

# SYDE 631 Course Project: LinkCity Customer Traffic Time Series

*Baoshi Sun, WatID# 20625524*

*Monday, November 30, 2015*

## Synopsis

In the course SYDE 631 (Time Series Modelling, 2015 Fall) by Professor Keith Hipel, the rich variety of time series models that are defined, explained and illustrated. This project attempts to apply the knowledge and skills acquired from the course to solve an essential problem for retail business: the customer traffic analysis. Fresh data from a shopping mall are adopted. Seasonal model is mainly used to fit and forecast the customer traffic, cross validation is conducted, and a number of further research topics are proposed as well.

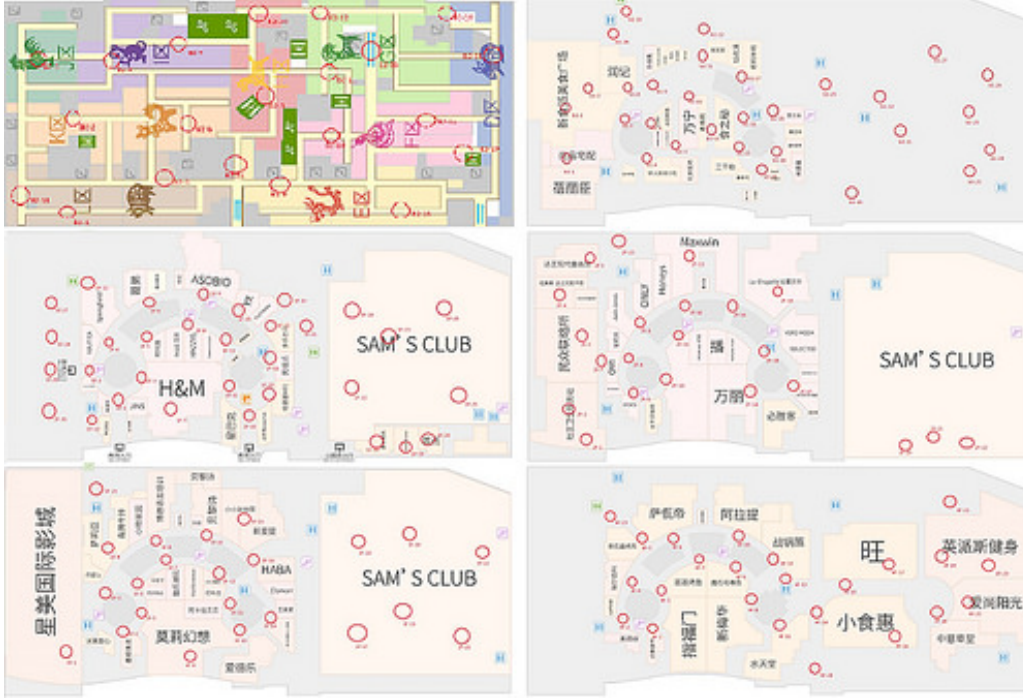
## Background

### About LinkCity

LinkCity is a middle-sized shopping mall in Suzhou, China. The 6-storeyed mall sized 600,000 square foot offers shopping, catering, cinema, gaming, supermarket, and many other retail services. In early 2015, LinkCity deployed more than 200 Wi-Fi access points, so that all business areas and parking lot are covered with Wi-Fi signal.



With the new Wi-Fi infrastructure, LinkCity offers high speed free Wi-Fi access to its customers. On the other hand, the Wi-Fi system can record a lot of useful data of users' mobile devices, such as device presence status, dwelling duration, shopping path, phone number, device type, customer geo-distribution, and so on.



To reveal the value of these data, a few systems, for example customer grading system and recommendation system, are developed. However, they are just a small fraction of what could be studied. Much more topics still wait to be explored. One of the most important topic is the customer traffic analysis and forecast.

If we assume the number of Wi-Fi users can reflect the population of customers, we can use the statistics of Wi-Fi users to estimate the overall customer traffic.

## Motivation of Study

Although it is readily comprehensible that “how understanding customer traffic patterns can help good retailers become great retailers” (Mark Ryski 2011), little empirical evidence exists on the insight of store traffic characteristics (Perdikaki, O., Kesavan S., & Swaminathan J.M. 2012). In addition, because of the lack of traffic data, many studies use the number of transactions as a proxy for store traffic (e.g., Walters and Rinne 1986, Walters and MacKenzie 1988).

As the Wi-Fi users can reflect the population of retail customers, it is reasonable to perform traffic analysis on the data of Wi-Fi users. In recent years, researchers and business operators began to employ Wi-Fi data over a variety of venues to highlight differences in traffic volumes and patterns (Ghosh, A. 2011).

On the other hand, few articles, which make use of time series to study the customer traffic, could be found. However, intuitively, given a specific venue, for example a shopping mall or a specialty store, the customer traffic should be able to be expressed as a function of time.

## Data and Methodology

In this study, the Wi-Fi user login records in LinkCity from May 1, 2015 to September 24, 2015 are obtained. For each record, we picked four data fields (variables) for traffic analysis, including:

- “loginid”: Wi-Fi user login ID, such as phone number or social media username
- “nasidentifier”: Wi-Fi Access Point Mac address, which can be used to determine user location
- “callingstationId”: Wi-Fi device Mac address
- “responsetime”: Wi-Fi user login time

During the project, the second batch of data (September 25 to the end of October) were received. These data are used for model validation. Of course, we can also combine the two batches of datasets together, re-train our models and examine the influence of the enlarged size of observations.

Considering the purpose of this project is to practise the learning in the course, time series modelling methodology, such as seasonal models, TFN, intervention analysis, etc., will be employed as much as possible.

## Reproducible Research

Besides the output article in PDF or HTML format, the work of this project can be examined and reproduced by running codes on raw data. The codes are embeded in a R markdown file (RMD), and the raw data are stored in csv files. The RMD writeup and data files can be found in SYDE631 directory at <https://github.com/sunbaoshi1975/UWStudy.git>.

## Exploratory Data Analysis

### Load the Wi-Fi raw data

Each row of the raw data represents a record of a Wi-Fi user's login operation. To narrow down the size of data table, irrelevant fields are filtered out except for the four variables those we mentioned before.

The local Weather condition data and the local CPI data are also loaded. The time spans of both datasets should be the same as the Wi-Fi raw data, say between May 1, 2015 and September 24, 2015.

