Time Series Analysis of 51 Chicago Upper Upscale Hotels

Yimei Tang
DePaul University
ytang30@mail.depaul.edu

***Abstract*— In the area of hospitality, we have three major players: hotel guests who are the end users of the hotel rooms, hotel sales and operation team who set the rate and deliver the best possible services for guests and hotel investors who make the final decisions whether to build or renovate hotels in certain areas. In order to leverage the use of data, I apply knowledge in time series analysis in providing answers for these three major players: how much will a guest expect to pay for a room on average? how many rooms will be sold in the next month? what is the hotel's market value?**

I. INTRODUCTION

In this report, I will explore how much it will cost to book an upper-upscale hotel room on average (among these 51 hotels, see Appendix 1), how many occupied rooms we will expect in these 51 hotels across the busiest areas of Chicago and how much an investor could expect to earn from every hotel room he invested in the following 12 months. My goal in this project is to find the time series relationship (if any) respectively in ADR(Average Daily Rate), Occupied Rooms and RevPAR(Revenue Per Available Rooms) and if there is time series relationships , what is the ideal model for each factor.

II. DATA AND PREPROCESSING

A. Dataset Description

These datasets are collected by STR company from 51 upper-upscale hotels in downtown Chicago and river north area.

The following three types of data are used in this report:

1) Monthly ADR: ADR stands for Average Daily Rate. Monthly ADR are calculated by adding that months daily rate of each hotel in each day and divide them by the number of days in the month. The daily rate of each hotel is calculated by dividing daily room revenue by rooms sold in transient, group and contract three segments (each segment has different daily rates).
2) Monthly Occupied Rooms: Monthly Occupied Rooms are calculated by adding all rooms that are sold each month.
3) Monthly RevPAR: RevPAR stands for Revenue Per Available Room. Monthly RevPAR are calculated by dividing total room revenue generated that month by total supply rooms that month.

B. Data Preprocessing

Although the datasets contain raw data, daily data, weekday and weekend data, monthly data, quarterly data and yearly data, we focus on the monthly data since it is more stable and is a large enough sample size for our project. Monthly data has 147 months values, starting from Jan 2005 to Mar 2017 (inclusively). We will use the first 143 months' data as training set and the rest 4 months' data as test set.

When we began our data cleaning process, we initially proposed to take out the data before 2010 since the economic crisis has impacted the industry of hospitality significantly. However, we found that we could still build dependable models by using the original dataset that has 11-year span. Therefore, we chose not to eliminate data that were observed before 2010.

Since the ADR range from 100 to 300, while the total occupied rooms and total supply rooms are both above 200,000 range, to read the data easier, we divided total occupied rooms and total supply rooms by 1,000.

Example:

TABLE I
DATA EXAMPLE

Date	ADR	Occupied Rooms	RevPAR
6/1/2006	205.81	453,135	174.212
...	...	...	...
11/1/2015	200.25	470,262	155.219

III. DATA ANALYSIS

A. Data Exploration

Step One: Create Time Series Plot We transformed three sets of data into Time Series data (Monthly ADR data, Monthly Occupied Rooms data and Monthly RevPAR data) to create time plots.

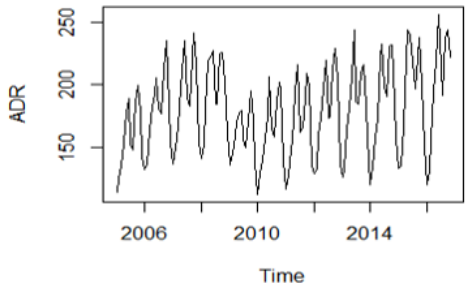

Fig. 1. Monthly ADR Time Series Plot

1) From Fig 1, we can tell that the ADR data has seasonality behavior. There was a growing trend from 2005 to 2008. Than the ADR drop between 2008 to

