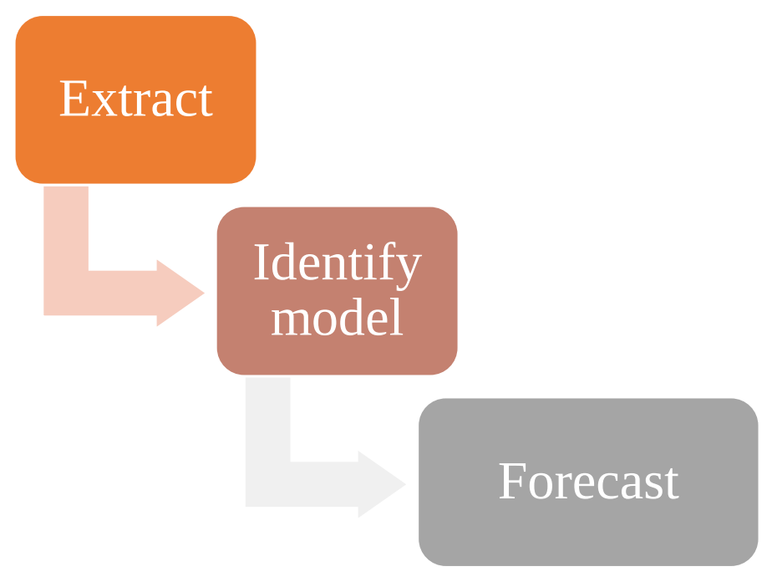
Sales Data Analysis with Time Series using Filtering and Adaptive Boosting

Nasheen Nur
Dept. Of Software and Information System, University of North Caroline AT Charlotte

Project Motivation & Objective

- In order to extract meaningful statistics and other characteristics of sales data.



- Identifying time expand of peak sales or sales drop by observing periodic data
- Predicting future values based on previously observed values
- Comparing values of a single time series or multiple dependent time series at different points in time
- Helping different agencies to allocate their time and resources more wisely based on the reliable predictions

