ARIMA- Time Series Forecasting

Time Series Analysis of
S&P/Case-Shiller Home Price Index

Prepared By: Jasmeet Sasan

Contents

Objective:

The study aims to find significant insights of S&P/Case-Shiller Home Price Index data by performing data exploration and forecasting.

Assumptions

- Index Data: Since we are dealing with indices, the prediction may not be able to provide great insights or become a real problem solver. Forecasting the index may not show the real picture because the index itself is made up of multiple components (can't decompose it into substituents):
  - Index is directly proportional to average change in House Prices
  - Index calculating process has not been changed for all the reported years in data.
  - **NO** New variables have been introduced in between.
- Composite-10 is joint index of 10 metro cities.
- Every city's data represents a univariate time series, with **no** correlation with other series.
- Dummy variables would be added to smoothen the prediction and explain unexpected behavior
  - It may or may not be able to completely explain time series components such as seasonality, cycles or trends.

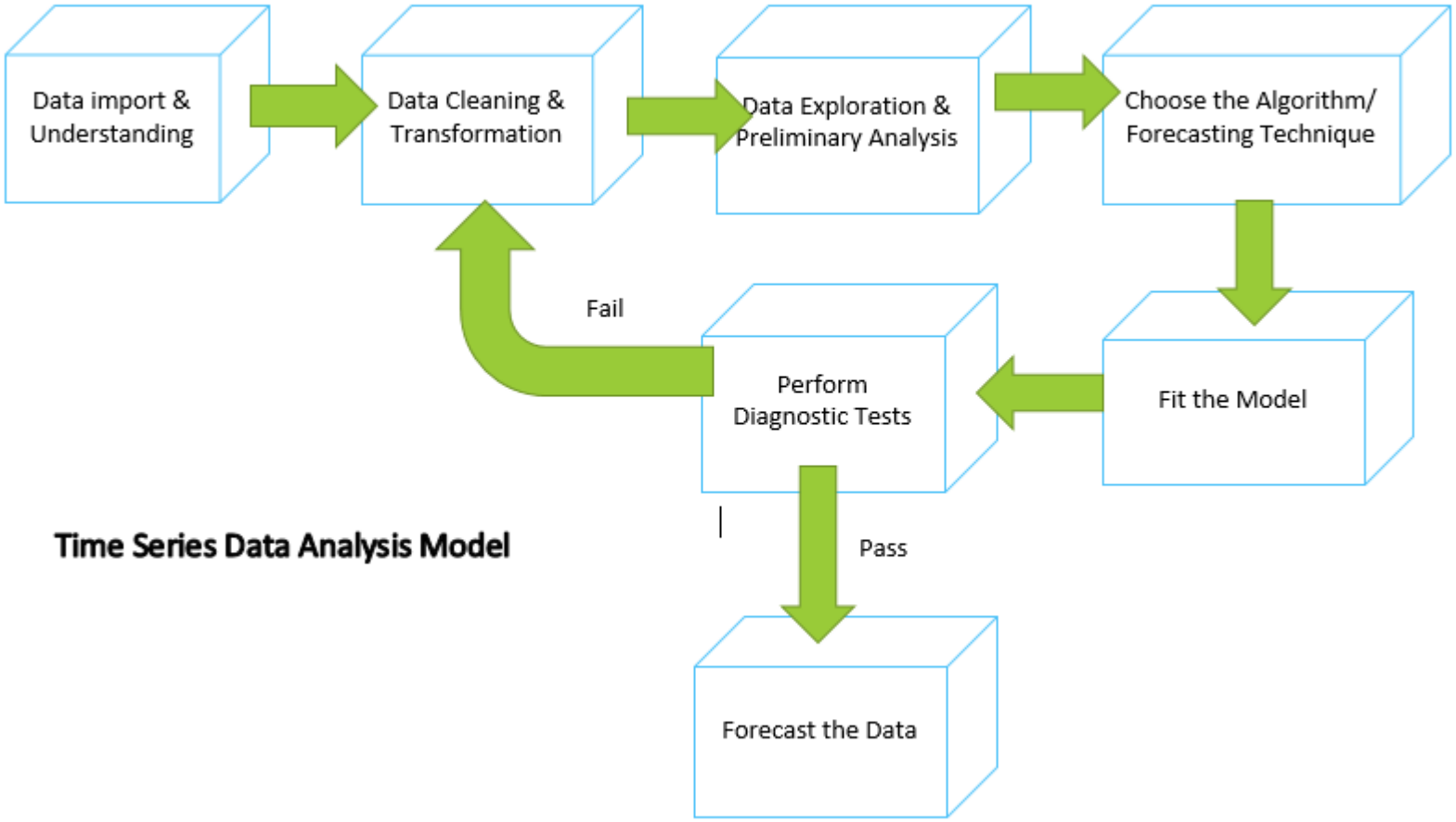
Data

- Given is monthly univariate time series data composed of housing price index for 19 cities and a composite index from 1991 to 2009 available monthly.
- New Variables would be added as forecasting demands.