### Preprocess

Since we intent to conduct a daily based time series analysis, the raw data should be preprocessed beforehand. The process consists of two major steps:

- To trim out the data points earlier than '2015-05-01 00:00:01' and later than '2015-09-24 23:59:59'
- To aggregate data on daily basis and count the number of user logins in each day

Calcualte the time span of the dataset.

```
## [1] 2015-04-08 16:15:24.233
## 273841 Levels: 2015-04-08 16:15:24.233 ... 2015-09-25 09:58:14.091
```

```
## [1] 2015-09-25 09:58:14.091
## 273841 Levels: 2015-04-08 16:15:24.233 ... 2015-09-25 09:58:14.091
```

Only keep data between 2015-05-01 00:00:01 to 2015-09-24 23:59:59.

Then the detailed records are aggregated on daily basis, so that each record in the processed dataset represent the number of logins during a specific date.

Samples of login records (raw data):

```
##      loginid  nasidentifier      responsetime  LoginDate
## 1 18852404253      B1-03-0061 2015-05-01 00:42:39 2015-05-01
## 2 15186075051      B1-19-0056 2015-05-01 01:22:52 2015-05-01
## 3 15186075051      B1-19-0056 2015-05-01 01:35:23 2015-05-01
## 4 13584833983 3F-SAM-17-0104 2015-05-01 05:50:07 2015-05-01
## 5 15995897097      1F-14-0198 2015-05-01 06:35:07 2015-05-01
```

##	loginid	nasidentifier	responsetime	LoginDate
## 270479	18915135559	B1-18-0052	2015-09-24 22:47:10	2015-09-24
## 270480	18913508781	B1-19-0056	2015-09-24 22:59:02	2015-09-24
## 270481	13801351389	3F-02-0026	2015-09-24 23:18:36	2015-09-24
## 270482	13801351389	3F-03-0217	2015-09-24 23:27:26	2015-09-24
## 270483	18699144339	3F-03-0217	2015-09-24 23:51:11	2015-09-24

Raw Wi-Fi login records summary:

```
## [1] " Total rows: 270,483 from 2015-05-01 00:00:01 to 2015-09-24 23:59:59"
```

“Samples of aggregated data on daily basis:”

##	LoginDate	count	"..."	LoginDate	count
## 1	2015-05-01	2437	...	2015-09-15	1438
## 2	2015-05-02	2439	...	2015-09-16	1607
## 3	2015-05-03	2169	...	2015-09-17	1527
## 4	2015-05-04	1316	...	2015-09-18	1763
## 5	2015-05-05	1369	...	2015-09-19	2395
## 6	2015-05-06	1240	...	2015-09-20	2062
## 7	2015-05-07	1139	...	2015-09-21	1276
## 8	2015-05-08	1268	...	2015-09-22	1294
## 9	2015-05-09	1754	...	2015-09-23	1301
## 10	2015-05-10	1756	...	2015-09-24	1390

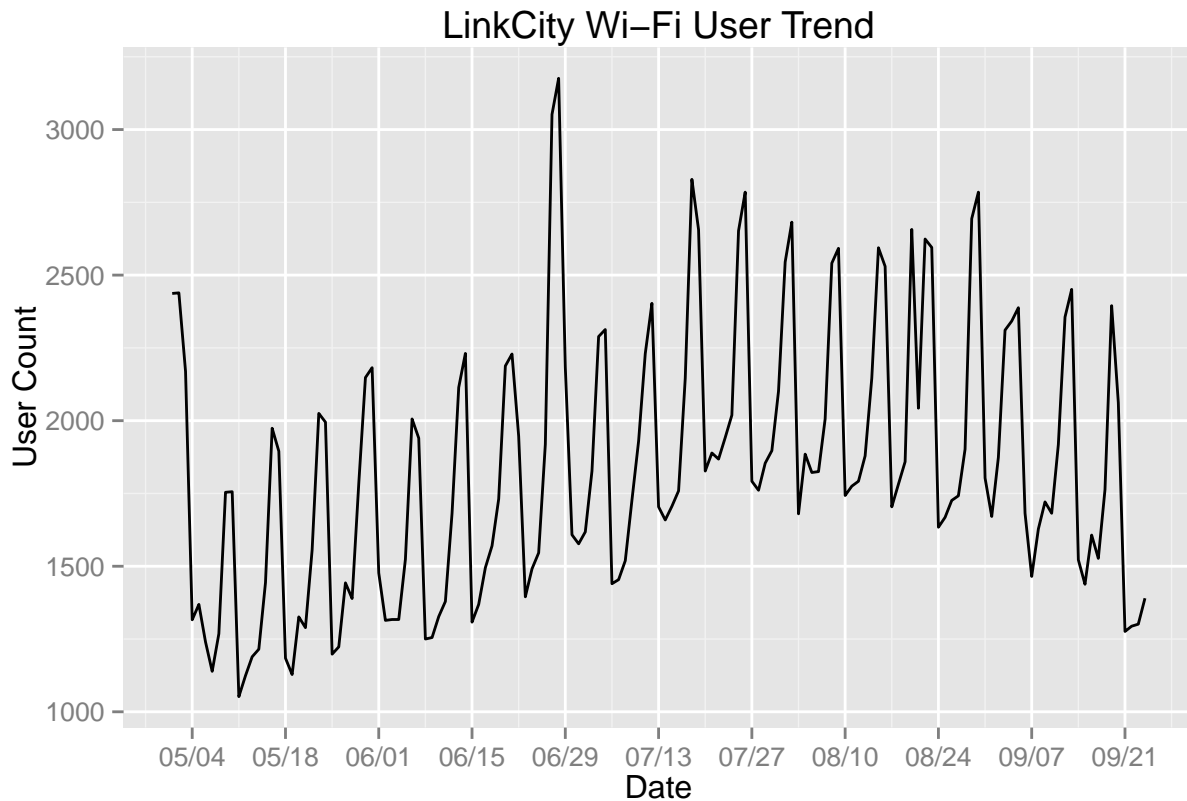
##	LoginDate	count
## Min.	:2015-05-01	Min. :1052
## 1st Qu.	:2015-06-06	1st Qu.:1484
## Median	:2015-07-13	Median :1775
## Mean	:2015-07-13	Mean :1840
## 3rd Qu.	:2015-08-18	3rd Qu.:2146
## Max.	:2015-09-24	Max. :3176

A rough summary of the daily traffic can be seen above. There are totally 270,483 login records throughout 147 days or 21 weeks.

## Plotting

From the plot of daily number of Wi-Fi users below, we can observe some characteristics:

- Nontationary, which can also be verified by seansonal Mann-Kendall test
- Periodic on weekly basis
- Some exterem values, e.g. around May 1 and June 27



Apply seasonal Mann-Kendall test to check the trend:

```
SeasonalMannKendall(ts.df)
```

```
## tau = 0.41, 2-sided pvalue =6.3027e-12
```

The test result indicates a significant upward trend.