Dataset Characteristics

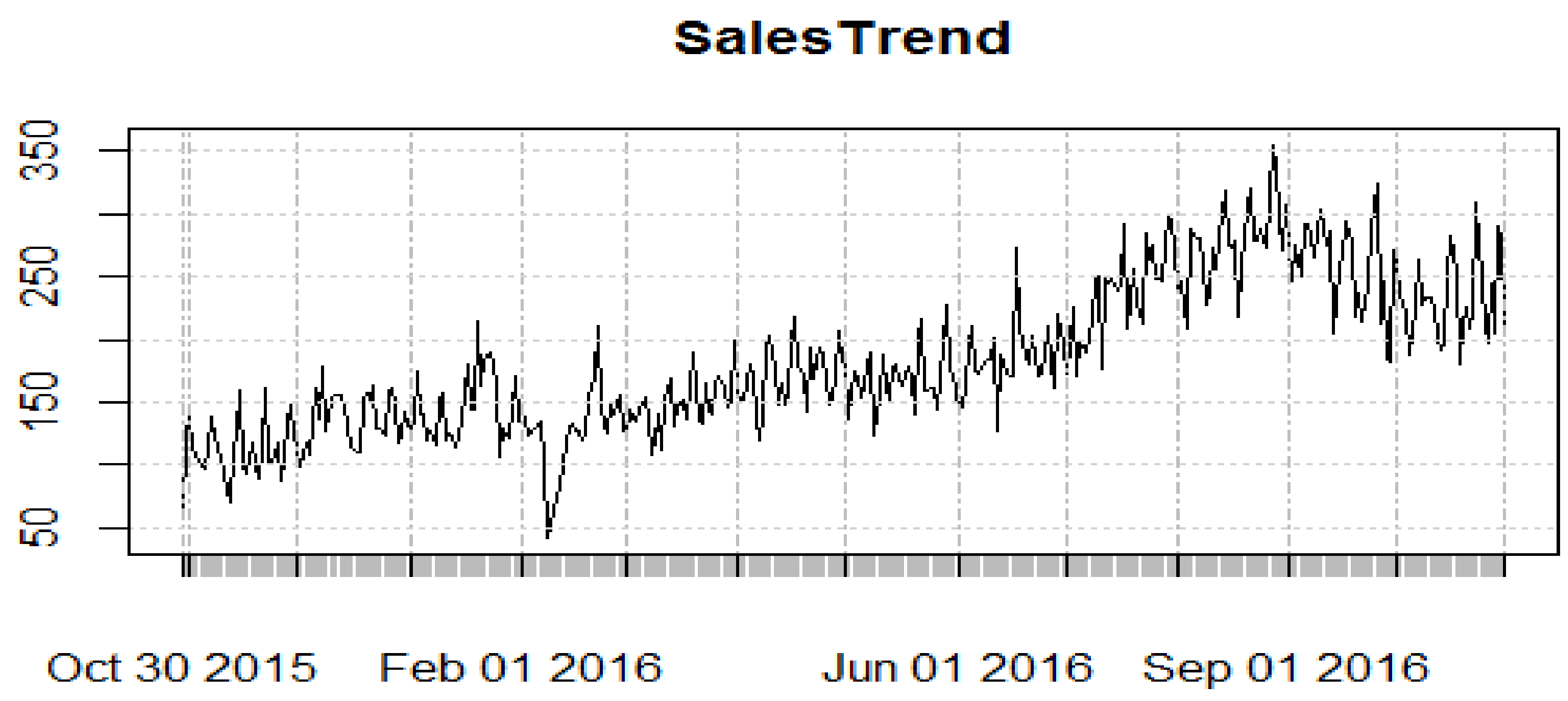
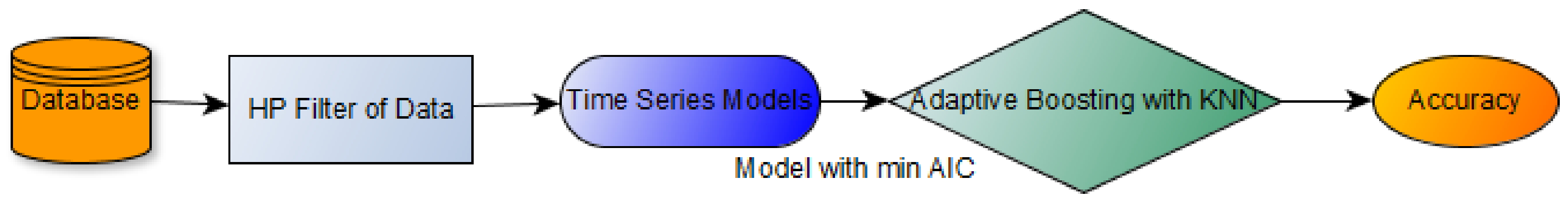


Figure 1: Sales data in original format for a particular shop

- We are using the IJCAI competitions (2017)[1] database
  - ➢ Database consists of 2000 sales data for 1 year (365 days)
  - ➢ Each shop's data consists of date and associated #sales for that day, so each shop has such data for 365 days
- Our task is to predict last 14 days sales data from the prediction model of 351 days data.
- The day on day trend clearly shows that the #sales have been increasing without fail.
- The variance and the mean value in July and August is much higher than rest of the months..
- Even though the mean value of each month is quite different their variance is small. Hence, we have strong seasonal effect with a cycle of 12 months or less

Boosting Best Fit Filtered Data

- Preprocessing– Hodrick-Prescott Filter of Data
- Finding Best fit based on AIC
- Adaptive Boosting (KNN boosting)



- The data is preprocessed by HP filtering to remove the cyclical component of the time series from raw data.
- The processed data are fed into several time series models (ARIMA,ARFIMA) and best model based on min AIC(Akaike information criterion) is selected.
- Five "bags" of data was taken and ran through the main algorithm to give diversity to the data and hopefully create a higher testing accuracy

Reference

[1] https://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100066.333.8.k2tLnj&raceId=231591.

