STAT 515 Applied Statistics & Visualization for Analytics

Final Report

Rongxin Zheng

Xiaotian Wu

Nadiya Nijiati

**Foreign Trade Analysis**

In our project, we are focusing on foreign trade data to analyze various aspects of foreign trade for the past 15 years. Foreign trade is the official source for U.S. export and import statistics and responsible for issuing regulations governing the reporting of all export shipments from the United States (Census.gov). We are able to collect data of import, export, yearly transaction total, trading partner countries, balance of payment, each state import and export statistics. The export and import indicates the economic situation of a country. Getting to know the influence factors of export and import will help people to know the performance of a specific area. Based on the data and analysis, we will be able to evaluate information on export regulations, commodity classifications, or a host of other trade related topics.

Our data is available on U.S. Census website. U.S. Census serves the public as the leading source of quality data about the nation's people and economy. Based on Each individual topic, the data is downloaded in different formats then converted into Excel in order to process in R. Overall, data is clean and contains numbers of variables based on each topic.

Topic 1. Checking Function

At the beginning of project, we programmed a function, in order to check the data set automatically because the scales of data sets are too large to check by eyes. We preprocessed data mainly manually, so it has a big possibility to raise errors. Through this function, every time when we read the data, R Studio compares the state name with data frame we will use, and check whether all the meta-data is in a right form. As an output, it will write the check result on the console.

For example, in some computer, excel use thousand separators, which R cannot read. So that, as for this kind of data, R will read them as factor. For this case, the checking function will warn us with this problem.

Topic 2

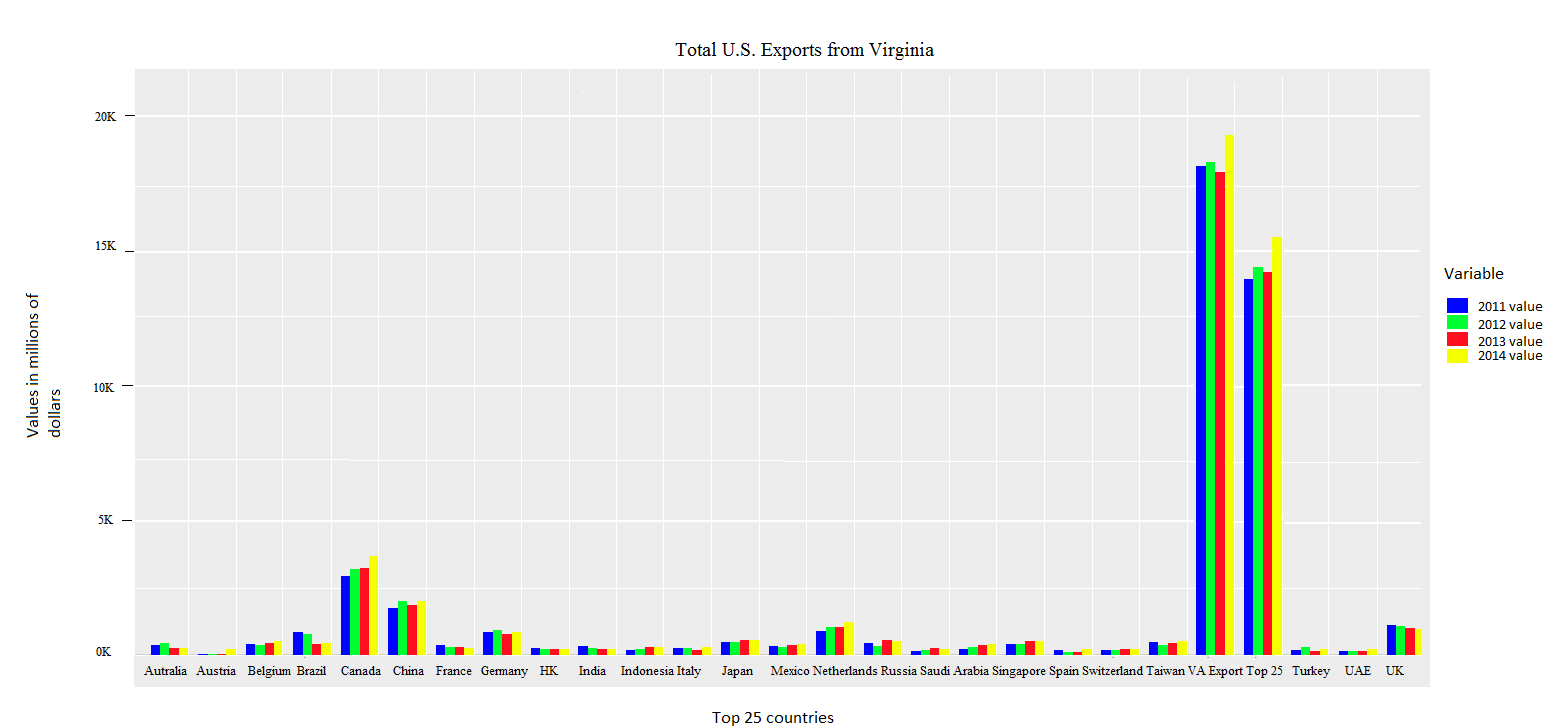
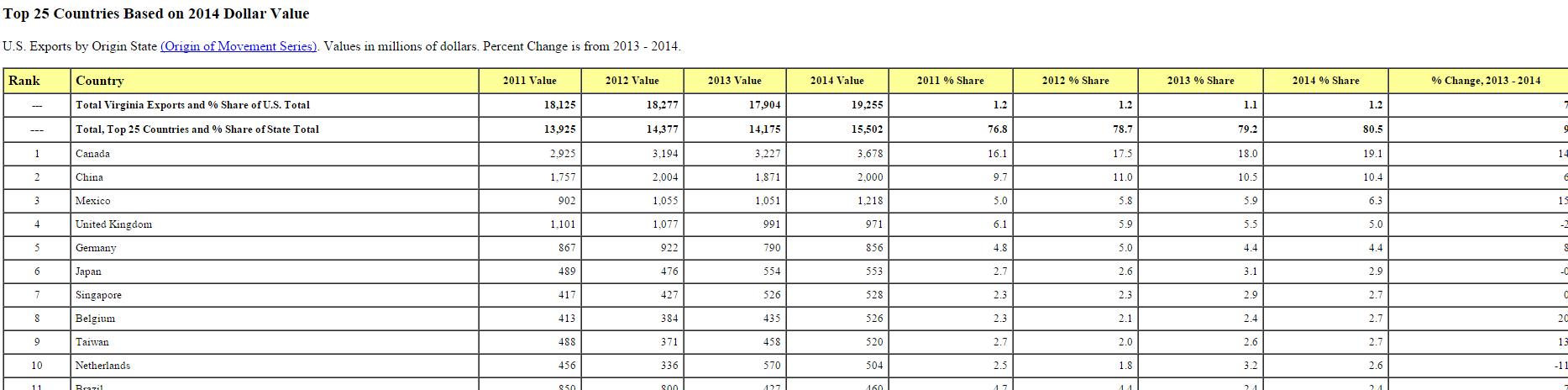
**Total U.S. Exports (Origin of Movement) from Virginia: Top 25 Countries Based on 2014 Dollar Value**

As Virginia is our home state, we are interested to find out how much of exports from total U.S. exports is coming from Virginia. In the other word, how much dollar value did Virginia produce from exports and what are the countries that Virginia provide the most exports.

