STAT 248 Class Project Report

Clustering Analysis of Airport Capacity-demand Scenarios for Strategic Air Traffic Management

Lei Kang

Department of Civil and Environmental Engineering

Institute of Transportation Studies

107 McLaughlin Hall, University of California, Berkeley

Berkeley, CA 94720, USA

Phone: (510) 693-9671

Email: kang119@berkeley.edu

Supervisor: **David Brillinger**

May 2015

University of California, Berkeley

1. Introduction

The question that is will be considering in this paper is to characterize and identify representative capacity-demand profiles based on an airport's historical operational outcomes through time series clustering analysis. Air traffic managers today are typically limited to personal experience to make Traffic Flow Management (TFM) decisions (Wolfe S.R. and Rios, 2011; Grabbe et al., 2013; Liu et al., 2014). These decisions include whether or not there is a need for Traffic Management Initiatives (TMIs), such as Ground Delay Program (GDP) and Airspace Flow Program (AFP), and how TMIs should be planned when they are considered necessary. Traffic managers with different experiences or different preferences may create different TMI plans for the same situation. This unpredictability in decision creates uncertainty for National Airspace System (NAS) users and may hinder them from taking effective proactive actions (Liu et al., 2014). To address this issue, systematic approaches should be developed to better inform and assist managers in TFM decision making. One way of achieving this goal is to provide representative capacity-demand profiles based on historical operations to TFM decision makers. By learning from historical performance (characterized by demand and supply/capacity), TFM decision makers can infer the expected performance of current situation. To this end, one fundamental issue is to characterize and identify representative capacity-demand profiles based on an airport's historical operations.

For a given day at an airport, I define a capacity profile as a series of declared airport acceptance rates (AARs) recorded for each quarter-hour. We could have a capacity profile from 0:00 am to 23:59pm with length equal to 96 (24 hours * 4 quarters). These rates are reported by traffic managers based on weather conditions, runway configurations, and arrival and departure traffic mix. Demand profile is defined in a similar manner which is a time series of intending-to-

land rates. Intending-to-land values not only include scheduled arrivals, but also reflect arrival traffic congestion. Some flights might hold in the air at destination airport due to arrival congestion, thus also being captured by intending to land rates. Both capacity and demand profiles defined here are realizations of actual traffic conditions under weather influence, and thus can be used to characterize an airport's operation.

2. Exploratory Data Analysis

In this study, I will focus on San Francisco International Airport (SFO), but similar analysis could be easily conducted for any other airports. Quarter hourly AARs and quarter hourly intending to land rates from January 2005 to December 2014 (3652 days) were obtained from the Federal Aviation Administration (FAA) Aviation System Performance Metrics database (ASPM). Based on ASPM records, the maximum quarter-hourly AAR for SFO is 16. Thus, I will use 16 as upper bound for AARs.

Before jumping into various fascinated clustering methods, let's look at the data in depth using exploratory data analysis (EDA). The purposes of EDA in this case are two folds:

1. Better understand the data and identify potential patterns and structures;
2. By conducting EDA, we can obtain a preliminary assessment on daily variation regarding airport capacity and demand. This could help test whether cluster based on daily profiles is appropriate or not.

Figure 1 presents capacity and demand profiles of SFO in 2005 with y axis denotes AAR/demand values and x axis denotes daily quarter-hour index ranging from 1 (0:00 am) to 96 (23:45pm). Due to space limitation, profiles with respect to 2006 through 2014 are listed in the Appendix A1-1. Based on the feedback from a previous interview[1] with FAA traffic managers who were in charge of recording AARs, we learn that AARs from midnight to 7:00am are

[1] This interview was conducted by my colleague Yi Liu in 2013.