Filtered Data Properties

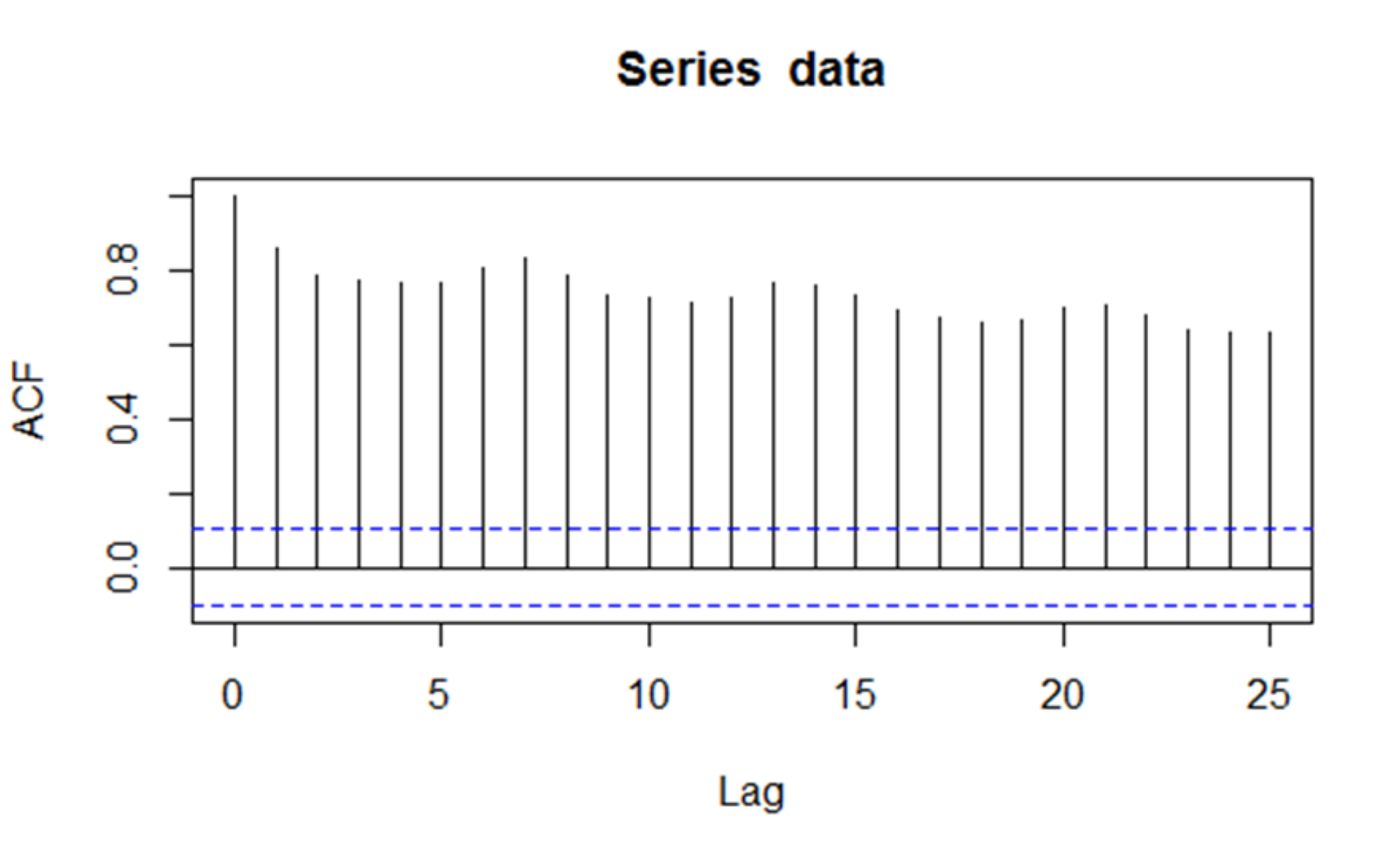


Figure 2(a): HP Filter of Data

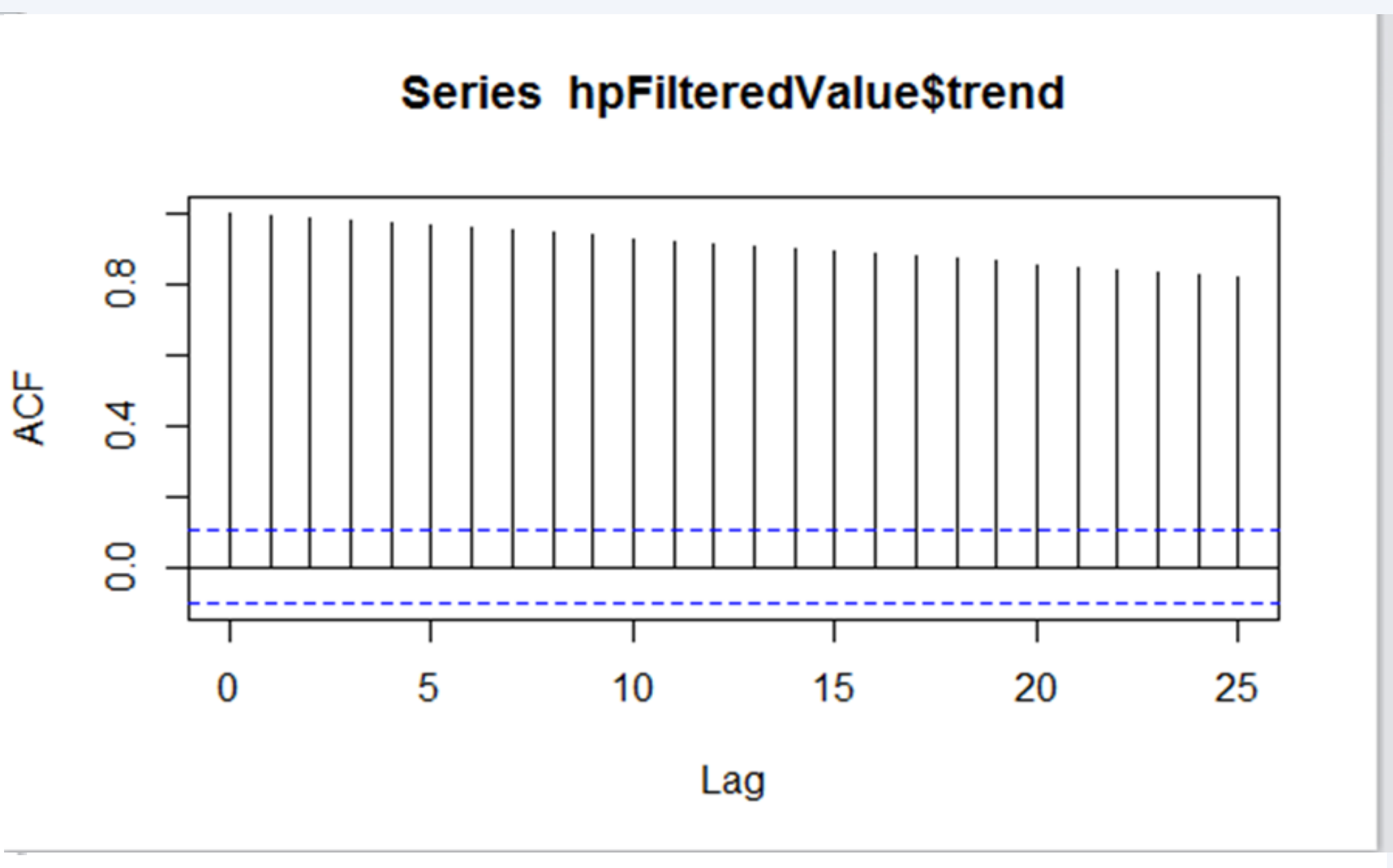


Figure 2(b): ACF plot after Data Processing

- ACF(Auto Correlation Function) is a plot of total correlation between different lag functions.
- The correlation is gradually going down without any cut off value after filtering the data

