Project – Rossmann drug store data analysis

Arthur: Liuqing Yang, Yinpu Li

May 10, 2017

Abstract

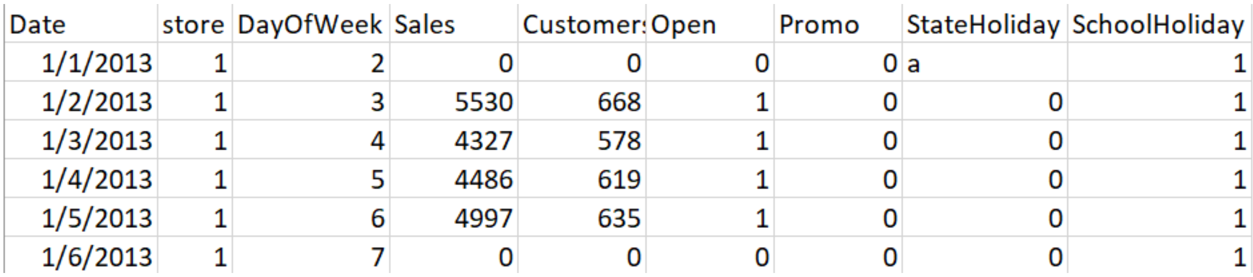
Daily Panel data analysis under Time Series scenario is now a attractive but also challenging topic for Statistics, Econometrician mainly for two reasons: 1. the highly varying frequency to model; 2. when objects is too many, it is impractical to work on each's corresponding time series. To investigate these obstacles, our group found a specific daily panel data to analyze, the Rossmann drug store data. We assume for all objects, their time series are highly similar and apply the same models and then choose the model by evaluating the prediction. We found that for this data, $VAR(p)$ with $p = 28, 30$ are strong models to fit and forecast. Also, by the result of clustering, over 95% objects are clustered into one group, which denote most time series for all the objects that we are working with is very similar or close to each other.

1 Introduction

Basically, our project aims to:
1. Apply linear model to panel data to exclude the effect of time-variant variables and check if corresponding residuals are stationary.
2. With assumption that the resulting time series are similar between each objects, apply time series techniques to analyze the residuals to one object and choose an appropriate model to fit and forecast one month's data.
3. Compare the predicting strength of each model and make choice of the final candidate with proper measure evaluation.
4. Check the former assumption by conducting clustering on the time series. 5. Find the potential reasons that lead to different behaviors of stores by extra information.

Method

The data is a part from a former Kaggle data competition. It is a large size panel data with 934 drug stores from a company named Rossmann in the country Germany. For each stores, the data is from 01/01/2013 to 07/31/2015, i.e. 942 days:



Date	store	DayOfWeek:	Sales	Customer:	Open	Promo	StateHoliday	SchoolHoliday
1/1/2013	1	2	0	0	0	0	a	1
1/2/2013	1	3	5530	668	1	0	0	1
1/3/2013	1	4	4327	578	1	0	0	1
1/4/2013	1	5	4486	619	1	0	0	1
1/5/2013	1	6	4997	635	1	0	0	1
1/6/2013	1	7	0	0	0	0	0	1

Two dependent variables $Sales$, $Customers$ that we focused on and other independent variables:

- Store: a unique Id for each store
- Sales: the turnover for any given day
- Customers: the number of customers on a given day
- Open: an indicator for whether the store was open: 0 = closed, 1 = open
- Promo: indicates whether a store is running a promo on that day
- StateHoliday: indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday: indicates if the (Store, Date) was affected by the closure of public schools

The data was split into two datasets, one for training models, containing data of all stores from Jan. 1^{st}, 2013 to June 30^{th}, 2015, the other one for testing models, containing data of all stores from July 1^{st}, 2015 to July 31^{st}, 2015.

1.1 Fixed Effect Models vs. Random Effect Models

We assume fixed effect models are proper to fit the data. The first thing we did was excluding the effect of these time-variant's effect on $Sales$, $Customers$ and get the corresponding residuals as time series that we would working on. Assumed the panel data linear models to be:

$$\begin{aligned} Sales_{it} &= \mu_i + \beta_1 open_{it} + \beta_2 promo_{it} + \beta_3 StateHoliday(a)_{it} + \beta_4 StateHoliday(b)_{it} \\ &\quad + \beta_5 StateHoliday(c)_{it} + \beta_6 SchoolHoliday_{it} + \epsilon_{it} \\ &:= \mu_i + X\vec{\beta} + \epsilon_{it} \end{aligned}$$

$$
\begin{aligned}
Customers_{it} &= \omega_i + \alpha_1 open_{it} + \alpha_2 promo_{it} + \alpha_3 StateHoliday(a)_{it} + \alpha_4 StateHoliday(b)_{it} \\
&\quad + \alpha_5 StateHoliday(c)_{it} + \alpha_6 SchoolHoliday_{it} + v_{it} \\
&:= \omega_i + X\vec{\alpha} + v_{it}
\end{aligned}
$$

where i denotes the store ID, μ_i and ω_i are the fixed effects of store i for its $Sales$ and $Customers$. Then get the residuals or time series:

$$S_{it} := \hat{\epsilon_{it}} = Sales_{it} - \hat{\mu_i} - X\hat{\vec{\beta}}$$

$$C_{it} := \hat{v_{it}} = Customers_{it} - \hat{\omega_i} - X\hat{\vec{\alpha}}$$

Alternatively, we assumed them to be random effected model, where μ_i is identically distributed from a distribution with variance σ_s^2 and where ω_i is identically distributed from a distribution with variance σ_s^2. The new residuals of time series are:

$$S^*_{it} := \hat{\mu_i} + \hat{\epsilon_{it}} = Sales_{it} - X\hat{\vec{\beta}}$$

$$C^*_{it} := \hat{\mu_i} + \hat{v_{it}} = Customers_{it} - X\hat{\vec{\alpha}}$$

To decide between fixed effect model and random effect model, we used Hausman test to check if random effect models were consistent. The results showed that, for $Customers_{it}$, random effect model is inconsistent but for $Sales_{it}$ it is consistent. However we didn't use the resulting residuals for random effect model this time because clearly: $Var(S^*_{it}) > Var(S_{it})$, which will lead to bad prediction. One of our goals was making as better prediction as possible, time series with smaller variance were preferred.

