Predicting Walmart Sales

MSA 8200

Meet the Team

Bennie Amani

Pranidhi Prabhat

Nimeelitha Akkiraju

Keerthi Bojja

Data Overview

Total Observations: 421,570

Feb 2010 to Oct 2012: 143 weeks

Stores: 45 | Departments: 90+

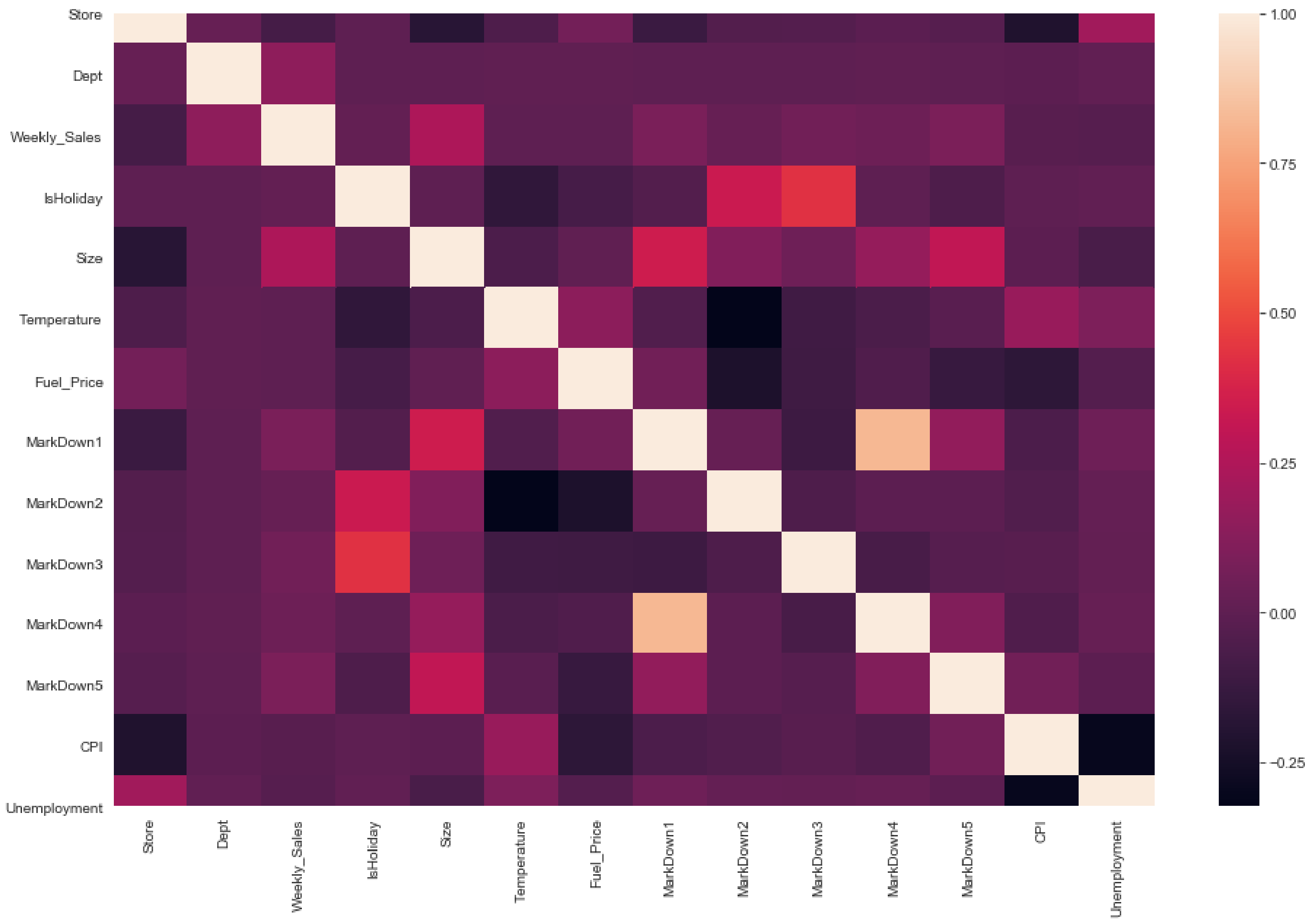
Store Types: A, B & C

Exploratory Data Analysis

Correlation Matrix

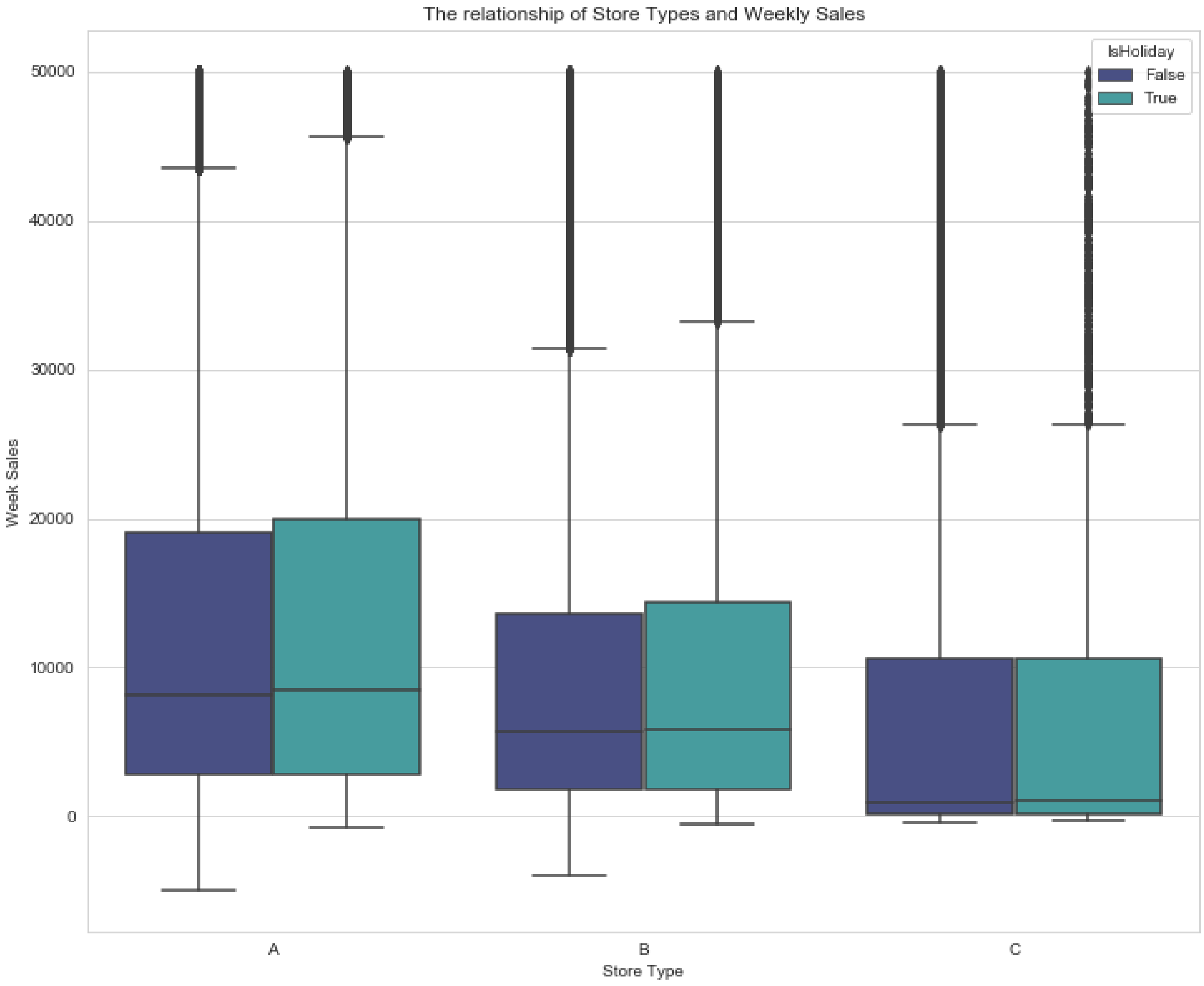
- None of the features have very strong correlation with Weekly Sales

Features	Correlation with Target Variable
Size	0.243828
Dept	0.148032
MarkDown5	0.090362
MarkDown1	0.085251
Store	-0.085195



Exploratory Data Analysis

- Avg Sales: $15,981
- Store Type A has higher range of Weekly Sales compared to Type B & C
- Variation in Weekly Sales is almost same irrespective of whether it is a Holiday week or not



Data Selection for Prediction

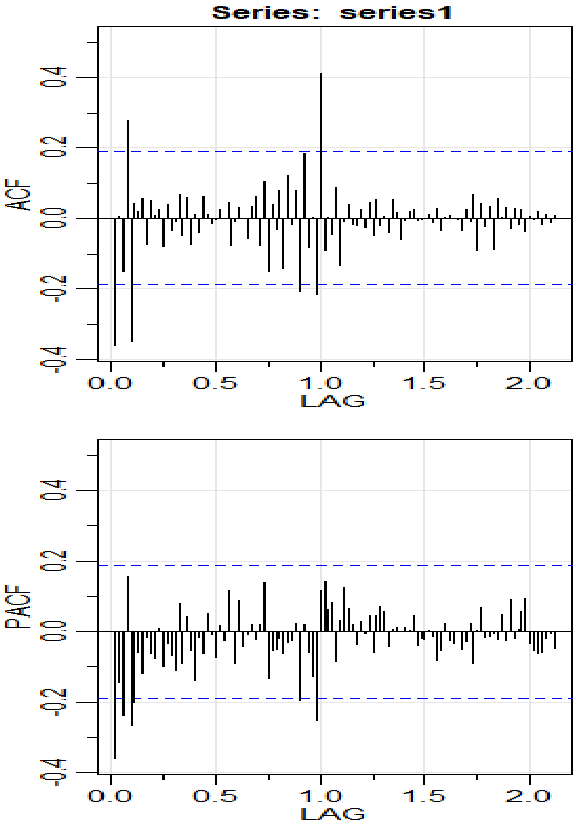
- Filtered data store wise
- Aggregated **weekly sales across department as there** are more than 90 **entries for** each store
- Out of 45 **stores, we randomly selected store** no's 3**, 20 & 30 from different store types for** prediction
- Split **the data into 80-20 for train and test**