Tools

- R
- Tableau

Analysis Model



Time Series Data Analysis Model

Data Import

- Data is available on an HTML page
- R's XML package handles the extraction job very well.

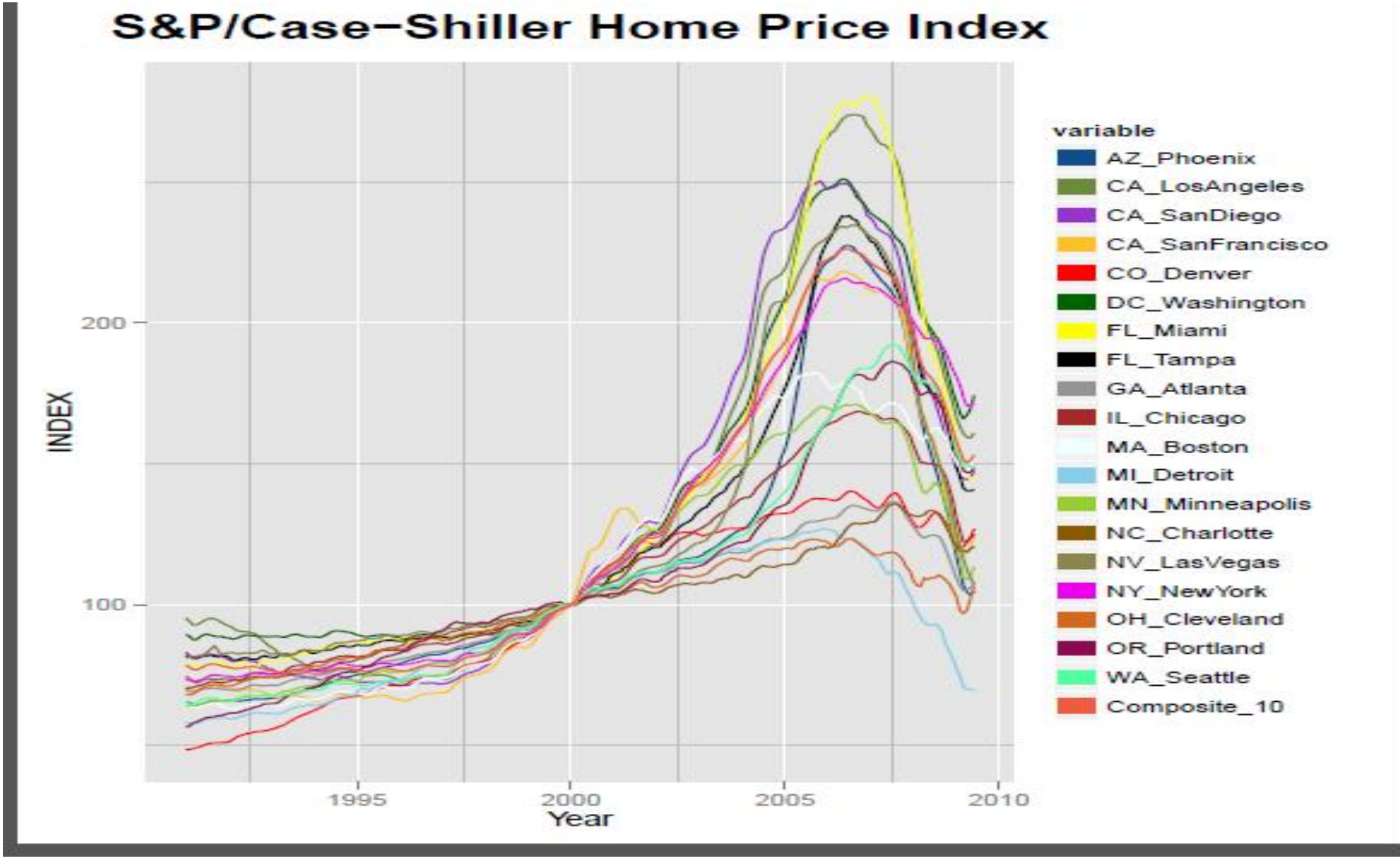
```
6  install.packages("XML")
7  require(XML)
8
9  sp.url<- "http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_091609_SnP_HomePriceIndex"
10 sp.extractedData<- readHTMLTable(sp.url, which =1, header= TRUE, stringAsFactors= FALSE)
```