1.2 Study of Residuals and VARs

Then we made an investigation into the residuals and applied time series techniques to the residual series. The coefficients of fixed effect model for Sales are given as follows:

```
 Coefficients :
                            Estimate Std. Error   t-value  Pr(>|t|)
Open                       5881.7544     5.7390 1024.8680 < 2.2e-16 ***
Promo                      2311.8470     4.0682  568.2694 < 2.2e-16 ***
as.factor(StateHoliday)a  -765.4787    14.3884  -53.2012 < 2.2e-16 ***
as.factor(StateHoliday)b -1301.0695    24.0838  -54.0227 < 2.2e-16 ***
as.factor(StateHoliday)c  -193.6304    29.1034   -6.6532  2.87e-11 ***
SchoolHoliday              230.2312     5.1450   44.7487 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:    9.2373e+12
```

```
Residual Sum of Squares: 2.5341e+12
R-Squared:      0.72567
Adj. R-Squared: 0.72536
F-statistic: 374705 on 6 and 849934 DF, p-value: < 2.22e-16
```

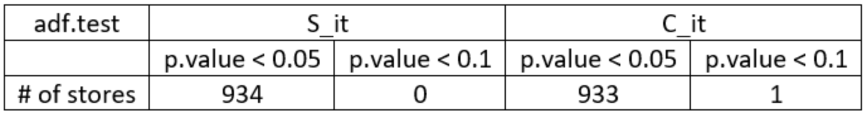
The coeffecients of fixed effect model for Customers:

```
Coefficients :
               Estimate Std. Error   t-value  Pr(>|t|)
Open          686.25834    0.60807 1128.5844 < 2.2e-16 ***
Promo         155.93788    0.43104  361.7697 < 2.2e-16 ***
StateHolidaya -48.40075    1.52450  -31.7487 < 2.2e-16 ***
StateHolidayb -93.15607    2.55175  -36.5067 < 2.2e-16 ***
StateHolidayc -25.68988    3.08360   -8.3311 < 2.2e-16 ***
SchoolHoliday  22.91653    0.54513   42.0389 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:    1.0143e+11
Residual Sum of Squares: 2.8448e+10
R-Squared:      0.71954
Adj. R-Squared: 0.71923
F-statistic: 363419 on 6 and 849934 DF, p-value: < 2.22e-16
```

Clearly, the effects of those time invariant variables on Sales and Customers are significant.

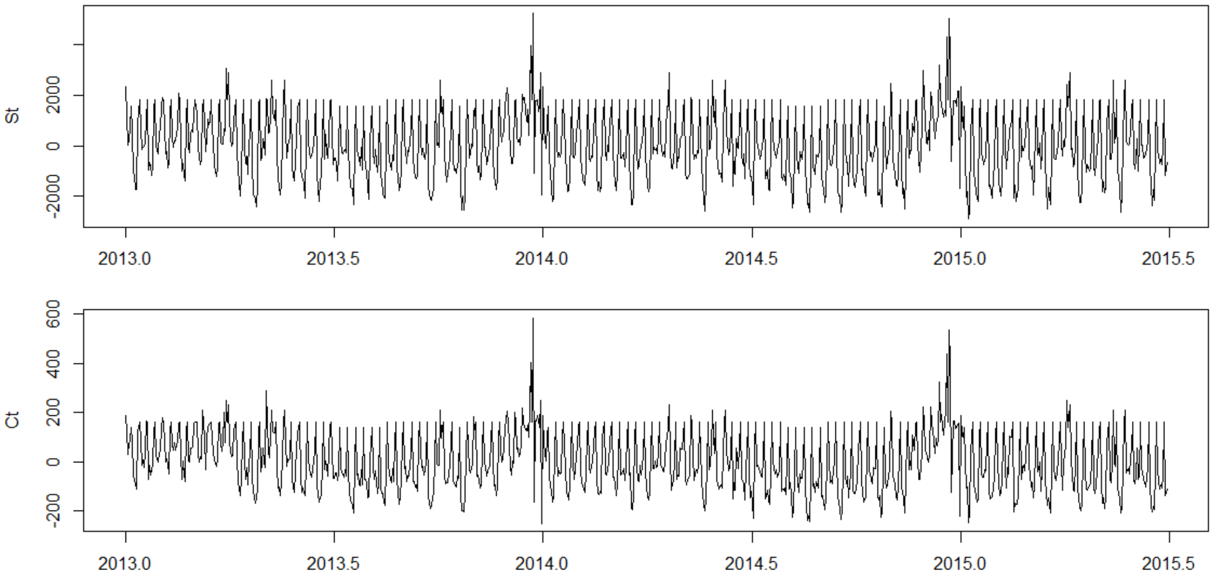
Then we had 943 × 2 = 1886 times series and relied on Augmented DickeyFuller Test to check if they were stationary:



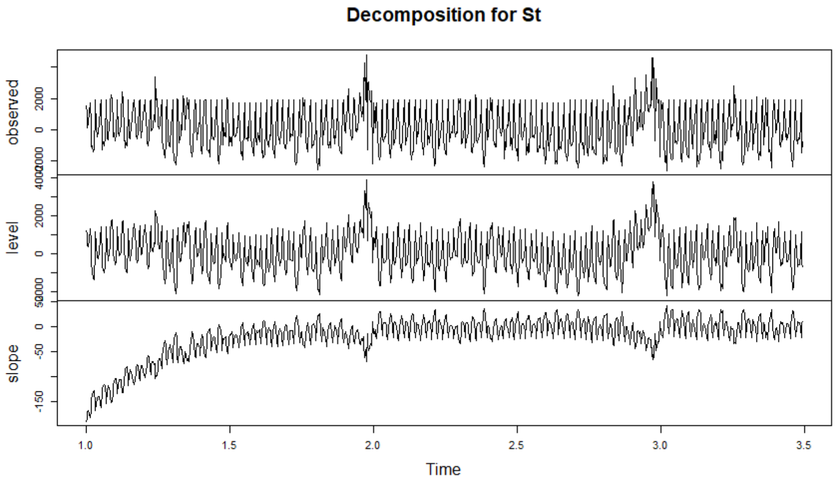
adf.test	S_it		C_it	
	p.value < 0.05	p.value < 0.1	p.value < 0.05	p.value < 0.1
# of stores	934	0	933	1

