Stat 153 (Spring 2018) Final Report
Hotel Occupancy Time Series Modeling and Prediction
Nate Huang & Dodo Qian

1. Introduction

Hotel business is an important component of leisure industry. Hotel demand and occupancy rates fluctuates widely based on the time of year. There are several motivations that a forecasting model for the hotel occupancy could be useful. First, the prediction of rooms demand in the future could generally help to improve the management of the business and maximize the profit. For example, the hotel can arrange some events and increase the price during the peak seasons to better serve the customers. During low seasons, the hotel can advertise for promotions ahead of time. Second, the prediction can assist in short-range hiring and lay-off decisions. Third, the prediction can also help the hotel to understand current situation better. If the occupancy is low, is it because the hotel become less attractive to travelers? Or it is just because a normal low season? Lastly, since the prediction gives a rough indication of the tourism in this area (especially in a tourist city), the city can plan for activities, regulate police and cleaners, adjust open times of scenic spots, etc. to accommodate the change in number of tourists.

In this project, we select the monthly occupancy of O'Donovan's Hotel in Ireland from January, 1963 to December, 1976. This hotel is located in Clonakilty, which is the tourism hub, so the analysis of this data set can give us a better understanding on the changes in hotel demand in a tourist region. This data set is obtained from the Times Series Data Library created by Rob Hyndman, Professor of Statistics at Monash University, Australia. For our analysis, we mainly focus on building the model in two ways: Spectral Analysis and ARIMA.

2. Data Analysis

2.1 Exploratory Data Analysis (EDA)

Based on the plot of the data (Fig.1), there is no obvious outliers. To verify this, an outlier test is performed by the *tsoutliers* package in R. It shows that there is no additive outlier, innovation outlier, level shift, temporary change in our data set. From the graph, the data have an increasing trend and slightly unequal variance, and thus is not stationary. We first take a log transformation to our data to stabilize the variance in the time series. Linear regression and Differencing will be used respectively to chase stationarity in Spectral Analysis and ARIMA for our convenience. From the graph, it is reasonable to expect an increasing trend and periodicity behavior for future prediction.

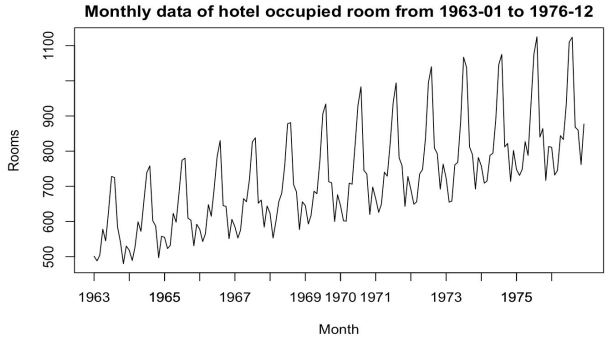

Fig.1. Plot of the data

In order to test the validation of the model, we take out the last two years' data (1975.1-1976.12) as the test set to verify the performance of the models; the remaining data (1963.1-1974.12) will be the training set to build the model.

2.2 Spectral Analysis Model

Spectral analysis is focusing on the frequency domain approach to analyze a time series. Any stationary time series might be thought of approximately as a random combination of sines and cosine waves oscillating at different frequencies and amplitudes. This fact can be utilized to study the periodic behavior of a time series. Our goal is to identify the key frequencies that contribute to the periodic behavior.

In this part of analysis, periodogram is used to identify the key frequencies of a time series. The periodogram is an estimation (sample version) of the spectral density, and it measures the relative importance of different frequency values that might explain the variance in the time series. It seems like there is an increasing linear trend in the log data, so we detrend it first through linear regression to make it stationary (Fig.2) for using periodogram.

The raw periodogram without any smoothing and tapering looks rough(Fig.3). This is because we only use the discrete fundamental harmonic frequencies for the periodogram, but the spectral density is defined over a continuum of frequencies.