Data Munging

- The imported data is not in the desired format
- Transformation of variable is performed to
  - Delete unwanted bad characters from data
  - Rename variable names
  - Transform data types
  - Create new variables
    - Dummy (Seasonality), Date etc.
  - Delete undesirable data
  - Difference the series
  - Create csv

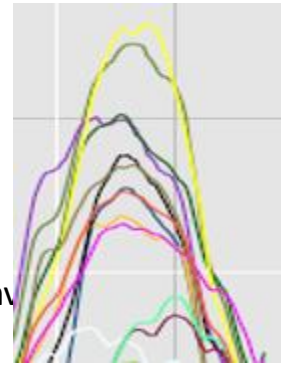
```
#renaming the column, unwanted characters
colnames(sp.extractedData)[23]<- "Composite-10"

#function to get rid off leading and trailing spaces
trim <- function (x) gsub("^\\s+|\\s+$", "", x)
sp.oldName<- names(sp.extractedData)
sp.newName0<- trim(sp.oldName)

#function to replace hyphens from variable names with _
repHyp <- function (x) gsub("\\-", "_", x)
sp.newName<- repHyp(sp.newName0)

#replacing dataset with new variable names
colnames(sp.extractedData) <- sp.newName
names(sp.extractedData)

#Function to convert character value to numeric
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}

sp.ds <- data.frame(lapply(sp.extractedData, as.numeric.factor), stringsAsFactors=FALSE)
any(is.na(sp.ds))

#converting numeric to factors (categorical variables)
sp.ds$Month<- as.factor(sp.ds$Month)
sp.ds$Year<- as.factor(sp.ds$Year)

#getting the values from older dataset
sp.ds$Month<- sp.extractedData$Month

#saving the dataframe in R format
save(sp.ds, file="sp.fullDataset")
# saving the dataset in csv format
write.table(sp.ds, file="fullDataset.csv", sep=",", row.names= FALSE)

#check for missing values
any(is.na(sp.ds))

######### Here ends the data cleaning and validation ###########
```

Exploratory Data Analysis



Overall Plot- Quick Insights

- The graph depicts that the house prices for all the region have been consistently increasing till a certain point.
- All the regions experiences a significant change (increase) in the Index beginning from year 2003, attend a peak in 2006, and then tails off very quickly to a new in a very short period till year 2009
- Most of the region have resurging indices at end of the series.
- Cities of states like Florida, IL, NY and California have →

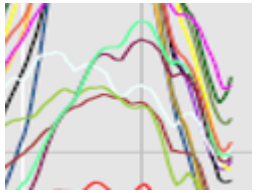


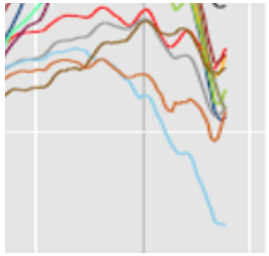
  experienced biggest jumps and bumps.
- The lower region of graph hasn't experienced significant peaks like the one discussed

above. The index change is not very drastic. These regions are Detroit, Cleveland, Atlanta, Denv