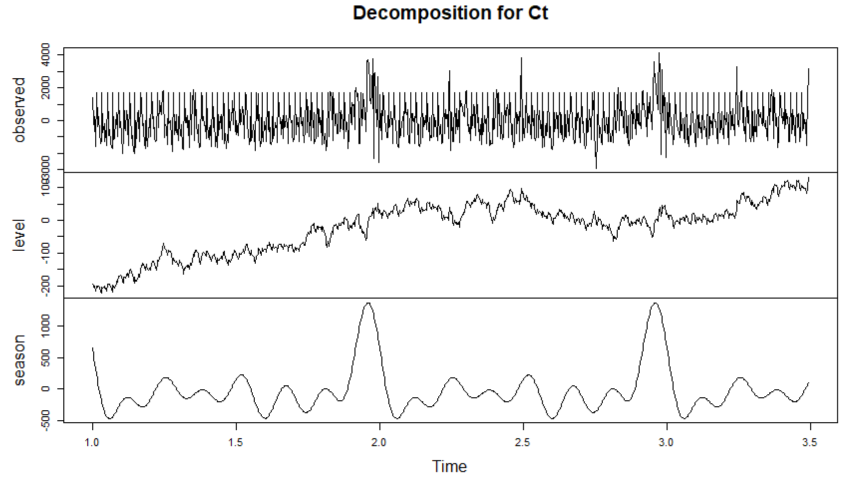
From the p-value from the table above, we should reject the null-hypothesis that the time series are non-stationary.

next we plotted the S_t, C_t for one randomly selected store:

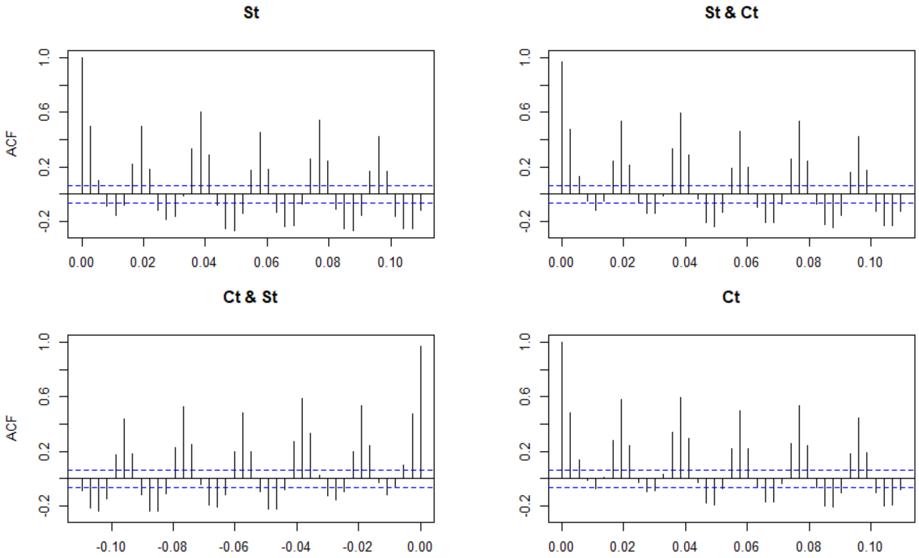


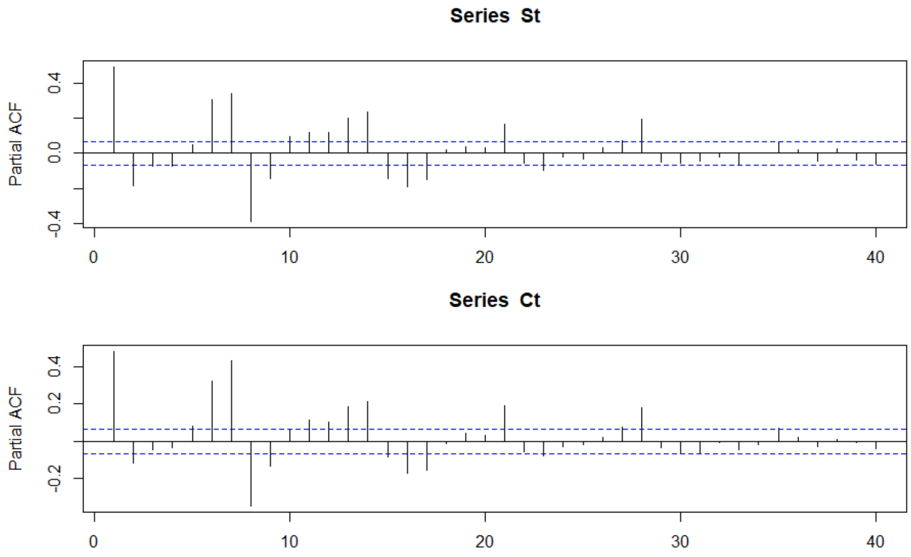
Both of S_t, C_t had weekly pattern and before the end of each year there were peaks around Christmas. The first model we tried was TBATS. TBATS model (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components) is proposed by De Livera et al. (2011), they modified the ETS (Exponential smoothing state space model) models in order to include a wide variety of seasonal patterns and solve the problem of correlated errors. However, when we apply TBATS model on the S_t, C_t it failed to decompose it into two seasonal patterns:



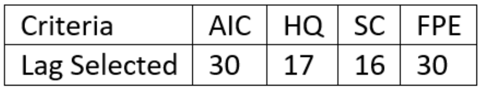


Also, TABTS can not consider the cross-correlation effect between S_t and C_t. And more information a model included, higher accuracy the prediction will be of. Thus, to better forecast, we change to another model, VARMA.