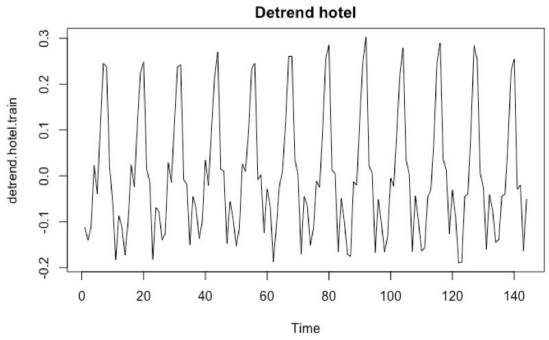

Fig.2. Detrend data

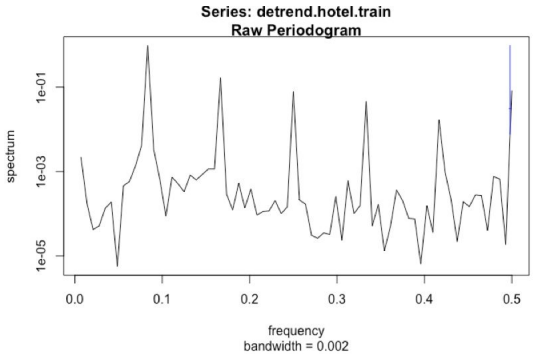

Fig.3. Raw periodogram

To better identify the key frequencies, some smoothing methods can be used to improve the periodogram. Daniell kernel smooths the data by moving averages. Tapering applies more weight in the center of the data than the ends of the data, so it can reduce spectral leakage. After trying these methods (Daniell kernel, modified Daniell kernel, convolution of two/three modified Daniell kernel, tapering) with various spans, a smoother periodogram with sharper peaks (less bias) is made. The key frequencies are indicated by the vertical dashed line (Fig.4).

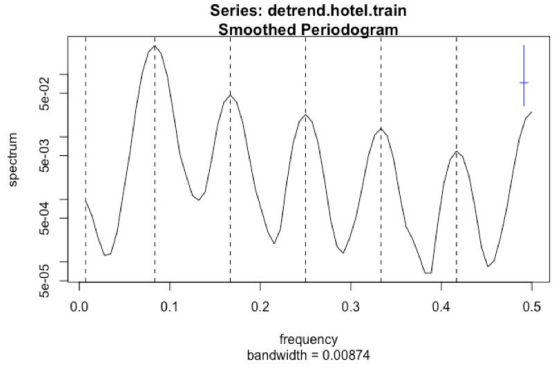

Fig.4. Key frequencies labeled

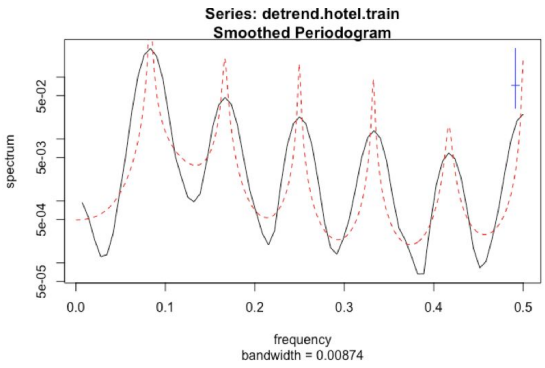

Fig.5. Nonparametric vs. Parametric spectrum

The nonparametric spectral (smoothed periodogram) is compared with the parametric spectral (red dotted line) (Fig.5). All the peaks overlap, so the two methods agree with each other. We use all the five peaks as the top frequencies to generate features (sine and cosine terms) for the assumed model:

$$x_t = \alpha + \beta t + \sum_{j=1}^{5} \left[c_j cos(2\pi\omega_j^* t) + d_j sin(2\pi\omega_j^* t) \right] + w_t$$

Where ω_j^*'s are the key frequencies, α, β, c_j, d_j are unknown parameters, and w_t is iid $N(0, \sigma^2)$.