In addition, using the `decompose()` function, we can roughly break the data series into three parts: the trend component, the seasonal component and the white noise. In other words, we assume:

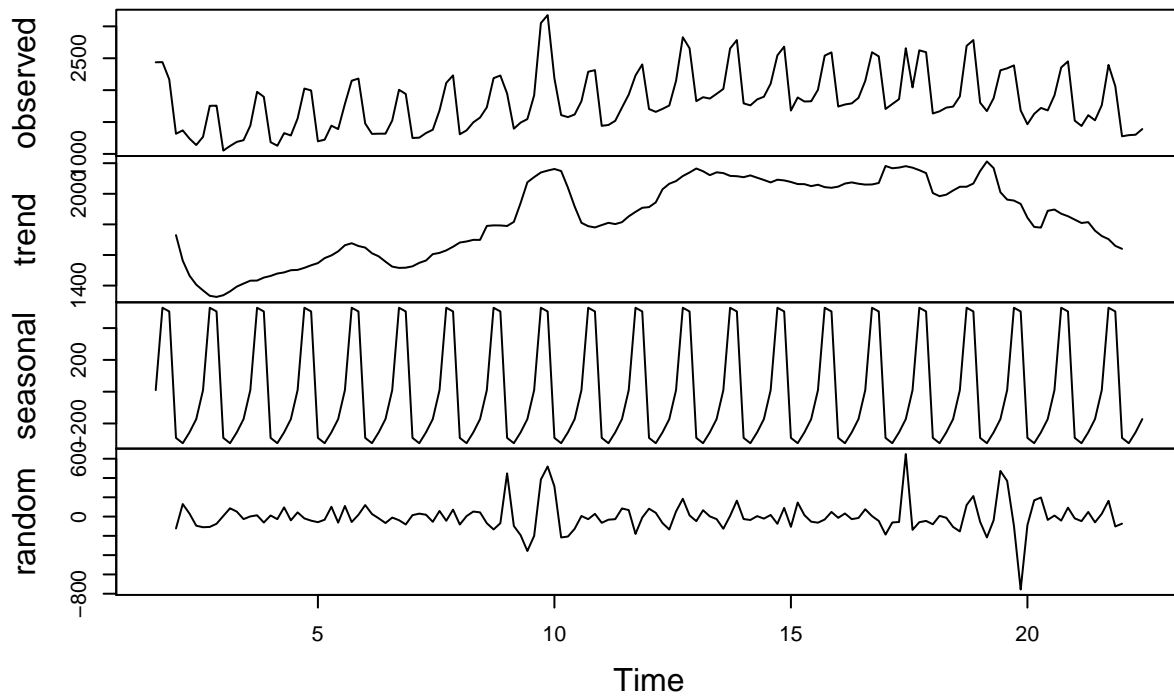
$$Output = Trend + Seasonal + Noise$$

where the output is the time series of Wi-Fi user data

By examining the plots of the decomposition of additive time series, the three characteristics get strong support.

Moreover, the nonstationary trend appears to be a curve, which may be assumed as a longer seasonal circle, e.g. monthly or quarterly. From the physical point of view, the monthly or quarterly pattern of traffic somehow holds its stand. With more observations collected in the future, we can perform seasonal analysis on monthly and quarterly basis.

## Decomposition of additive time series

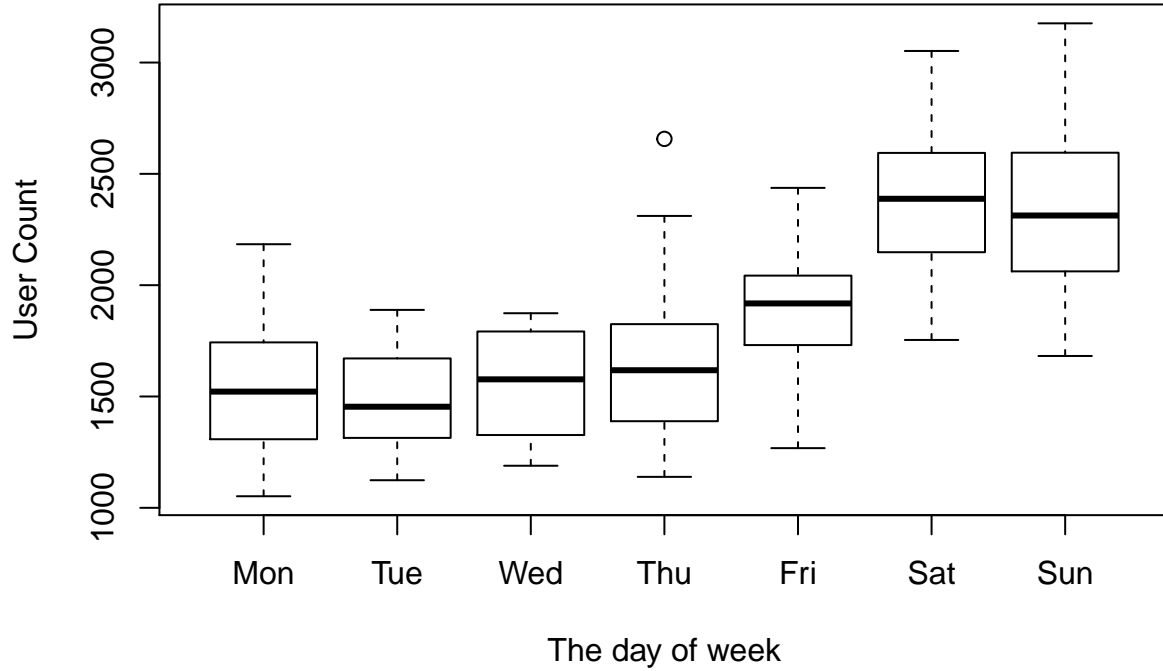


When we take a look at the ACF and PACF, the possible modelling directions may be pointed out.

1. The ACF and PACF of overall time series indicate seasonal models, weekly pattern in this case, should be required.
2.  $AR(p)$  process may be needed apparently.
3.  $MA(q)$  component might also be needed.

To examine the data of each day throughout a week, Box-and-whisker graphs are plotted out. From the plots, we can assume that the distributions of all means are normal and logarithmic transformation is not necessary.