Evaluation

- **Performance of Filtered Data :**

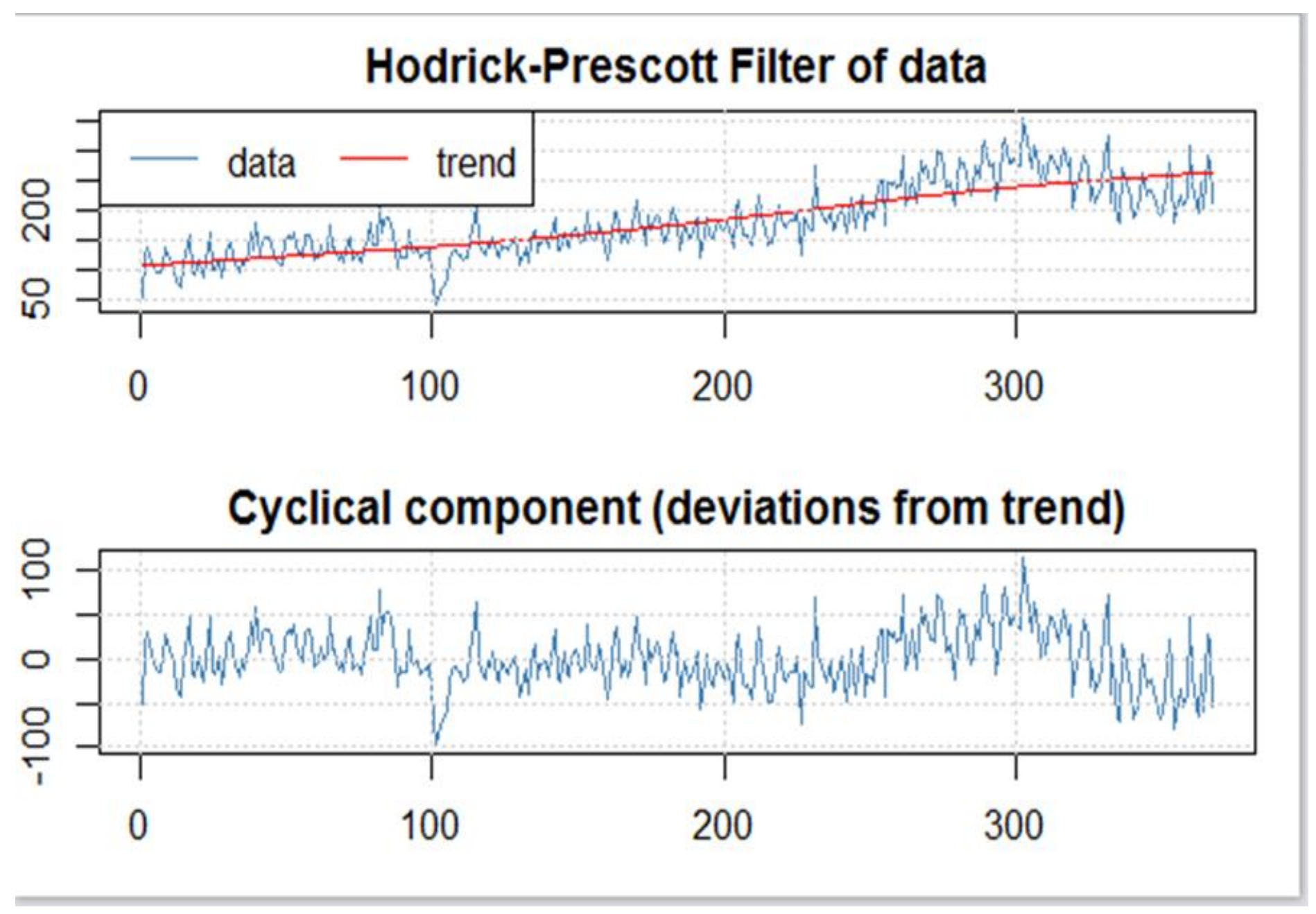


Figure 3(a): HP Filter of Data

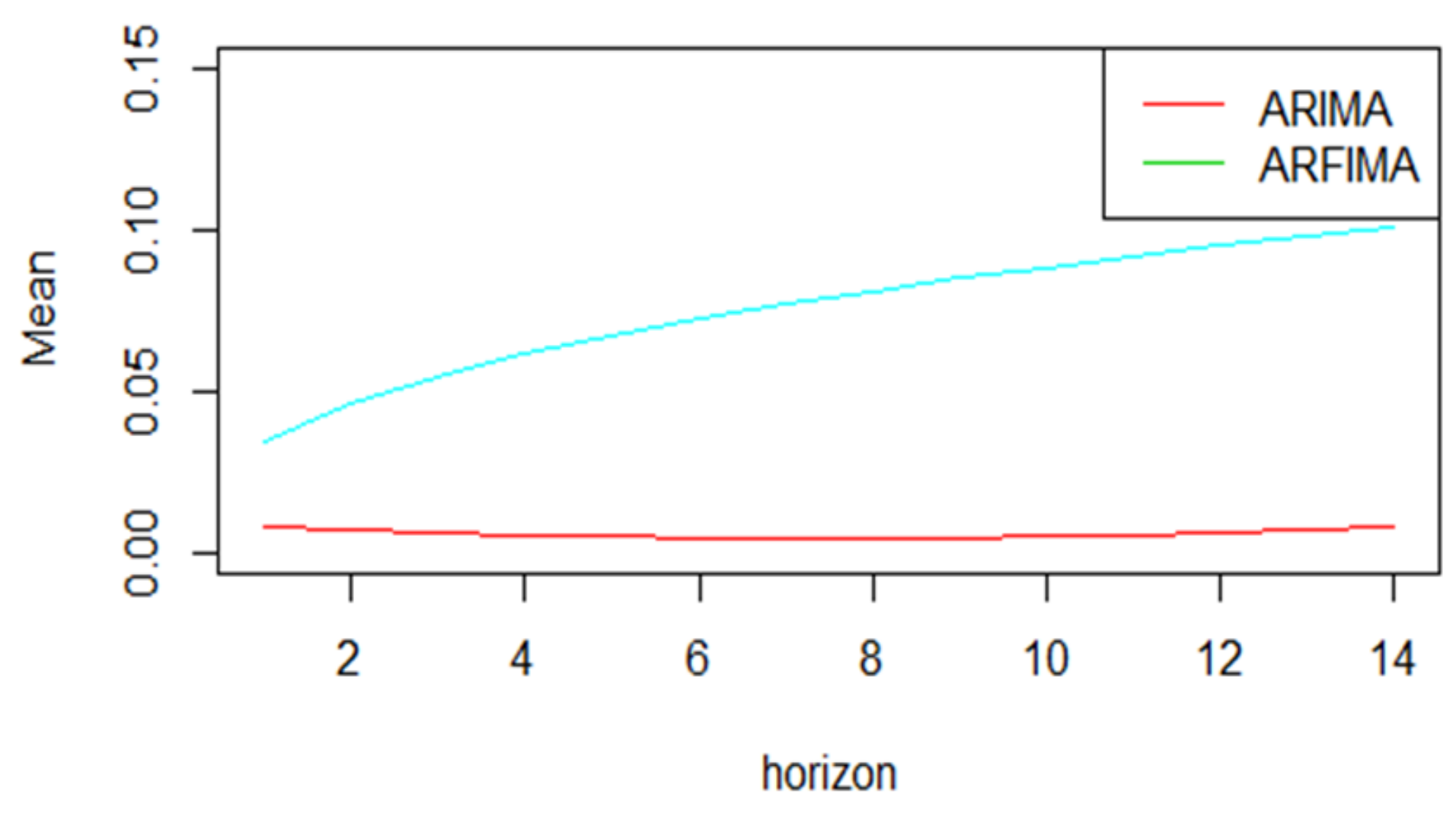


Figure 3(b): MAE of TS models

- ➢ HP filter is used to decompose the time series in different components i.e trend and deviation from trend. We worked with the trend part.
- ➢ Later the trend component output from HP filter is used to define forecasting models (ARIMA,ARFIMA)
- ➢ The model with min AIC is taken and fed in to the next step