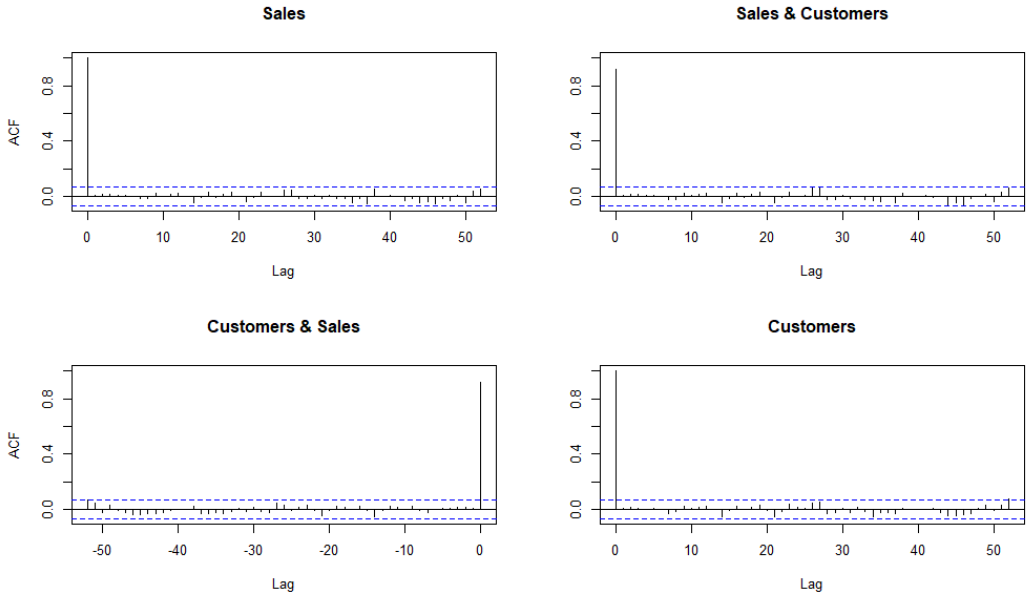


From the ACF plots, clearly for both S_t and C_t the weekly pattern was obvious and also the cross-correlation was significant. From the PACF plots, both of them cut off at lag=28. Thus, we fitted the data with VAR(p) and large p for following reasons: 1. It considered the cross-correlation effect; 2. It would take care of the weekly, bi-weekly or even monthly pattern; 3. Our data had quite long time period this time didn't need to worry about over-fitting. By the recommendation from several criteria, we choose VAR(30) to first fit the data:

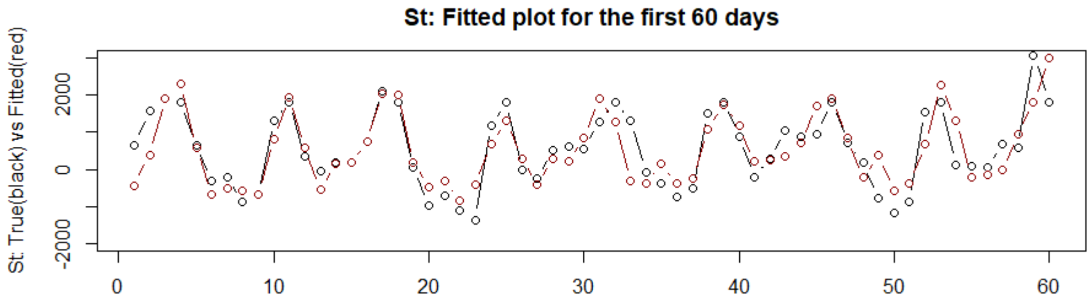


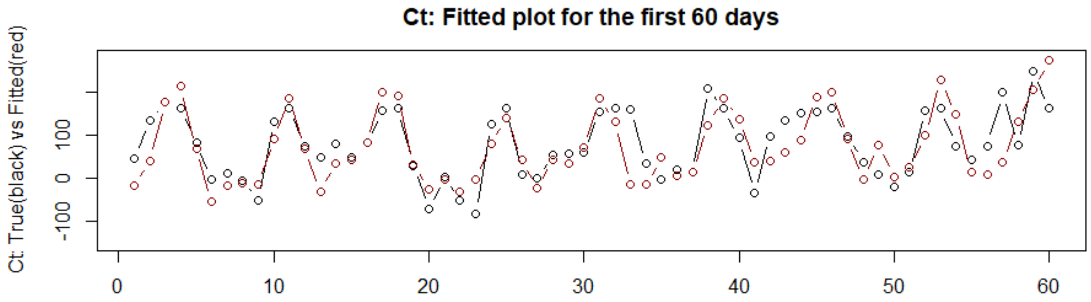
Criteria	AIC	HQ	SC	FPE
Lag Selected	30	17	16	30

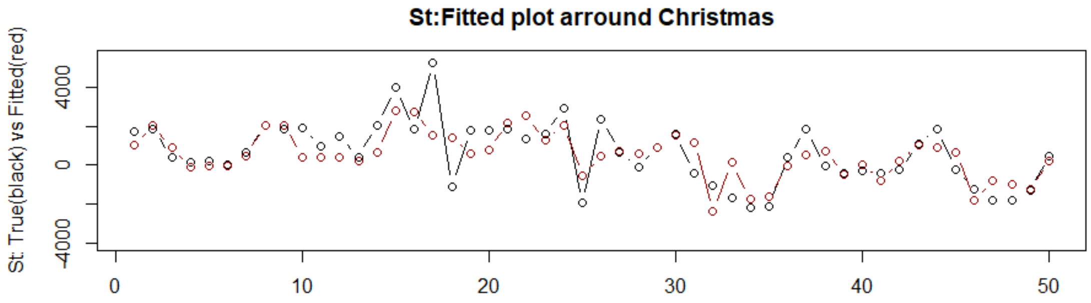
After fitting with VAR(30), we plotted the ACF for the residuals:

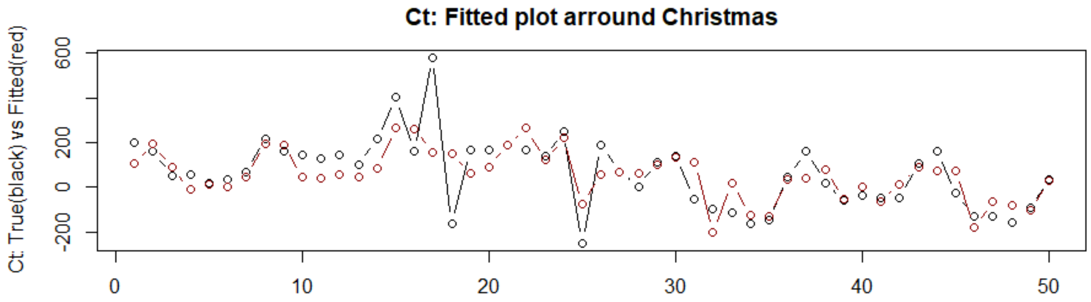


It looked like the VAR(30) had successfully excluded both the auto-correlation and cross-correlation of S_t and C_t. That is, as we pre-assumed before, VAR(30) took care of the seasonal pattern at the same time excluded the cross-correlation effect. Also plotted the fitted value:









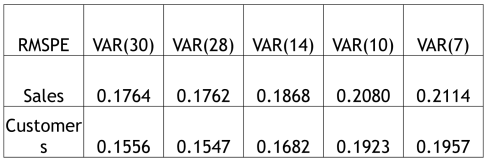
The fitted values capture the pattern of the true values for both S_t and C_t for most of the time. However, at the dates around Christmas, VAR(30) failed to capture the sudden peak.

With the assumption that the residuals from the former models of each store behaves similarly, we used the the same lag in VARMAs to all of the stores. But did we made a reasonable assumption? Is it correct to apply the same lag to most of the stores? Actually, if we expand our sample stores, instead of one single choice, we were provided with $lag = 7, 14, 28, 30$ for different stores by the same criteria. Thus, with the aim of achieving optimal overall accuracy, we used root mean square percentage error to evaluate these lag candidates:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2}$$

, where y_i represents the true value, $\hat{y}_i$ represents predicted value from our fixed effect model and VAR models and n represents the number of stores. Thus, this statistic computes the average percentage regression loss from each model candidates.

VAR models with different lags based on the two residual series after extracting the fixed effects were assessed via $RMSPE$ criteria, the results are shown in the following table. Eventually, the optimal parameter is $lag = 28$, with root mean square percentage error equals to 17.62% for $Sales'$ residuals prediction and 15.47% for $Customers'$ residuals prediction. However, we chose $lag = 30$ eventually, as the accuracy of prediction of VAR(30) did not make too much difference from that of VAR(28) and it also corresponds to the natural season of one month.



RMSPE	VAR(30)	VAR(28)	VAR(14)	VAR(10)	VAR(7)
Sales	0.1764	0.1762	0.1868	0.2080	0.2114
Customers	0.1556	0.1547	0.1682	0.1923	0.1957