## LinkCity Wi-Fi User Count vs. Weekday

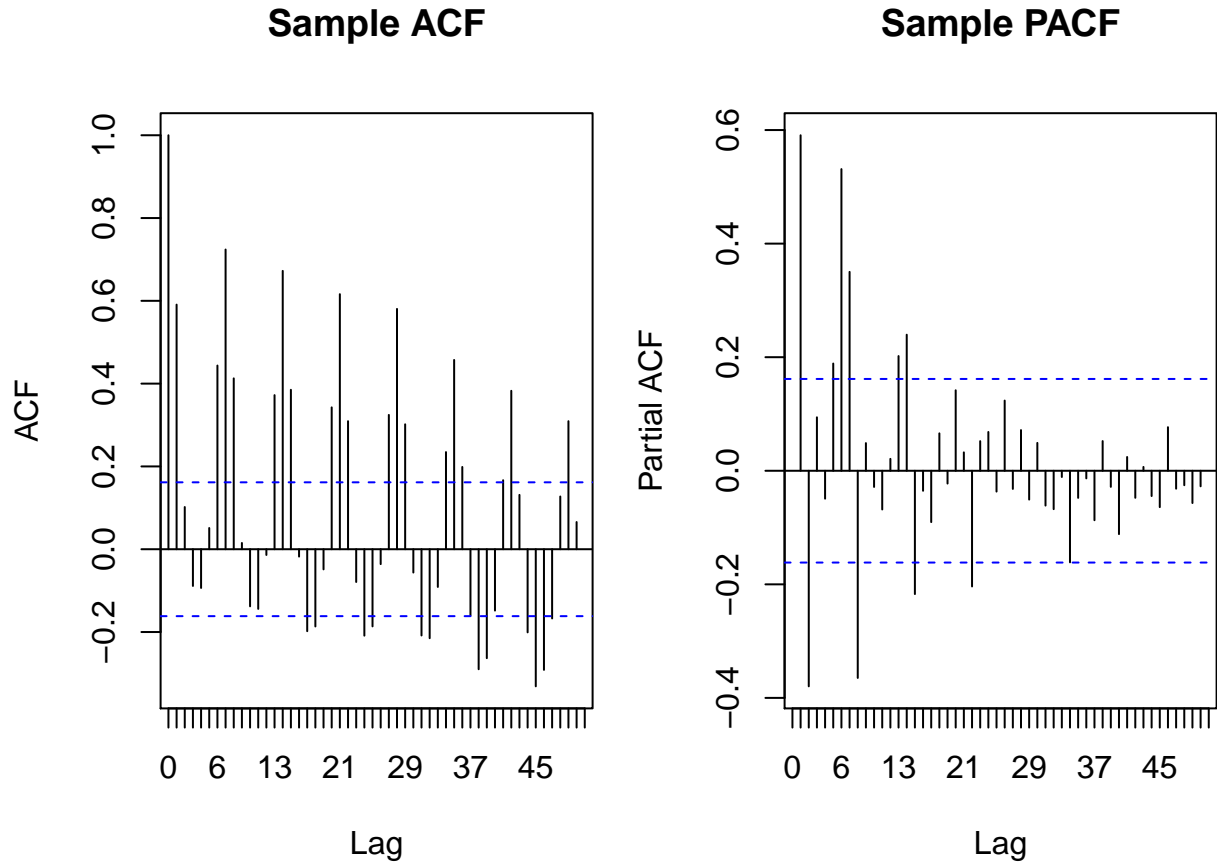


## Hypotheses

By putting the exploratory data analysis and the empirical understanding of retail customer traffic together, the following hypotheses are proposed.

1. Traffic is a seasonal ARMA on weekly basis
2. Traffic = Day of week + Weather Condition + CPI + Noise, where the day of week could be considered as a pulse intervention,
3. Intervention analysis can be employed for special events, like marketing promotion, major holidays and extreme weather.
4. The nonstationary part could be caused by external invention or a larger seasonal factor (monthly or quarterly).

In this project, we are going to focus on the first hypothesis. As the test of the other hypotheses require extra external data and more observations, we will save them for future study once the necessary data are collected.



## Seasonal Model

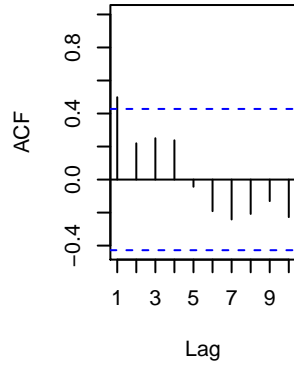
### Model identification

As introduced in the course, there are a couple of candidates of seasonal models, including SARIMA, deseasonalized models, periodic models, etc.

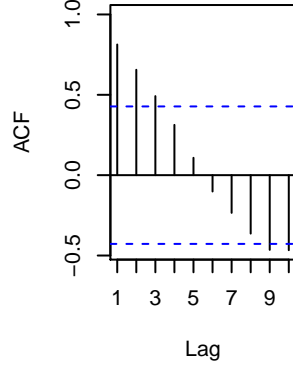
Since it is obvious that the time series in this project follows a periodic pattern on weekly basis, in order to figure out the suitable seasonal method, the dataset is divided into seven subsets corresponding to the data on Monday to Sunday respectively. The plots of their ACF and PACF are illustrated below.