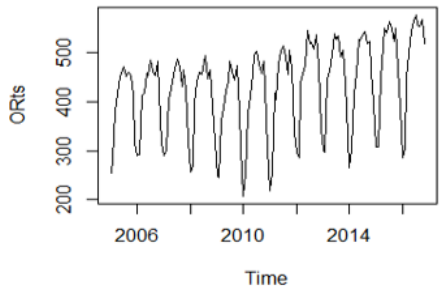

Fig. 2. Monthly Occupied Rooms Time Series Plot

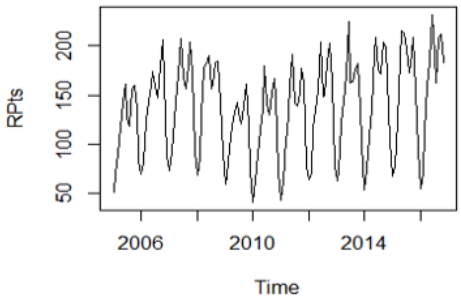

Fig. 3. Monthly RevPAR Time Series Plot

2010. After 2010, the ADR starts to grow again till 2016.

2) From Fig 2, we can see that the number of occupied rooms kept stable between 2005 and 2010 and has a growing trend after 2010.
3) From Fig 3, we can see the revenues per room has similar trend as ADR.

Step Two:Analyze Normal Distribution We create histograms and QQ plots to see the distribution of three data sets, and use JB test to test the data normality.The P-values of JB tests for three data sets are 0.05, 0.004, and 0.04. All P-values are significant in = 0.05 level, which means all three time series data sets are normal distributed.

Step Three: Plot ACF Since we saw seasonality behavior and trends for all three data sets on the time plots, we decide to use auto-correlation function to analyze the stationarity of ADR, Occupied Rooms and RevPAR and their first difference. And we would like to check whether there is evidence for seasonality. Since three sets of data have similar patterns, we only describe the ACFs of ADR data set.

1) From the ACF plots of ADR data set (Fig 4 to Fig 6), we can tell that after de-trending and de-seasonalizing, the data is still not stationary.
2) Both Occupied Rooms and RevPAR data set are also not stationary after de-trending and de-seasonalizing.

B. Model Fitting & Residual Analysis

During the data exploration, we can tell that all three sets of data have similar features. It is explainable since ADR, Occupied Rooms, and RevPAR are all collected in same time

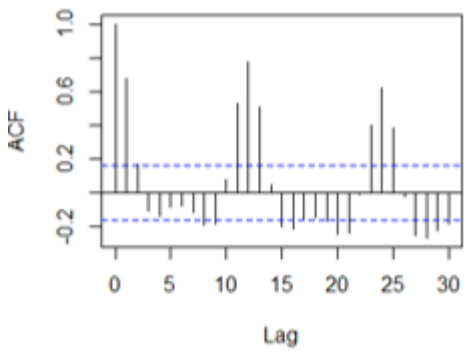

Fig. 4. ACF of Monthly ADR Time Series Plot

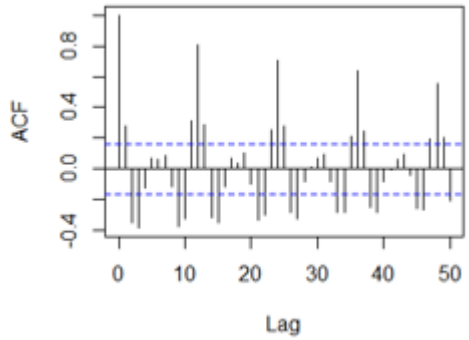

Fig. 5. ACF of 1st diff Monthly ADR Time Series Plot

range. So we will explain our model fitting processes by taking ADR data set as example.