SARIMA
with no additional variables

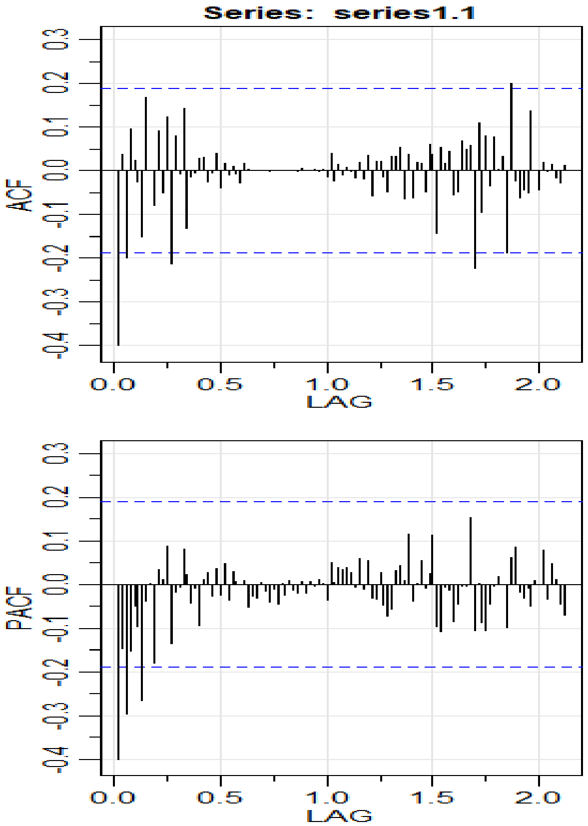
Stationarity and Model Selection

Type A /Store 20 – ARMA(1,0,1)

ACF and PACF (seasonality)



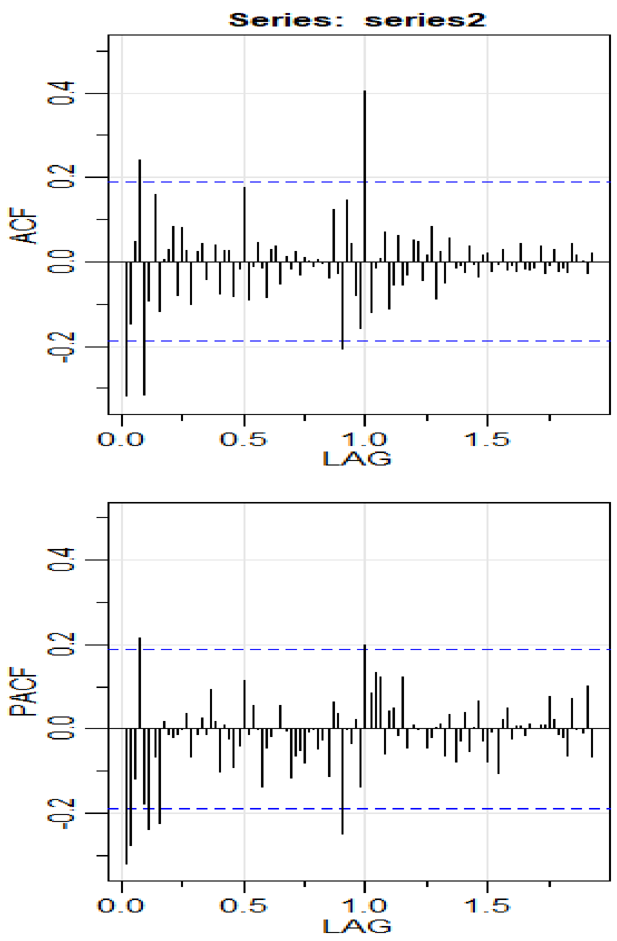
ACF and PACF (no seasonality)



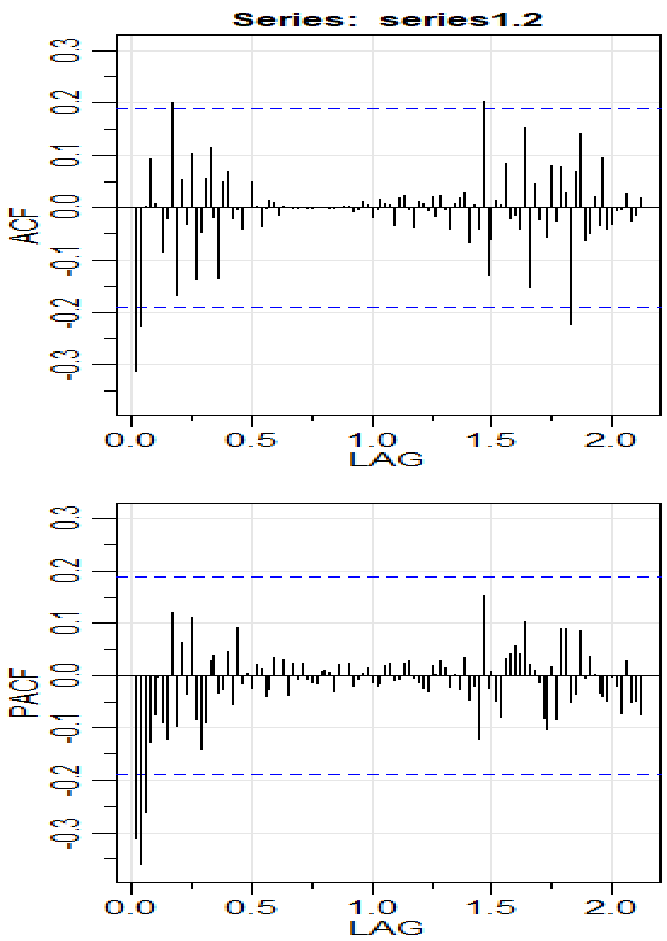
Stationarity and Model Selection

Type B/Store 3 - ARMA(3,0,2)

ACF and PACF (seasonality)



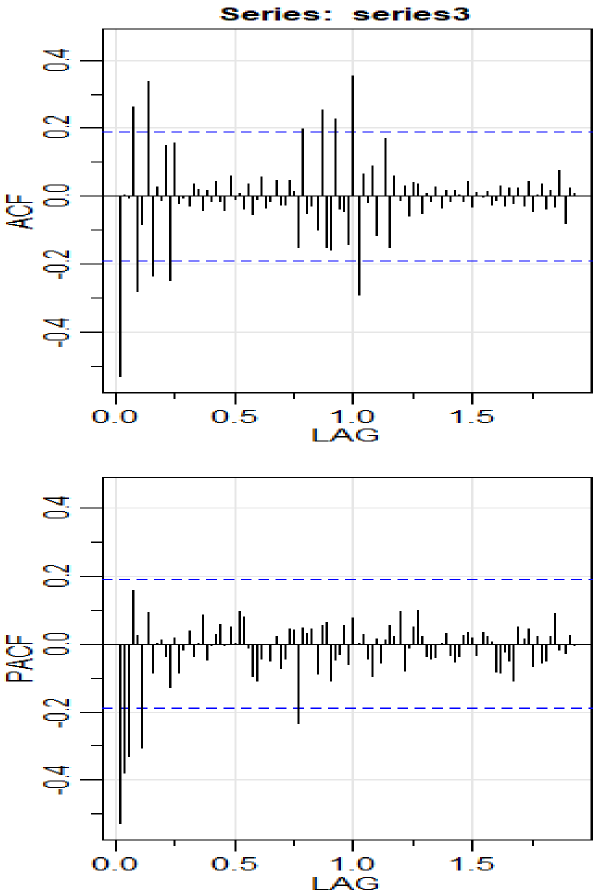
ACF and PACF (no seasonality)



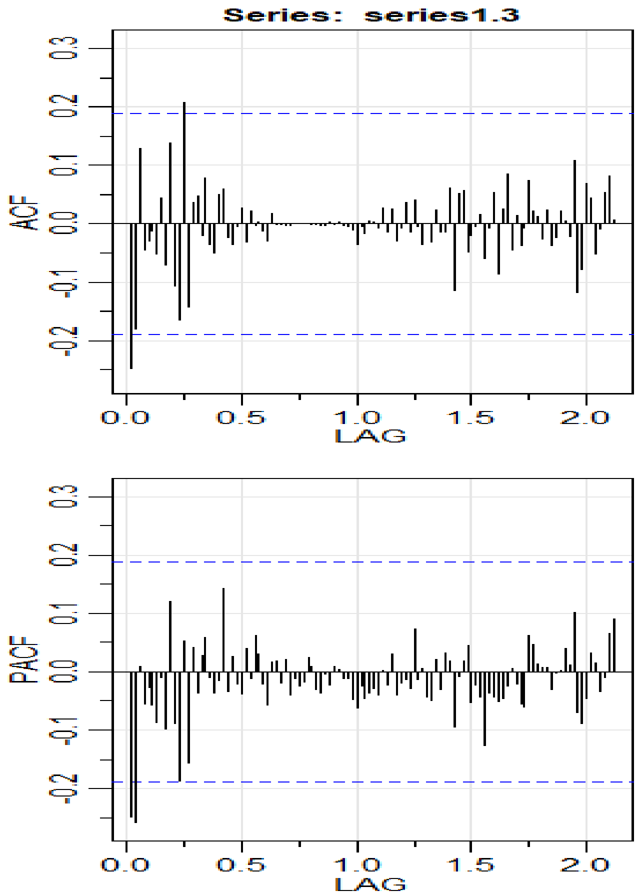
Stationarity and Model Selection

Type C/Store 30 – ARMA(2,0,1)