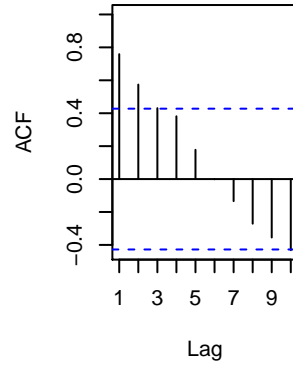
**Sample ACF on Mond**



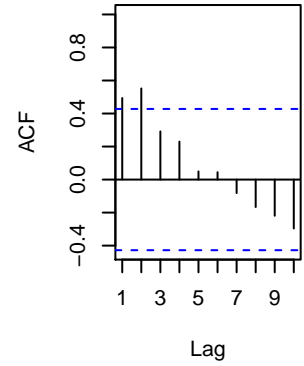
**Sample ACF on Tues**



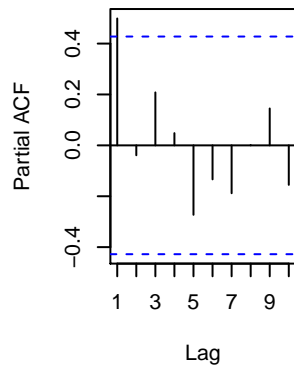
**Sample ACF on Wedne**



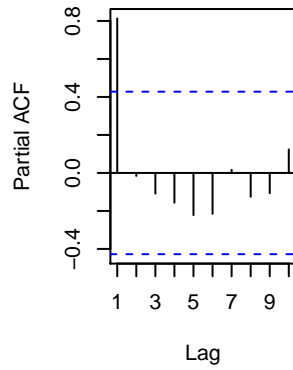
**Sample ACF on Thurs**



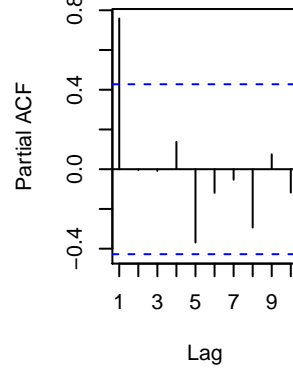
**Sample PACF on Mond**



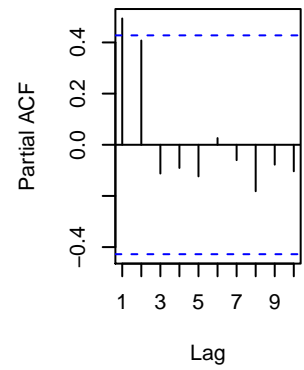
**Sample PACF on Tues**

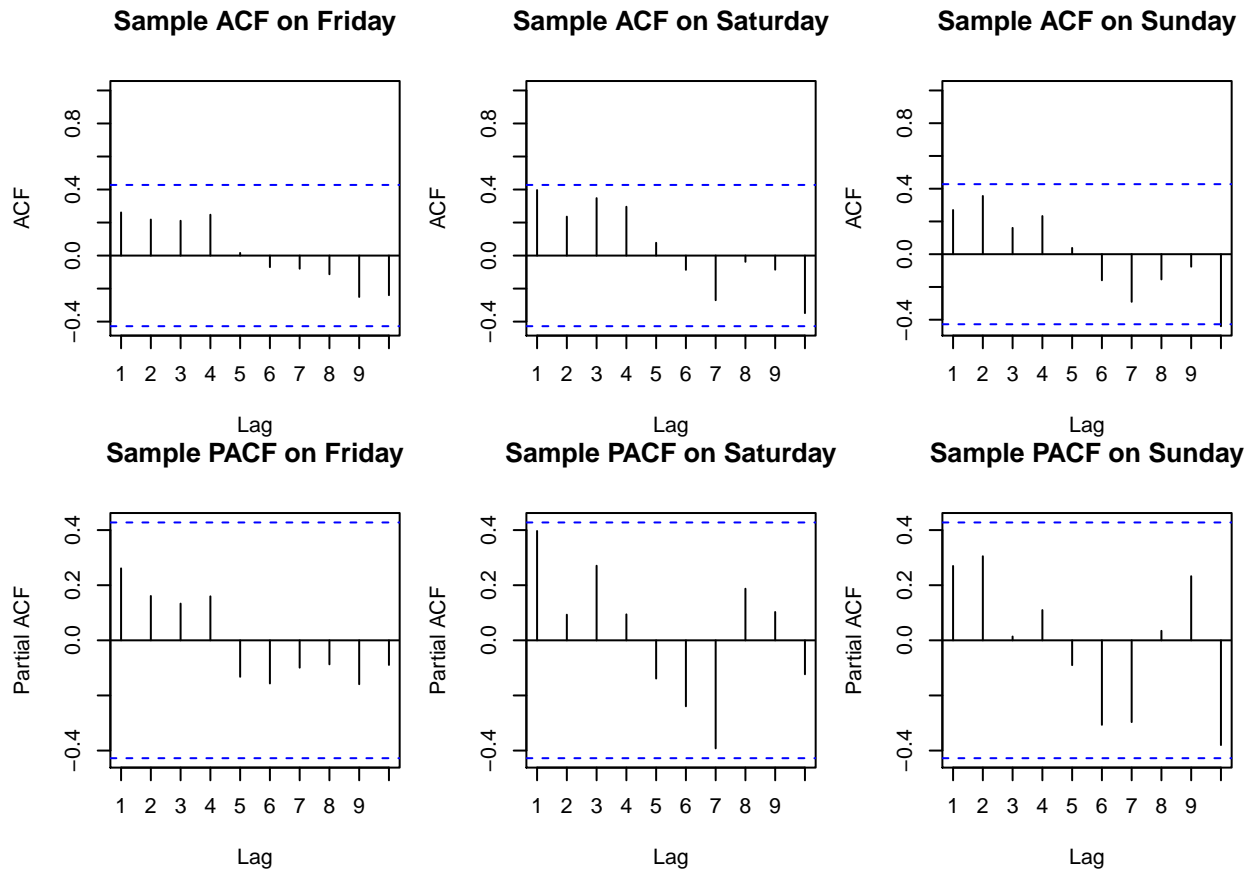


**Sample PACF on Wedne**



**Sample PACF on Thurs**





As we can see, for Monday to Thursday, the PACFs cut off at lag 1 and the ACFs die off from order 1 to 3. For Saturday, both ACF and PACF at lag 1 are only close to the significant confidence interval. The large value of PACF at lag 7 can be interpreted as external intervention. However, for Friday and Sunday, we can not observe any significant correlation. If we are not able construct separated models for Firday and Sunday, it is hard to apply periodic models. One possibe interpretation to this problem is that the correlation between the day and the day before it is much stronger than the correlation between the day and the day in previous week.

Furthermore, although deseasonalized models can reduce the number of paramters, it may not a good option in this case either considering the data points are not sufficent enough. Nevertheless, as the data are kept bringing in, we can apply deseasonalized model in the future.

On the other hand, the SARIMA turns out to be a proper choice. From the exploratory data analysis we performed, it is reasonable to assume the correlation within a week or across seasons is the same.