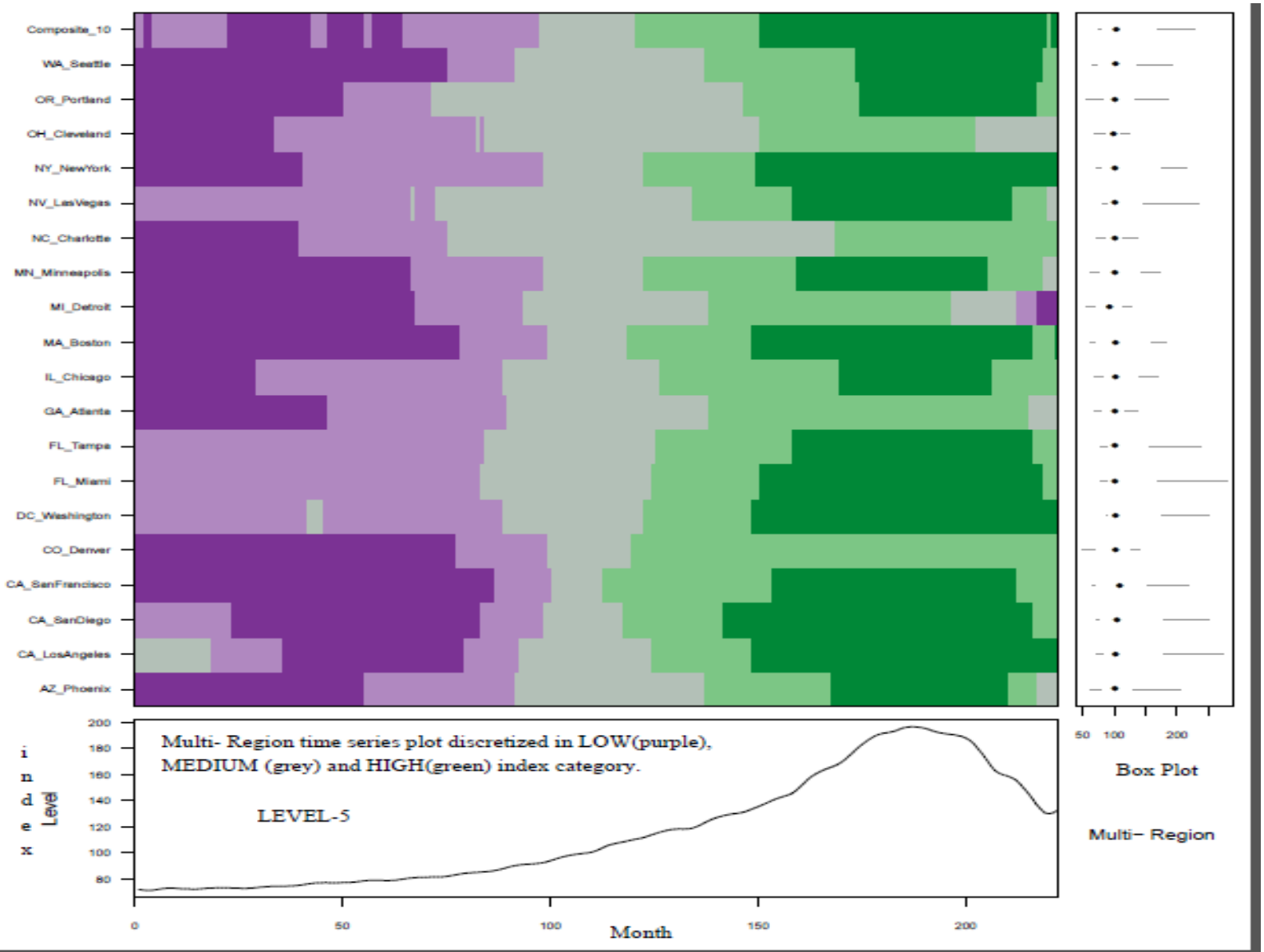
- The middle region of graph also shows some peaks and

  they also tailed of quickly, includes Seattle, Minneapolis, Portland and Boston





Index Change



5- Level Index Change

- The multi-level time series graph on the previous slide explains the index difference in terms of levels from low (dark purple) to high (green). It bolsters the line plot discussed earlier.
- Here on the x-axis, we have Month numbers for years starting from 1991
- The box plot on its right hand panel explains the non constant variance in all the cities
- The line plot on its bottom depicts the average index for all the region and again we saw a similar trend as noted earlier.
- It is clear from the graph that how indices have been changing over the time.

FORECASTING

- Forecasting is required to predict the indices of next 18 months.
- The time series data is Stochastic in nature so differencing, Moving Average and AutorRegression models are implemented for forecasting.
- Am I working on correct dataset?
    - Do I have stationary series? Tests Performed:
        - Standard Mean and Constant Variance Test
        - Dickey Fuller Tests
        - KPSS Test
        - Shapiro Test
- ACF-PACF plots do not tail off quickly for any time series. Click Here .
- Create stationary time series
    - Transformation of existing data
        - Differencing (1^{st} and 2^{nd} Order)

Timer Series Stationarity

- Differencing (Stationarizing)(1st and 2nd order)
  - ACF and PACF plot tails off and dies at regular lags. Click here for plots
  - Reduced the mean and brought some constancy in variance
  - Click here for 1st Order and here for 2nd order Differencing Decomposition Plots
- Why Data Cut
  - Plot here shows that there is significant change in the price index from end of the year 2002 and 2004
  - Trend and remainder of the Decomposition plots also suggest that the variation begins near 2003.
- No seasonality component has been observed in the data
  - Decomposition Plot confirms this. Click here.
- Forecasting of next 18 months of price index has been done using the data from 2003 to 2009

Forecasting Models used in the Study

- Exponential Time Series using filtering techniques
  - Dropped because of poor performance, increasing complexity and time constraint.
- ARIMA (Autoregressive Integrated Moving Average)
  - auto arima
    - Automated the R Code (can be scaled)
    - Poor Forecasting performance
  - Manual Modeling by fitting parameters(p,d,q)
    - Optimized the model selection, on the basis of ACF-PACF cut-off and residual diagnostics
    - Best model achieved with lowest AIC value is used for prediction