ACF and PACF (seasonality)

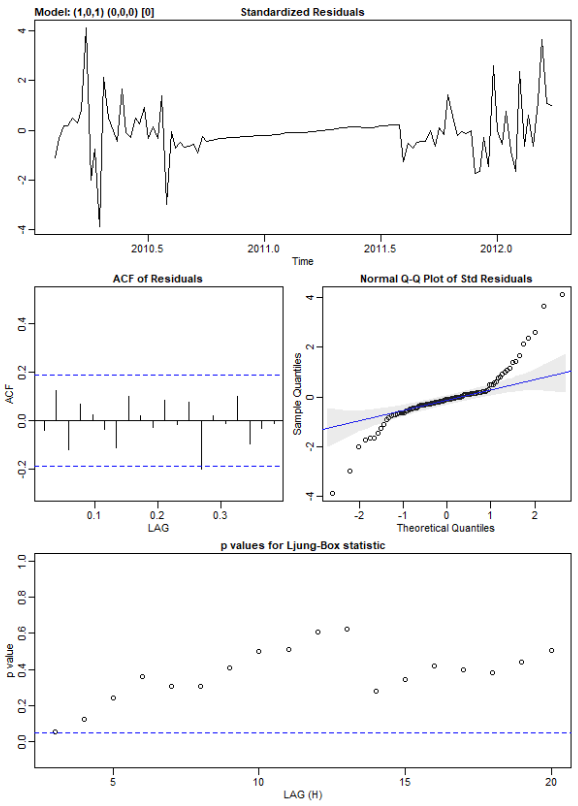


ACF and PACF (no seasonality)

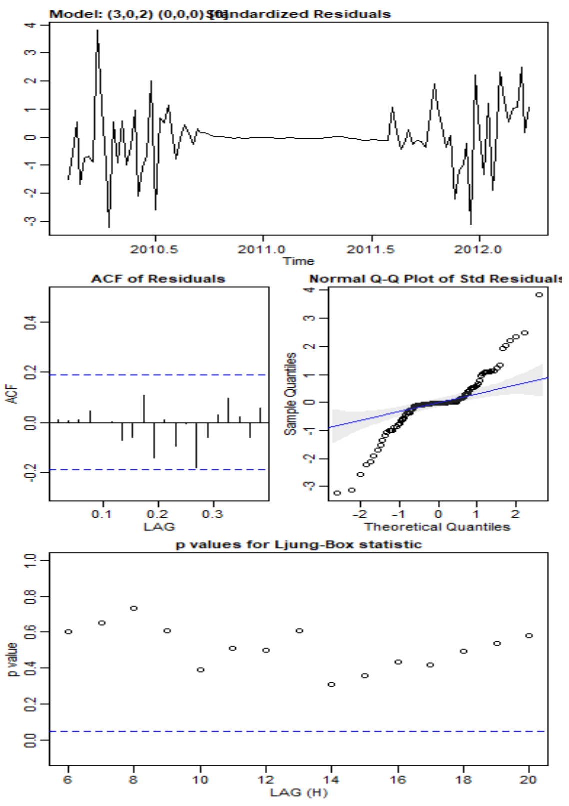


SARIMA model output

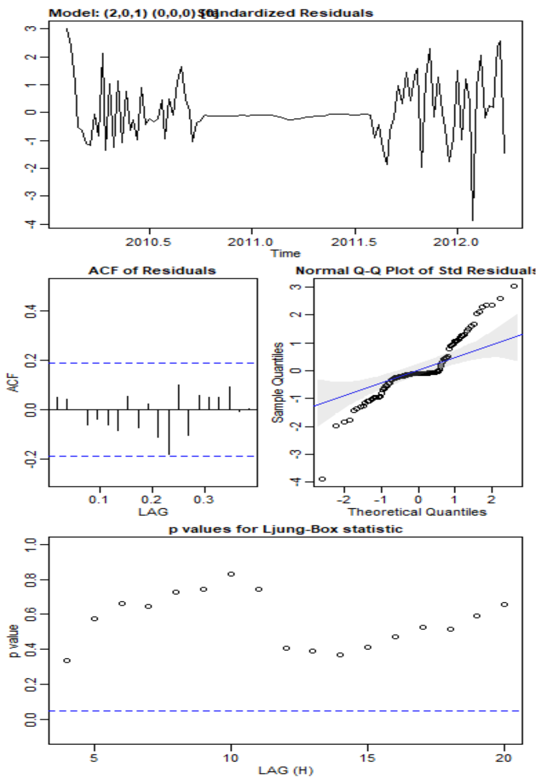
Type A



Type B



Type C



Forecast Results: MAPE

- Type A: 4.34%
- Type B: 3.68%
- Type C: 2.24%

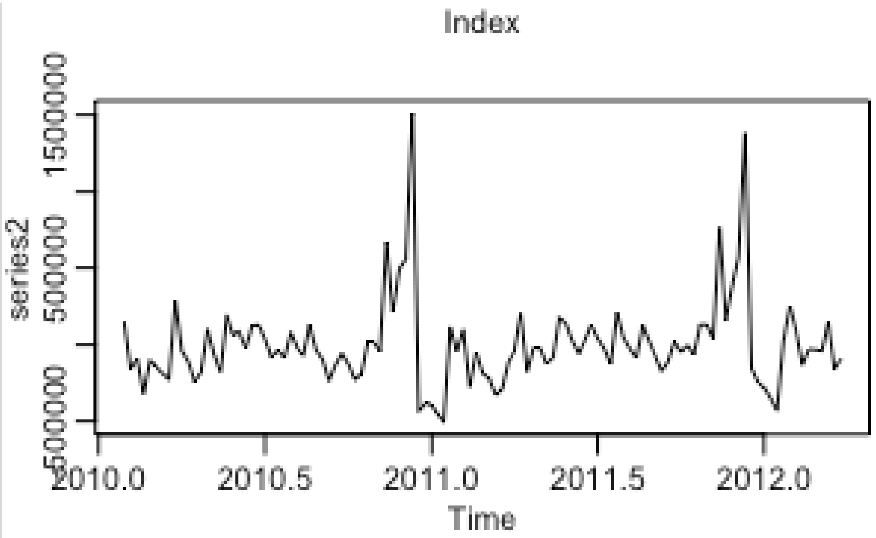
SARIMA

with additional variables

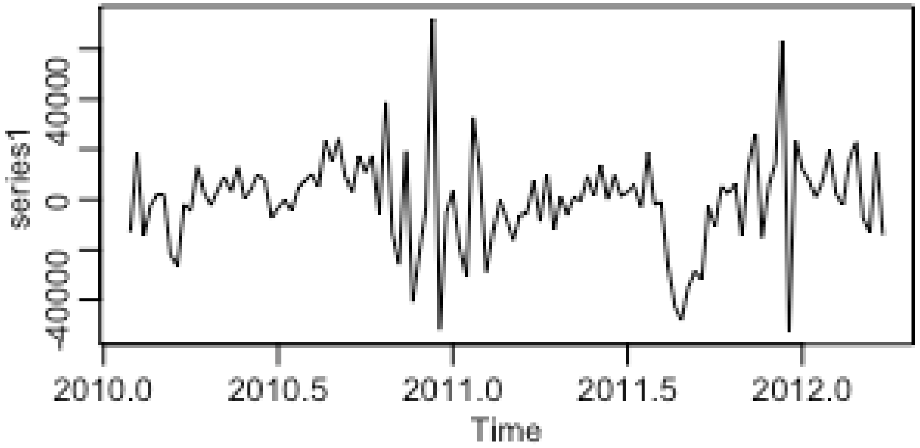
Regression Model

- Store Type A - 20
  - fit1 <- lm(Total_Sales ~Temperature + MarkDown3+ month, data = train_data)
- Store Type B – 3
  - fit <- lm(Total_Sales ~Temperature , data = train_data)
- Store Type C – 30
  - fit <- lm(Total_Sales ~Temperature + month + year, data = train_data)

