MSBA7013

Forecast and Predictive Analysis

PASSENGER VOLUME FORECAST AND ADVERTISING FEE PRICING

January 2020

Group B2

Name	UID
Chen Jiaojiao	3035675579
Li Jiatao	3035675153
Luo Xinyue	3035232353
Wu Yingxian	3035674795
Xie Siyang	3035675323
Yuan Wengao	3035676183
Zhang Xinyi	3035674446

1. Executive Summary

Our company profits by providing advertising spaces on China's high-speed railway (CRH) trains and at railway stations in China. Advertising fee of the former is relatively fixed and wholesale is the mainstream whilst that of the latter varies with historical exposure estimated by independent market survey companies. Given the rapid development of high-speed railway and coming equilibrium, customers may be upset by the undeserved growing price. On the other hand, in case of unexpected volume change, price adjustment only applies to later clients and our company earns less than it should have. These are mostly attributed to excessive reliance on intuition and difference between present and past. Thus, our company needs a fair pricing mechanism, an impression-based strategy combined with current demand-based strategy. In this report, we focused on CRH advertisements due to time limit.

We proposed representing CRH advertisement impression with the average CRH passenger volume per CRH (ACP)[1] and derived the impression-based price by assuming same year-on-year change for price and ACP. We collected time-series data of railway passenger volume and six candidate predictors from sources including National Bureau of Statistics of China, Wind and official reports. After comparison, we eventually adopted a dynamic regression model with GDP and the air passenger volume as predictors and predicted a 3.2% increase in ACP in 2020.

With the forecast, we recommended to increase the price by 3.2% in 2020 if the market deteriorated or adjust current demand-based price with 3.2% increase in mind. If given specifics of current pricing and past prices, we would suggest further and quantified integration.

Our strategy could justify prices with quantitative proof of advertisement performance, retain clients and stabilize revenue, especially when the market is pessimistic in the near future. With our precise prediction, our company could also gain customer trust and future profit by providing impression as their expectation and our promise. This model might be applicable to other areas such as air tickets pricing and flow control. During the next half year, we plan to build another model to predict impression of advertisements at railway stations to facilitate fair pricing and recovering lost revenue from inaccurate and untimely data.

2. Introduction

2.1. Problem

Our company's business model is to exchange advertising spaces on CRH and at railway stations for advertising fee. For advertisements on CRH, same price is charged across months, trains and routes and clients able to afford them prefer bundling all advertisements on several trains for the entire year. This practice allows lower discount for clients and less sales efforts from our company. On the contrast, advertising fee of advertisements at railway stations is determined based on the historical impression (times an advertisement viewed by audience) provided by independent marketing companies. Due to unexpected fluctuation of passenger volume, the fee could be adjusted anytime yet only applies to clients who sign contracts after the adjustment.

Generally, the price increases by 25-50% each year based on advertisers' willingness to pay. The relatively high percentage is mainly caused by previous shortage of high-quality advertising places and optimistic expectation of exposure. However, the major completion of high-speed railway construction plan and moderate increase in CRH quantity may suggest an approaching market equilibrium and a yield for fair pricing. Particularly, it refers to impression-based pricing as implied by the prevalent payment structure of digital advertisements, cost-per-million-view, and intensive competition of developing approaches to monitor actual

[1] $ACP = \frac{Annual\ railway\ passenger\ volume \times CRH\ passenger\ percentage}{Number\ of\ CRH}$

impression of outdoor advertisements. Consequently, indifference to the trend could result in unsatisfying customers, deteriorated reputation and decreasing revenue.

2.2. Proposal

To tackle the problem, an objective pricing mechanism based on forecast rather than past performance should be established in addition to the current one which depends on experts' intuition and experience. Moreover, forecast timeliness and accuracy are critical to shifting advertisers' surplus to our company and earning the amount we should have earned.