- **Performance of Ada Boost with KNN**

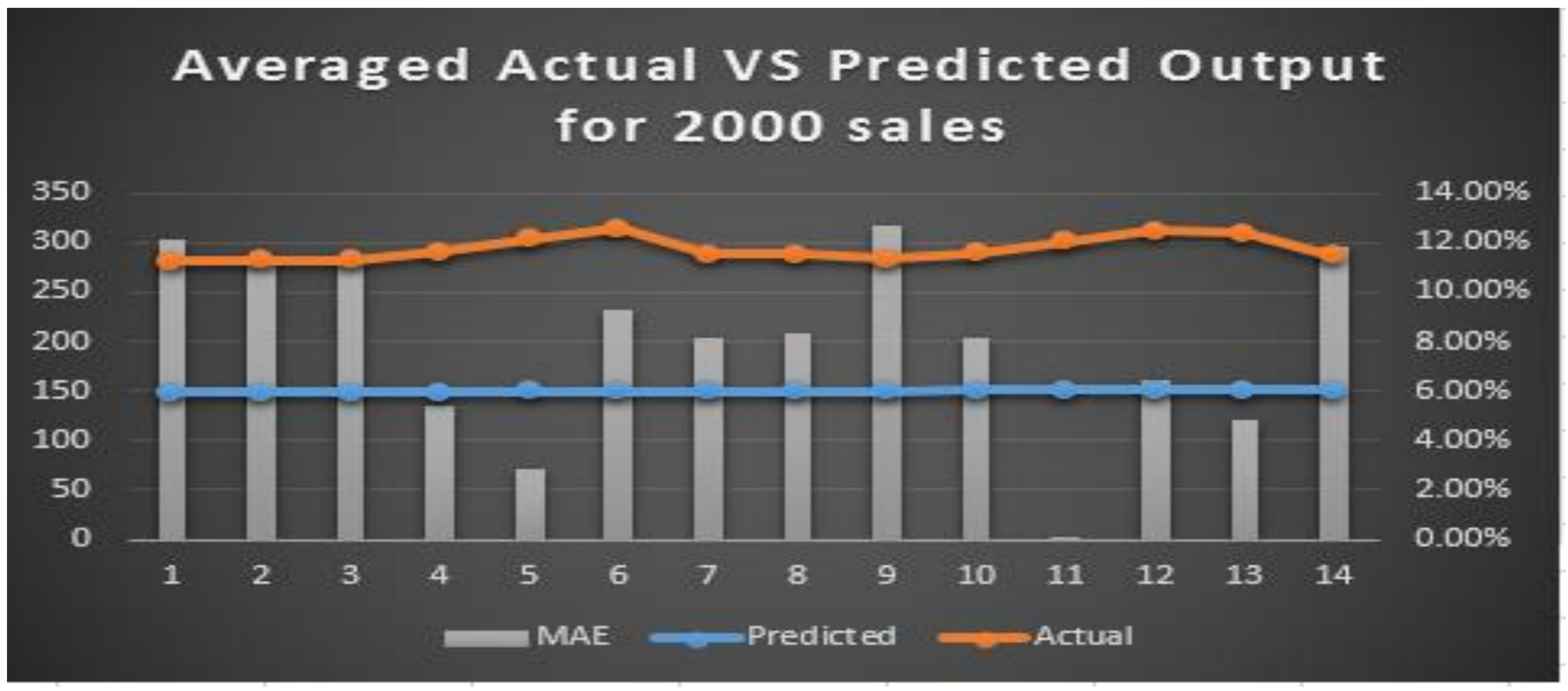


Figure 4: Averaged Actual vs Predicted Output of 14 days for 2000 sales

- ➢ Due to how time series was already manipulating the data in order to fit a certain time period, boosting did not enhance the final accuracy ratings by much. That's why we kept it's improvement as one of our future works.

- **Future Work :**
  - ➢ Working with data imputation.
  - ➢ Working and analysis on the other components of the filtered data other than the trend
  - ➢ Considering seasonal peak and bottom sales values
  - ➢ Using some other prediction models.
  - ➢ Implementation of adaptive boosting with other algorithms such as Naive Bias, Decision tree, Random forest.