Diagnostic Used in Modeling

- QQ Plot on residuals
- ACF-PACF on residuals
- Box-Pierce statistics
- Lowest AIC-BIC

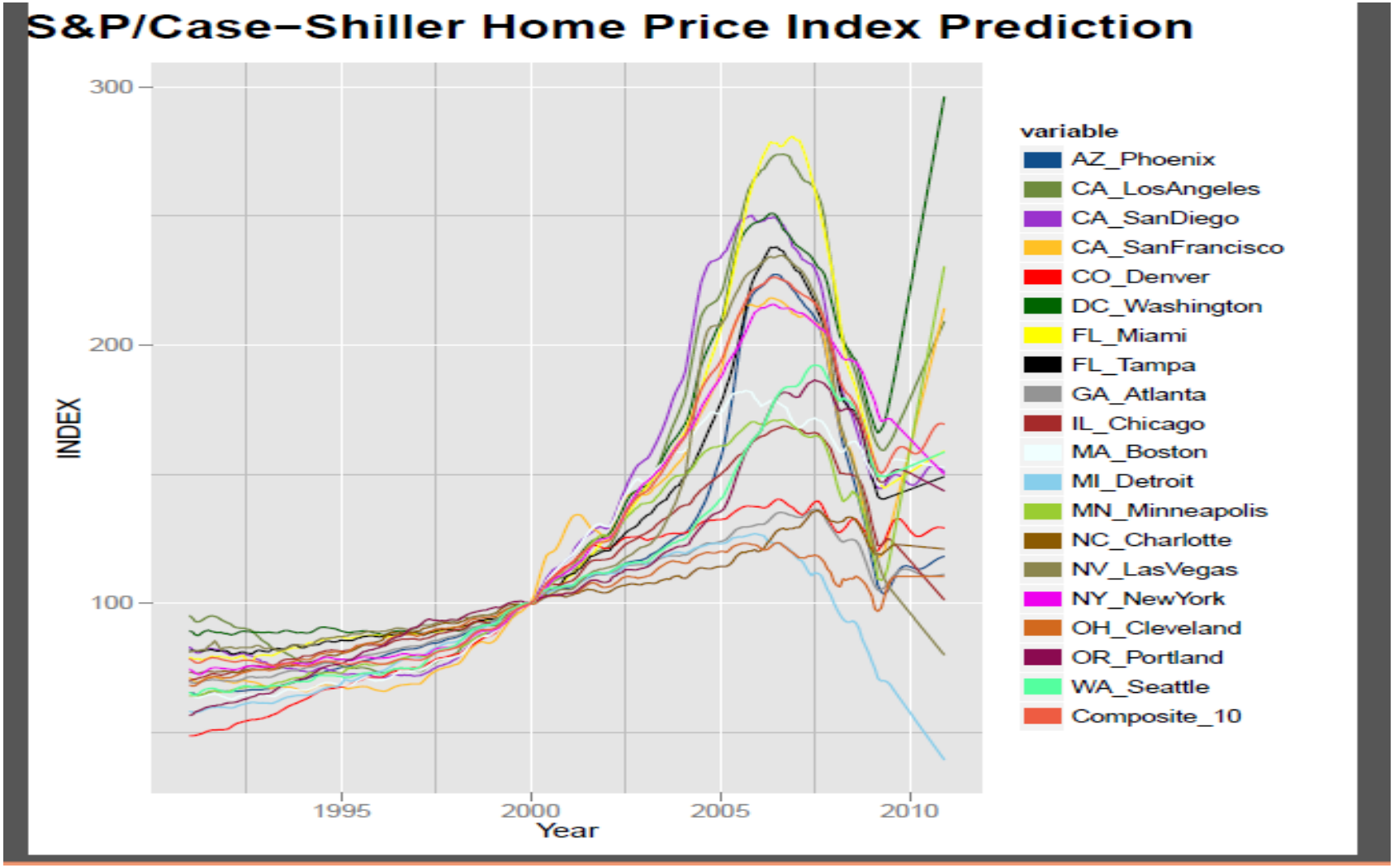
Code is available at :

