMSE because of its penalization on deviations of large forecasts, and MAPE because of its allowance of comparing through diverse scales. By evaluating the validation sets and the test sets, an assessment was allowed to be done regarding the stability of the models and their robustness instead of relying on a single measure of accuracy.

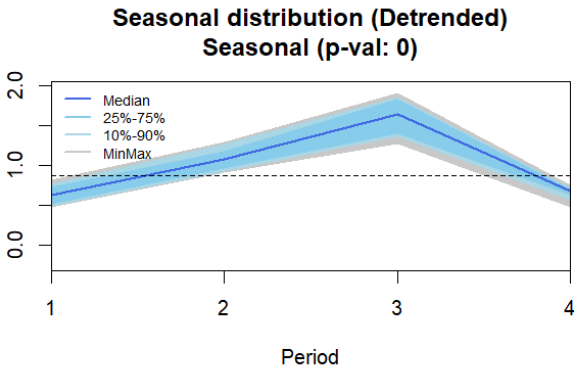
The results for the validation set did show that ETS did slightly outperform the other models, whereas ARIMA did achieve the lowest error values overall for the test set. However, there was a consistency within the rankings among the periods of validation and test which showed that the process of modelling was seen as robust and did not overfit to a certain sample. The regression model did also perform well which suggests that the inclusion of macroeconomic variables like GDP had a positive contribution on the performance of forecasting in the short term. Through all the models, the time series did display a regular and strong pattern that is seasonal with a peak that was consistent during the third quarter, and a decline during the fourth quarter. For 2019, the forecasts indicated a continued growth for the arrivals in Canada from Mexico. There was a drop around 2008 to 2009 which could be correlating with the global financial crisis, where this reflects the tourism flows' sensitivity to events of economics in a large scale. The limitations was that even though the GDP is good for estimating the strength of economy, it is usually calculated as annual which needed to be based on quarters of repetitions, where this might have got the fluctuations in the short term smoothed out. The models did not account for other shocking events such as changes in the policy and restrictions of traveling which could have had an influence on the demand of travel. Since the analysis did exclude any data after 2019, disruptions like COVID-19 was not being considered. Recommendations for the future would be integrating more variables of explanatory such as the rates of exchange and the capacity of flights in order to enhance the responsiveness to changes in short term. The models should also be updated regularly with the latest data and then revalidated in order of ensuring their robustness and reliability over time.

Appendix for Executive Summary

Centered moving average



Seasonal distribution



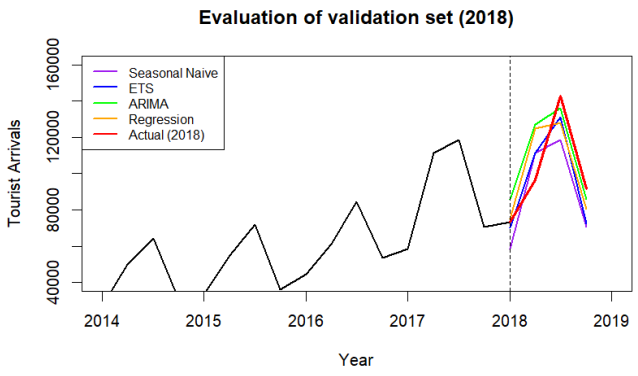
Validation results

Model <chr>	MAE <dbl>	RMSE <dbl>	MAPE <dbl>
ETS	12092	13396	12.11
ARIMA	13883	17094	14.93
Regression	13893	16862	13.45
Seasonal_N...	18735	19159	18.89

Test results

Model <chr>	MAE <dbl>	RMSE <dbl>	MAPE <dbl>
ARIMA	19300	21334	14.94
Regression	24226	27702	18.30
ETS	28267	30439	22.11
Seasonal_N...	34916	36406	28.19

Validation set visualization



Test set visualization