In this report, we concentrated on impression-based pricing of CRH advertisements. Specifically, we proposed using the ACP[2] to represent CRH advertisement impression and deriving the impression-based price by assuming the same year-on-year change for price and ACP. The validity of the translation lies on that CRH advertisements are exposed to passengers and staff only (the latter is neglectable compared with the former) and passengers are unlikely to ignore all the advertisements. Moreover, an average is meaningful since advertisers usually purchase all advertisement space on trains for a year and our company charges the same for all routes and trains. Given the time constraint, we focused on predicting annual railway passenger and assumed number of CRH and CRH passenger percentage continue growing stably until reaching 2020 goal.

2.3. Literature Review

Scholars mainly use three methods to predict railway passenger volume. Among them, qualitative methods refer to prediction with expertise and experience. They are the easiest to implement yet subjective due to over-dependence on experts' intuition (Liu, 1999; Guo, 2007). Quantitative methods mostly include time series prediction model and regression analysis (Hao, Cui & Han, 2015; Liu, Shao & Li, 2010) along with their combination, which is not uncommon. They are easy to execute and more accurate than qualitative methods in spite of requirement for abundant data and inadaptability to patterns absent before. ARIMA, for example, is regarded a good fit for the monthly railway passenger data that shows regular seasonality. (Liu, 2010). Lin (2016) also proves that time series model overperforms simple grey prediction model in terms of accuracy. The third methods, soft-computer methods, are newly-developed and capable of handling high uncertainty and non-linearity despite complexity and limited interpretability (Wu, 2018; Liu, 2016; Li et al., 2014; Wang, 2014).

As for predictors, data indicating economy condition, substitution effect and supply-demand theory are widely used among others. For instance, Sun (2018) includes six factors, China's GDP, national population, tourist volume, consumption level of rural residents, railway mileage, civil aviation passenger volume and the number of cars, in a combined model using correlation coefficient method. Similarly, Lin (2016) builds ARIMAX model with China's GDP, tourist volume and railway mileage.

2.4. Approach

Considering accuracy and interpretability, we chose simple ARIMA and dynamic regression. We first built models using simple ARIMA and dynamic regression with six predictors, namely, GDP, airline passenger volume, the proportion of service industry GDP, migrant worker volume, tourist volume and total population in China respectively. The candidate predictors were selected for they covered all three aspects and were proved to have significant impact on railway passenger volume (Sun, 2018; Lin, 2016). They were also abundant and accessible. Eventually, we used AICc to select among these candidate models.

In the rest of the report, Section 3 presents collection, processing and summary statistics of railway passenger volume, six candidate predictors and two figures used for ACP calculation. Section 4 demonstrates model

[2] $ACP = \frac{Annual\ railway\ passenger\ volume \times CRH\ passenger\ percentage}{Number\ of\ CRH}$

development and selection. A simple ARIMA model was built followed by a dynamic regression model and the latter was selected for lower AICc. Forecast railway passenger volume and ACP in 2020 was generated with the latter model and forecast of candidate predictors obtained by ARIMA. Section 5 explains the forecast's application and the further plan to price advertisements at railway stations according to predicted impression.

3. Data

We collected data mentioned in Section 2 and details are shown in *Table 1*.

Variable Name	Description	Unit	Frequency	Seasonaltiy	Trend	Source
Railway	Number of passengers travelling by railway in China	Tens of thousands people	Monthly	12	Upward	"National Bureau of Statistics of China" (2019)
Air	Number of passengers travelling by air in China	Tens of thousands people	Monthly	12	Upward	"National Bureau of Statistics of China" (2019)
GDP	Gross domestic production in China	Hundreds of millions RMB	Quarterly	4	Upward	"National Bureau of Statistics of China" (2019)
ServiceIndustry	The percentage of GDP service industry accounts for in China	%	Quarterly	4	Upward	"National Bureau of Statistics of China" (2019)
MigrantWorker	Number of workers from rural areas in China	Tens of thousands people	Yearly	NA	Upward	"National Bureau of Statistics of China" (2019)
Tourist	Number of domestic tourist trips in China	Tens of thousands people times	Yearly	NA	Upward	"National Bureau of Statistics of China" (2019)
Population	Total population in China	Tens of thousands people	Yearly	NA	Upward	"National Bureau of Statistics of China" (2019)
CRH passenger percentage	The percentage of railway passenger volume CRH passenger accounts for in China	%	Yearly	NA	Upward	"The CRH Passenger Volume in 2008 to 2018" (2019), "A Notice on Issuing the 13th Five-Year Railway Development Plan" (2017)
Number of CRH	Number of CRH trains in China	Train	Yearly	NA	Upward	"Wind Database" (2019), "A Notice on Issuing the 13th Five-Year Railway Development Plan" (2017)