usually unreliable due to careless reporting behaviors of airport controllers. This statement is verified by observing little fluctuation of AARs from 0-7am (the dash line in Figure 1 depicts 7:00am). Moreover, from demand side, usually there are very few scheduled arrivals from midnight to 7:00am which makes this period less informative to traffic managers. Thus, I decide to only consider time period from 7:00am to 12:00am in the sample days. There is a major limitation of such truncation. In Figure 1, we do observe traffic from midnight to 2:00am in demand profiles. This is due to congestion spill-over effect which means for a particular day, due to accumulated congestion (i.e. under convective weather) traffic delays will spill over to the next day. In such cases, an effective day for demand should no longer be 7:00am to 23:59pm, but 7:00am to 2:00am (next day). However since we don't have reliable capacity values from midnight to 2:00am and our objective is to identify capacity-demand bundle profiles as a whole, I will ignore spill-over effect in this study. But we should bear in mind that spill-over phenomenon might influence later clustering analysis.

From now on, I will work on truncated series (N=68 from 7:00am to 23:59pm), and in total there are 248336 points for each capacity and demand series from 2005 to 2014 (3652 days * 68 points/day), see Figure 2 for series representation for supply and demand. To be brevity, I only present the first 6 days with y axis denotes AAR/demand values and x axis denotes quarter-hour index (6*68= 408) and different color represents different days. AARs sometimes fluctuate within a small range (as indicated in day 2, indicated by the yellow line). When AARs are low, intending to land rates are generally high due to congestion (see corresponding demand profile in day 2).