## Parameters estimation

By taking the advantage of the `auto.ariam` function in the `forecast` package of R, which provides a shortcut of seasonal ARIMA model identification, we can quickly try and test quite a few combinations of SARIMA parameters and pick a proper the one with the MLE or minimum AIC.

```
## Series: ts.df
## ARIMA(2,1,1)(2,0,0)[7]
## Box Cox transformation: lambda= 0.1
##
## Coefficients:
```

```
##          ar1      ar2      ma1      sar1      sar2
##          0.5451 -0.0232 -0.9595  0.4356  0.4421
## s.e.    0.0868  0.0919  0.0396  0.0760  0.0787
##
## sigma^2 estimated as 0.04767:  log likelihood=15
## AIC=-16.38  AICc=-15.78  BIC=1.52
```

The result show a suitable model may be

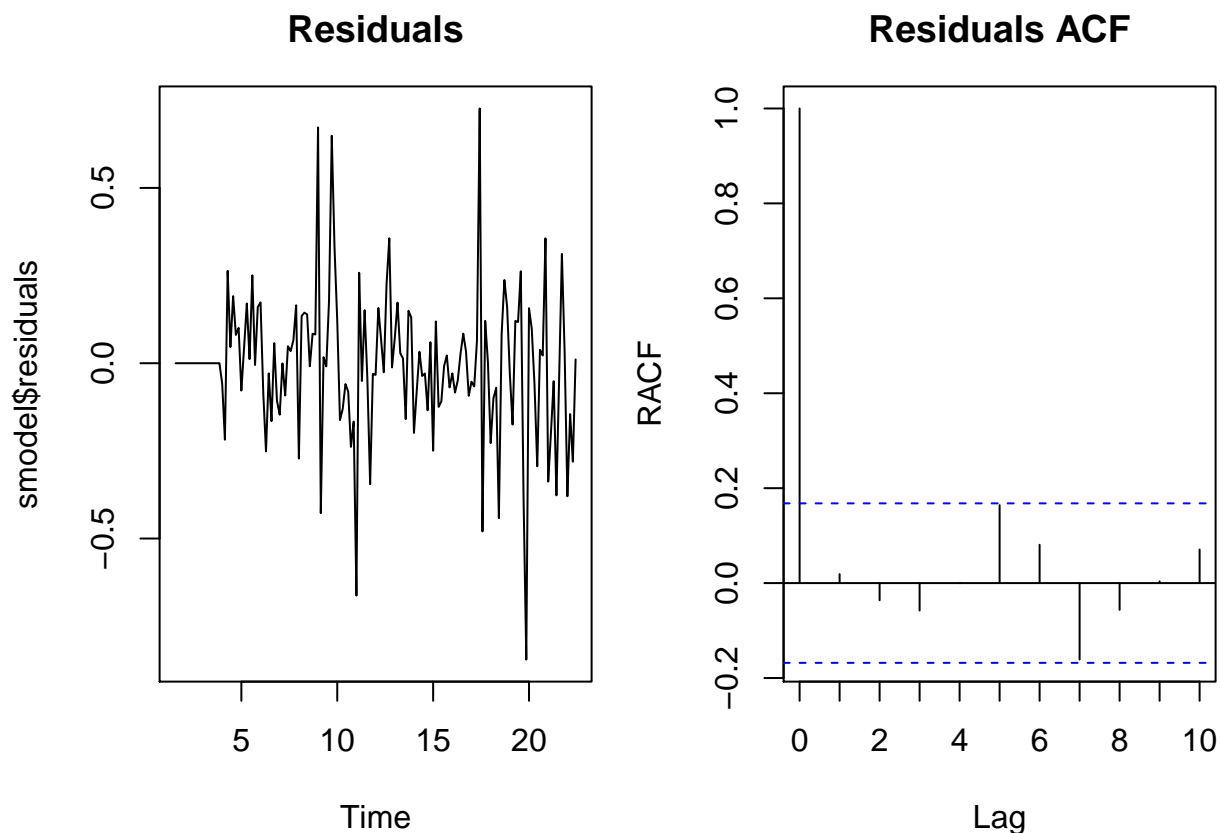
$$SARIMA(2, 1, 1) \times (2, 0, 0)_7$$

The lambada of Box-Cox transformation is 0.1, and the value of AIC equals to -16.38. The coefficients and their standard errors are shown above. In addition, the number of processes for both seanoal and nonseanoal components apparently conform to the ACF and PACF plots.

## Diagnostic checks

### Whiteness Check

The residual plot and the residual autocorrelation function (RACF with 95% confidence limits) plot are drawn below.



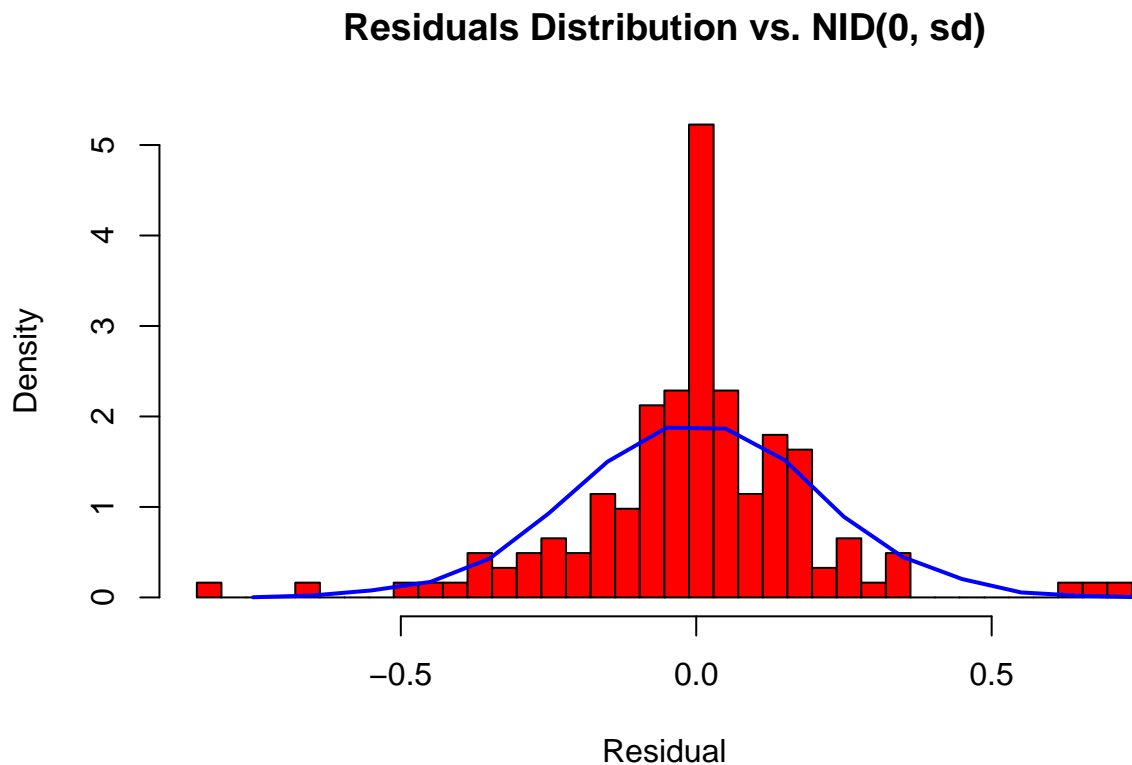
Ljung-Box test, a.k.a portmanteau test, is conducted to check the whiteness.

```
Box.test(smodel$residuals, type="Ljung-Box", lag=10)
```

```
##
## Box-Ljung test
##
## data: smodel$residuals
## X-squared = 11.303, df = 10, p-value = 0.3344
```

The p-value is larger than 0.05, which means the residual can be considered as whiteness.