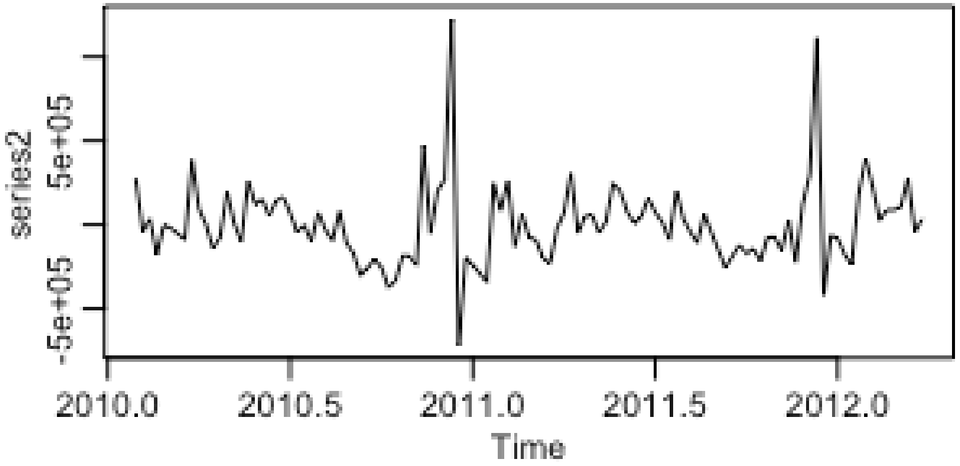
Residual Plots



Store 3

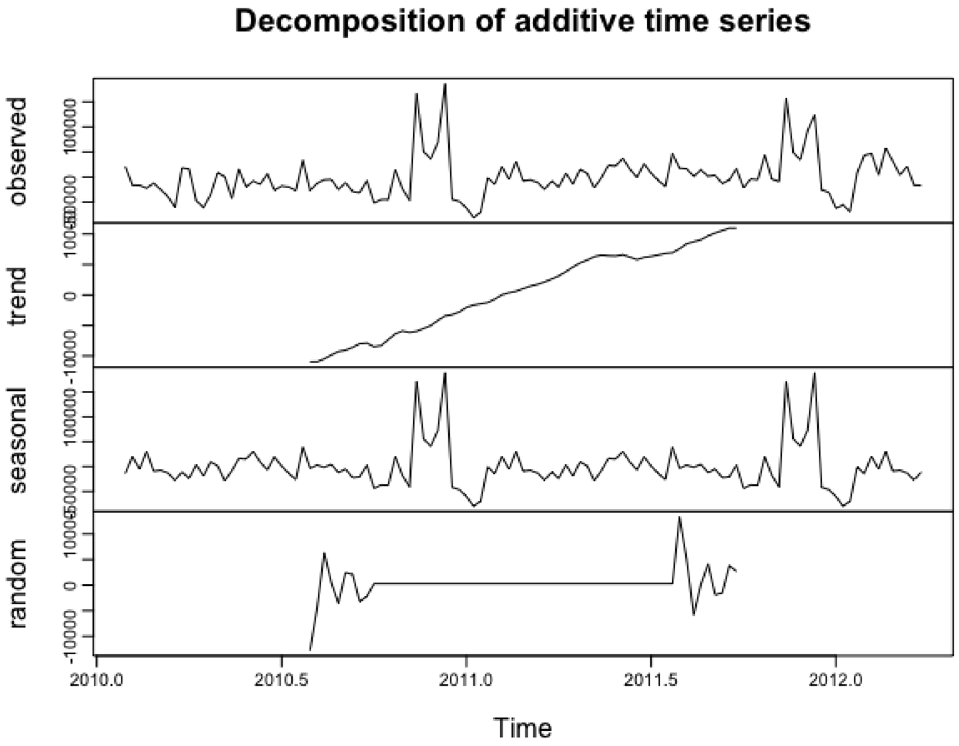


Store 30

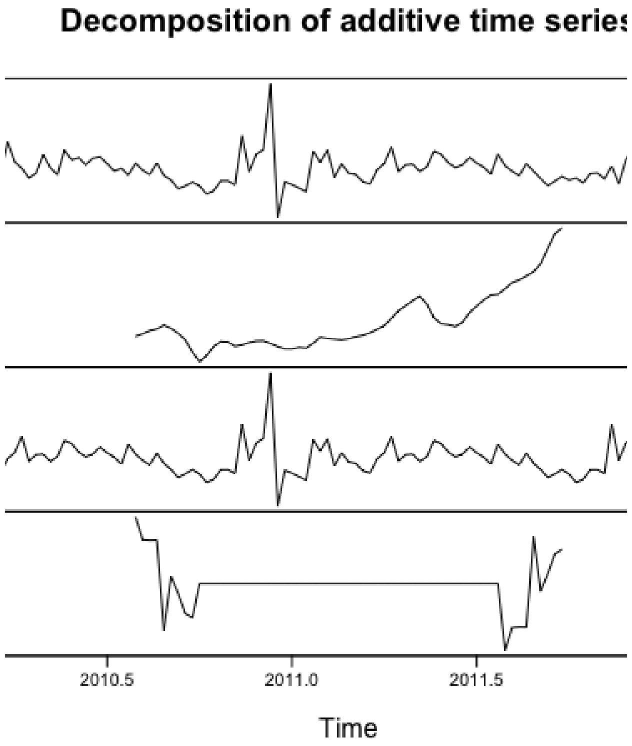


Store 20

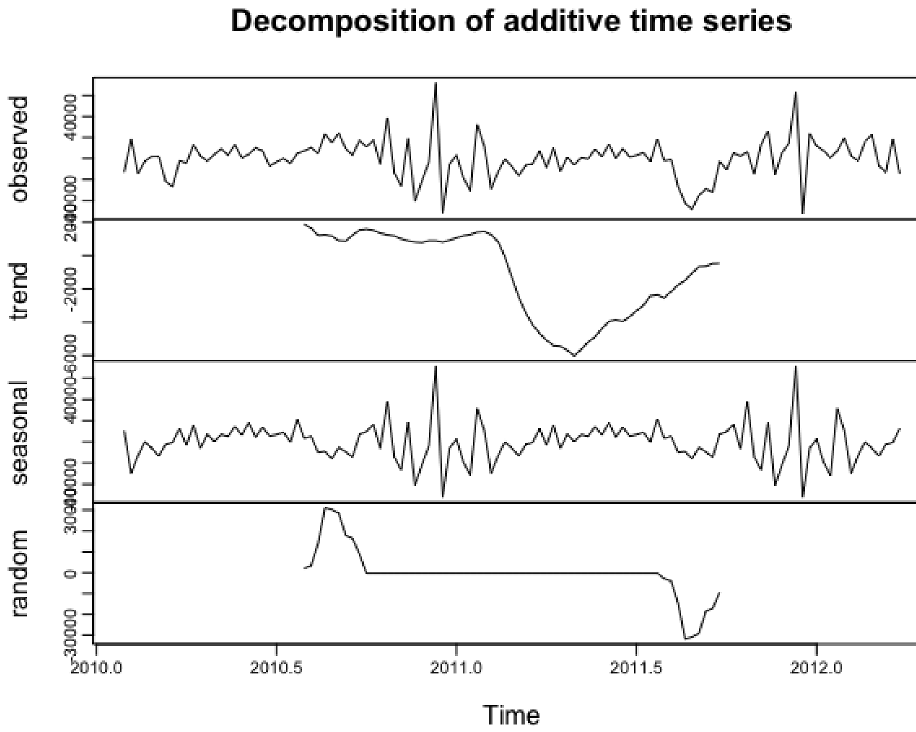
Decomposition



Store 3

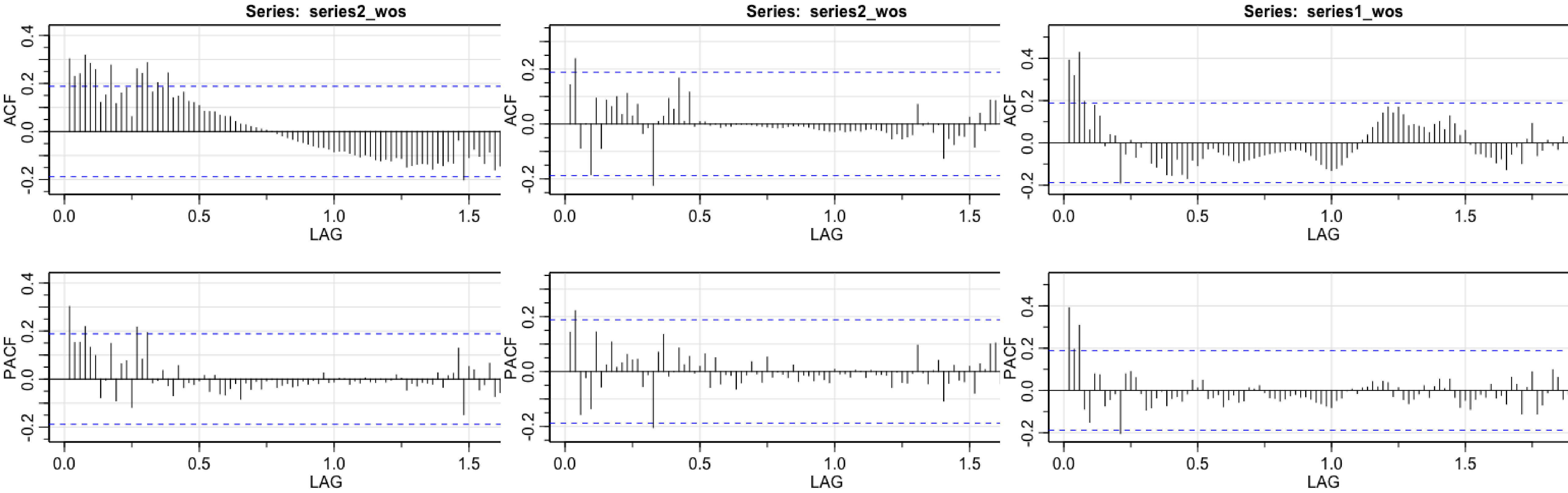


Store 20



Store 30

ACF/PACF

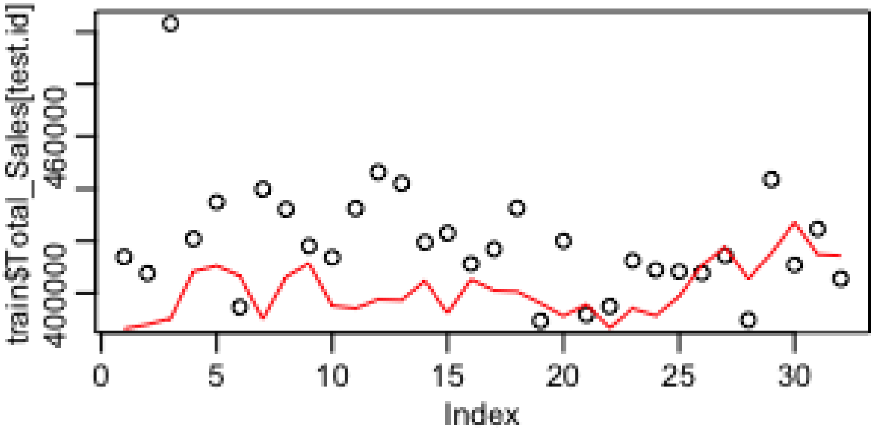


Store 3
ARIMA(1,0,1)