The data is available under State Data on the U.S. Census website. Data set is fair clean with adequate information. We converted the information from webpage table into excel. We excluded the percentage of share and ranking of countries in our excel table because bar graph does not present correlation between more than two variables.

The bar graph we produced below simply shows exports amount in dollar value that is originated from Virginia to top 25 countries in between the year of 2011-2014. Variables are colored in blue, green, red and yellow to present year 2011, 2012, 2013 and 2014. Countries receiving exports from Virginia are represented on x-axis. Amount of exports in dollar value are represented on y-axis.

As we can visualize from the graph, from year 2011-2014, top three countries received the most exports from Virginia are Canada, China and Netherlands with the highest dollar value. Overall, exports from Virginia has been increasing from 2011. In 2014, total Virginia exports reached 19,255 millions of dollars, and 15,502 millions of dollars from the total value is coming from these top 25 countries.



Topic 3

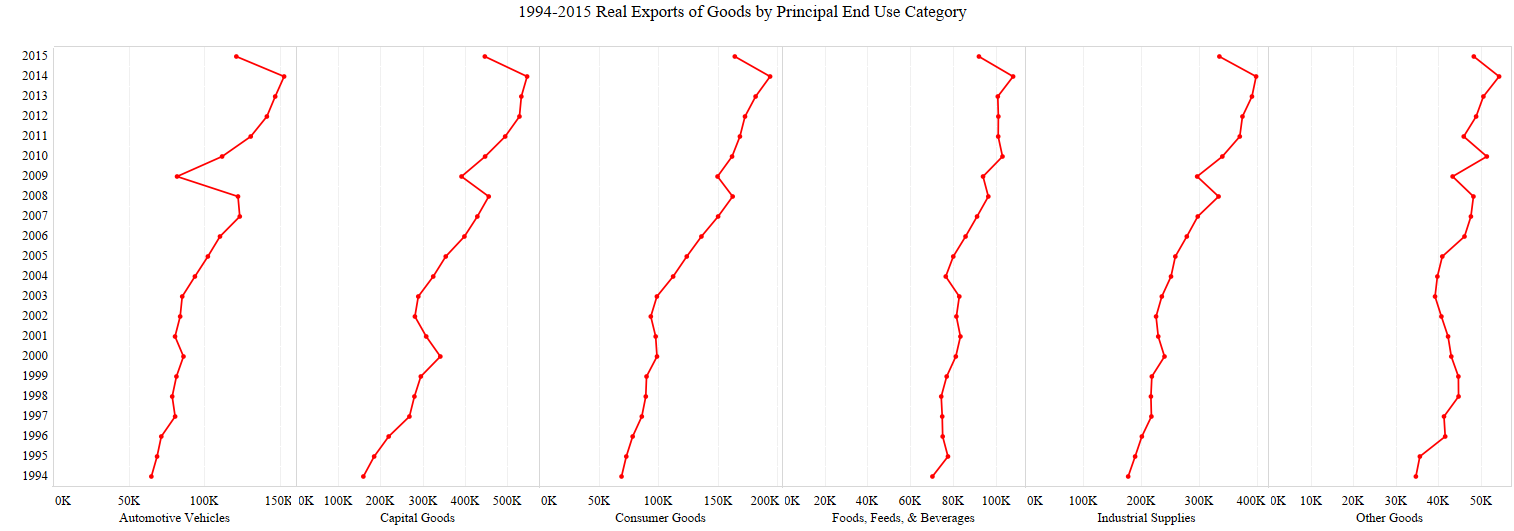
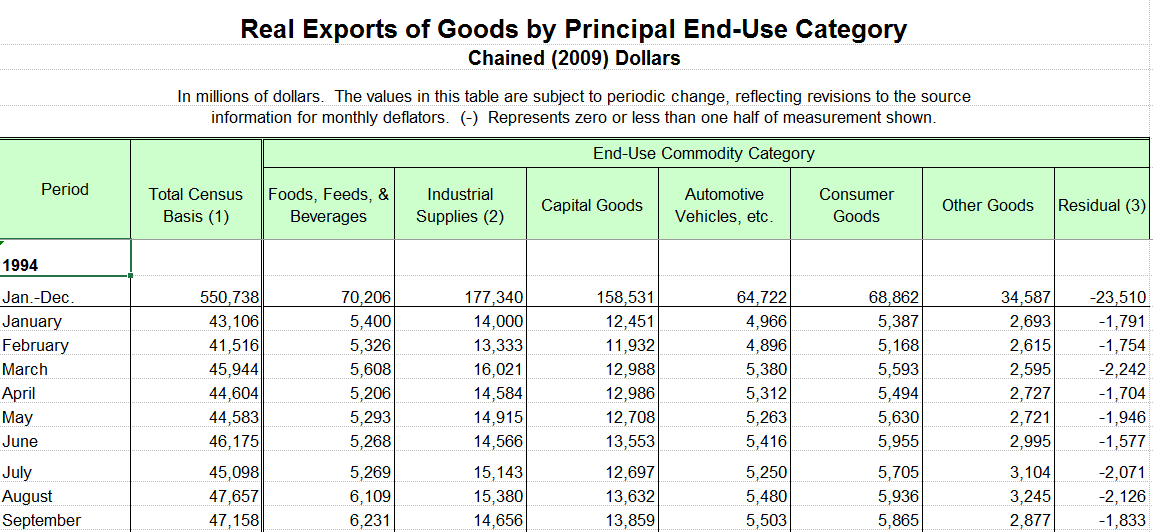
**Real Exports of Goods by Principal End-Use Category from 1994-2015**

From all U.S. exports, every raw material has its own use. In this topic, we are analyzing real exports of good by principal end-use in categories from 1994-2015.

Data for this topic is available under historical series on U.S. Census website. Data is stored in excel format and the data is fairy clean. Data is grouped into 9 variables: year, total amount of goods by principal end-use categories, food beverages, industrial supplies, capital goods, automotive vehicles, etc.; consumer goods, other goods and residual.

We applied time series model to better represent the table. We excluded total census basis and residual good in our graph. In our time series graph, we are focusing on represent the changes of amount of exports in different principal end-use categories from year 1994 to 2015.

As we can see from the graph, exports of goods by principal end-use categorized in, food beverages, industrial supplies, capital goods, automotive vehicles, consumer goods, and other goods have been increasing in amount. Exports for these 6 end-use categories has decreased in 2015.