### Normality Check



```
ad.test(smodel$residuals)
```

```
##
## Anderson-Darling normality test
##
## data: smodel$residuals
## A = 3.1868, p-value = 5.04e-08
```

The graphical method displays an approximate normal distribution. However, the Anderson-Darling test rejected the hypothesis of normality. One of the possibility is the size of observations is not large enough.

### Homoscedasticity Check

The residual plot shows the variances are nearly constant over time.

In summary, the fitted SARIMA model can be considered to have passed the diagnostic check.

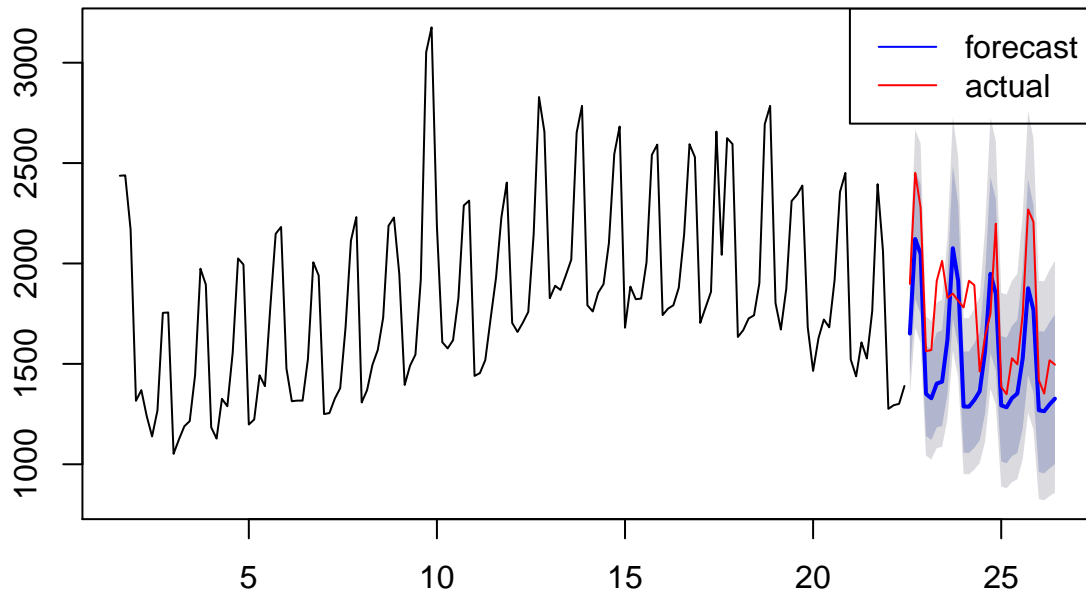
## Forecasting and Validation

### Forecasting

With the fitted SARIMA model, we performs a 4-weeks-ahead forecast (from Sep 25 to Oct 22, 2015) and plots the results with 80% and 95% confidence intervals. The inversed Box-cox transformation is conducted automatically.

##	Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
##	22.57143	1650.849	1443.4123	1884.740	1343.3816	2020.261
##	22.71429	2122.242	1823.1891	2464.740	1680.7498	2665.496
##	22.85714	2047.449	1746.8391	2393.841	1604.3514	2597.796
##	23.00000	1350.325	1141.3747	1593.094	1042.9810	1736.906
##	23.14286	1328.700	1121.4132	1569.856	1023.9058	1712.852
##	23.28571	1401.792	1183.2950	1655.952	1080.5027	1806.644
##	23.42857	1411.147	1190.6615	1667.721	1086.9663	1819.888
##	23.57143	1622.171	1352.0512	1939.926	1226.1049	2129.880
##	23.71429	2076.265	1728.3960	2485.925	1566.3397	2731.016
##	23.85714	1911.815	1584.9393	2298.159	1433.1061	2529.919
##	24.00000	1287.661	1057.7471	1561.614	951.6427	1726.929
##	24.14286	1286.574	1055.6839	1561.957	949.2128	1728.254
##	24.28571	1320.240	1082.9031	1603.408	973.4877	1774.446
##	24.42857	1363.412	1118.1438	1656.082	1005.0843	1832.876
##	24.57143	1563.974	1253.7322	1941.678	1113.0114	2173.198
##	24.71429	1948.590	1554.3670	2430.621	1376.1786	2727.023
##	24.85714	1849.612	1467.1712	2319.546	1294.9944	2609.542
##	25.00000	1293.306	1014.5179	1639.244	889.9968	1854.240
##	25.14286	1283.601	1004.8680	1630.085	880.5508	1845.697
##	25.28571	1329.359	1040.0325	1689.211	911.0484	1913.232
##	25.42857	1352.114	1056.8643	1719.629	925.3255	1948.553
##	25.57143	1527.264	1178.7477	1965.926	1024.8719	2241.363
##	25.71429	1876.761	1445.8933	2419.927	1255.9014	2761.369
##	25.85714	1768.155	1354.8439	2291.673	1173.2980	2621.891
##	26.00000	1268.762	960.8127	1662.859	826.6788	1913.287
##	26.14286	1264.125	955.1994	1660.232	820.8514	1912.284
##	26.28571	1298.368	980.0227	1706.935	841.6844	1967.091
##	26.42857	1326.804	1000.4030	1746.107	858.6736	2013.280

## 4 weeks ahead forecast from SARIMA(2,1,1)(2,0,0)[7] with Lambda 0



### Validation

To verify the accuracy of the forecast, we collected the actual data points from Sep 25 to Oct 22, 2015. The actual data is depicted in red color on the plot. As can be seen, the accuracy of the forecasts is not quite good and the ACF of errors at lag 1 is larger than the confidence limits, which means the model can be improved. The errors may come from both model uncertainty and parameter uncertainty.