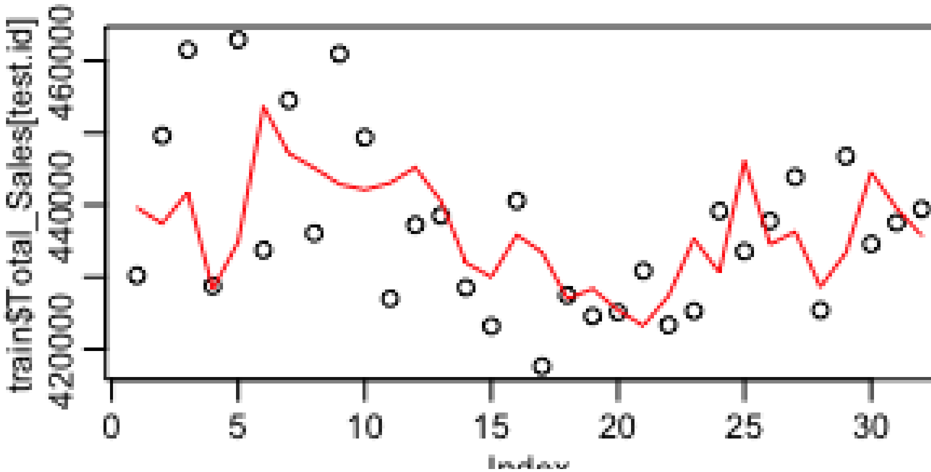
Store 20
ARIMA(1,0,1)

Store 30
ARIMA(3,0,3)

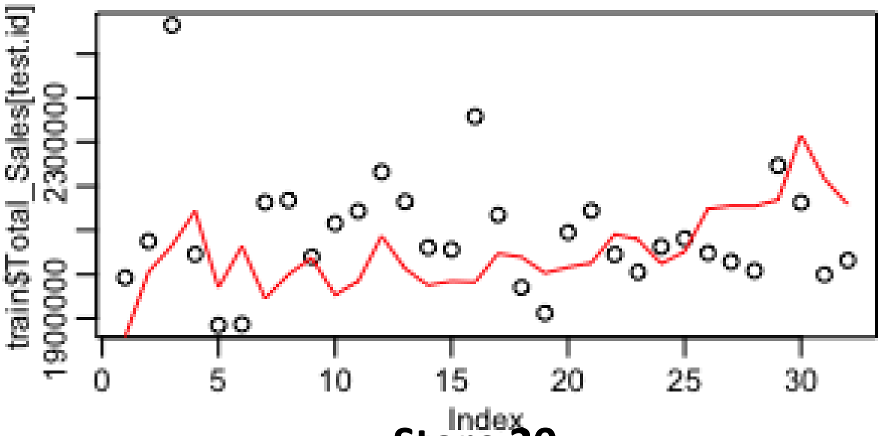
Forecast Results



Store 3
ARIMA(1,0,1)



Store 30
ARIMA(3,0,3)



Store 20
ARIMA(1,0,1)

Forecast Results: MAPE

- Store 3: 4.34%
- Store 20: 3.68%
- Store 30: 2.24%

BSTS

Why BSTS?

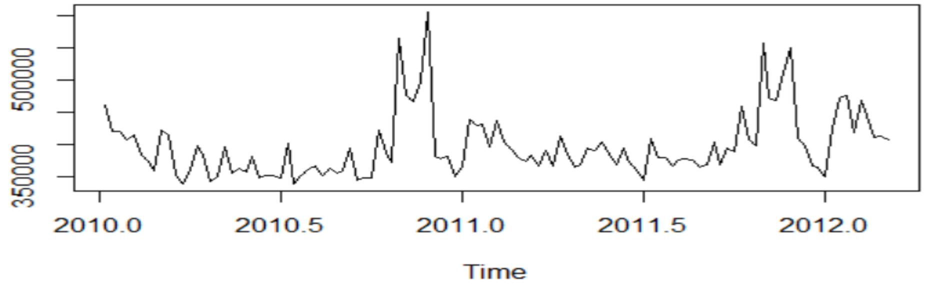
Handling shortage of data points

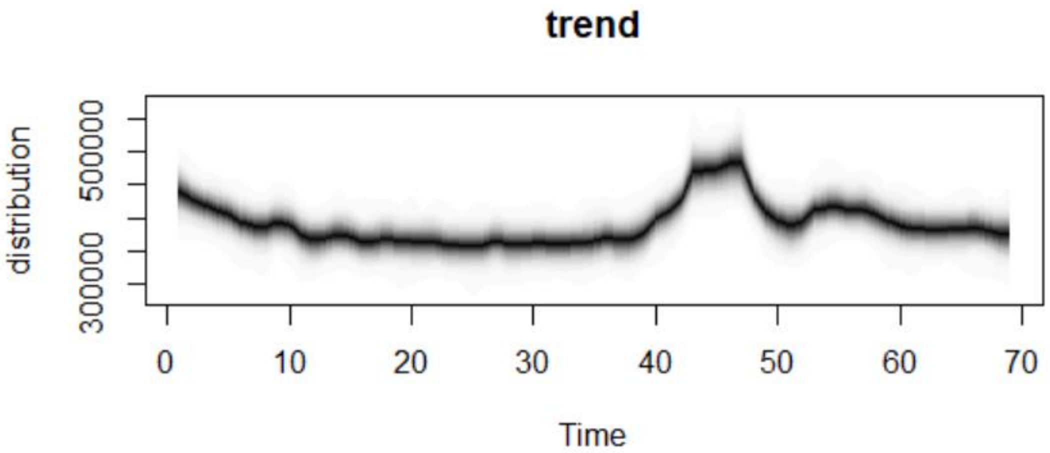
Handling Uncertainty

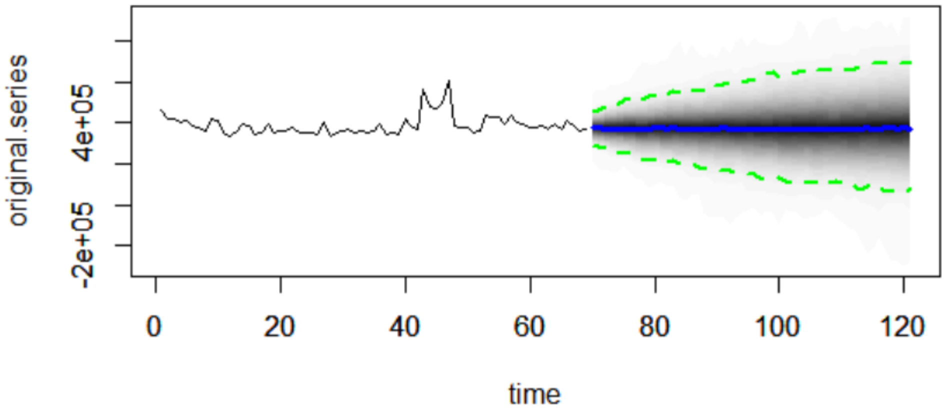
- Doesn't rely on differencing, lags and MA
- Combines prior and estimate parameters from posterior probability distributions of each co-variate

Store Type B – 3
Local level model

```
llm <- ts(Store3_train$Total_Sales, start = c(2010,2,5),
         end = c(2012,10,26), frequency = 30)
llss<-AddLocalLevel(state.specification = llss, y=llm)
llfit<-bsts(llm, state.specification = llss, niter = 1000)
```