This model includes the linear trend as well as the periodic terms. We know the values for t and all the sine and cosine terms, so we can fit a linear regression to those values, and the coefficients can be calculated. After the model is identified, we can plot the fitted values of the data along with the original data (Fig.6). From the graph, the fitted values capture the trend and periodicity. Visually, we can see that the fit is good. To better analyze the fit, the residuals are plotted (Fig.7), and they looks randomly scattered around zero, indicating that the model describes the data well.

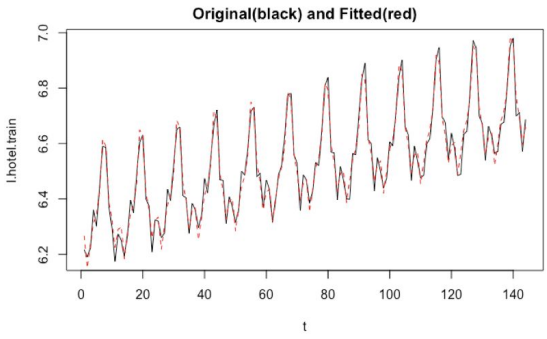

Fig.6. Original(black) vs. Fitted(red)

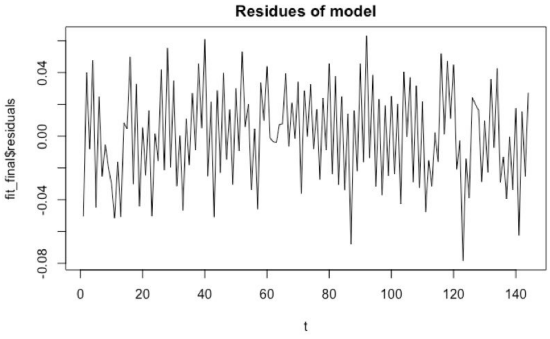

Fig.7. Residues of the model

Then we can predict the test set (last two years) using the above model from the train set. The prediction (in blue) fits the original data (in black) very well (Fig.8). All the original data lie inside the confidence interval (purple dotted line) except one point (Fig.9). The mean square error(MSE) is calculated to be 1173.

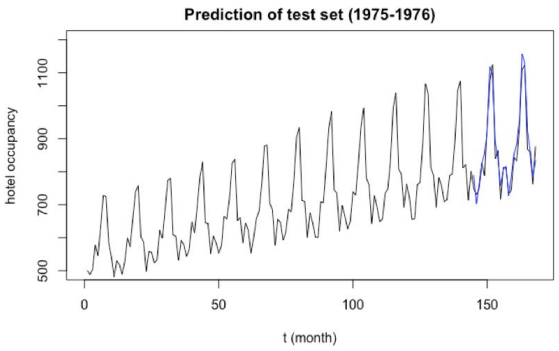

Fig.8. Prediction of test set (1975.1-1976.12)

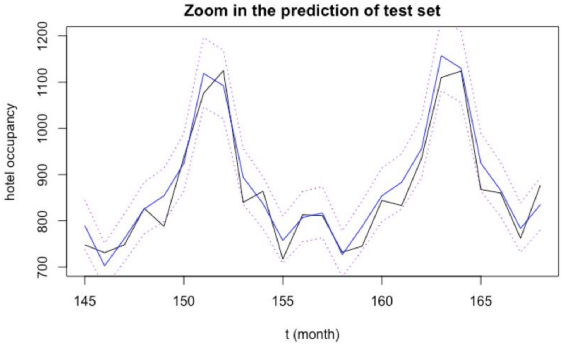

Fig.9. Zoom in of prediction of test set

2.3 ARIMA Model

ARIMA is focusing on the time domain approach to analyze a time series. ARIMA model assumes that the data at a time point depend on the previous data and random noises. AR(Autoregression) means the data is regressed on it own lagged values; MA(Moving Average) means the random noises from the past to now contribute to data now. The "I'(Integrated) cooperates the differencing feature, which is sometimes necessary for nonstationary data, of the model. Because we observe seasonal trend in our data, a seasonal ARIMA model can be used to describe and predict the data. In order to satisfy the stationarity requirement, first order differencing is used to eliminate the trend for the log time series. Notice that the first differencing eliminates the stationarity of the data(Fig.10). We once again observed a yearly seasonal behavior from the graph.

