STAT 443 Term Project Report

Xinwei Kuang
Shiyang Li
Yichun Liu
Yubin Lyu

Summary

Tourism has always been one of the most important industry for BC. In 2016, tourism exports generated revenue of $4.9 billion. Therefore, how many tourists will travel to British Columbia is essential to be known.

Goal: In this project, the goal is to forecast the flow volume of non-resident travellers to British Columbia from December 2018 to November 2019.

Conclusion: The data relevant to non-resident travellers to British Columbia per month from January 1992 to November 2018 are analyzed. Among methods of persistence, average of all past, simple/linear exponential smoothing, ARIMA, ARMAX with explanatory variables, and ARIMAX with explanatory variables, MA(2) turns out to be the best model, since it leads to the smallest out-of-sample holdout set forecast root mean square error (RMSE).

Data

The data were obtained from Statistics Canada website and Bank for international settlements(Appendix). The cleaned data contain year, month, the number of non-residential visitors to British Columbia, nominal exchange rate and real exchange rate of Canada. The data range from January 1992 to November 2018.

Description of data

The dataset consists of five columns, year, month, value(number of travellers), nominal (nominal CAD exchange rate indices) and real (real CAD exchange rate indices).
The first 24 years are used as the training set and the last three years are used as the holdout set. For prediction, the model is trained on all 27 years of data and gives one year of predicted value.

Summary Statistics

```
    value            nominal           real       
 Min.   :417610   Min.   : 69.03   Min.   : 69.26  
 1st Qu.:504119   1st Qu.: 76.13   1st Qu.: 78.61  
 Median :576611   Median : 82.13   Median : 84.17  
 Mean   :576796   Mean   : 84.41   Mean   : 85.84  
 3rd Qu.:641586   3rd Qu.: 93.00   3rd Qu.: 94.09  
 Max.   :761174   Max.   :105.57   Max.   :104.94  
```

Figure 1: summary statistics

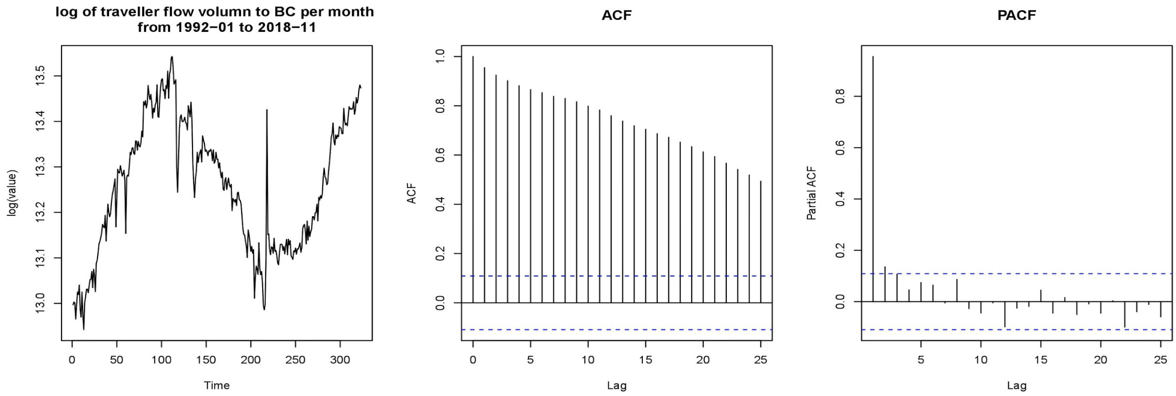

Figure 2: plot, acf plot and pacf plot of logged data

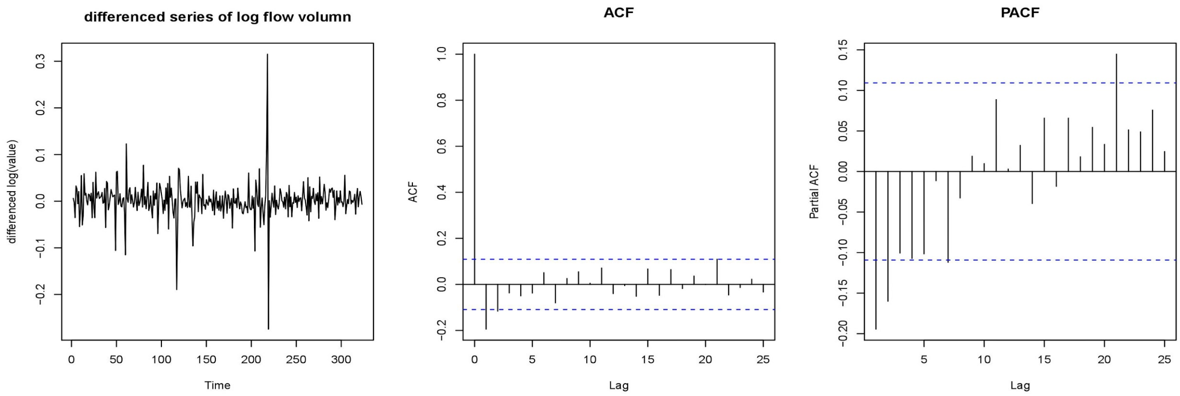

Figure 3: plot, acf plot and pacf plot of differenced logged data

Result

As can be seen in Figure 1, the variable values are quite large, therefore, taking the log of the data is necessary. This process is shown in formula (1).