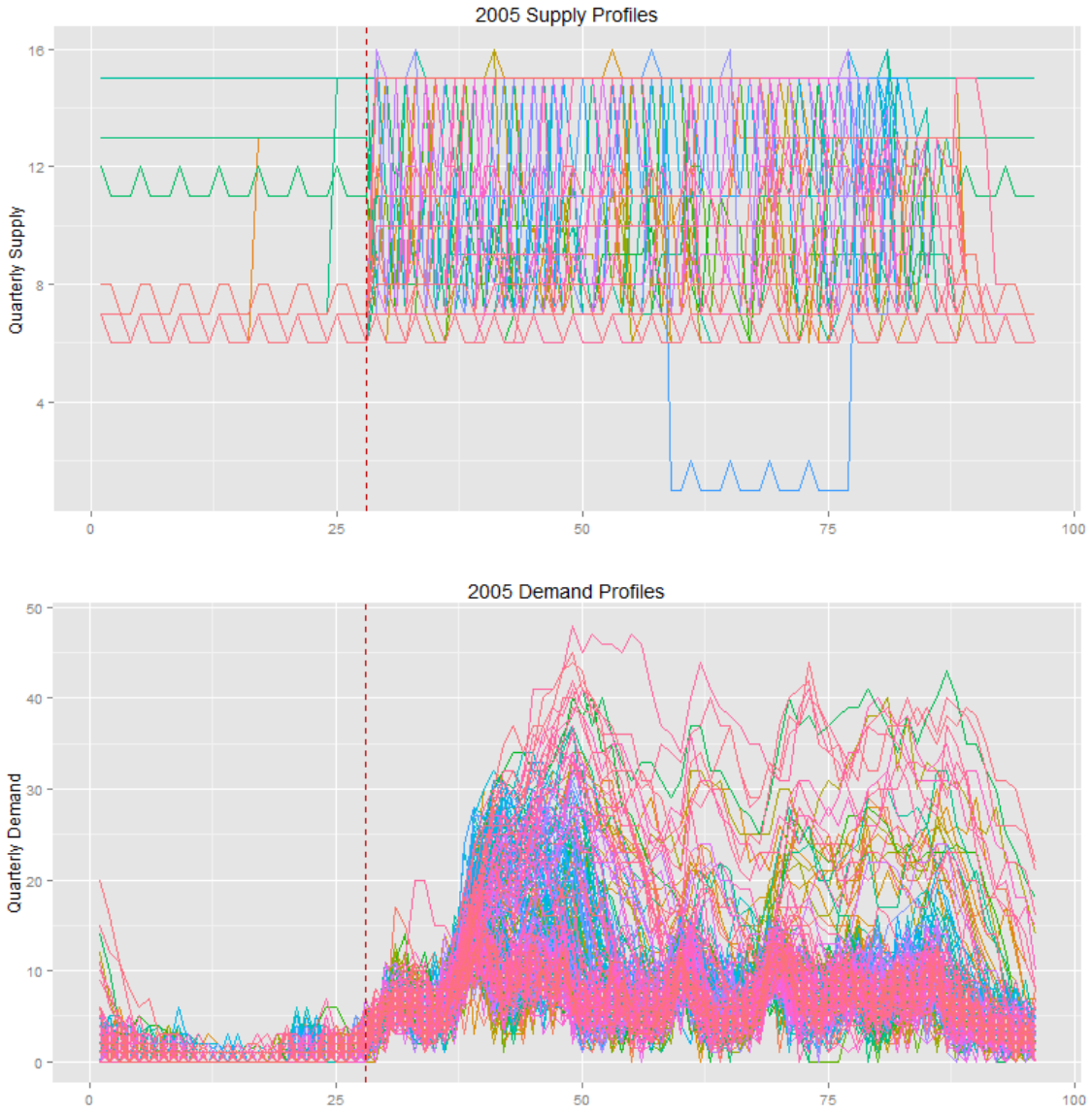


Figure 1 2005 Supply and Demand Profiles

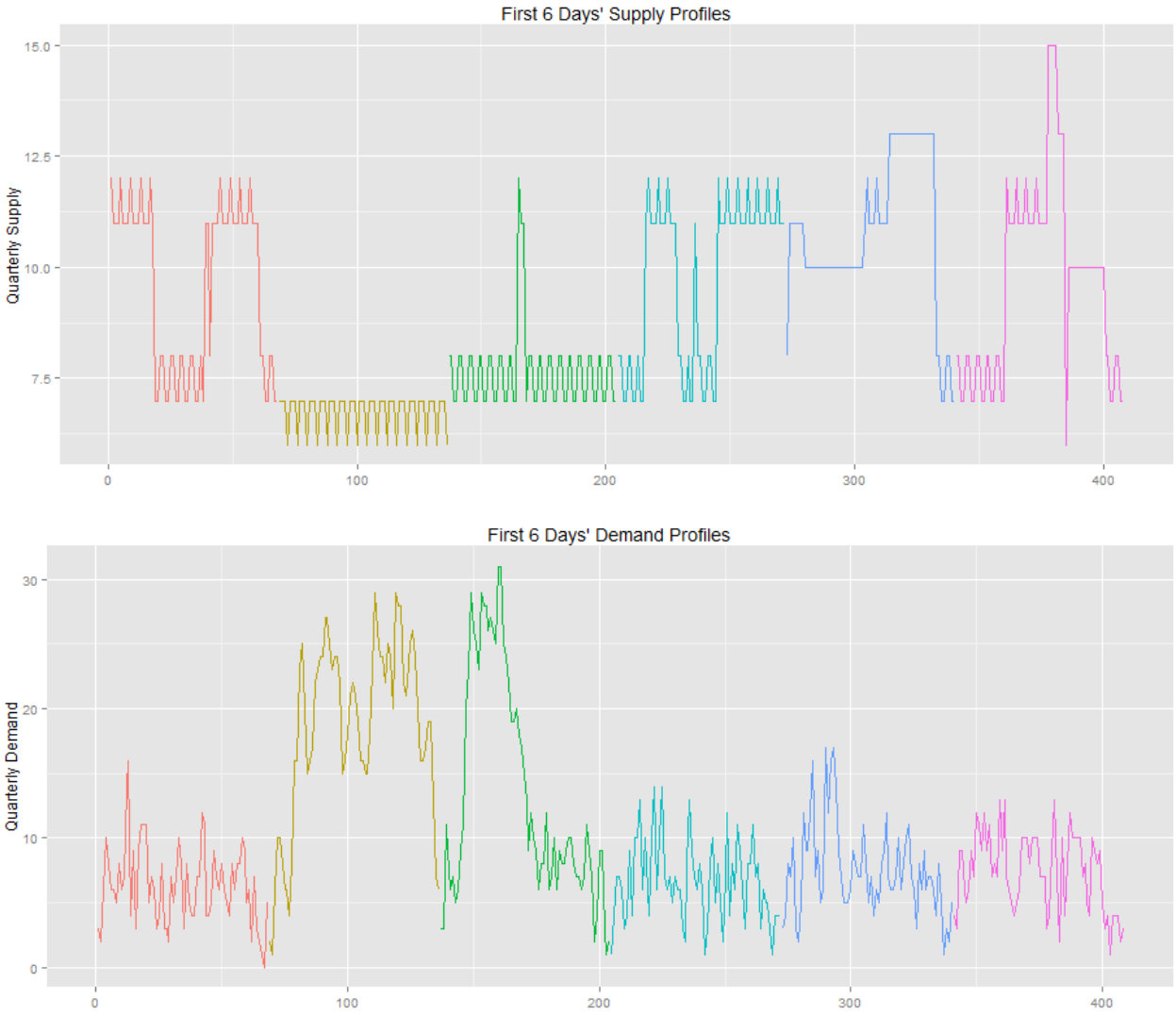


Figure 2 First 6 Days' Profiles

The autocorrelation functions (ACF) for total supply and demand series as displayed in Figure 3 indicate strong daily periodic patterns (with lag equals to 68 which represents 1 day indicated by the dashed lines in Figure 3). Such strong seasonality will generally make whole series non-stationary. In order to capture seasonal effects, regression models have been developed for demand and capacity profiles separately by including year fixed effects, month effects, day fixed effects, hour fixed effects, and quarter hour fixed effects as control variables. However, even though controlling for seasonal effects, residuals still don't behave like white noise, see Appendix A1-2. Moreover, long-lag (i.e. lag=68) differencing has also been tested, but the ACF

and PACF of differenced data still present seasonality. This result suggests the complexity of periodic patterns in the data cannot be easily captured. That's why we also need to look at frequency domain and later clustering analysis is based on frequency domain distance measure.