Table 1. Data summary

We predicted monthly railway passenger volume to retain sufficient data and shorten time required to examine prediction accuracy. To convert frequency, we assumed same GDP and ServiceIndustry for months in the same quarter. We also assumed MigrantWorker, Tourist and Population follow a linear a trend and obtained monthly data with linear interpolation. The last two were unchanged since they were used to calculate ACP only. Graphs of candidate predictors are shown in Appendix 2.

4. Analysis

To predict railway passenger volume, we developed two types of models, ARIMA on railway passenger volume itself and dynamic regression model with six candidate predictors and ARIMA residuals. Given that the former is a special case of the latter with all predictor coefficient being zero, it is valid to compare them using AICc. Thus, we chose the best model of its type followed and used the one with lower AICc to forecast.

To build a simple ARIMA model, we first applied Box-Cox transformation ($\lambda = 1$) and differencing ($D = 1, d = 1$) to stabilize variance and mean (*Figure 1*).

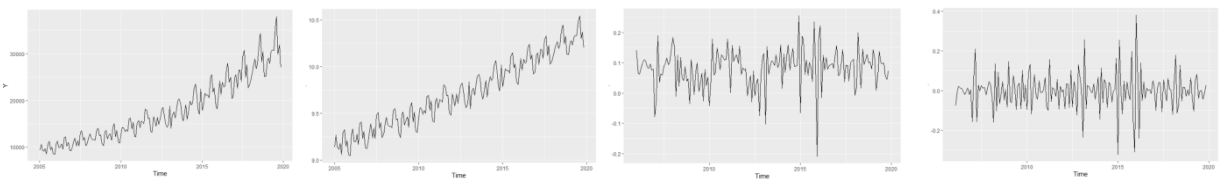


Figure 1. Process of obtaining stationary times series of Railway

Next is parameter selection. We proposed MA(2) for seasonal ARIMA based on ACF and PACF plot (*Figure*

2) and checked residuals of seasonal MA(2) (*Figure 3*). Similarly, ARIMA(2,1,1)(0,1,2)[12] and ARIMA(0,1,3)(0,1,2)[12] were suggested by ACF and PACF plots of residuals of seasonal MA(2). With white-noise residuals and no redundancy, **ARIMA(2,1,1)(0,1,2)[12]** was selected.

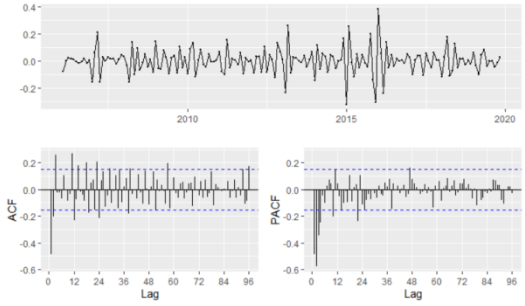


Figure 2

```
Series: log(Y)
ARIMA(0,1,0)(0,1,2)[12]

Coefficients:
          sma1     sma2
      -0.4807  -0.3707
s.e.   0.1052   0.1009

sigma^2 estimated as 0.00621:  log likelihood=180.5
AIC=-354.99   AICc=-354.84   BIC=-345.65

        Ljung-Box test

data:  Residuals from ARIMA(0,1,0)(0,1,2)[12]
Q* = 98.336, df = 22, p-value = 1.258e-11
```

Figure 3

The other type of model is dynamic regression with ARIMA residuals. Box-Cox transformed data was also employed to transform the original data into time series with constant variance (*Figure 5*). Two differencing strategies were conducted to ensure constant mean, specifically, (1) seasonal differencing once and regular differencing twice, and (2) regular difference three times. To guarantee interpretability, all predictors were differenced using the same strategy. Otherwise, the same index of different variables might have different lags with the original data resulting in difficulties in explaining predictors' impact on response.