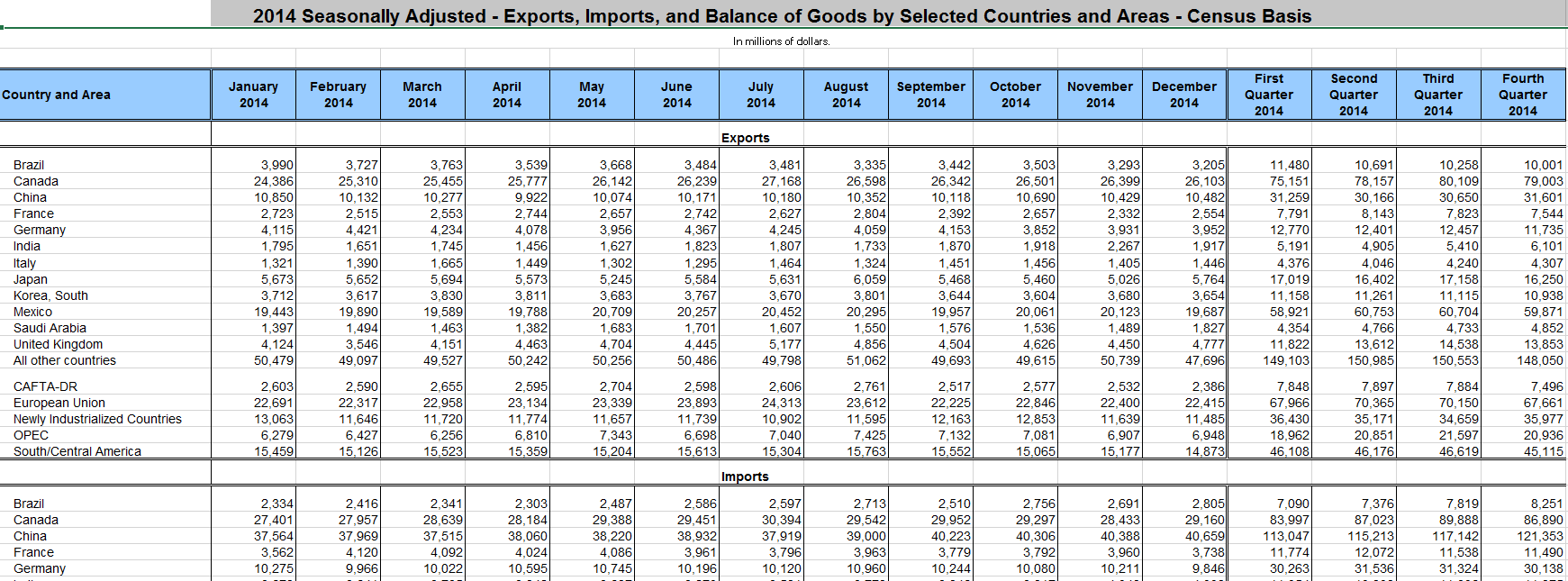
Topic 4

**2014 Seasonally Adjusted - Exports by Selected Countries and Areas**

In this topic, we are focusing on finding top countries that have the highest value of exports from 2014.

## Data is downloaded in excel format from U.S. Census website under Country and Product Trade Data. The data table includes both monthly exports and imports value for 2014. In the analysis, we are focusing on exports only.

## In the plot, we are using countries name for y-axis and amount of exports in billions of dollars for x-axis. The last four dots in the graph represent the amount of exports of four quarters in 2014. Countries are ranked in the order of who has the highest value of exports in 2014. On the other hand, it is obvious that the value of exports has been increasing from January for each other in 2014. Canada, European Union, and Mexico have the highest value of exports in 2014.



Topic 5

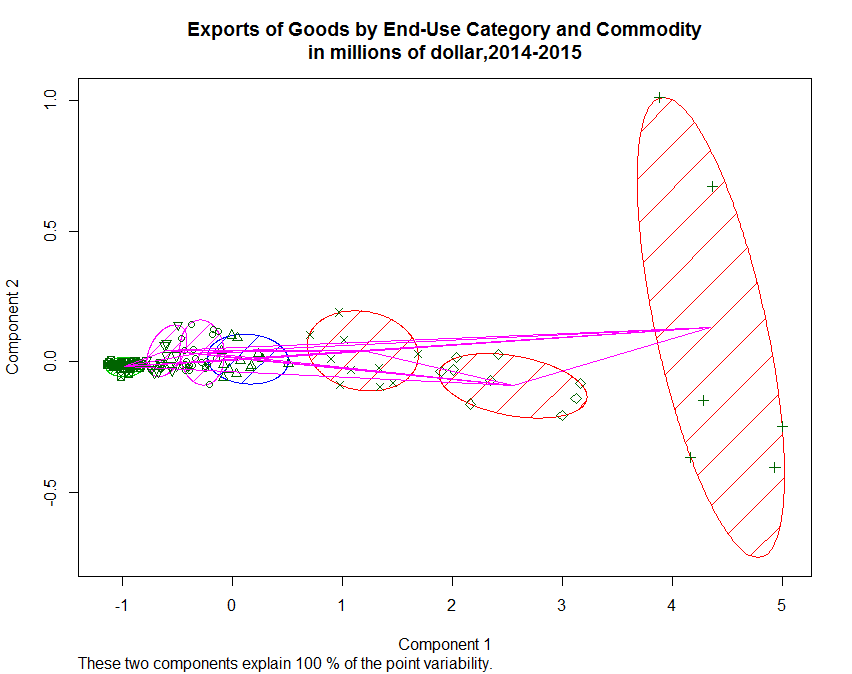
**Trade in Goods by End-Use Category and Commodity**

U.S Trade in Goods by End-Use Category and Commodity shows all kinds of good and the export value of two years. The question we have for this data is that how many groups of commodities are in the general based on exports?

We downloaded the table in excel format and read table in R. We clean the table by deleting some item name. After scaling the data, we applied clustering analysis to generate groups. The observations are represented by points in the plot.

We also applied principal components scaling. We assumed there are 6 clusters, but some of the data overlapped each other and hard to read. Then we changed to 7 clusters, the graph is much easier to read.

As a result, that there are 7 groups of commodities in export. More than half of the items have lower volume for international trade. There are 5 goods to be considered as high value of trade.



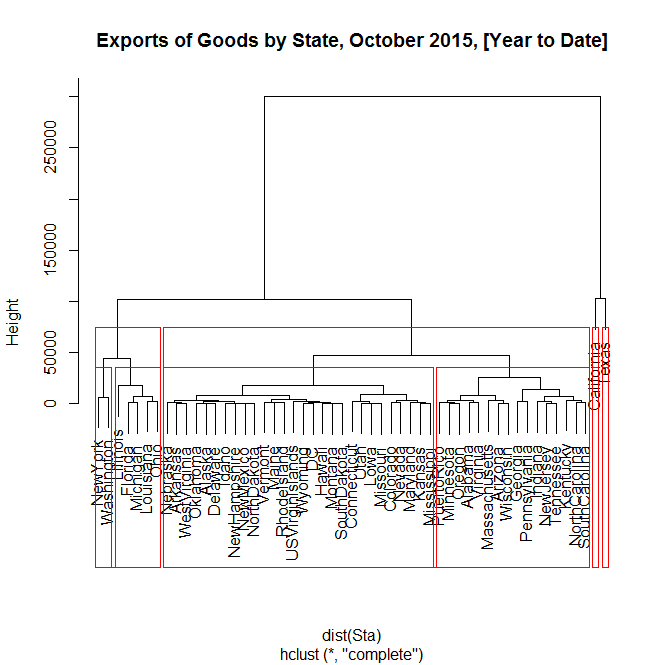
Topic 6

**Exports of Goods by State 2015**

In Exports of Goods by State 2015, we can read the value of export and import based on each state level. By analyzing the data, we are trying to find out which state has the highest amount of export.