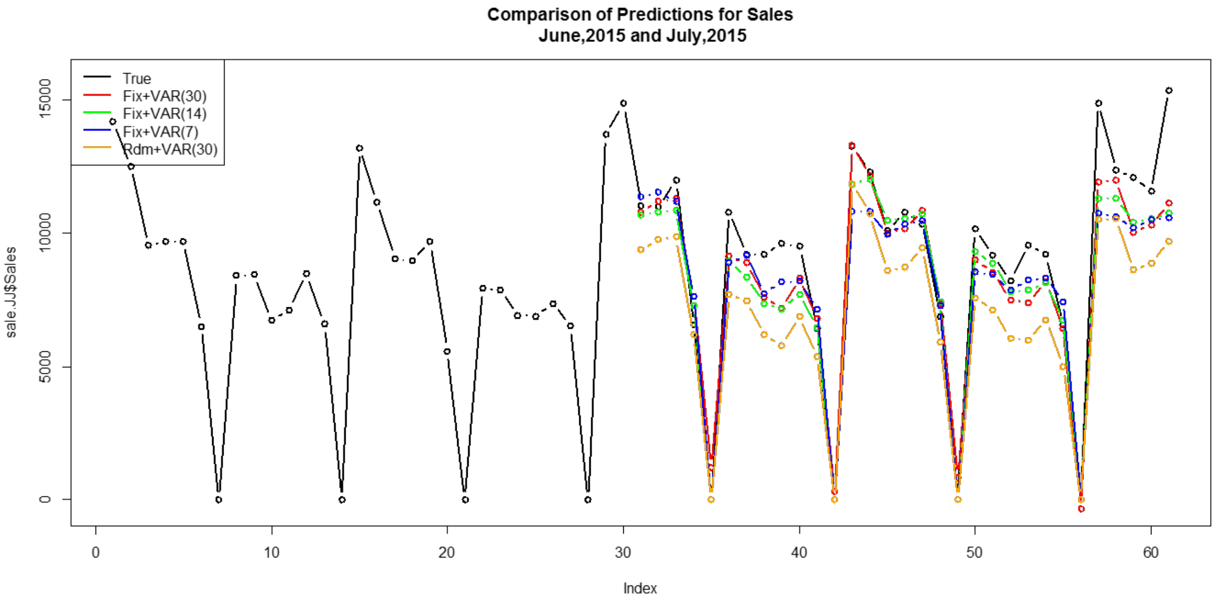
2 Predictions

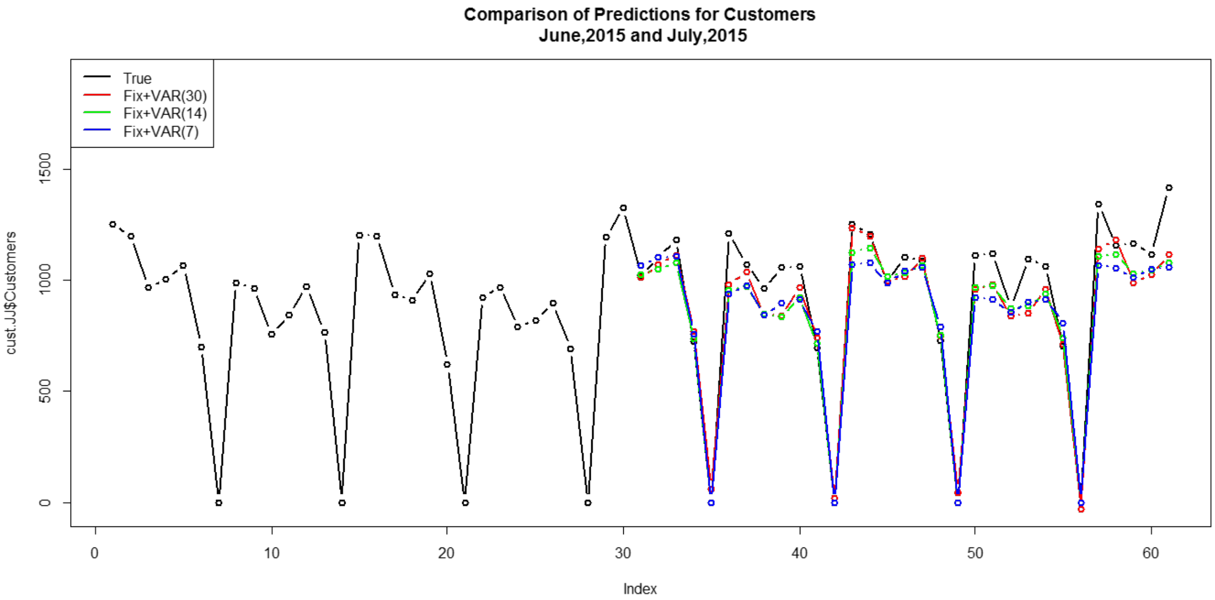
By the analysis from the first part, we have chosen our final model, which is a linear regression of panel data with fixed effect plus a VAR(30). Thus, the predicted value could be computed as follows:

$$\begin{aligned} Sales_{it} &= \mu_i + \hat{\beta}_1 open_{it} + \hat{\beta}_2 promo_{it} + \hat{\beta}_3 StateHoliday(a)_{it} + \hat{\beta}_4 StateHoliday(b)_{it} \\ &\quad + \hat{\beta}_5 StateHoliday(c)_{it} + \hat{\beta}_6 SchoolHoliday_{it} + \epsilon_{it} \\ &:= \mu_i + X\vec{\hat{\beta}} + \epsilon_{it} \end{aligned}$$

$$\begin{aligned} Customers_{it} &= \omega_i + \hat{\alpha}_1 open_{it} + \hat{\alpha}_2 promo_{it} + \hat{\alpha}_3 StateHoliday(a)_{it} + \hat{\alpha}_4 StateHoliday(b)_{it} \\ &\quad + \hat{\alpha}_5 StateHoliday(c)_{it} + \hat{\alpha}_6 SchoolHoliday_{it} + v_{it} \\ &:= \omega_i + X\vec{\hat{\alpha}} + v_{it} \end{aligned}$$

Then we forecast *Sales* and *Customers* in the following month, July 2015. To better show the results, we randomly selected one store and plot the true value as well as the predicted values for *Sales* and *Customers* respectively. The prediction comparisons among different candidate models are shown in the following two figures. As we can see in the both of the two figures, the red lines, which represent predicted values, are closest to the true value compared to the rest, which correspond to the former assumption that a universal lag parameter could be applied to all the stores.





To be more convincing, we conduct the time series clustering to show most of the stores belong to a single cluster and thus behave alike. The idea of clustering is to concern itself as the creation of groups of objects, where each group is called a cluster.

In detail, clusters are formed by grouping objects that have maximum similarity with other objects within the group, and minimum similarity with objects in other groups. As time series clustering is based on similarity, we used Dynamic Time Warping(DTW) distance as our dissimilarity metric to conduct hierarchical clustering:

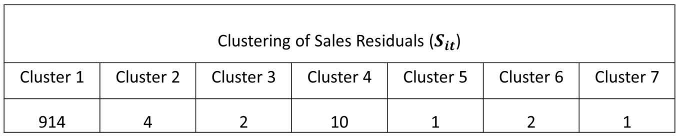
$$DTW_p(x, y) = (\sum \frac{m_\phi lcm(k)^p}{M_\phi})^{1/p}, \forall k \in \phi$$

, where

$$lcm(i, j) = (\sum_\nu |x_i^\nu - y_i^\nu|^p)^{1/p}$$

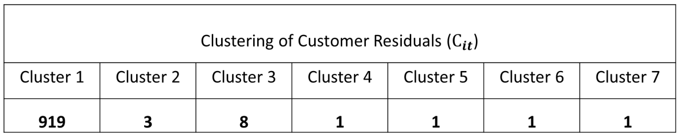
is the local cost matrix and it computes the l_p norm between two series x_i and y_i, explicitly denoting that multivariate series are supported. And then the DTW algorithm finds the path that minimizes the alignment between the two series ,x and y, by iterating stepping through the lcm and aggregating the cost. We define $\phi = \{(1, 1), ..., (n, m)\{$ as the set containing all the points that fall on the optimum path, then the final distance would be computed with the first equation, where m_ϕ is a per-step weighting coefficient and M_ϕ is the corresponding normalization constant.

As a result, which is shown in the following tables and figures, most of the data points are clustered into one group, even if we expand the number of cluster to be 7.