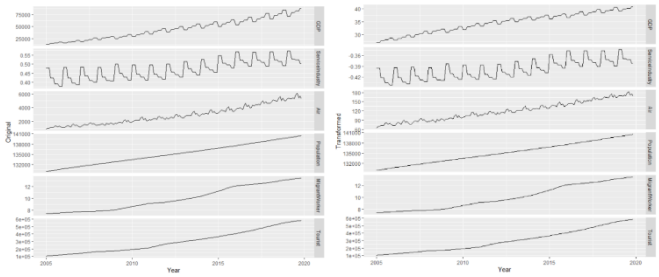


Figure 5. Box-Cox transformation of six candidate predictors

Regardless of differencing strategy, we removed the most insignificant predictor one at a time and arrived at same predictors, GDP and Air. Although there were two candidate models has lower AICc, we regarded the decrease in AICc not worthy of increase in predictors.

Since the response and the two selected predictors only required regular differencing once and seasonal differencing once, we refit the model and obtained stationary residuals. Then we chose seasonal MA(2) for seasonal ARIMA model based on residuals plots (*Figure 6*) and proposed ARIMA(2,1,2)(0,1,2)[12] and ARIMA(0,1,3)(0,1,2)[12] as candidate models based on ACF, PACF and EACF. The second model was selected for white-noise residuals, lower AICc and no redundancy. The best of dynamic regression model includes GDP and Air on Railway as predictors and residuals following **ARIMA(2,0,1)(0,0,2)[12]**.

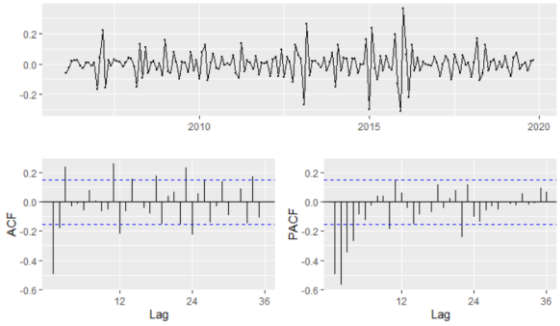


Figure 6. Residual plots of seasonal MA(2)

Comparing best models generated by two methods, the dynamic regression model with GDP and Air being predictors and residuals following **ARIMA(2,0,1)(0,2,2)[12]** has lower AICc and was thus selected.

Before forecasting response, we first used ARIMA to obtain future values of each predictor (*Figure 4*). Box-Cox transformation was applied to GDP, Air, and SI and differencing to all predictors for stationarity purpose. Fitting candidate models based on ACF and PACF plots, we checked residuals and selected those with the smallest AICc (*Appendix 3*).

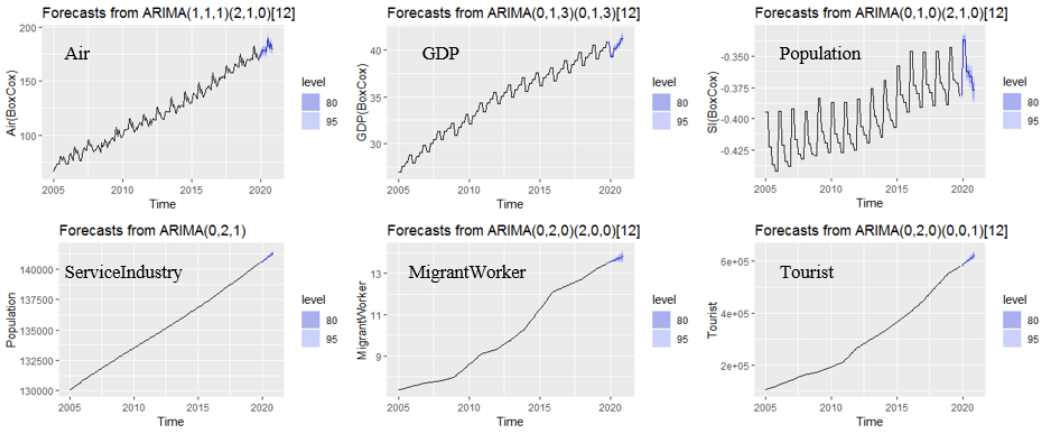


Figure 4. Forecast of six candidate predictors

Then, we forecast railway passenger volume for the coming 13 months and obtained annual railway passenger volume in 2019 and 2020. ACP change in 2020 is expected to be **3.2%** and the rest are illustrated in *Appendix 4*.