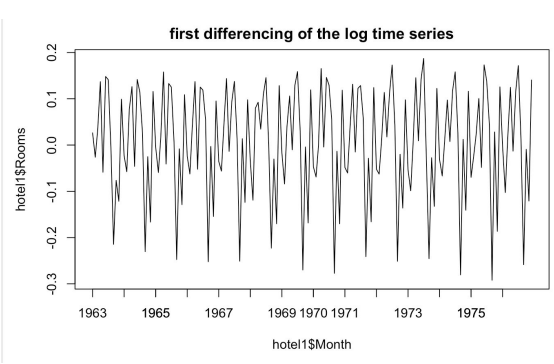

Fig.10. First differencing of the log t.s.

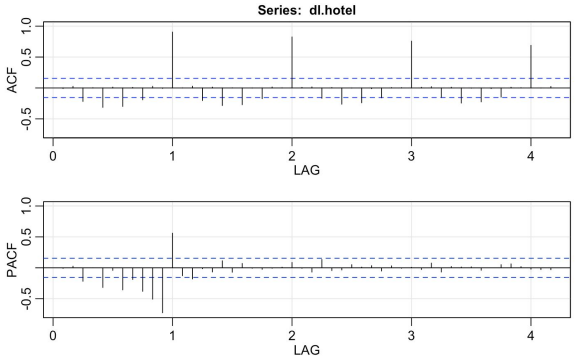

Fig.11. ACF & PACF of the first differencing data

ACF(Fig.11) informs similar behavior. From the ACF graph, there is a high correlation every 12 months, which shows a 12-month seasonality in the data. Comparing the correlation at Lag 1, 2, 3 and 4 in the ACF graph, we can see the correlation decreases slowly in every year. To reduce the yearly correlation, a first 12-month seasonal differencing is used. Therefore, a Seasonal ARIMA model with a 12-month seasonality is a reasonable choice.

In the process of picking parameters for the model, *eacf* function in R was used, which provides little information. The ACF/PACF of the first differencing data/seasonal first differencing data(see appendix) gives little information either. So we pick some sets of parameters and run *sarima* function to see the model's performance. After trying a considerable amount of models and comparing their AIC and residues, we find the model $ARIMA(0,1,6)\times(0,1,1)_{12}$ yields the best performance(Fig.12). The AIC of this model, -7.26553 is reported the lowest among the many models we have tried. From the residue analysis, we cannot reject the model: the standardized residuals are randomly scattered around 0 with constant variance, which shows no clear pattern, indicating that the residuals are very likely to be random noise; the ACF of residuals in different lags are mostly insignificant, which means the residuals are uncorrelated; the normal Q-Q plot roughly follows normal distribution; the p-values for Ljung-Box statistic are above the 0.05 line, showing that the model is adequate.