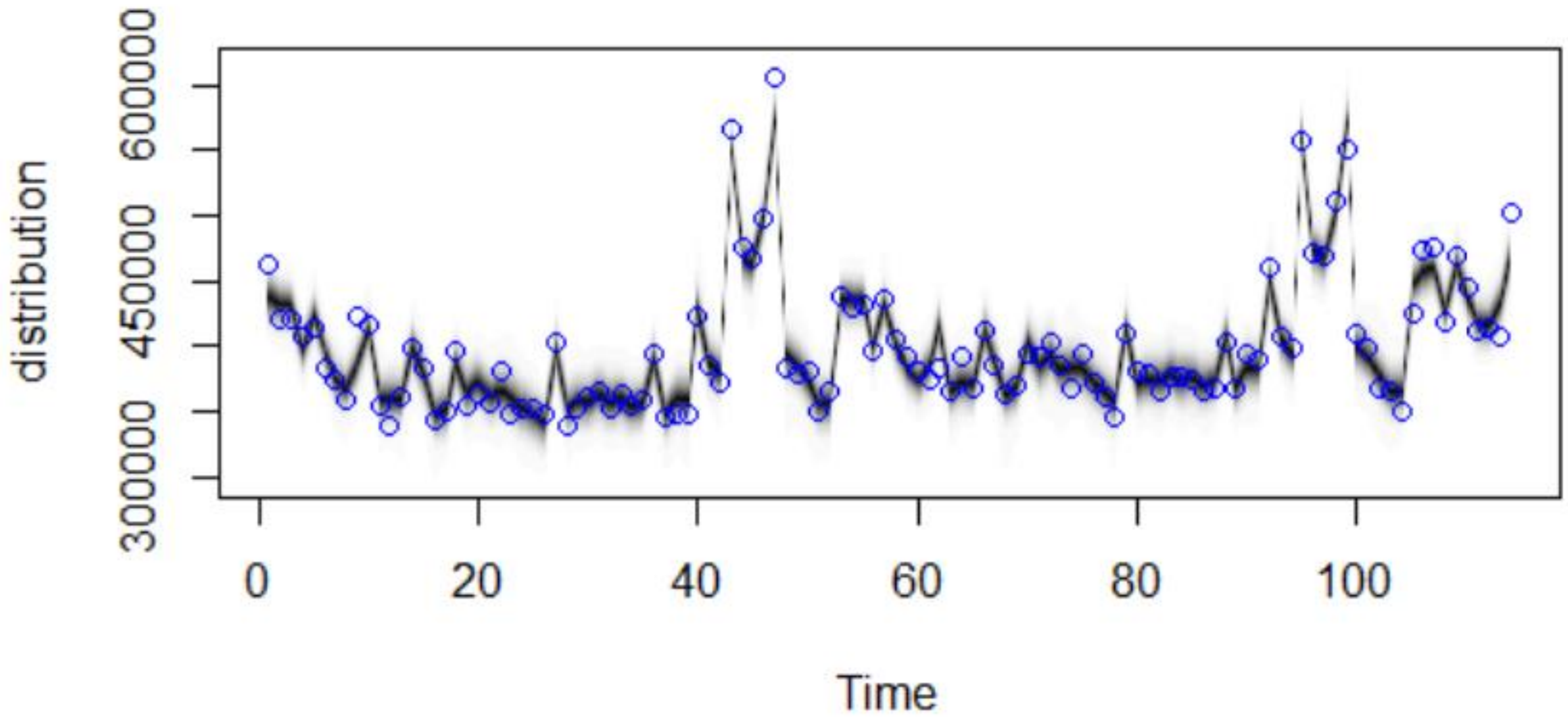




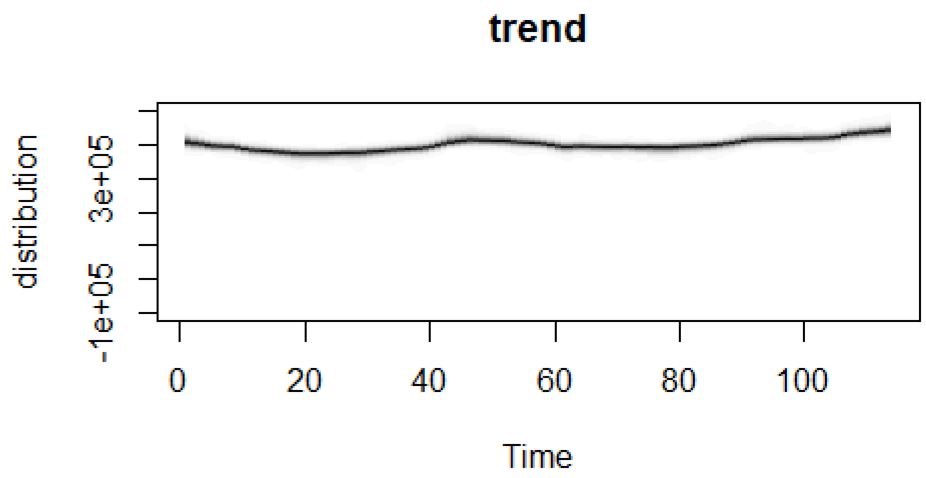
Store Type B – 3

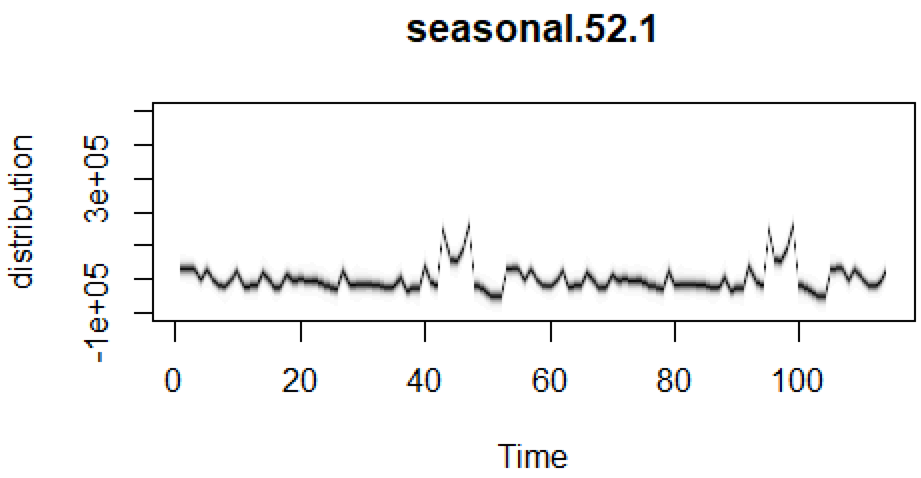
Local trend with seasonality

```
ss <- AddLocalLinearTrend(list(), Store3_train$Total_Sales)
ss <- AddSeasonal(ss, Store3_train$Total_Sales, nseasons = 52)
model1 <- bsts(Store3_train$Total_Sales,
               state.specification = ss,niter = 1000)
```

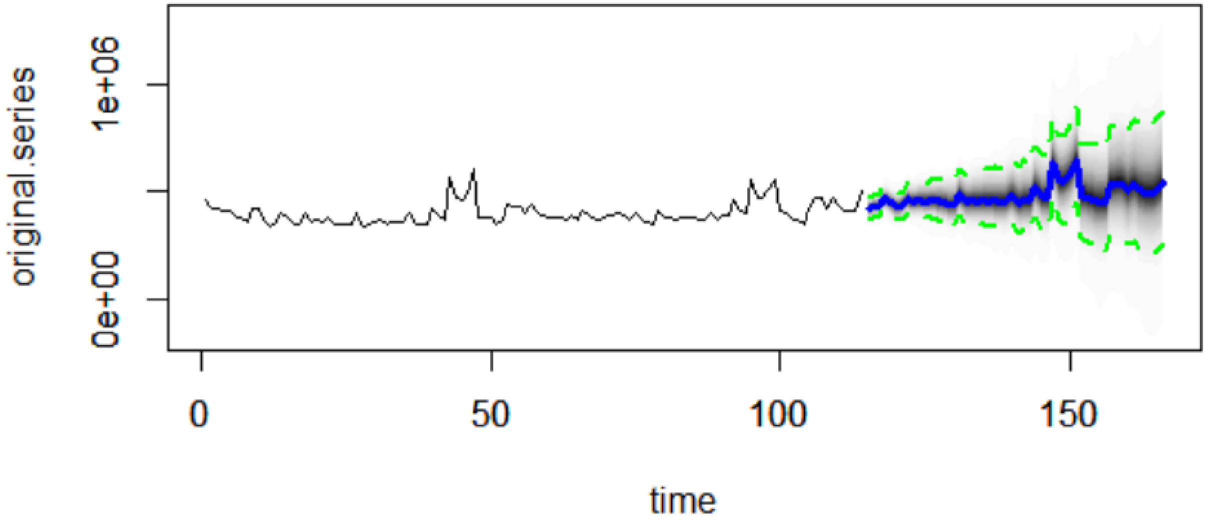


Local trend with seasonality – Components & Prediction






```
pred1 <- predict(model1,newdata = Store3_test,
                 horizon = 52)
plot(pred1, plot.original = 156)
```