1) We used autoarima function and BIC criterion in R to help primarily select a model for each data set. The output showed that ARIMA (1,0,1) (0,1,1) [12] is the best model. The model is denoted as M1.
   From the figure 7, we can see that the auto-correlation at some lags are still big. So I went further conduct JB test and Ljung Box test on the first models residuals. The result shows that the residuals are independent and normal distributed. Therefore, I think the model is appropriate to fit the ADR data.
   Based on the first model, I tried a lot of other models and selected out following two models as good ones: ARIMA (0,1,1) (0,1,1) [12] and ARIMA (0,1,1) (1,0,1) [12]. Both models residuals were all white noises tested through JB test and Ljung Box test.
2) For Monthly Occupied Rooms, autoarima function selects ARIMA (1,1,1) (0,1,2) [12]. However, Since the residuals are not independent, this model is not appropriate. After several experiments, ARIMA (0,1,6) (0,1,2) [12] was the only model with white noise residuals.
3) For Monthly RevPAR, autoarima function selects ARIMA (2,0,0) (2,1,1) [12]. The residuals are independent without normal distribution. Also, there are two additional models: ARIMA (2,0,1) (0,1,1) [12](the residuals are independent without normal distribution)

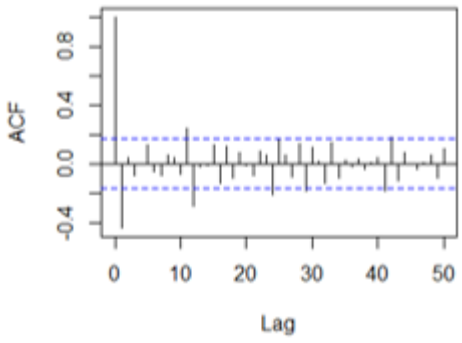

Fig. 6. ACF of 1st diff & seasonal diff Monthly ADR Time Series Plot

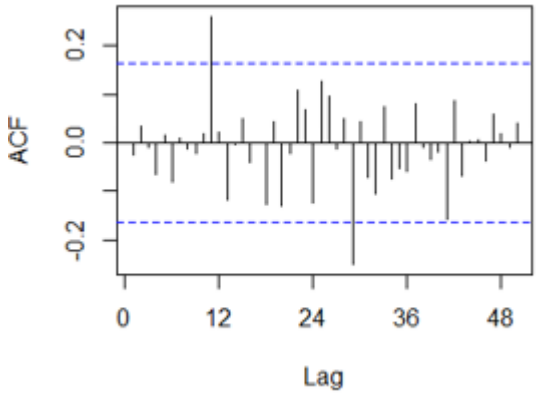

Fig. 7. ACF of M1 Model Residuals

and ARIMA (1,0,1) (1,1,0) [12] (The residuals are white noise)

C. Forecast Analysis

After building the models, we computed 12 months forecast for ADR, Occupied Rooms, and RevPAR. We have computed forecasts for every fitted model. We select three best models and their forecasts plots are shown in Fig 8 to Fig 10.