To solve this questions, we applied ward hierarchical clustering to generate cluster tree. At first, we loaded and read the columns “state” and “export of October, 2015” into R. We are applying tree diagram to cluster by using Euclidean distance as the distance metric. Since R can’t read the category variable, we needed to set up the labels otherwise it will only show the order number. To better present the groups, we used rectangle to highlight the clustering.

We created four clusters. The groups are divided according to the height which is the value of export. As the graph shows, California and Texas have the highest value of export. In addition, the difference between other states and these two states is very obvious. To better cut these states into more groups, we applied lower value of 100000 to cut the tree. As a result, there are 6 clusters. However, by looking at the states in each cluster, we can find that the value of export is not based on geographic locations and it is related to the specific economic situations of each state.



Topic 7

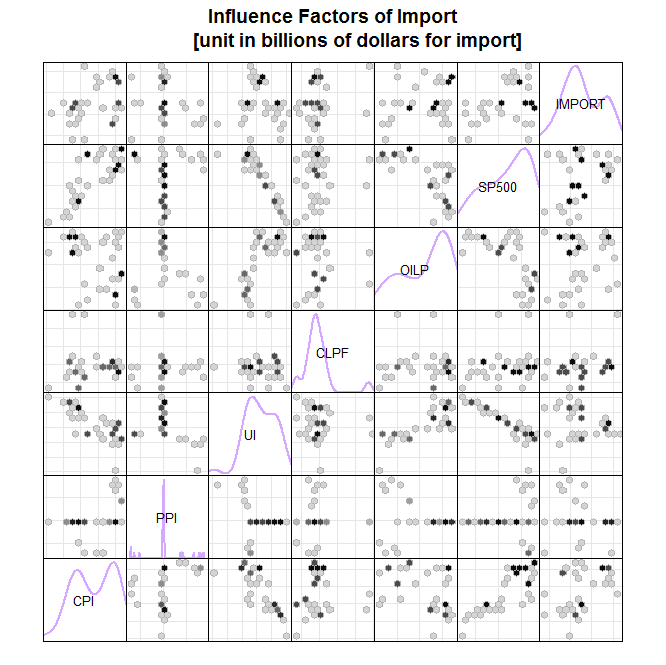
**Influence Factor of Imports**

Clustering can provide general information about the relationship between variables. We also need regression to better predict the relationship between variables. In this topic, we are analyzing the factors that will influence overall imports of U.S.

We picked the following indicators of economic performance to show what will affect export: index (CPI) whichmeasures changes in the price level of a market basket of consumer goods and services purchased by households; unemployment rate (UI), inflation rate (RI); producer price index(PPI) whichmeasures the average change over time in the selling prices received by domestic producers for their output; SP500 which many consider it one of the best representations of the U.S. stock market, and clothes domestic retail price(CLPF) and oil price.

Data is sorted monthly from 2013 to November, 2015. Variables include many areas, we collected data from various website. They are Bureau of Labor Statistics, Economic Research of Federal Reserve Bank of ST LOUIS, U.S Energy Information Administration. We used monthly data for exports. The original data for export is counted in millions of dollars. The x scale can’t accept the huge range of data. Therefore, we changed the unit to billions of dollars to adjust data.

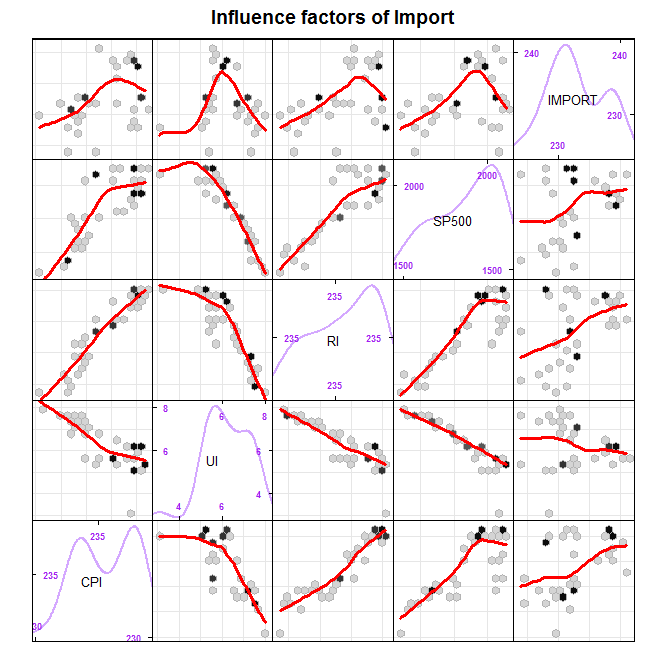
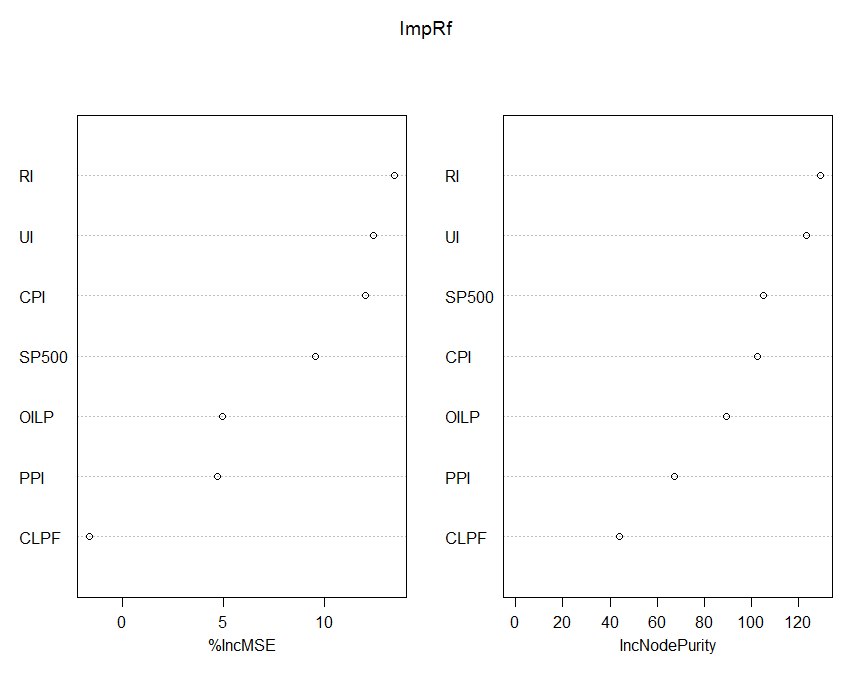
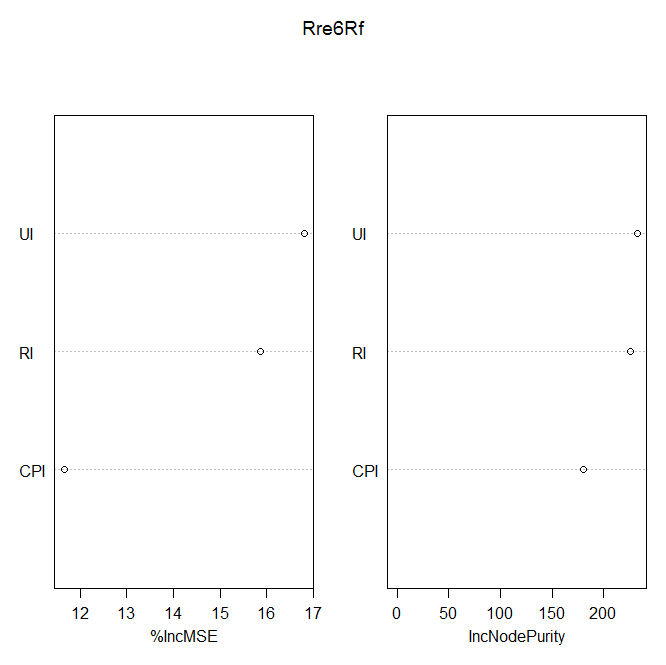
Applying random forest, we first generated a splom to show the linear regression between import and other variables. R considers some of the variables as factors, we set them to numeric variables and set the data to the same level.



