1 Introduction

Within weeks in March 2003, severe acute respiratory syndrome (SARS) spread from the Guangdong province of China and infected some 37 countries around the world. Beijing is one of the most affected cities. SARS also had a big blow on Beijing’s tourism. In this project, we built a prediction model for the number of overseas tourists to Beijing. The data we have are the monthly number of overseas tourists to Beijing, from year 1997 to year 2005. Our model considers the seasonality of the data and also the effect of SARS. The intervention prediction model works well. Our results showed that the intervention model considering SARS has much better prediction accuracy than the model without intervention. The monthly data we have are from the official website of Beijing Tourism Bureau [Web09].
2 Tourist Number Model

2.1 ARIMA Model

The monthly data for the number of overseas tourists to Beijing, from year 1997 to 2005 is plotted in Fig. 1. First, note that the tourist number clearly have a seasonal pattern. The non-stationarity of the series is confirmed by the augmented Dickey-Fuller (ADF) test [Wei05], which has p-value 0.04176. Then we take a monthly difference of the series, and the p-value of the ADF test becomes smaller than 0.01, hence the difference series is stationary. The difference series, its autocorrelation function (ACF) and partial autocorrelation function (PACF), are shown in Fig. 2. Hence we postulate a seasonal ARMA(0, 0, 0)(1, 1, 1)[12] model for the tourist number. The fitted coefficients are given by the following table:

<table>
<thead>
<tr>
<th></th>
<th>AR(1)</th>
<th>SAR(1)</th>
<th>SMA(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.7903</td>
<td>0.2195</td>
<td>-1</td>
</tr>
<tr>
<td>p-value</td>
<td>4.551138e-37</td>
<td>5.596678e-02</td>
<td>1.130978e-11</td>
</tr>
</tbody>
</table>

Figure 1: Number of overseas tourist to Beijing, in unit $\times 10^4$ people.
This model has AIC = 550.94, BIC = 561.2, and $\sigma_a^2 = 13.5$. To check the whiteness of the residuals, we use the Ljung-Box test, which has a p-value of 0.5319348. Hence the residuals are white. Note that all the roots of $f(B) = (1 - 0.7903B)(1 - 0.2195B^{12})$ are outside the unit circle, so the model for the difference of the series is stationary. Denote the monthly tourist number as $x(t)$. The model for tourist data is

\[(\text{Tourist Model 1}) \quad (1 - B^{12})(1 - 0.7903B)(1 - 0.2195B^{12})x(t) = (1 - B^{12})a(t). \quad (1)\]

where $a(t) \sim WN(0, 40.09)$, and $WN$ means white noise.

We also used `auto.arima` command in R, which automatically selects model for series. It gave a different result. The fitted model is ARIMA(1, 1, 1), with the coefficients summarized in the following table:
<table>
<thead>
<tr>
<th></th>
<th>AR(1)</th>
<th>MA(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.7083</td>
<td>-0.9652</td>
</tr>
<tr>
<td>p-value</td>
<td>5.057000e-17</td>
<td>3.087335e-140</td>
</tr>
</tbody>
</table>

For this model, AIC = 655.15, BIC = 655.38, and $\sigma^2_a = 40.09$. Hence we have another tourist number model suggested by `auto.arima`:

\[(\text{Tourist Model 2}) \quad (1 - B)(1 - 0.7083B)x(t) = (1 - 0.9652B)a(t). \quad (2)\]

where $a(t) \sim WN(0, 25)$. The p-value of the Ljung-Box test statistics is $2.753607e-06$. Note that all the roots of $1 - 0.7083B$ are outside the unit circle, so the model for the difference of the series is stationary.

To diagnose these models, we plot the ACF and PACF of the Tourist Model 1 and 2 in Fig. 3. Note that the residuals of Model 1 is white; not white for Model 2.

**Figure 3:** Upper: ACF and PACF of Tourist Model 1; Lower: ACF and PACF of Tourist Model 2.

**Remark:** Note that Tourist Model 1 has smaller AIC and BIC values than Tourist Model 2. Also, the residual of Tourist Model 2 is non-white. Later on we will show that Tourist
Model 1 have better prediction mean-squared-error (MSE) than Tourist Model 2. The model suggested by `auto.arima` is not the best.

### 2.2 Tourist Model with Intervention

Due to SARS starting from March 2003, we can see a clear drop in the number of tourist in Fig. 1. To model this effect, we built an intervention model. Model the starting of SARS as an indicator

\[
I(t \geq 77) = \begin{cases} 
1, & t \geq 77; \\
0, & \text{Otherwise.}
\end{cases}
\]  

(3)

since SARS at Beijing starts in March of 2003, which is the 77th data point.

![CCF of tourist number and indicator.](image)

Figure 4: CCF of tourist number and indicator.

Next we plot the cross correlation function (CCF) of the indicator and the tourist number in Fig. 4. Note that delay 12 has the largest value. So we include a regressor of the delayed indicator function with delay 12. The fitted result is shown in the following Table. The p-value of the coefficient of indicator is 1.09e-01, not very significant, but we can include it at a level of 0.1. This means the effect of SARS on the number of tourist is temporary. The plot for the ACF and PACF of the fitting residuals are given in Fig. 5. Note that there
Figure 5: ACF and PACF of the residual after we fit seasonal ARIMA(1, 0, 0)(1, 1, 1) plus a multiple of indicator. Note that the residual is non-white and we can fit an seasonal ARIMA model of order 1 frequency 12.

is a large spike at delay 12, which means the noise is non-white and hence we can build a seasonal ARIMA(1, 0, 0) for the residual. Finally, the coefficients of the interventional model for tourist is given in the following Table:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>AR(1)</th>
<th>SAR(1)</th>
<th>SMA(1)</th>
<th>I(t ≥ 89)</th>
<th>noise SMA(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.3809</td>
<td>0.3809</td>
<td>-0.9780</td>
<td>5.4338</td>
<td>0.6319</td>
</tr>
<tr>
<td>p-value</td>
<td>NA</td>
<td>NA</td>
<td>1.243301e-41</td>
<td>1.096e-01</td>
<td>1.892211e-07</td>
</tr>
</tbody>
</table>

And the intervention model we have is

(Tourist Intervention Model)

\[ (1 - B^{12})(1 - 0.3809B)(1 - 0.3809B^{12})x(t) = (1 - 0.9780B^{12})(1 + 0.6319B^{12})a(t) + 5.4338I(t ≥ 89). \] (4)

where \( a(t) \sim WN(0, 15.64) \). The ACF and PACF of the residual is shown in Fig. 6, which is white.
2.3 Prediction Using Tourist Number Models

We set aside data in year 2005, use data from 1997 to 2004 to fit our three models: Tourist Model 1, Model 2, and the Intervention Model, and use them for a one-step-ahead prediction. The predictions are shown in Fig. 7, and the prediction MSE is summarized in the following table. Note that the interventional model has the smallest MSE, though it has higher complexity.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tourist Model 1</td>
<td>24.91515</td>
</tr>
<tr>
<td>Tourist Model 2</td>
<td>27.41734</td>
</tr>
<tr>
<td>Tourist Intervention Model</td>
<td>18.57392</td>
</tr>
</tbody>
</table>

References

**Figure 7:** Black Line & Dots: data, Red Line & Dots: Tourist Intervention Model, Blue: Tourist Model 1, Green: Tourist Model 2. The lower plot is the zoom-in of upper plot for the prediction years.