5. Conclusion

With impression-based price and demand-based price, our company can adopt the higher of the two or adjust one with the other. For instance, we suggest adopting adjusted demand-based price which is modified based on the 3.2% rise in ACP at present because of the promising market whilst we advise the other way around if the situation deteriorates soon. We will specify the figures provided with access to specifics of demand-based pricing strategy and past prices.

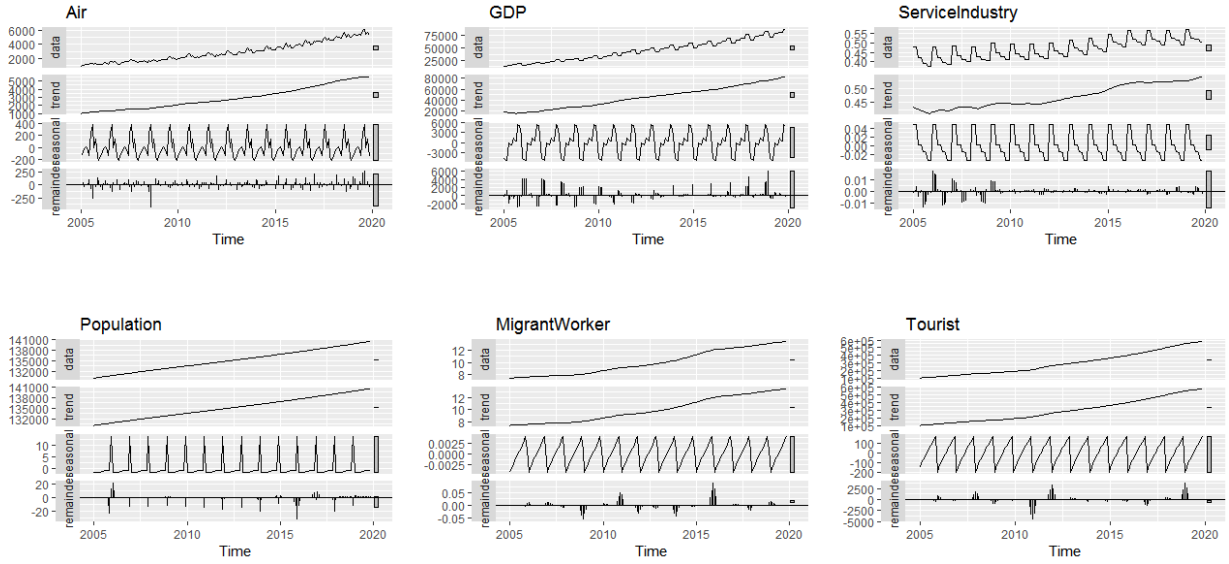
Of limited use as it seems currently, our proposed strategy is of considerable value during the coming slowdown as mentioned in Section 2. At that time, the impression-based pricing, which provides quantitative proof of advertisement performance and justifies the price, is expected to facilitate pricing and negotiation, retain clients and stabilize sales. Admittedly, our model requires constant review to update number of CRH which relies more on government plan rather than objective circumstances and ensure assumptions' compliance with company's selling strategy. We could also include other predictors in the existing model and develop another forecasting model for CRH passenger percentage to improve our results.

Within three to six months, our company can verify our monthly forecasts. Meanwhile, we plan to forecast impression of advertisements at railway stations. Given that the impression varies with time, location and city and is equal to passenger volume times visibility, we will derive impression-based prices after predicting passenger volume and estimating visibility. With timely and accurate prediction, our company is able to shift surplus from advertisers who sign contracts before price change. Our company can also gain clients' trust by delivering their expected impression precisely.