From the graph, we can observe some of the variables are highly related. However, the summary shows that random forest model only explains 45.33% of the variance. The number is relatively low. One of the problem we considered is the data set is not big enough to generate model. The other reason might be that the factors are just not relevant.

We decided to keep the import variables, so we used random forest classification to generate varImport. From the graph, we can see CPI, UI, RI and SP500 have higher MSE which means they have higher variable importance. To produce a new splom with CPI, UI, RI and SP500, we can see the variance increased to 48.36%.

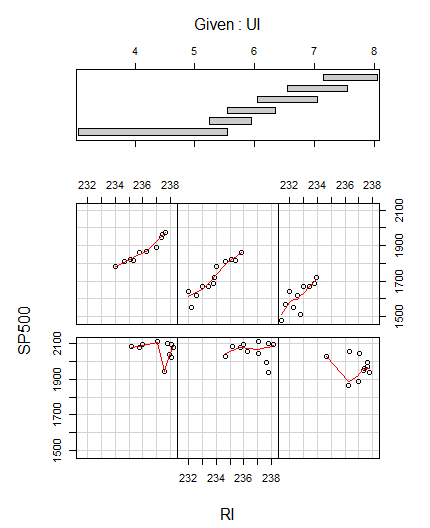
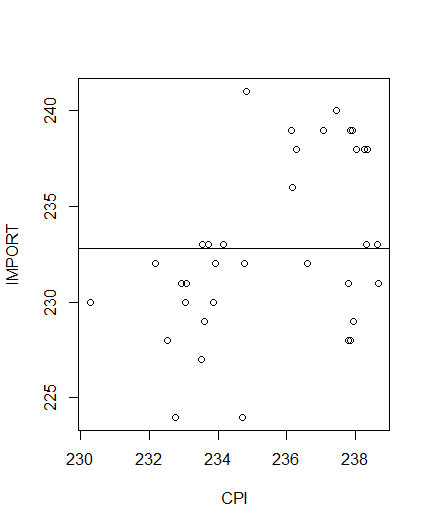
For the slight increase, we concluded that the model is not reliable. For the variable important plot, we choose to look more about the %incMSE. The reason is that Node purity is measured by Gini Index which is the difference between RSS before and after the split on that variable.



We observed an unexpected result that UI, RI and UI seems to have a close relationship with SP500. Although, the data set is not big enough, we can see a clear trend for these relationship. Also, the variable importance increased a lot. As the summary told, mean of squared residuals is 3559.339 and variance explained is 90.25%. The result shows that the performance of stock market has a positive relationship with CPI and RI. It indicates that higher inflation rate can also has a good performance for stock market. The negative relationship with unemployment rate shows lower the unemployment rate better a stock market, for unemployment rate indicates the performance of company operation.

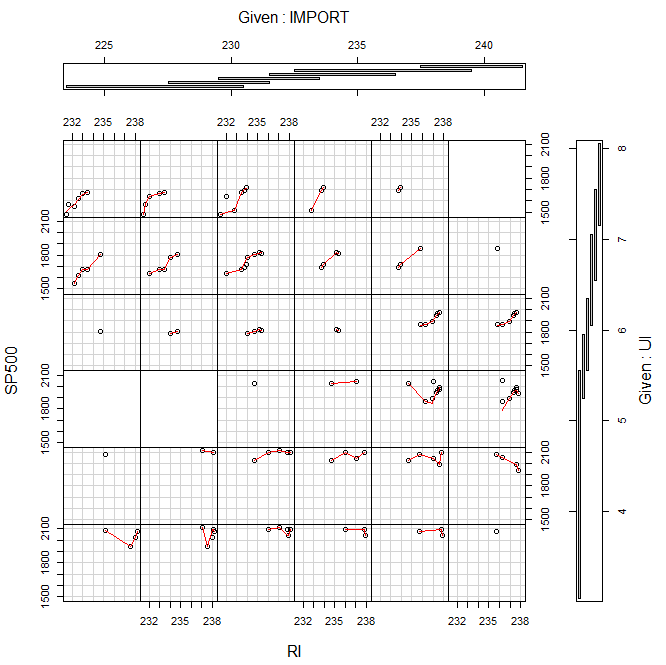
We also applied multiple regression for analyzing import. However, after we made several single linear regressions, I found the relationship is still weak. Then we did the multiple regression about SP500 as just mentioned above.

We used coplot to generate multiple regression. Coplot shows the effect of a third variable in a step-by-step fashion on scatterplots of two variables together. According to the basic knowledge of multiple regression using SAS, we want to find which independent variable weighted more than other variables. And we have some methods to drop less important variables. Unlike that, coplot give people a chance to show the relationship when variable change from low to high.



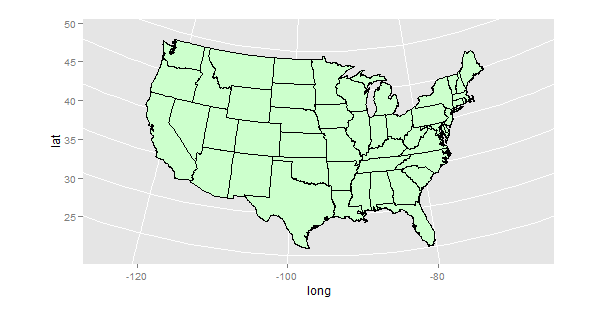
The p-value of UI and RI are both below 0.0001. They are significantly important. The bottom and middle panels correspond to each bar in the top panel. For the regression graphs, the order to read is from left to right and from bottom to top. It shows as inflation rate increase, the change of stock market. The result shows SP500 are high when inflation rate increase. Adding the variable of unemployment rate, there is more variability in level of UI and SP500 when unemployment rate is lower. That is when company’s performance is very good, it is not only caused by inflation rate. When the unemployment rate is higher, we can find the data are more gathered together. Also the stock performance are worse than before. We can see a clear positive relationship between SP500 and inflation rate. We assume that when the general economic is worse, unemployment rate is high because companies laid off more employee. At the same time, the inflation rate is high so that people feel the burden of life is heavier than before. We added import to the regression function. We can see the data is less than before. We can’t have a strong conclusion about their relationship. Expect for the import, p-value of RI and UI are still below 0.05.