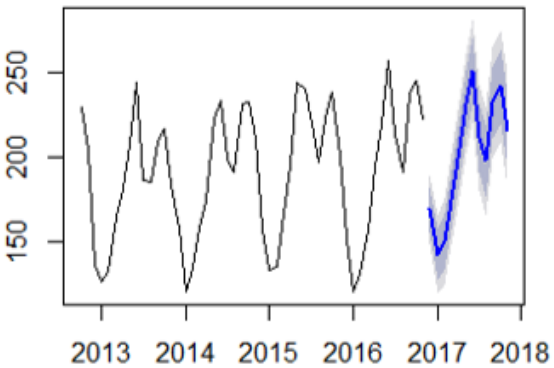

Fig. 8. 12 months Ahead Monthly ADR Forecast

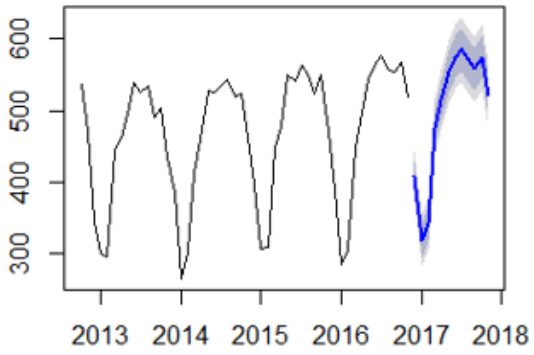

Fig. 9. 12 months Ahead Monthly Occupied Rooms Forecast

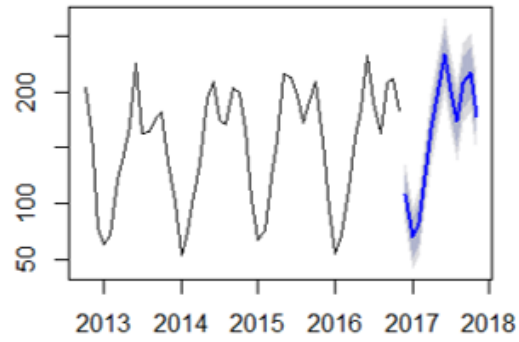

Fig. 10. 12 months Ahead Monthly RevPAR Forecast

D. Model Validation

1) From the forecast, we can tell that our models are catching the time series trends. To test the goodness of fit of our models, we used Back testing to validate models. We fitted models with 80% of our data and tested with the other 20%. After computing the Mean Absolute Percentage Errors between all candidate models for one data set, we chose a best model for ADR data and Occupied Rooms data. ADR Model: ARIMA (1,0,1) (0,1,1) [12] The MAPE for this model is 4.96%
2) Occupied Rooms Model: ARIMA (0,1,6) (0,1,2) [12] The MAPE for this model is 2.6%
3) For RevPAR data, though the lowest MAPE (6.73%) model is ARIMA (2,0,0) (2,1,1) [12], this models residuals are independent without normal distributed. So we decided to choose the model of ARIMA (1,0,1) (1,1,0) [12] with white noise residuals whose MAPE is lightly higher at 7.12%.

E. Model Performance

1) For monthly ADR model, the forecast error is bad during the first three months' forecast. However, the fourth month's forecast error is getting better. Comparing the outcome with monthly occupied rooms, I believe the ADR model's poor performance is due to other dependent factors that we should explore and take into consideration when building our time series model.

TABLE II

MONTHLY ADR FORECAST ERROR

Month	Forecast	Actual	MAPE
Dec 2016	168.90	139.92	-20.71%
Jan 2017	142.25	126.00	-12.89%
Feb 2017	149.62	130.53	-14.63%
Mar 2017	174.51	161.35	-8.16%

Possible reasons might be the increase of lodging tax or surplus of available lodging options.

2) For monthly occupied rooms model, the forecast error is relatively small. This means that the model generalizes well to the test data.

TABLE III

MONTHLY OCCUPIED ROOMS FORECAST ERROR

Month	Forecast	Actual	MAPE
Dec 2016	409,211	384,819	-6.34%
Jan 2017	318,778	311,462	-2.35%
Feb 2017	342,548	330,884	-3.53%
Mar 2017	476,226	470,345	-1.25%

3) For monthly RevPAR model, the forecast error shows the same pattern as that of monthly ADR. This is due to that monthly RevPAR is a dependent variable of monthly ADR.

TABLE IV

MONTHLY REVPAR FORECAST ERROR

Month	Forecast	Actual	MAPE
Dec 2016	108.17	82.38	-31.30%
Jan 2017	69.66	60.46	-15.21%
Feb 2017	82.33	73.67	-11.76%
Mar 2017	125.09	116.92	-6.99%

IV. CONCLUSION

SARIMA Model:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t$$

for

$$Z_t \sim WN(0, \sigma^2)$$

A. For Hotel Guests

How much does a hotel guest expect to pay for a hotel room on average in these 51 hotels?