https://github.com/jssasan/ARIMA-Time-Serie-Analysis

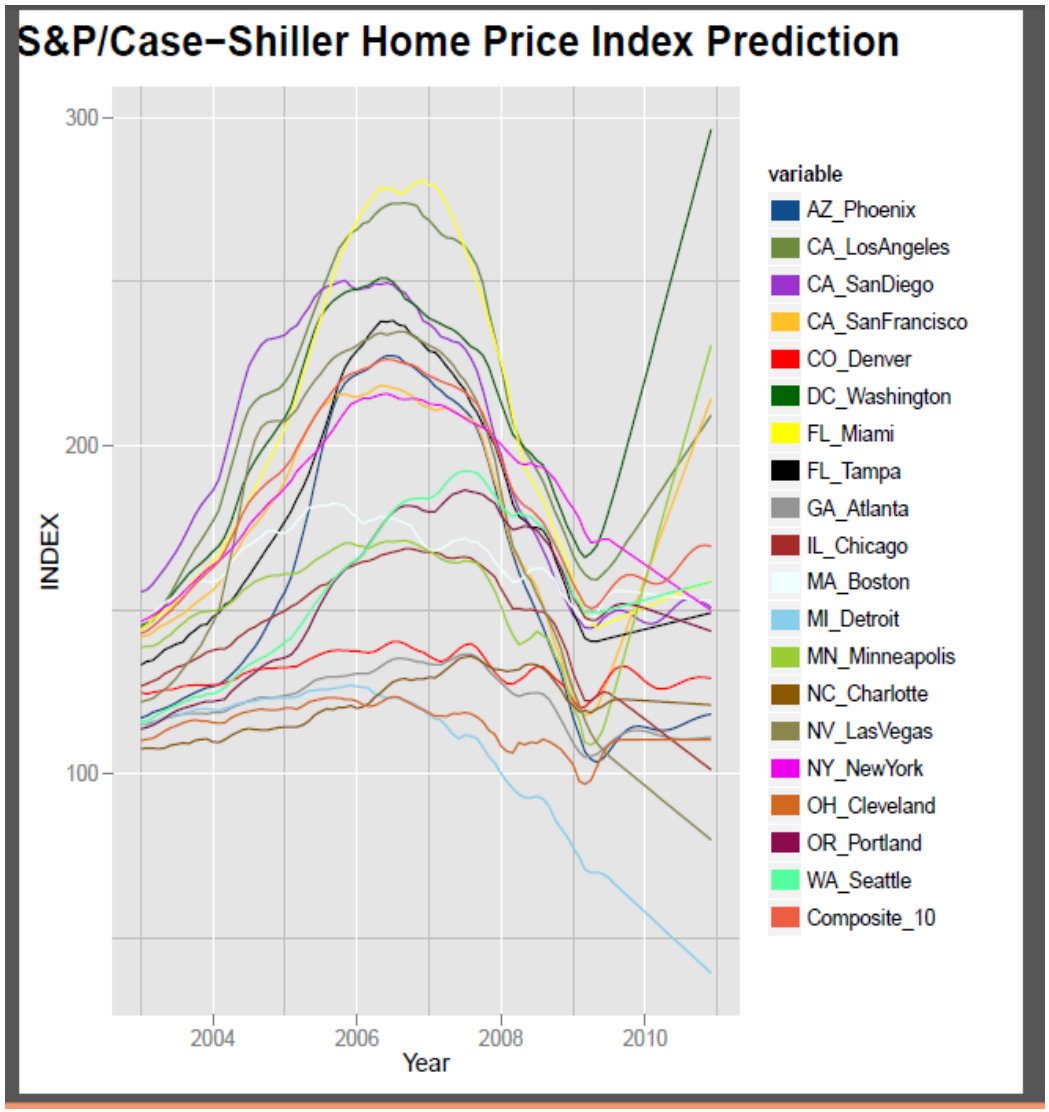
FORECASTING Jul 2009 – Dec 2010

Click here to see forecasting plots for all the regions.

Below is snapshot of all region predictions:

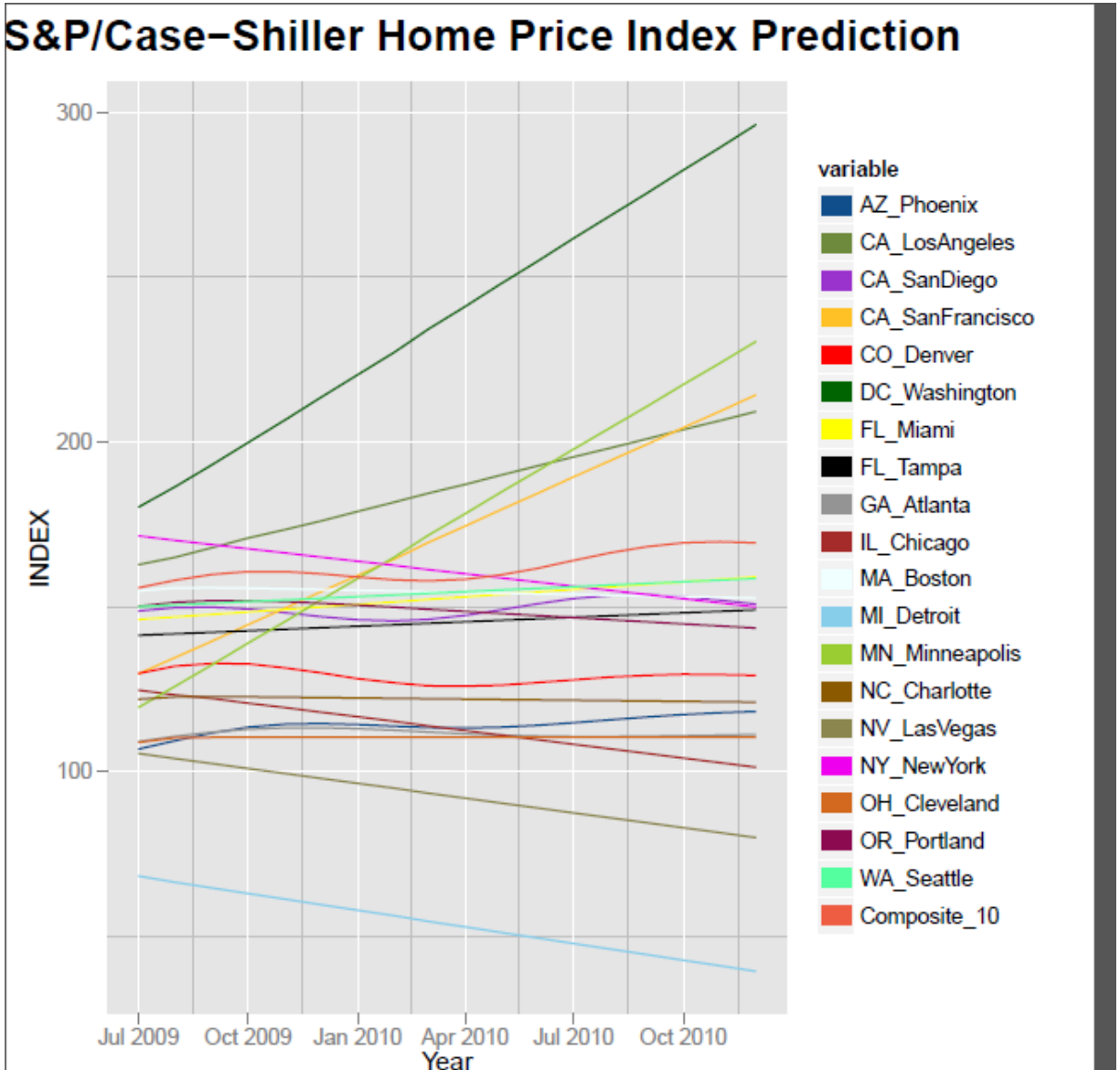


Predicted Index Analysis



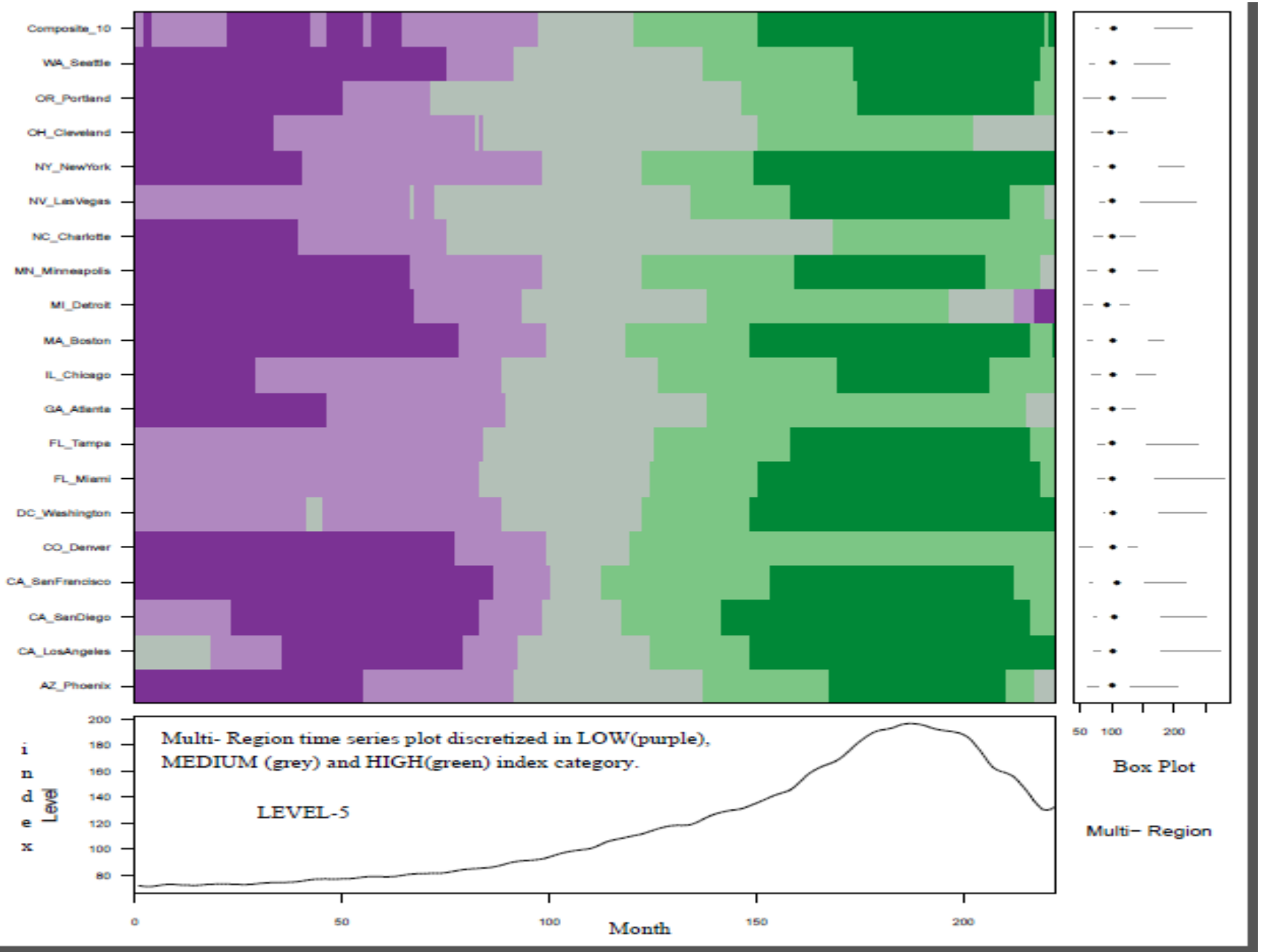
- For next 18 months, regions like LA, DC, Minneapolis, SF are expected to see a sharp rise in the index and hence house prices are expected to resurge
- Other regions do not see much variation in their index levels with few local peaks.
- Detroit and Vegas house prices are expected to fall constantly in coming year too.

Predicted Index from Jul 2009 to Dec 2010



Housing Bubble

Index Change



Housing Bubble contd..

- The economy has survived World-War II, so increasing house prices with time was never a grave concern.
- An unusual INCREASE in change has been observed in the Index data points in the first quarter of 2000 suggesting one of the possible start point of bubble building.
- 5-Level Plot on the previous slide also confirms that price indices, of all the cities, have started moving to another level (light green) around the month number120-130 i.e. year 2000 only.
- The dark green level indicates that most of the cities saw a major peak in the initial quarter of 2006
- The line plot suggests that the index started plummeting in the same year which was continued for rest of the period.
- The plots here gives clear forecasting picture of every city.

Buy a house or not?

- Buying a house depends on individual's financial situations, and faith in economy.
- A lot of people usually sell their old house prior buying a new one. Since the prices were already at all time low. There is no point in buying a new one.
- As an investor, it is lucrative to invest in the real estate but there is a forecasted danger of further decline in the indices, and hence prices.