We can conclude that when the import is lower, the level of unemployment rate, inflation rate and SP500 have a high level of variability. Considering unemployment rate is low, inflation rate is high and SP500 is also high but the number is more steadily.

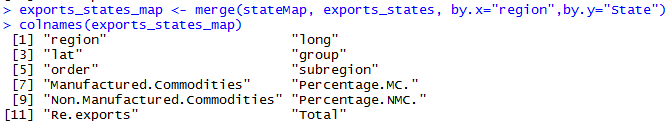
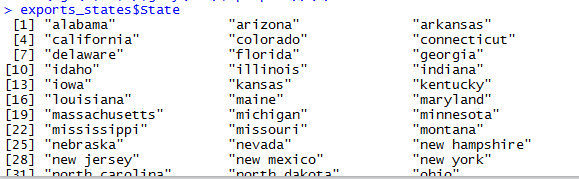


Topic 8. Albers equal area conic projection

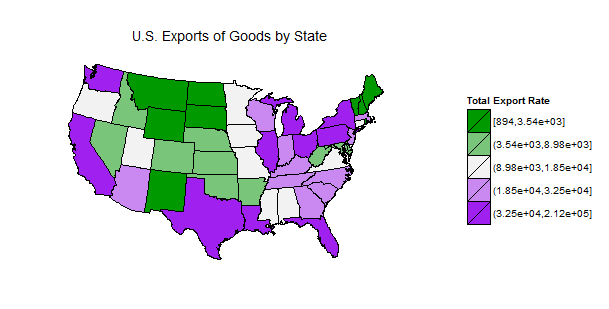
* 1. Get map coordinates for the conterminous U.S. states and make an empty map using an Albers projection.



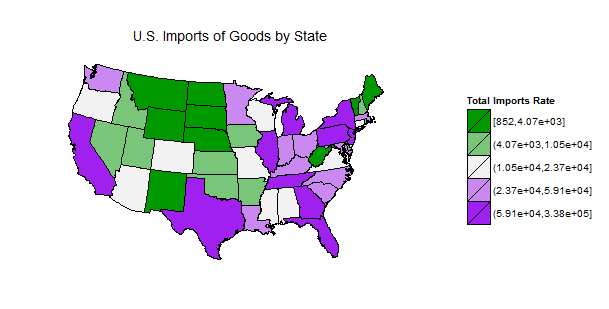
* 1. In the data file, row names are state names that start with a capital letter, so we converted the state names to lower case and add this in a new column, in order to match the region column of stateMap boundary file data. Then we could merge our own data frame with boundary file using lower case column.



* 1. In order to color all the states in the map with different colors, we used the quantile function to obtain quintile classes, and sorted the states into 5 classes with roughly the 20% states falling.



* 1. In the same way, we also drew choropleth map, using the data of U.S. imports of goods by States.



* 1. Transition

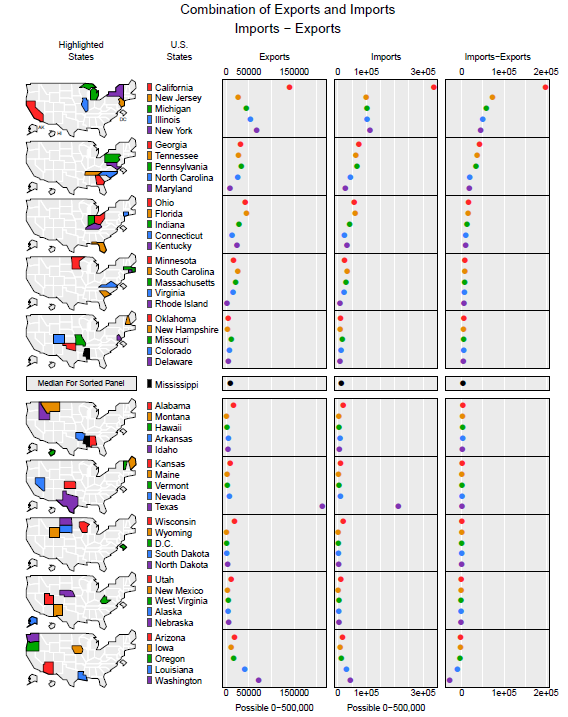
All we did above are about single map on the data frame of exports and imports. We cannot see the difference and connection through only these two plots. Somehow, assume that one rich place is supposed to have a much bigger value in imports than exports because people are so rich that they can consume more goods than people do who are living in poverty. Therefore, this time, we put the data of exports and imports together and sort them to see which of states seem relatively rich in this project.

Topic 9 MicromapST

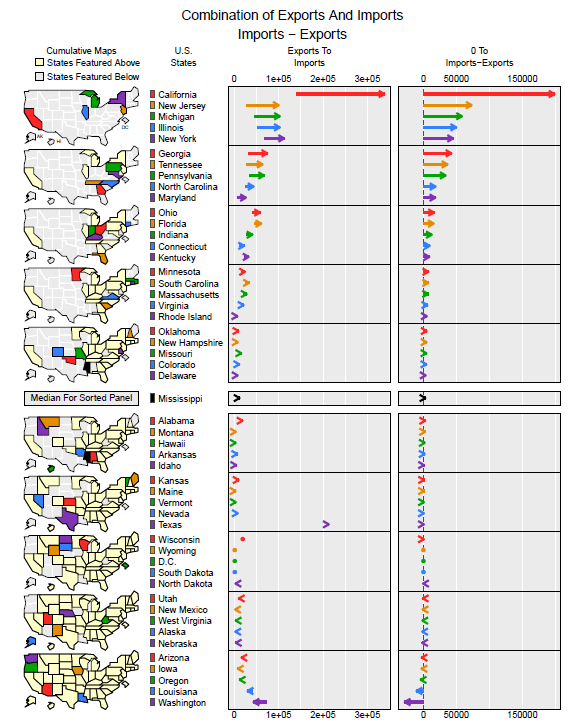
* 1. Re-add the dropped data, and make a new data set

As I said above, in the package of micromapST, the states of Alaska, Hawaii and D.C. are included, so I re-added them into the data frame. Moreover, I merged the data sets of “exports” and “imports” through “State” column, and added a new column to calculate the difference between imports and exports.