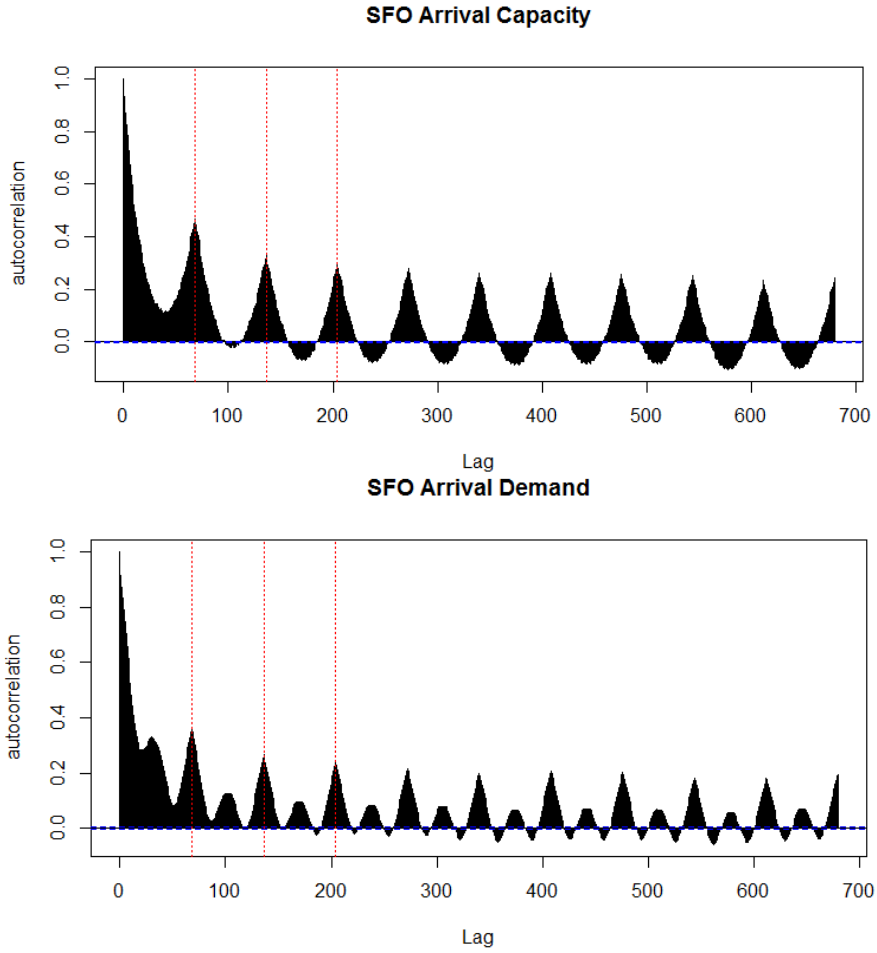


Figure 3 Series Autocorrelation Functions

If we switch to frequency domain, as indicated in Figure 4, smoothed periodograms suggest that for capacity series, the most influential frequency occurs at 1/68 which corresponds to 1 day. For demand series, besides significant daily frequency (occurs at 1/68), the most influential frequency occurs at 2/68 which corresponds to about 8.5 hours (=34/4). This location is about 15:30pm. This result indicates for demand profiles, there are two major periodic patterns: daily pattern and morning/afternoon pattern. Generally, morning traffic would differ from afternoon traffic starting from 15:30pm. Also need to point out that small frequencies (around zero) also

account for big share in explaining variance in the original series. This is due to small range of data. For example, regarding AARs, the maximum possible value is 16, whereas mean AAR value in this sample is around 11. Thus, we would expect to see a large number of small frequency sinusoid functions.