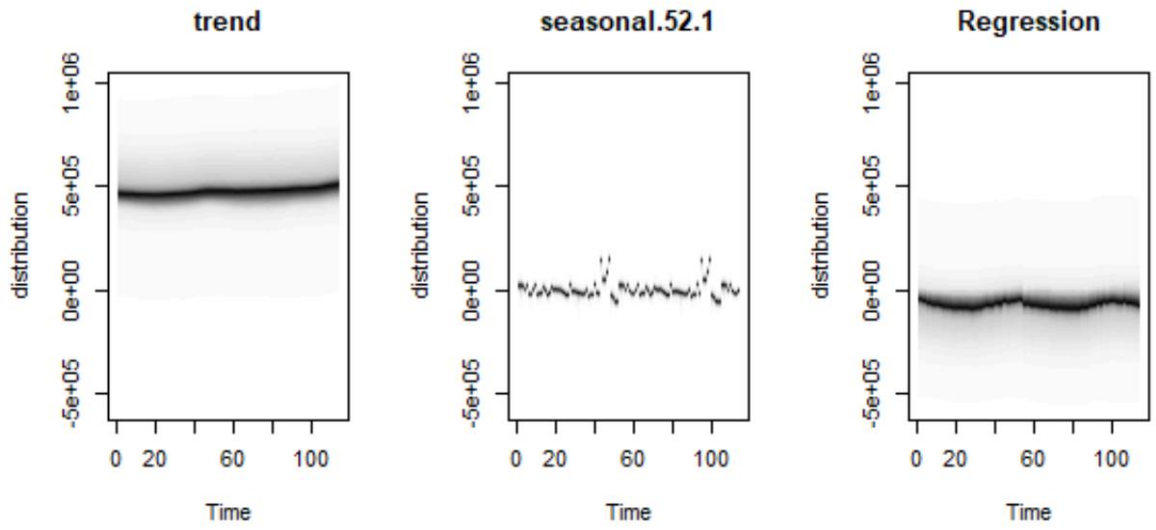


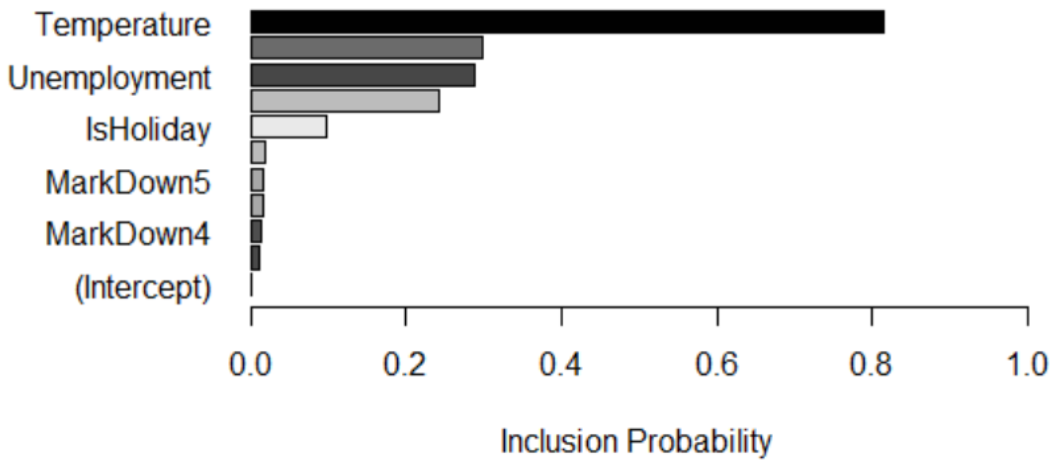
Store Type B – 3

Local trend with seasonality and Regression

```
model3 <- bsts(Total_Sales ~ ., state.specification = ss,
               data = Store3_train, niter = 5000,
               expected.model.size = 5)
```

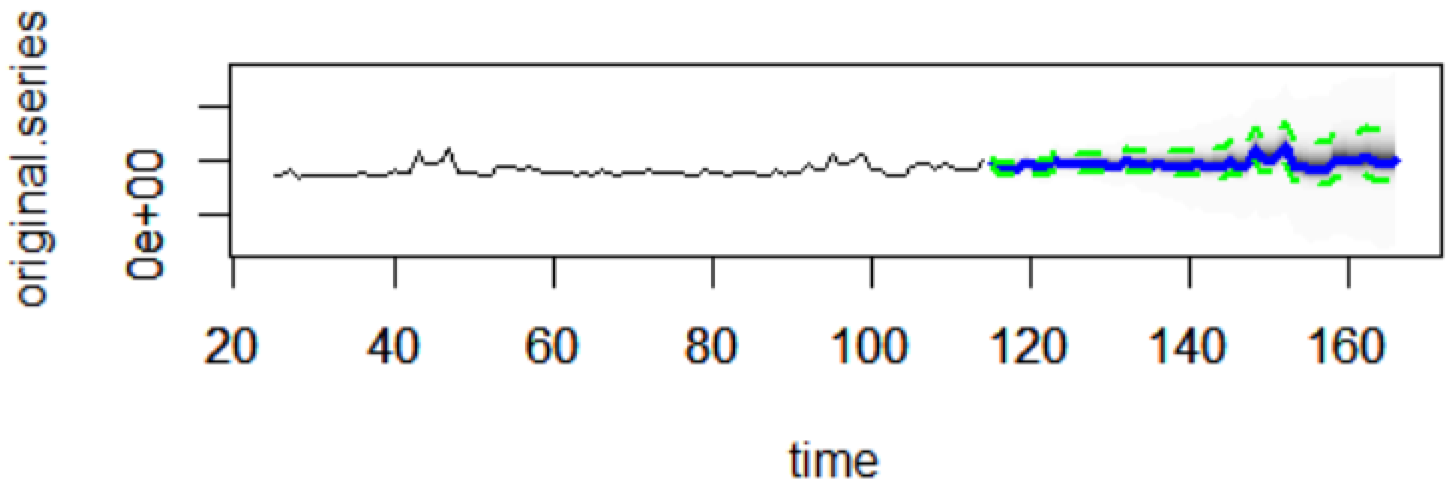
$$y_t = \mu_t + \tau_t + \beta^T \mathbf{x}_t + \epsilon_t$$



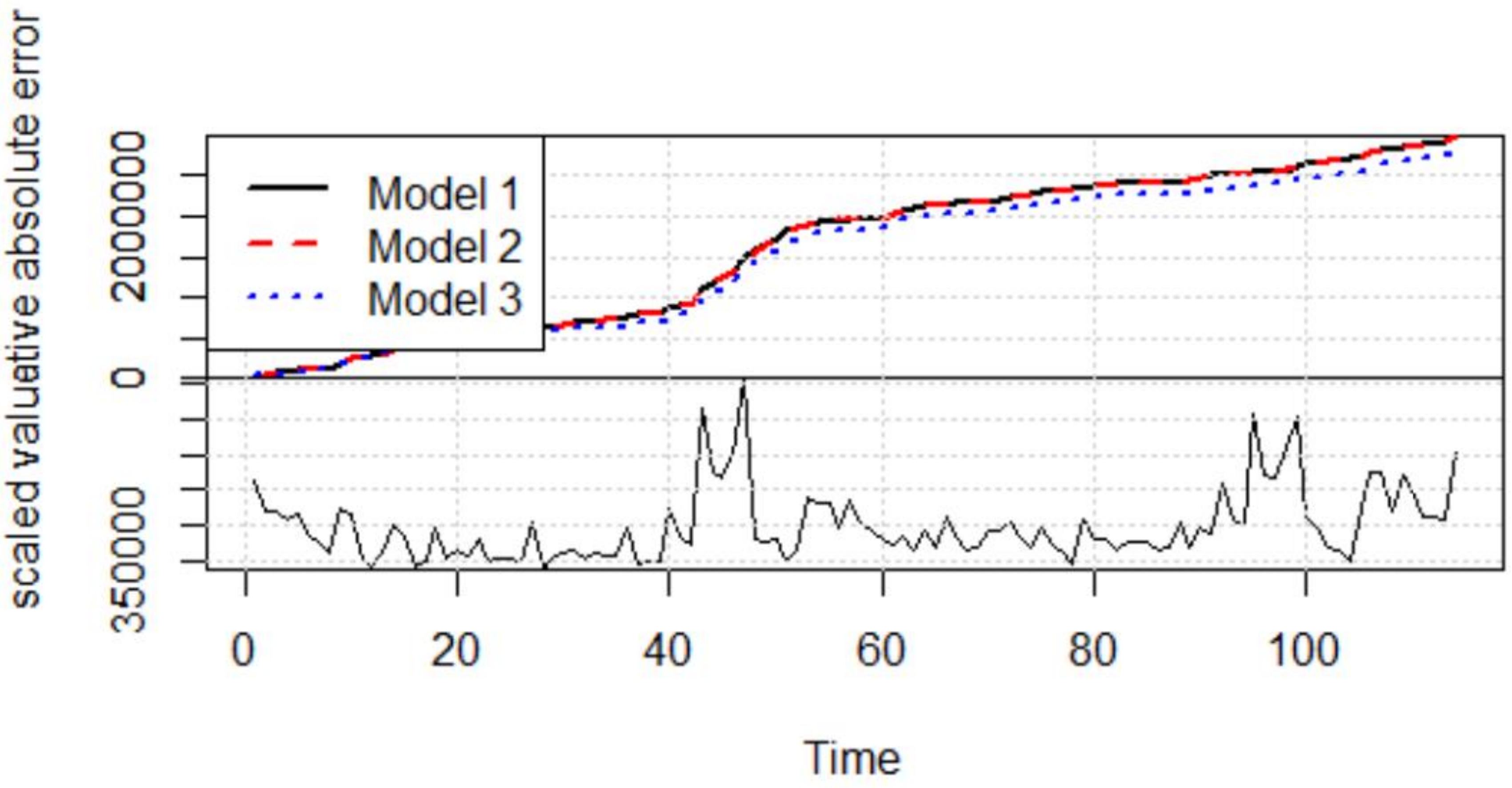


Prediction

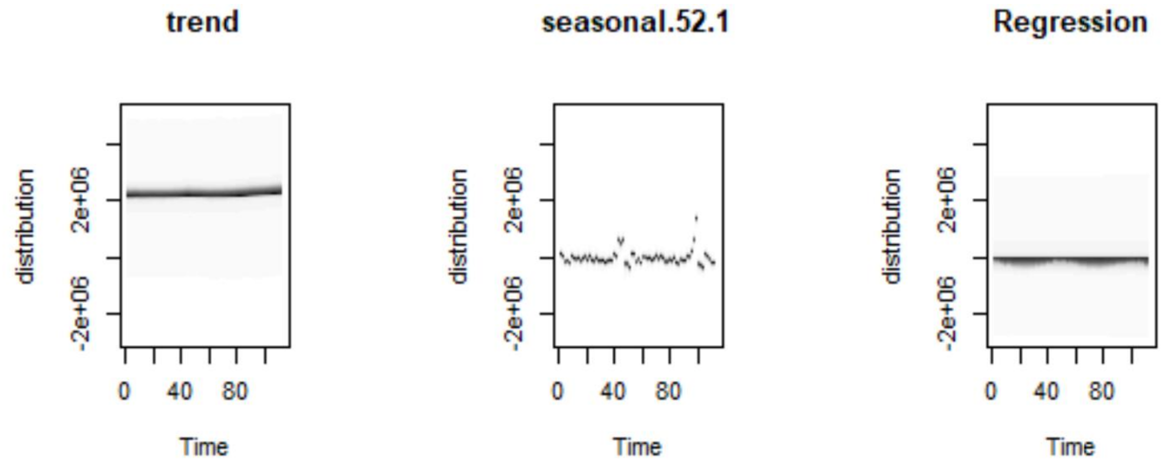
```
newpred<- predict(model3, newdata =Store3_test, horizon = 52)
plot(newpred, plot.original =90, main = 'trend')
```