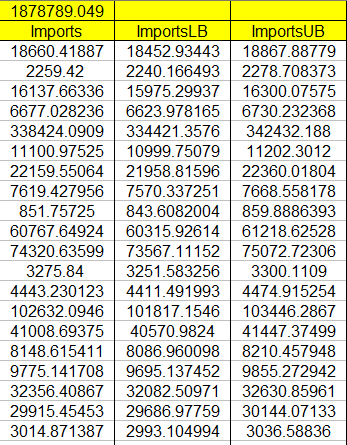
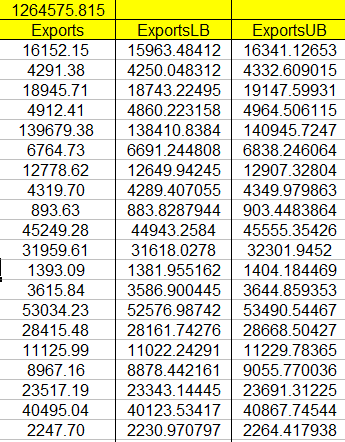
* 1. Get a MicromapST dot plot



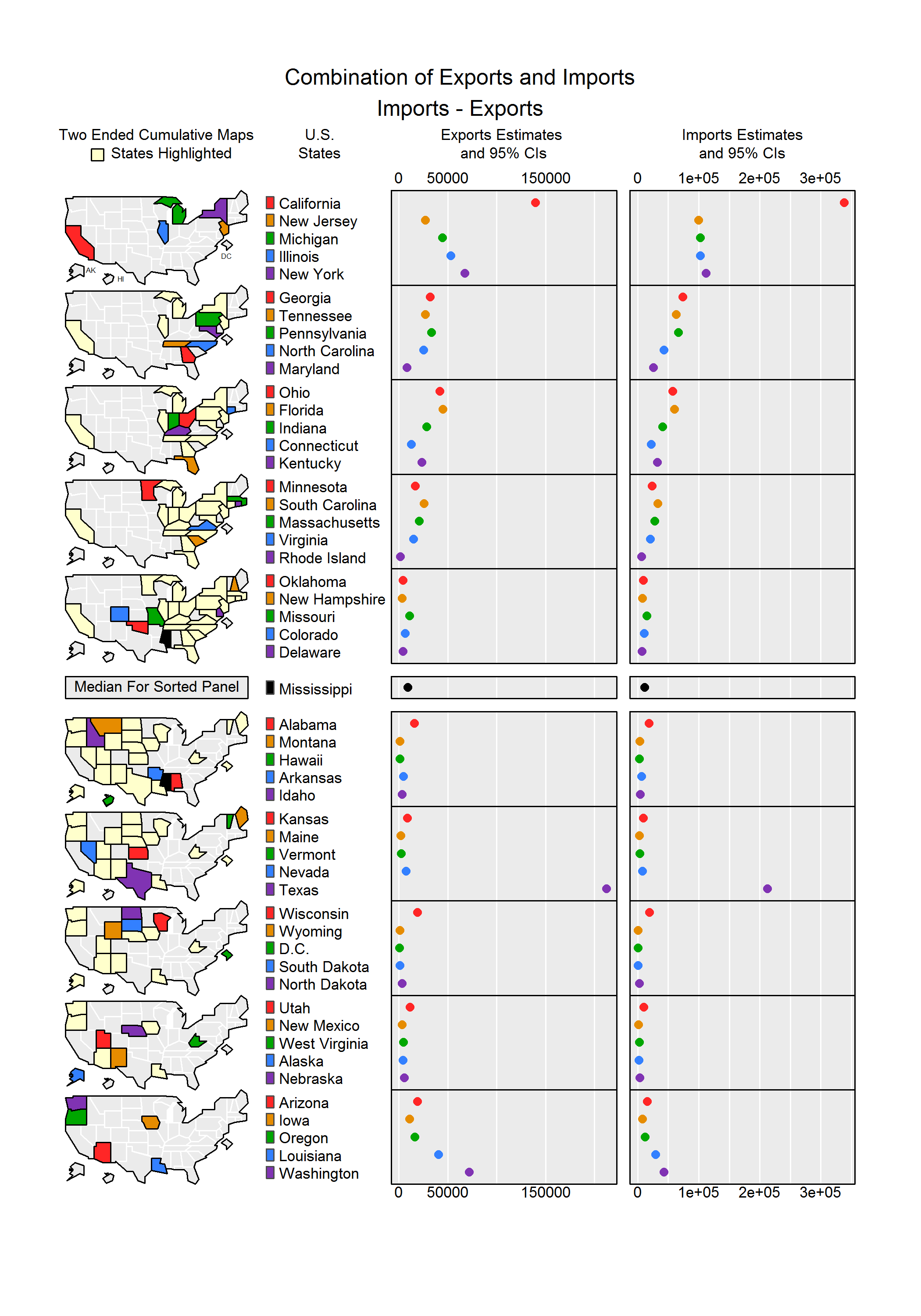
* 1. Cumulative maps, arrows, reference values and two variable transformations



* 1. Calculate the confidence interval



* 1. Maptail and dots with confidence Intervals



* **Improvements:**

1. There are many columns we preprocessed, but we didn’t use it to analyze. (Like percentages, manufactured and non-manufactured goods.)
2. In the Albers equal area conic projection, we can add the states of Alaska, Hawaii, Puerto Rico and US Virginia Island into Map\_States.
3. We can change the scale and limitation of axis, and set the breaks in each axis.
4. If the engineer add more boundary data into data frame, then we could set the filling gradient. Because the data set is not accurate enough, and pixel is not enough either, I tried once before, but the result is blur.

5. We can add more boundaries of other nations into data frame, and then we could also analyze other countries in that way.

Appendix

Multiple Regression

coplot(SP500~RI|IMPORT\*UI,panel=panel.smooth,data=Rre)

mean(Rre$SP500)

mean.SP500 = mean(Rre$SP500)

abline(h=mean.SP500)

model2 <- coplot(SP500~RI|UI,panel=panel.smooth,data=Rre)

model2=lm(SP500~RI+UI,data=Rre)

abline(model,col="red")

model3 <- coplot(SP500~RI|IMPORT\*UI,panel=panel.smooth,data=Rre)

model3=lm(SP500~IMPORT\*UI\*RI,data=Rre)

dendrogram

hc <- hclust(dist(Sta), "ward.D")

hc = hclust(dist(Sta))

plot(hc,labels=Sta$State,main="Exports of Goods by State, October 2015, [Year to Date]")

rect.hclust(hc,h=100000)

clusplot

fit <- kmeans(Exp,7)

clusplot(Exp, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=2)

clusplot(Exp, fit$cluster, color=TRUE, shade=TRUE,

labels=0, lines=1,main="Exports of Goods by End-Use Category and Commodity,in millions of dollar")

# 1. U.S. Exports of Goods by State

# 1.0. Setup

setwd("F:/乔治梅森大学/2015 fall/STAT-515/Final\_Presentation/final project/Linked MicromapSt(1. 画单图)");

source("theme\_clean.r");

library(ggplot2);

library(maps);

# 1.1. Get map coordinates for the conterminous U.S. states and make a map using

# an Albers projection

stateMap <- map\_data("state")

colnames(stateMap)

head(stateMap)

is.data.frame(stateMap) # TRUE

ggplot(stateMap,aes(x=long,y=lat,group=group))+

geom\_polygon(fill=rgb(.8,1,.8),

color="black")+coord\_map("albers",29.5,45.5)

# 1.2. A export data set for U.S. states

exports\_states <- read.table("clipboard",sep='\t',header=T);exports\_states$Total

rownames(exports\_states) <- exports\_states$State

exports\_states$State <- tolower(rownames(exports\_states))

exports\_states

exports\_states\_map <- merge(stateMap, exports\_states, by.x="region",by.y="State")

head(exports\_states\_map)

colnames(exports\_states\_map)

library(plyr)

exports\_states\_map <- arrange(exports\_states\_map, group, order)

head(exports\_states\_map)