Appendix 1. Work Division

Name	Duties	Contribution
Chen Jiaojiao	Data collection and literature review, drafting literature review and references in slides and report	12.8%
Li Jiatao	Data and model processing, drafting data summary and predictor forecast in slides and report	15.4%
Luo Xinyue	Strategy formulation, drafting strategy in slides and report, report arrangement,	16.7%
Wu Yingxian	Data collection, data and model processing, drafting model processing in slides and report	16.7%
Xie Siyang	Strategy formulation, draft future plan in slides and report	12.8%
Yuan Wengao	Drafting introduction and executive summary in slides and report, slides arrangement	12.8%
Zhang Xinyi	Data collection and literature review, drafting literature review and references in slides and report	12.8%

Appendix 2. Trend, Seasonality, and Variance of Six Predictors

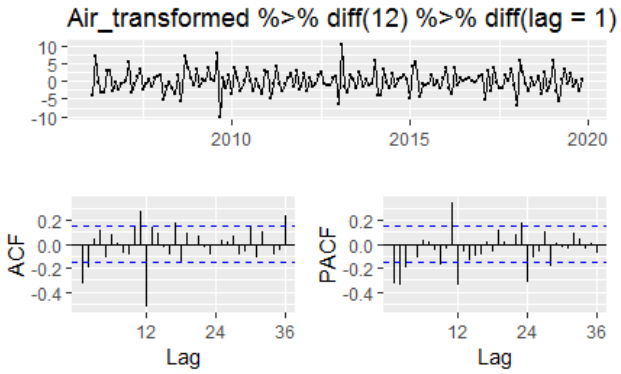


Appendix 3. Details of Predicting Predictors

Air

1. BoxCox transformation $\lambda = 0.524$ was applied to stabilize the variation.
2. nsdiffs() and ndiffs() show the necessity of a first order differencing(lag = 1) and a seasonal differencing(lag = 12) and such differencing was conducted.
3. According to Acf and Pacf, lag 1, lag 2 and lag 12 are significant in Acf while lag 1, lag2, lag 12, and

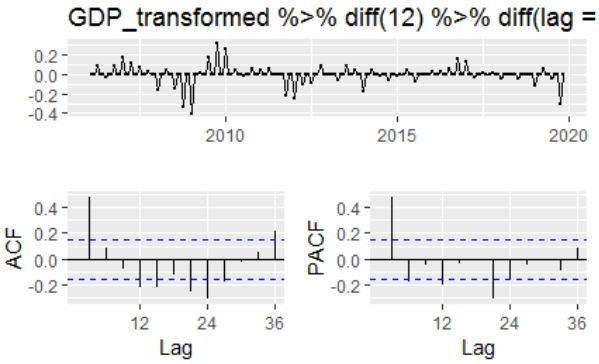
lag 24 are significant in Pacf. We choose ARIMA(2,1,0)(2,1,1)[12].



4. Add or reduce one order to p, q, P and Q, we had 7 candidate models and selected the one with smallest AICc, i.e. **ARIMA(1,1,1)(2,1,0)[12].**

GDP

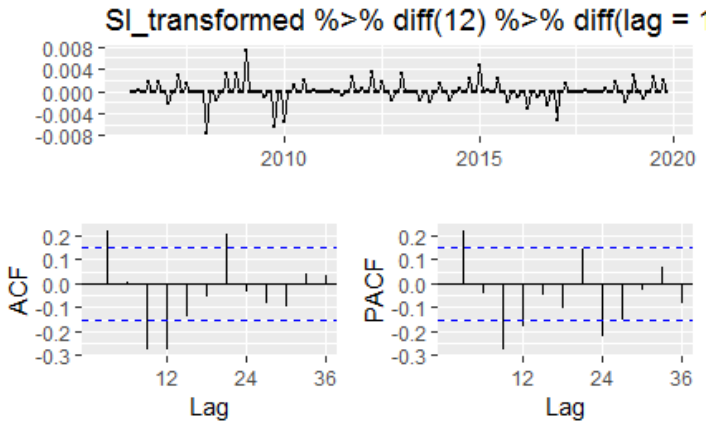
1. BoxCox transformation $\lambda = 0.191$ was applied to stabilize the variation.
2. nsdiffs() and ndiffs() show the necessity of a first order differencing(lag = 1) and a seasonal differencing(lag = 12) and such differencing was conducted.
3. According to Acf and Pacf, lag 1 is significant in PACF while lag 1, lag 12, and lag 24 are significant in ACF. We choose ARIMA(1,1,1)(0,1,2)[12].