##	Date	Forecast	Actual	diff%
## [1,]	"2015-09-25"	"1650.84855730869"	"1898"	"13.0%"
## [2,]	"2015-09-26"	"2122.24190975323"	"2451"	"13.4%"
## [3,]	"2015-09-27"	"2047.44873246204"	"2279"	"10.2%"
## [4,]	"2015-09-28"	"1350.32506160239"	"1563"	"13.6%"
## [5,]	"2015-09-29"	"1328.70013696719"	"1569"	"15.3%"
## [6,]	"2015-09-30"	"1401.7917587342"	"1913"	"26.7%"
## [7,]	"2015-10-01"	"1411.14659132319"	"2013"	"29.9%"
## [8,]	"2015-10-02"	"1622.17118672945"	"1829"	"11.3%"
## [9,]	"2015-10-03"	"2076.26456882833"	"1850"	"12.2%"
## [10,]	"2015-10-04"	"1911.81482813616"	"1814"	"5.4%"
## [11,]	"2015-10-05"	"1287.66128575741"	"1781"	"27.7%"
## [12,]	"2015-10-06"	"1286.57350475672"	"1914"	"32.8%"
## [13,]	"2015-10-07"	"1320.24035934444"	"1892"	"30.2%"
## [14,]	"2015-10-08"	"1363.41221216393"	"1462"	"6.7%"
## [15,]	"2015-10-09"	"1563.97439832318"	"1636"	"4.4%"
## [16,]	"2015-10-10"	"1948.58985772269"	"1750"	"11.3%"
## [17,]	"2015-10-11"	"1849.612334295"	"2198"	"15.9%"

```
## [18,] "2015-10-12" "1293.30636846278" "1385" "6.6%"
## [19,] "2015-10-13" "1283.60078524992" "1349" "4.8%"
## [20,] "2015-10-14" "1329.3590275533" "1528" "13.0%"
## [21,] "2015-10-15" "1352.11410624038" "1498" "9.7%"
## [22,] "2015-10-16" "1527.26438067683" "1747" "12.6%"
## [23,] "2015-10-17" "1876.761049216" "2269" "17.3%"
## [24,] "2015-10-18" "1768.15539883257" "2207" "19.9%"
## [25,] "2015-10-19" "1268.76210618876" "1419" "10.6%"
## [26,] "2015-10-20" "1264.12525882463" "1352" "6.5%"
## [27,] "2015-10-21" "1298.36794513729" "1518" "14.5%"
## [28,] "2015-10-22" "1326.80443547246" "1496" "11.3%"
```

The overall forecast accuracy inclusive of the golden week holiday:

```
##           ME      RMSE      MAE      MPE      MAPE      ACF1 Theil's U
## Test set 230.3058 314.4064 267.6393 12.4641 14.53344 0.3521027 1.021222
```

On the other hand, considering the forecasted period is very close to the golden week for Chinese National Day, intervention factors should have been introduced in the reality. In order to fit a better model, more training data those cover similar situation should be included. So if we screen off the data around the golden week (between Sep 30 and Oct 7), the forecast accuracy of the rest date should be better.

The forecast accuracy excluded the golden week holiday:

```
normalDateIndex <- c(1:5, 14:28)
accuracy(smodel$forecasts$mean[normalDateIndex], test.ts[normalDateIndex])
```

```
##           ME      RMSE      MAE      MPE      MAPE
## Test set 188.0113 232.0323 207.8703 10.39937 11.53417
```

```
tResult <- t.test(smodel$forecasts$mean[normalDateIndex], test.ts[normalDateIndex]);tResult
```

```
##
## Welch Two Sample t-test
##
## data: smodel$forecasts$mean[normalDateIndex] and test.ts[normalDateIndex]
## t = -1.8259, df = 36.522, p-value = 0.07606
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -396.74296 20.72037
## sample estimates:
## mean of x mean of y
## 1540.689 1728.700
```

As we can see, the accuracy of the forecast from 2015-09-25 to 2015-09-29, and from 2015-10-08 to 2015-10-22 is acceptable. The t-test of 95% confidence interval shows no significant difference ( $p\text{-value} = \text{format}(tResult$p.value, digits=3)' > 0.05$ ) in means between the forecasts and the true values.

## Conclusion

Customer traffic analysis is very important to retailers. The customer traffic presents strong correlation with time. But the time series modelling methods are rarely used in the customer traffic research before.

By applying basic time series methodology on the traffic data from a shopping mall, this course project demonstrated the feasibility of this approach. Although the forecast accuracy is not perfect at this moment, the results are acceptable in the off-peak-week days. We are confident of the output can be improved by bringing in more data later.

In addition, a lot of future research directions are also discussed. We believe that time series modelling has enormous application opportunities in retail industry.

## Future Study Topics

The project is just a debut of customer traffic analysis for LinkCity. In fact, there are plenty of topics need to be researched in the future. Some of them are listed below.

1. Model refining with more observations
2. Monthly trend analysis by accumulating more data
3. Intervention analysis, including weather, CPI, etc.

The weather condition and CPI data come from public sources. The historic weather data of Suzhou contain temperature, wind, precipitation and air quality, which can be found at <http://lishi.tianqi.com/suzhou/>. And the CPI data of Suzhou was excerpted from the URL at <http://lishi.tianqi.com/suzhou/>. But only monthly CPI data are available.

4. Traffic distribution analysis by floor and zone
5. Traffic versus sales volume by zone and store
6. Customer in-store dwelling duration (may also have seasonal character)
7. Trend of revisiting customers in one month or one week
8. Multi-location Traffic Analysis. LinkCity has several other shopping malls in the same city. If we can also collect the data from another shopping mall, CARMA model probably can be used for multi-location analysis.

## References

Ghosh, A., Jana, R., Ramaswami, V., Rowland, J., & Shankaranarayanan, N. K. (2011). Modeling and characterization of large-scale Wi-Fi traffic in public hot-spots. In 2011 Proceedings IEEE INFOCOM (pp. 2921-2929). <http://doi.org/10.1109/INFCOM.2011.5935132>

Ark Ryski. (2011). When retail customers count. Bloomington, Indiana: Author House

Perdikaki O, Kesavan S, Swaminathan JM. Effect of traffic on sales and conversion rates of retail stores. Manufacturing Service Oper. Management (2012) 14(1):145-162 Abstract

R in Time Series: Holt-Winters Smoothing and Forecast. (n.d.). Retrieved from <http://www.quantilego.com/howto/holt-winters-smoothing-and-forecast/>

Retail Customer Traffic Counters In Retail Analytics | SenSource. (n.d.). Retrieved December 1, 2015, from <http://sensourceinc.com/blog/retail-customer-traffic-counters-plays-vital-role-in-retail-analytics/>

Time Series Analysis: Building a model on non-stationary time series. (n.d.). Retrieved from <http://www.r-bloggers.com/time-series-analysis-building-a-model-on-non-stationary-time-series/>