# 1.3. Defining export classes

head(exports\_states)

TotalExportQuints<- quantile(exports\_states$Total,seq(0,1,by=.2))

exports\_states$TotalClass = cut(exports\_states$Total,

TotalExportQuints,include.lowest=TRUE)

ExportsColors <- colorRampPalette(

c(rgb(0,.6,0),gray(.95),'purple'))(5)

# 1.4. Finally we produce the plot and clean up the background

library(grid);

ggplot(exports\_states,aes(map\_id = State,fill=TotalClass))+

geom\_map(map=exports\_states\_map, color=gray(0),size=.3)+

scale\_fill\_manual(values=ExportsColors)+

expand\_limits(x=stateMap$long,y=stateMap$lat)+

coord\_map("albers",29.5,45.5)+

labs(fill="Total Export Rate",title="U.S. Exports of Goods by State")+

theme\_clean()

# 2. U.S. Imports of Goods by State

# 2.0. Setup

setwd("F:/乔治梅森大学/2015 fall/STAT-515/Final\_Presentation/final project/Linked MicromapSt(1. 画单图)");

source("theme\_clean.r");

library(ggplot2);

library(maps);

# 2.1. Get map coordinates for the conterminous U.S. states and make a map using

# an Albers projecton

stateMap <- map\_data("state")

colnames(stateMap)

head(stateMap)

is.data.frame(stateMap) # TRUE

ggplot(stateMap,aes(x=long,y=lat,group=group))+

geom\_polygon(fill=rgb(.8,1,.8),

color="black")+coord\_map("albers",29.5,45.5)

# 2.2. A Imports data set for U.S. states

imports\_states <- read.table("clipboard",sep='\t',header=T);imports\_states$Total

rownames(imports\_states) <- imports\_states$State

imports\_states$State <- tolower(rownames(imports\_states))

imports\_states

imports\_states\_map <- merge(stateMap, imports\_states, by.x="region",by.y="State")

head(imports\_states\_map)

library(plyr)

imports\_states\_map <- arrange(imports\_states\_map, group, order)

head(imports\_states\_map)

# 2.3. Defining import classes

head(imports\_states)

TotalImportQuints<- quantile(imports\_states$Total,seq(0,1,by=.2))

imports\_states$TotalClass = cut(imports\_states$Total,

TotalImportQuints,include.lowest=TRUE)

ImportsColors <- colorRampPalette(

c(rgb(0,.6,0),gray(.95),'purple'))(5)

# 2.4. Finally we produce the plot and clean up the background

library(grid)

ggplot(imports\_states,aes(map\_id = State,fill=TotalClass))+

geom\_map(map=imports\_states\_map, color=gray(0),size=.3)+

scale\_fill\_manual(values=ImportsColors)+

expand\_limits(x=stateMap$long,y=stateMap$lat)+

coord\_map("albers",29.5,45.5)+

labs(fill="Total Imports Rate",title="U.S. Imports of Goods by State")+

theme\_clean()

# 3. micromapST\_Combination Data

# 3.0. Setup

library(micromapST)

# 3.1. The micromapST package has a stateNamesFips data.frame That has official

# states names as row names. (The District of Columbia name has been shorted

# to "D.C.")

stateNamesFips

# 3.2. Defining a data preparation function

micromapSTprep <- function(stateDF,stateId=NULL,ref=stateNamesFips)

{

if (is.null(stateId))

nam <- row.names(stateDF)

else

nam <- stateDF[,stateId]

nam <- ifelse(nam=="District of Columbia","D.C.",nam)

check <- match(nam,row.names(ref))

bad <- is.na(check)

good <- !bad

nbad <- sum(bad)

if (nbad>0)

{

warning(paste(nbad,"Unmatch Names Removed",nam[bad]))

stateDF <- stateDF[!bad,]

nam <- nam[!bad]

check <- check[!bad]

good <- good[!bad]

}

ngood <- sum(good)

if (ngood < 51)

warning(paste("Only",ngood,"State Ids"))

row.names(stateDF) <- ref[check,2]

return(stateDF)

}

# 3.3. Reading and preparing a state data.frame

setwd("F:/乔治梅森大学/2015 fall/STAT-515/Final\_Presentation/final project/MicromapSt(2.Combination)");

combination <- read.table("clipboard",sep="\t",header=TRUE,as.is=TRUE)

head(combination)

ImportsMinusExports <- micromapSTprep(combination,"State")

head(ImportsMinusExports)

# 3.4. A quick look at micromapST dot plots

# 3.4.1. Making a panel description data.frame

colNumbers <- 1:ncol(ImportsMinusExports)

names(colNumbers)=colnames(ImportsMinusExports)

colNumbers

panelDesc <- data.frame(

type=c('map','id','dot','dot','dot'),

lab1=rep("",5),

lab2=c('','','Exports','Imports','Imports-Exports'),

lab3=c('','','Possible 0-500,000','Possible 0-500,000',''),

col1=c(NA,NA,"Exports","Imports",12))

t(panelDesc)

# 3.4.2. The micromapST() function

pdf("Combination.pdf",width=7.5,height=10)

micromapST(ImportsMinusExports, panelDesc,sortVar=12,ascend=FALSE,

title=c("Combination of Exports and Imports","Imports - Exports"))

dev.off()

# 3.5. Cumulative maps, arrows, reference values and two variable transformations

ImportsMinusExports$Zero <- rep(0,nrow(ImportsMinusExports))

panelDesc <- data.frame(

type=c('mapcum','id','arrow','arrow'),

lab1=c('','','Exports To','0 To'),

lab2=c('' ,'','Imports','Imports-Exports'),

col1 = c(NA,NA,4,13),

col2 = c(NA,NA,9,12),

refVals=c(NA,NA,NA,0))

pdf("Combination Arrows.pdf",

width=7.5,height=10)

micromapST(ImportsMinusExports,panelDesc,

sortVar=12,ascend=FALSE,

title=c("Combination of Exports And Imports","Imports - Exports"))

dev.off()

# 3.6. Maptail and Dots with confidence Intervals

panelDesc <- data.frame(

type=c('maptail','id','dotconf','dotconf'),

lab1=c('','','Exports Estimates ','Imports Estimates'),

lab2=c('' ,'','and 95% CIs','and 95% CIs'),

col1 = c(NA,NA,4,9),

col2 = c(NA,NA,5,10),

col3 = c(NA,NA,6,11))

png(file="Combination of Exports and Imports DotConf.png",

width=7,height=10,

units="in",res=300)

micromapST(ImportsMinusExports, panelDesc,

sortVar=12,ascend=FALSE,

title=c("Combination of Exports and Imports","Imports - Exports"))

dev.off()

# 4 time series

data<- read.table("clipboard",header=TRUE)

data

library(ggplot2)

colnames(data)

ggplot(data = data)+

geom\_line(aes(x=Month,y=Exports,colour =as.character(Year)))+

geom\_point(aes(x=Month,y=Exports),data = data)

Reference

United States Census Bureau

[http://www.census.gov/foreign-trade/Press-Release /](http://www.census.gov/foreign-trade/Press-Release%20/)current\_press\_release /index.html#ft900

Bureau of Labor Statistics

http://www.bls.gov/cpi/