Based on our statistical analysis, the best model (least MAPE) we found for forecasting ADR is SARIMA (1,0,1) (0,1,1) [12]

$$(1 - B^{12})(1 - \Phi(B))X_t = (1 - \theta_1(B))(1 - \theta_2 B^{12})a_t$$

for

$$a_t \sim WN(0, \sigma^2)$$

After replacing the parameter with actual value, the model is:

$$(1 - B^{12})(1 - 0.93B)X_t = (1 - 0.38B)(1 - 0.68B^{12})a_t$$

```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1   0.931728   0.040832 22.8185 < 2.2e-16 ***
ma1  -0.384822   0.099908 -3.8518 0.0001173 ***
sma1 -0.681488   0.088885 -7.6671 1.759e-14 ***
```

Fig. 11. Coefficients of Monthly ADR Model

Therefore, if you are going to stay at one of the 51 hotels, your average daily rate will greatly depend on the same month of last years rate, previous month of last year and the previous month of this years rate.

If you are a savvy customer, you will probably pick January, February, March and December. If you are visiting Chicago in June, September and October, you should expect to spend more in hotel rooms than other months. However, July and August will be better choices than June if you want to visit Chicago in the summer. (Fig 12)

For hotel operators, it is not worthwhile to have price war during winter since the price is already low enough. However, hotels that want to capture more market share could lower their ADR within a reasonable range in June, September and October to attract price-sensitive guests.

On the other hand, as we could tell from the ACF plot and the model itself, if the previous months ADR is above its average, then next months ADR will increase above its average. However, if last years month ADR is above its average, then the months ADR this year will be below its average. Therefore, if we have a higher than average ADR in March last year and February this years ADR is lower than average, we will expect March this years ADR will below average as well.

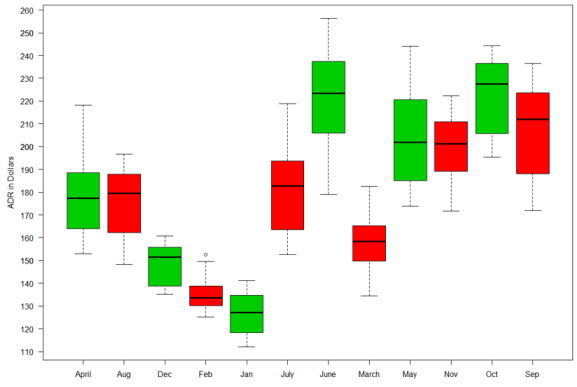

Fig. 12. Boxplot of Monthly ADR

B. For Hotel Operations

How many occupied rooms will hotels expect in the next 12 months? Based on our statistical analysis, the best model (least MAPE) we found for forecasting Occupied Rooms is ARIMA (0,1,6) (0,1,2) [12].

After replacing the parameter with actual value, the model is:

$$(1 - B^{12})(1 - B)X_t =$$

$$(1-0.84B)(1-0.31B^6)(1-0.46B^{12}-0.23B^{24})a_t$$

```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ma1  -0.840143   0.099331 -8.4580 < 2.2e-16 ***
ma6  -0.307137   0.064581 -4.7558 1.976e-06 ***
sma1 -0.461982   0.099224 -4.6559 3.225e-06 ***
sma2 -0.227047   0.095711 -2.3722   0.01768 *
```