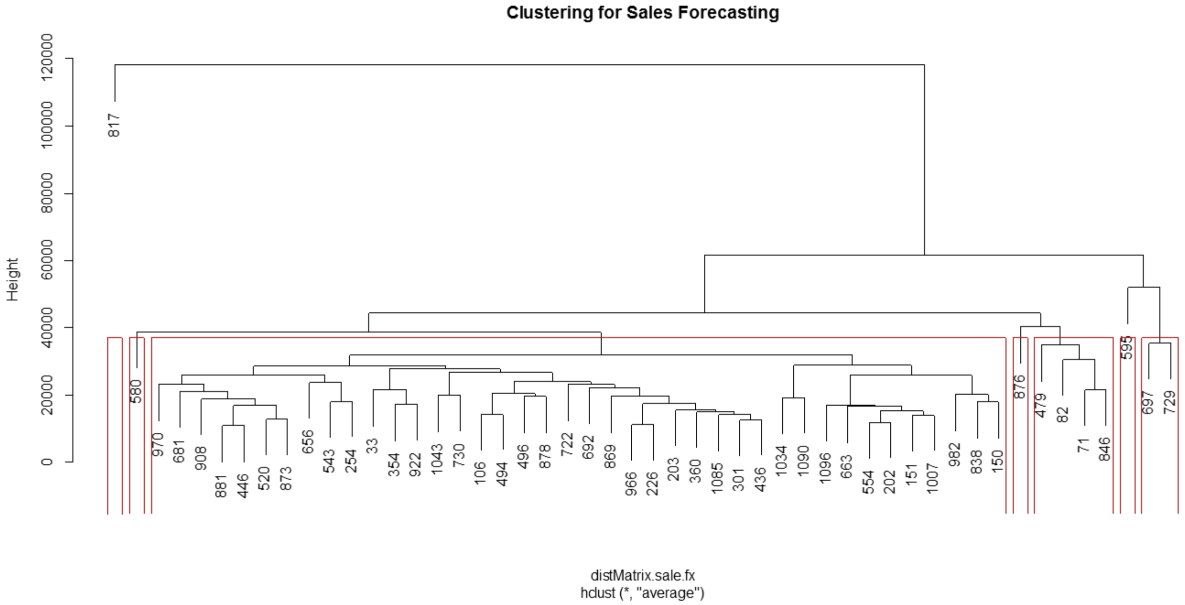
Clustering of Sales Residuals ($\boldsymbol{S_{it}}$)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
914	4	2	10	1	2	1

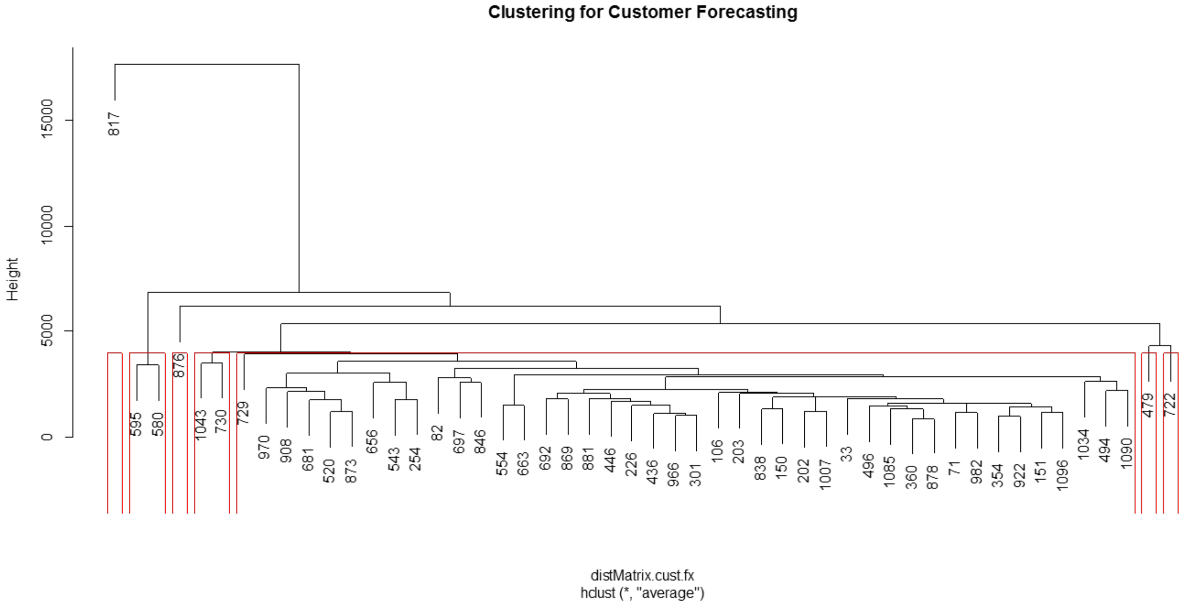


Clustering of Customer Residuals (C$_{it}$)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
919	**3**	**8**	**1**	**1**	**1**	**1**