- Subprime-ing also might not be a good idea since the index is forecasted to decline further for most of the metropolitan cities.

A Thought!!

- It would be interesting to analyze the House Price Index jointly with few other indices like CCI and S&P 500 in order to derive more generic insights
  - Consumer confidence index Data which includes several factors like family income, business condition, employment condition and so on.
  - S&P 500 Stock Index Data which contains a lot of information related to the growth of businesses and investments. It gives good indicators for both businesses and consumers.

Project Highlights

- Rapid Automation of ARIMA models and diagnostics in R.
- Rstudio's incapability to accommodate more than 1000 line on console. Therefore, script was divided in multiple parts which stimulated a more organized way of modeling in R.
- Stationarity of the time series should have been checked with multiple measures to avoid rework.
- Data Exploration and Model Diagnostics could have been automated to a certain extent at an early stage.
- Components of time series should have been checked at the very first place.

Further Improvements

- Prediction accuracy of fitted models can be enhanced.
- Few diagnostics depicted poor model performance which can be improved with the iterative process.
- Better time series models can be build for forecasting given more related data and time; and by tuning ARIMA model parameters (p,d,q).
- Unobserved Component Models (UCM) could be a good replacement of ARIMA, because UCM is
  - Time efficient (Automates greatly)
  - Higher Model Interpretability (great benefit)
  - Equivalent forecasting accuracy
- Running R within the Tableau could have generated great plots and provide better insights


THANK
YOU