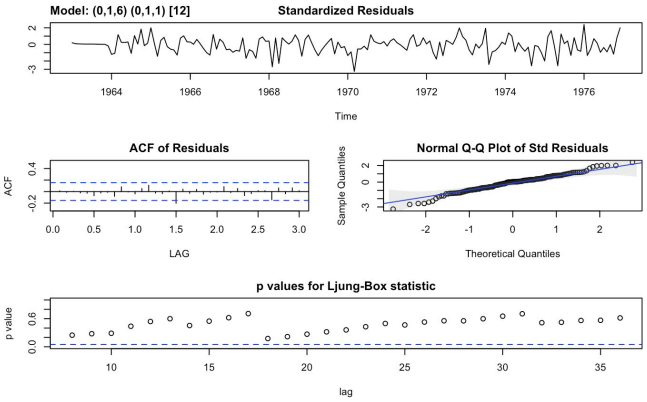

Fig.12. Residuals analysis

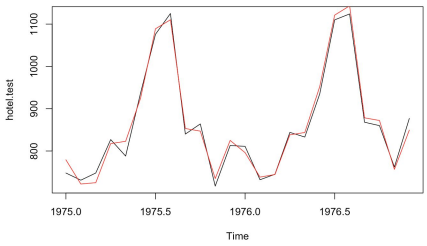

Fig.13. “predicted”(red) vs. real(black) for test set

To compare with the model we obtained from spectral analysis, it is reasonable to evaluate the mean square error for the model reserve some of the data as the training set and use the model to “predict” the test set. From the graph(Fig.13), we can see that the model did a good job on predicting the test set. The mean square error, which is 275, provides a straightforward view of how well the model “predict” the data from 1975 to 1976. More importantly, mean square error will be used in the next section to evaluate the performance between the models provided by spectral analysis and ARIMA.

2.4 Prediction using ARIMA

After comparing the mean square error of Spectral Analysis model and ARIMA model, we think ARIMA model might be a better model for prediction since it has a smaller mean square error in model testing. In the prediction(Fig.14), the seasonal pattern and the upward trend are retained, which is reasonable based on the data. This agrees with our hypothesis at the end of EDA.

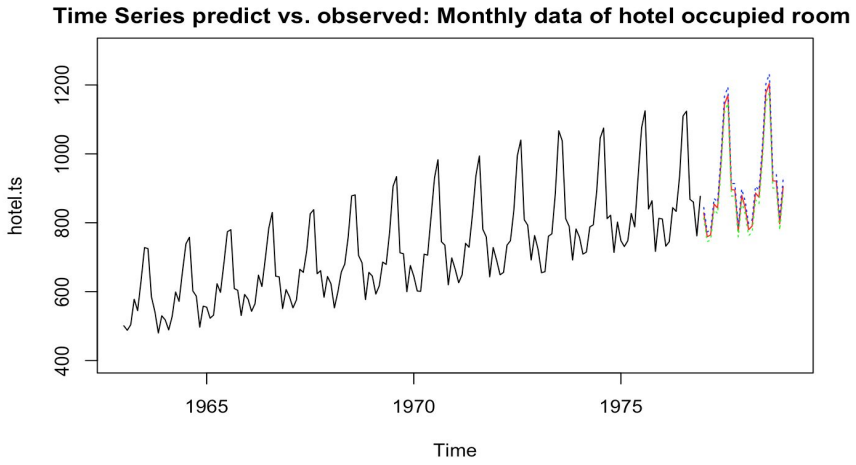

Fig.14. Prediction the next two years of room occupied after 1976

3. Conclusion

As it is known, hotel industry has seasonal pattern in room booking: in some months, the demand of room increases. The observed data in this project captures this pattern. Not surprisingly, the models we built from both Spectral Analysis and ARIMA pertain this feature as well. In this sense, both models are reasonable for the purpose of understanding the data or forecasting. We choose to use the ARIMA model for forecasting because it has a lower mean square error. Based on our prediction, the hotel room demand will keep increasing in trend and vary in different month in a year. The increasing trend will imply that people spend more time on traveling and pay more attention on the living condition. The hotel should expect more customers in general and prepare to serve even more customers in the summer seasons (the hotel is near the beach) in the next two years.

The analysis in this project is only limiting to one hotel in this region. To better understand the hotel performance and supply, further analysis can be done by using more hotels in this region. This information is very useful for cities famous for tourism since a large part of their economy depends on the income from tourism.