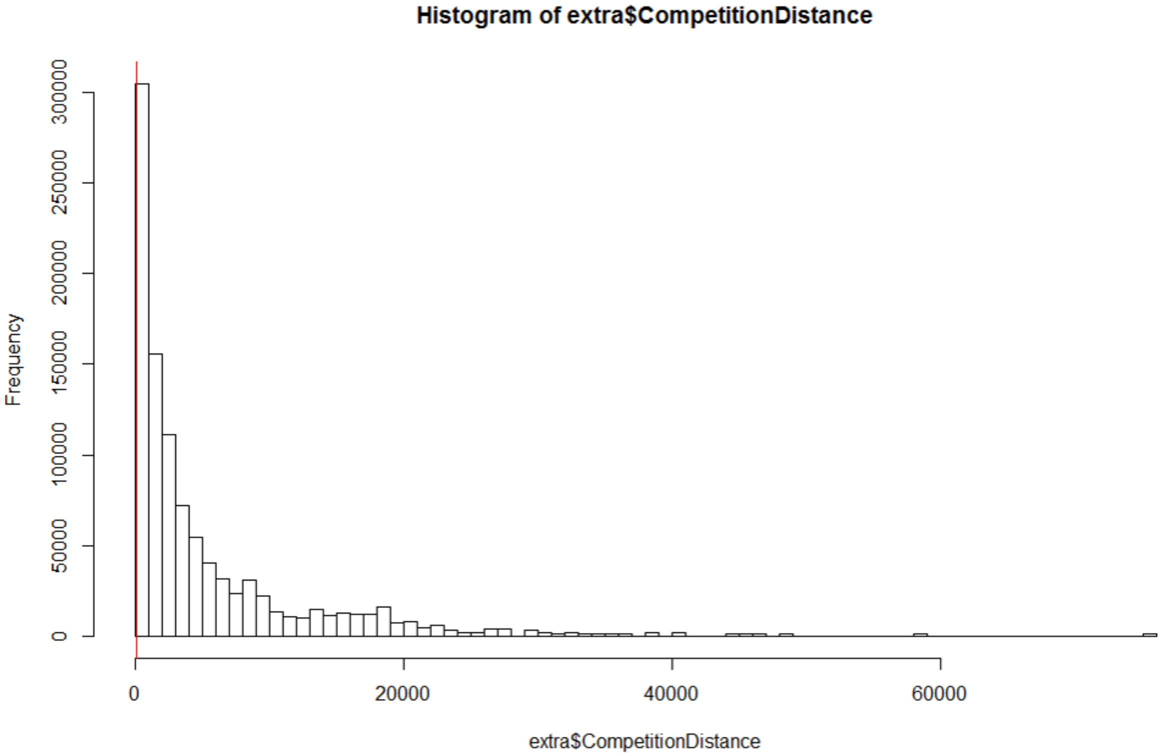


Clustering for Sales Forecasting

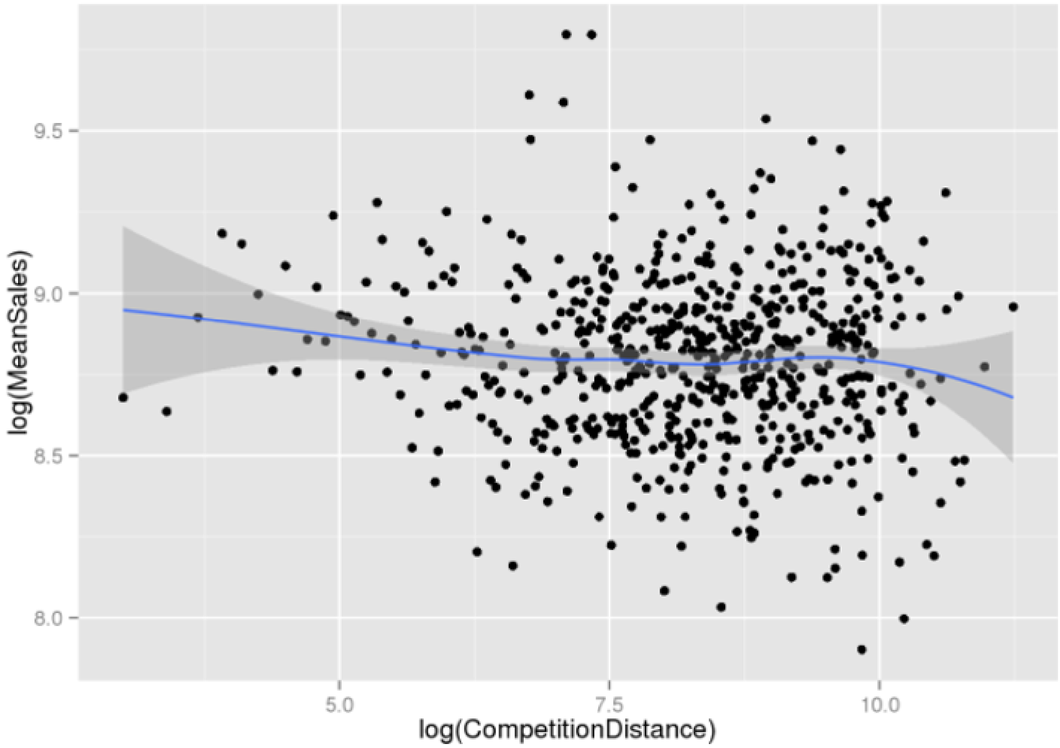


Clustering for Customer Forecasting

Due to the size of the data, we only showed a sample of 50 stores that were randomly selected from the data. As is shown in the cluster tree, only store 817 stood alone in one branch, which implies our model performs well on most of the stores. If we took a close look at the data of store 817, we found that it was closer to the next competitor than most of the rest stores, which seemed to be counterintuitive at first glance that lower distance to the competitors implies higher sales or more customers.



However, it may happen as our data were taken from drug stores in Germany, a store with low distance to the other competitors may locate in the inner cities or crowded areas, thus with higher flow of visiting customers and hence with higher sales. If we plot the $log(MeanSales)$ $v.s.$ $log(CompetitionDistance)$, where $MeanSales$ represents the average sale data for a stores with non-NA data and $CompetitionDistance$ represents the distance between the store to the nearest competitor, we can see there is a negative linear relationship between sales and competition distance.



3 Conclusion and Discussion

We analyzed a panel data for Rossmann Store in Germany and made predictions of the number of customers and daily sales for 30 days ahead. With the aim of higher accuracy, fixed effect model was used to extract fix effects in the data and then VAR models were applied to the residuals with lag 30. The model we chose captures the features of the data but performed poorly when it comes to Christmas. We could consider making adjustment of seasons provided that we get more information about the stores. Clustering is a useful approach to identify structures of the unlabeled residual data by organizing data into similar groups. As most of the data are clustered into one single cluster, we conclude that using one single lag for all of the data is reasonable and efficient. And with evaluation criteria RMSPE, our model reached acceptable inaccuracy of 17.64% for *sales* and 15.56% for *customers*. On the other hand, we could also consider the leading relationships of stores to their surrounding competitors, which is beyond our scope due to the lack of sufficient information.

References

[1]. Alexis Sarda-Espinosa, Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package, 2016

[2]. Yves Croissant, Giovanni Millo, Panel Data Econometrics in R: The plm Package, 2008