4. Add one order to p, q, P and Q, we had 7 candidate models and selected the one with smallest AICc, i.e. **ARIMA(0,1,3)(0,1,3)[12]**.

ServiceIndustry

1. BoxCox transformation $\lambda = 1.912$ was applied to stabilize the variation.
2. nsdiffs() and ndiffs() show the necessity of a first order differencing(lag = 1) and a seasonal differencing(lag = 12) and such differencing was conducted.
3. According to Acf and Pacf, lag 1 and lag 12 are significant in ACF while lag 12, and lag 24 are significant in PACF. We choose ARIMA(0,1,1)(2,1,1)[12].

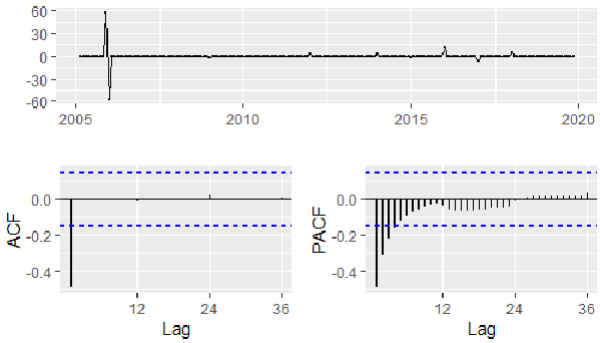


4. Add or reduce one order to p, q, P and Q, we had 7 candidate models and selected the one with smallest AICc, i.e. **ARIMA(1,1,0)(2,1,0)[12]**.

Population

1. nsdiffs() and ndiffs() show the necessity of twice first order differencing(lag = 1) and such differencing was conducted.

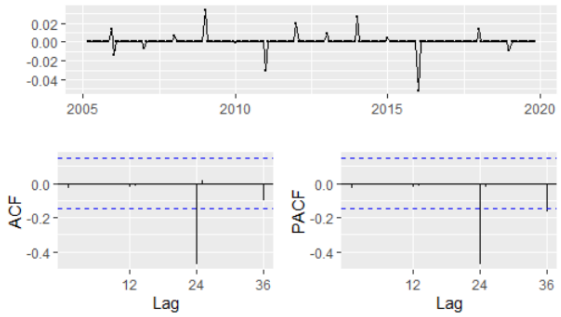
2. According to Acf and Pacf, lag 1 is significant in ACF while lag 1, lag 2, and lag 3 are significant in PACF. We choose ARIMA(0,2,1) or ARIMA(3,2,1). The second model is selected because of white noise residuals, i.e. **ARIMA(3,2,1)**.



Migrant Worker

1. nsdiffs() and ndiffs() show the necessity of twice first order differencing(lag = 1) and such differencing was conducted.

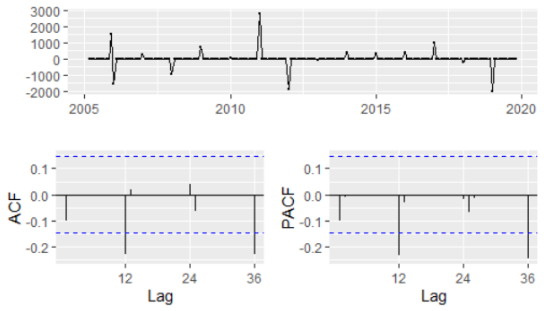
2. According to auto.arima, ARIMA(0,2,0)(2,0,1)[12] is a good model, including white noise residuals and all significant coefficients.