Fig. 13. Coefficients of Monthly Occupied Rooms Model

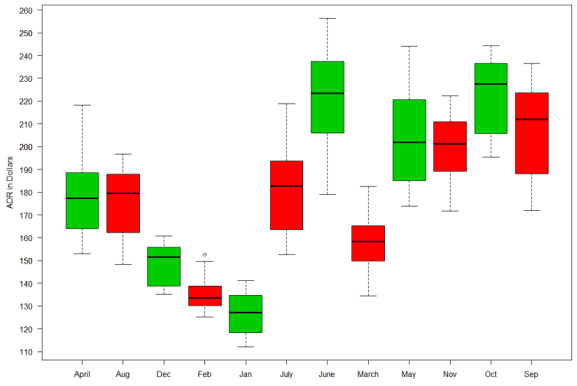

Fig. 14. Boxplot of Monthly Occupied Rooms

Therefore, the occupied rooms per month is influenced by occupied rooms from the same month last year, previous month last year, previous month this year and 6 months ahead this years occupied rooms. For hotel guests, just like the trend of ADR, in January, February and December, the demand (which is quantified in the occupied rooms) is lowest and therefore you will usually enjoy more comprehensive service in these three months compared to other busier months. For hotel operations, these are the slow seasons so it might be a good time for hotel renovation and conduct hotel staff training. Also, these three-months low occupancy should be taken into consideration for marketing strategy and pricing strategy, as well as hotel staffing guidance in doing yearly budget. We know that if the previous months occupied rooms are below its average, we will expect to see current months occupied rooms to be above its average. It is reasonable in real life. If some tourists are not visiting Chicago in a month, they might choose the month before that month or after than month (usually within 3 months span). What interesting is that if the month last year has below average occupied rooms, there will be above average occupied rooms in the same month this year. Our assumption for this behavior is that tourists usually dont visit Chicago two years in a row.

C. For Hotel Investors

How many revenues will a hotel room generate in the next 12 months?

Based on our statistical analysis, the best model (least MAPE) we found for forecasting RevPAR is: ARIMA (1,0,1) (1,1,0) [12]

$$(1-B^{12})(1-0.94B)(1+0.36B^{12})X_t=(1-0.57B)a_t$$

```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1   0.940279   0.036643 25.6609 < 2.2e-16 ***
ma1  -0.578007   0.081020 -7.1341 9.739e-13 ***
sar1 -0.359425   0.086502 -4.1551 3.251e-05 ***
```

Fig. 15. Coefficients of Monthly RevPAR Model

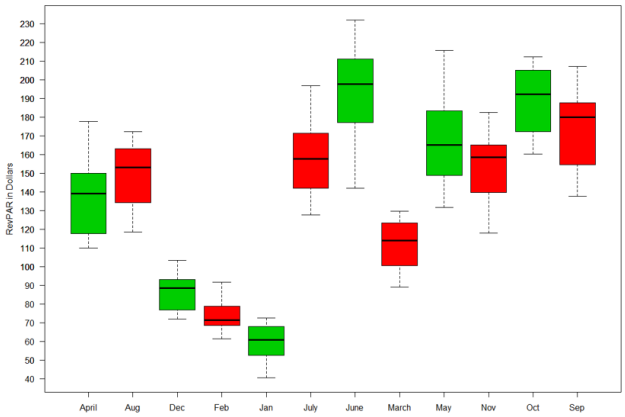

Fig. 16. Boxplot of Monthly RevPAR

Similar to our results from ADR and Occupied rooms, hotel investors are not generating much revenue from hotel rooms they invested in winter. Therefore, hotel investors should consider close its hotels for renovations or major upgrades in winter. Hotel investors should also consider the time it takes to construct a hotel and open new hotels in June, September or October to generate enough cash flows for future operations. It might worth postpone opening dates from February to March since the upside is much bigger.

Reading the ACF and PACF chart in addition to the model itself, we know that if the previous months RevPAR is below its average, this month s RevPAR will fall below its average as well. This will be a useful sign for hotel operator team: if they fail to meet their monthly revenue goal this month, there is great chance that they cant meet next months revenue goals. Therefore, they should be alert and take actions to remedy the situation right away. However, if previous month last years RevPAR fell below its average point, the RevPAR of the month this year might be above the average.

APPENDIX

1) List of 51 Hotels could be found here: GitHub Link
2) R code could be found here: GitHub Link

ACKNOWLEDGMENT

I am very grateful for Professor Lisa Thomas from School of Hospitality Leadership of DePaul University in providing datasets for this project.