$$Z_t = log(X_t) \tag{1}$$

Figure 2 shows the time series plot, acf, and the pacf of the processed data. This figure implies that the logged data is not stationary with trend, because the time series plot has dramatic changes and the acf decreases slowly. A common way of removing a trend is to difference the logged data. This process is shown in formula (2).

$$Z_t = log(X_t) - log(X_{t-1}) \tag{2}$$

As shown in Figure 3, the data is now stationary. The acf cuts off after lag 2, therefore, the data has a MA(2) pattern. In addition, all three plots have no seasonal pattern, hence there is nothing to do with the seasonal effect. Based on this finding, multiple related methods are applied and their out-of-sample RMSEs are stated below.

Methods and corrsponding RMSEs

method	out-of-sample RMSE	rank of RMSE
average	0.16284	12
persistence	0.01892	10
Simple exponential smoothing	0.01929	11
linear exponential smoothing	0.01857	9
MA(2)	0.01686	1
ARIMA(0,1,2)	0.01743	5
ARMAX(0,0,2) nominal	0.01714	3
ARMAX(0,0,2) real	0.01709	2
ARMAX(0,0,2) both	0.01714	4
ARIMAX(0,1,2) nominal	0.01769	7
ARIMAX(0,1,2) real	0.01765	6
ARIMAX(0,1,2) both	0.01774	8

Table 1: methods and RMSEs

Based on the result of Table 1, the best model which has the smallest RMSE 0.01686 is MA(2) without additional explanatory variables.

```
Coefficients:
          ma1      ma2    mean
      -0.2809  -0.1887  0.0015
s.e.   0.0543   0.0558  0.0011

sigma^2 estimated as 0.001357:  part log likelihood=607.62
```

Figure 4: R code output of MA(2) model

MA(2) Model

$$Z_t - 0.0015 = A_t + 0.2809A_{t-1} + 0.1887A_{t-2}$$

Prediction of using MA(2)

The prediction of future differenced data on logged values was obtained using MA(2).

2018 Dec	**2019 Jan**	**2019 Feb**	**2019 Mar**	**2019 Apr**	**2019 May**
-0.0016591	0.0014028	0.0014721	0.0014721	0.0014721	0.0014721
2019 Jun	**2019 Jul**	**2019 Aug**	**2019 Sep**	**2019 Oct**	**2019 Nov**
0.0014721	0.0014721	0.0014721	0.0014721	0.0014721	0.0014721

Table 2: The predicted values of difference of logged data

In Table 2, the predicted values after February 2019 remain the same. This is because the prediction of MA(2) uses the previous two innovations, and the predicted ones cannot be used as the inputs of the model. In the first step, the previous two innovations in the dataset are used while there is only one innovation is available in the second step. After step 2, no more innovations in the dataset can be used, and 0 is used to replace the missing value.

After reversing the transformations of the data, the following predicted values of the number of non-residential visitors to British Columbia are calculated. They are considered to be good enough because the number of visitors is increasing as expected.

2018 Dec	**2019 Jan**	**2019 Feb**	**2019 Mar**	**2019 Apr**	**2019 May**
709089	710084	711130	712178	713227	714278
2019 Jun	**2019 Jul**	**2019 Aug**	**2019 Sep**	**2019 Oct**	**2019 Nov**
715330	716384	717439	718496	719554	720614

Table 3: The predicted the number of non-residential visitors to British Columbia

Contributions

Our team is a team of four friends. The members are Xinwei Kuang, Shiyang Li, Yichun Liu and Yubin Lyu (order alphabetically by surname).
The work is evenly distributed to all team members. Major contributions are listed as following:

1. Topic: Yichun Liu
2. Coding and analysis: all team members
3. Writing: all team members
4. Organizer of discussion meetings: all team members

Appendix

Data source:

1. Statistics Canada: Table 24-10-0041-01 International travellers entering or returning to Canada, by type of transport
2. Bank for international settlements: https://www.bis.org/statistics/eer.htm

A table of all the methods that have tried and their out-of-sample RMSEs

method	out of sample RMSE	rank of RMSE
average	0.16284	20
persistence	0.01892	18
Simple exponential smoothing	0.01929	19
linear exponential smoothing	0.01857	17
MA(2)	0.01686	1
ARMA(1,1)	0.01716	5
ARIMA(1,1,1)	0.01780	12
ARIMA(0,1,2)	0.01743	6
ARMAX(0,0,1) nominal	0.01765	9
ARMAX(0,0,1) real	0.01758	7
ARMAX(0,0,1) both	0.01789	14
ARIMAX(0,1,1) nominal	0.01793	15
ARIMAX(0,1,1) real	0.01787	13
ARIMAX(0,1,1) both	0.01819	16
ARMAX(0,0,2) nominal	0.01714	3
ARMAX(0,0,2) real	0.01709	2
ARMAX(0,0,2) both	0.01714	4
ARIMAX(0,1,2) nominal	0.01769	10
ARIMAX(0,1,2) real	0.01765	8
ARIMAX(0,1,2) both	0.01774	11