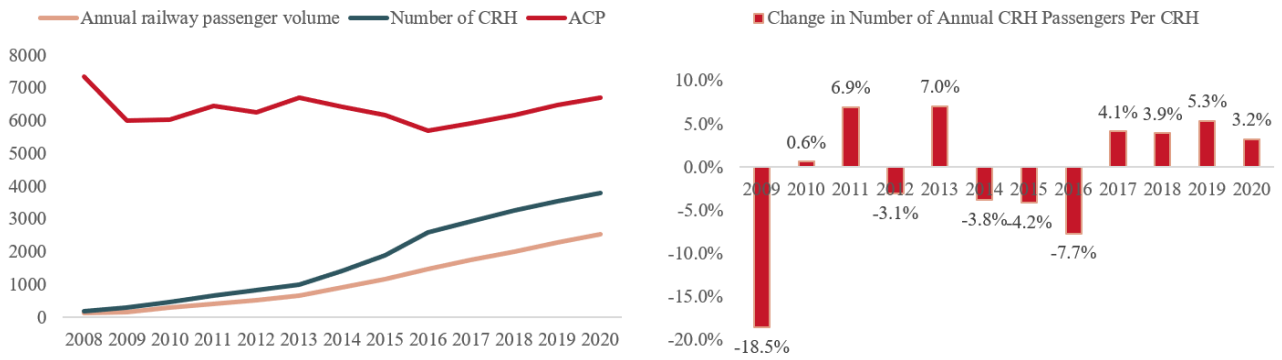
3. To check redundancy, we introduced several candidate models: ARIMA(0,2,0)(2,0,0)[12], ARIMA(0,2,0)(0,0,2)[12], and ARIMA(0,2,0)(1,0,1)[12]. Then based on coefficient significance, white noise residuals, invertibility, AICc, and result of cross validation, we selected **ARIMA(0,2,0)(2,0,0)[12]** as our final model.

Tourist

1. nsdiffs() and ndiffs() show the necessity of twice first order differencing(lag = 1) and such differencing was conducted.
2. According to auto.arima, ARIMA(0,2,0)(0,0,2)[12] is a good model, but sma2 is not significant, so **ARIMA(0,2,0)(0,0,1)** is selected, with white noise residuals and all significant coefficients.



Appendix 4. ACP (2008-2020) and ACP Change (2009-2020)



References

A Notice on Issuing the 13th Five-year Railway Development Plan. (2017). Retrieved from http://www.gov.cn/xinwen/2017-11/24/content_5242034.htm.

G, Z.Z. (2007). Hierarchy Analysis on the Railway Passenger Flow Factors. Journal Of Transportation Engineering And Information, 5(4), 68-71.

H, J.Z., C, Y.J., H, J.X. (2015). Prediction of the Railway Passenger Volume Based on the SARIMA Model. Mathematics in Practice and Theory, 18(45), 95-104.

Li, M., Ji, X., Zhang, J., & Ran, B. (2014). FA-BP Neural Network-Based Forecast for Railway Passenger Volume. Applied Mechanics and Materials, 641(3), 3-8.

L, L.Y. (2016). Research of Railway Passenger Volume Forecast Model Based on PCA-BP Neural Network. Comprehensive Transportation, 38(8), 43-47.

L, Q. (1999). The Influencing Factors on Passenger Flow and Analysis on Its Incidence Coefficients. Journal of Shanghai Tiedao University, (20)2, 41-44.

L, Y.P., S, Y.R., L, W.D. (2010). Combination Forecasting of Railway Passenger Transport Volume in China. Logistics Technology, 29(7), 58-59.

National Bureau of Statistics of China. (2019). Retrieved from http://data.stats.gov.cn/easyquery.htm?cn=A01.

S, L. (2018). Study on the Forecast of High-speed Railway Passenger Traffic Volume Based on Combined Model. (Master thesis). Lanzhou Jiaotong University, China.

The CRH Passenger Volume in 2008 to 2018. (2019). Retrieved from http://www.yte1.com/passenger-ou/d6b7943ea6604651905ea2a4b922634b.

Wang, C. (2014). Dynamic Mechanism of Effect Factors for Railway Passenger Volume. Applied Mechanics and Materials, 644-650, 5848-5852.

Wind Database. (2019).

W, H.W. (2018). Analysis of grey theory in railway passenger traffic volume forecast. Railway Computer Application. 27(8), 7-9+18.