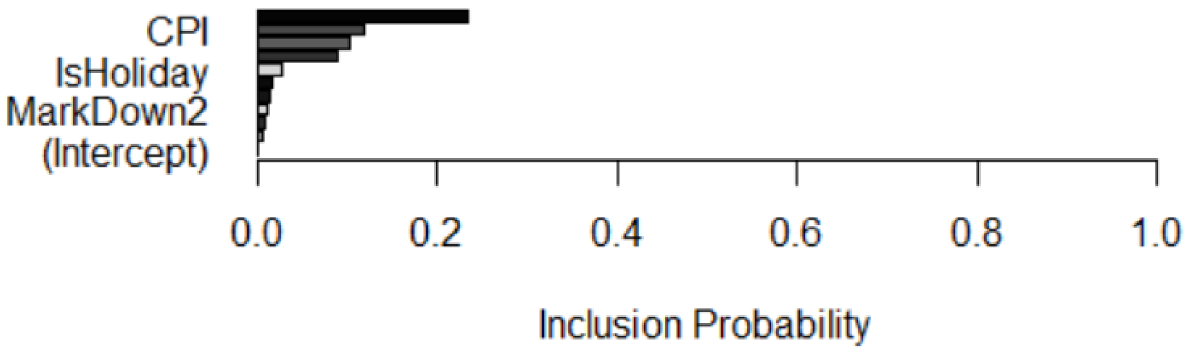


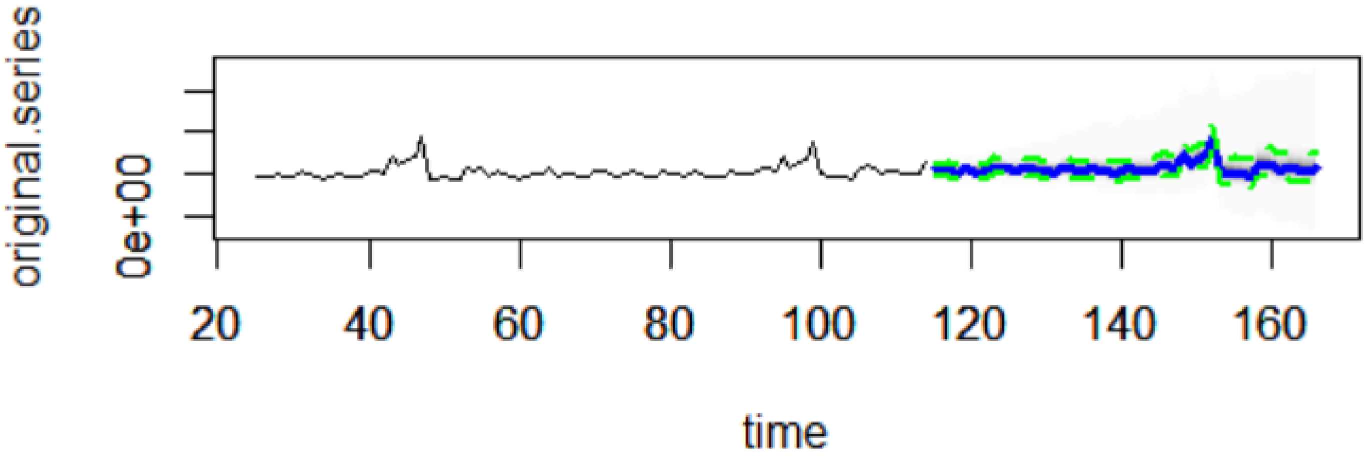
Comparison of models



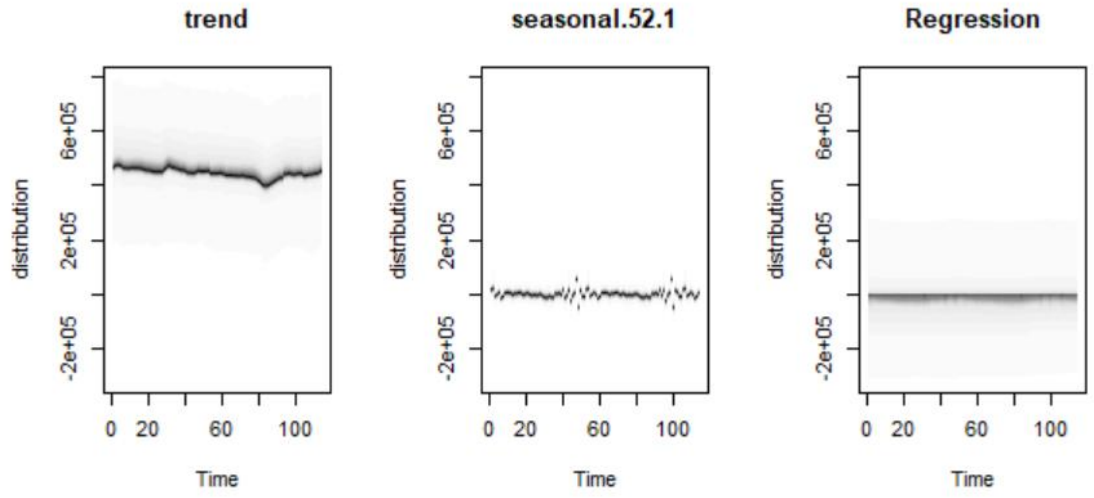
Store Type A - 20

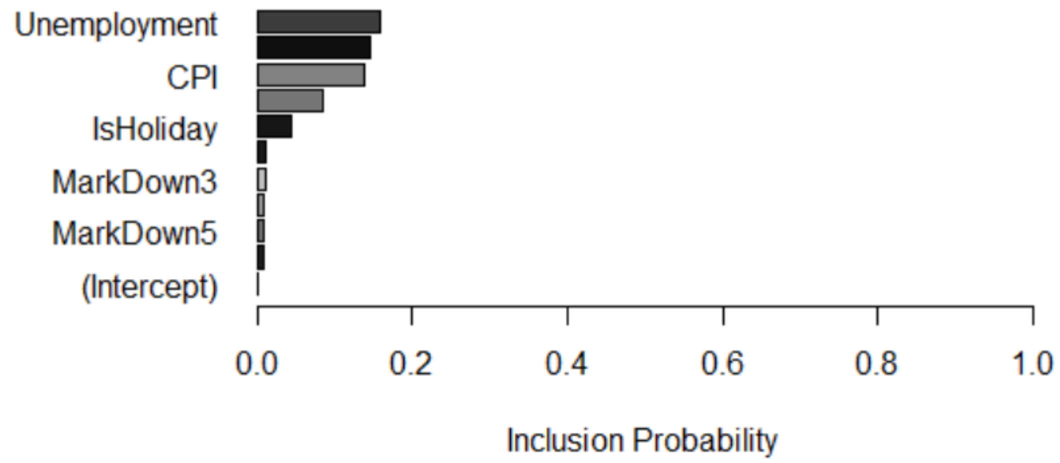


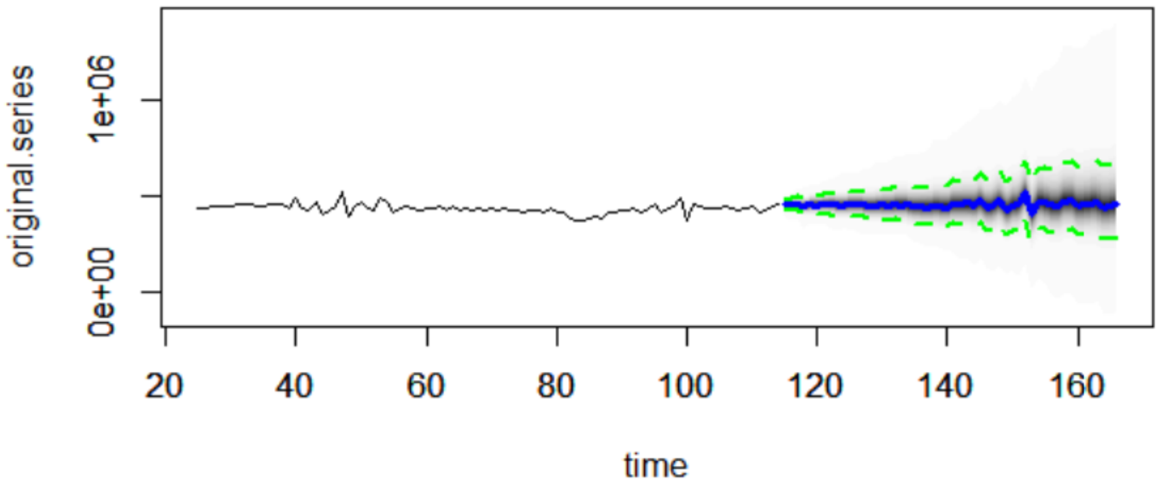




Store Type C - 30







Forecast Results: MAPE

- Store 3: 14.4%
- Store 20: 16.6%
- Store 30: 5.04%

Conclusion

- Comparing all the three methods
- Store wise conclusion

Next Steps

- Final Model Tuning
- Test one store each from the same story type category and fit the model to test accuracy
- Extend the analysis for across all stores
- Develop department wise forecast model