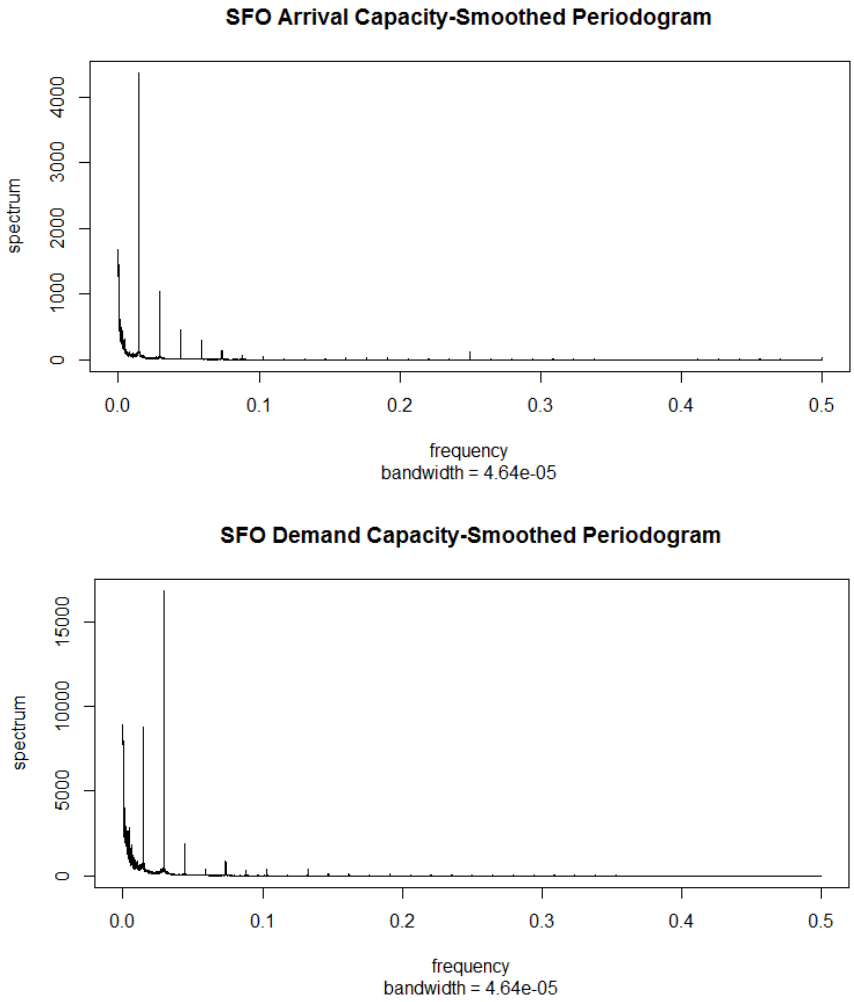


Figure 4 Smoothed Periodogram

Based on EDA, we can find that after removing possible seasonal effects, residuals still don't behave as white noise. Besides complicated periodic patterns, this result also suggests the presence of significant within-day variation for both demand and capacity profiles which makes clustering based on daily profiles a valid choice. Because if there is no within-day variation, then there will be no need to perform clustering analysis since every day behaves similar.

3. Literature Review

Time series clustering usually involves two choices: **distance measure** and **clustering algorithm**.

Regarding distance measures, three major categories have been widely adopted: raw data-based, feature-based, and model-based (Agrawal et al., 1993; Liao, 2005; Mori et al., 2014; Montero and Vilar, 2014). Raw data-based distance measure is based on directly comparing the raw values and the shape of the series in different manners. One common choice would be Euclidean distance. However, clustering based on raw data implies working with high-dimensional space in the context of time series clustering. It is sometimes not desirable. Instead, feature-based distance measure focuses on extracting a set of features from the time series and calculating the similarity between these features instead of using the raw values of the series. One example would be using Pearson's correlation coefficient as distance. The third alternative is model-based distance measure. Basically, this approach assumes the existence of the underlying data generating process (DGP) and develop parametric/non-parametric model to uncover the unknown DGP. Time series are considered similar when the models characterizing individual series are similar. For example, calculate the Euclidean distance between two sets of estimated coefficients of two ARMA models.

Regarding clustering algorithms, two popular options are k-means (or k-centroids) and hierarchical clustering. A hierarchical clustering method works by grouping time series into a tree of clusters. It (agglomerative hierarchical) starts by placing each time series in its own cluster and merge these time series into larger and larger clusters until all the objects are in a single cluster. K-means algorithm requires the input of number of clusters, and then tries to

minimize of an objective function (i.e. total distance between two clusters). Its solution relies on an iterative scheme, which starts with arbitrarily chosen initial cluster centers (Liao, 2005).

4. Method and Results

In this study, for any two profiles, we are not only interested in their relative magnitudes but also their trends and shapes. Sum of raw distance could not capture shape of a time series which makes variety of raw data based dissimilarity measures inappropriate in this case. Model-based approach usually assumes an ARIMA type of model for a time series. However, in this case, due to small variation of AARs for some days, distance measure derived from ARIMA type of model might not be robust. For example, the estimated ARIMA coefficients might be very sensitive to small change in AARs. Furthermore, simple feature-based distance measures like correlation coefficients can only track trend of a time series but not its relative magnitude. Therefore, for a given time series, in order to preserve as much information as possible, I employ J-divergence measure (one type of feature-based measure) suggested by Shumway and Stoffer (2011) which is a quasi-distance between estimated spectral matrices $\hat{f}_i$ and $\hat{f}_j$ for two sample time series. J-divergence measure is defined as:

$$J(\hat{f}_i, \hat{f}_j) = \frac{1}{n} \sum_{0<w_k<0.5} [tr\{\hat{f}_i(w_k)\hat{f}_j^{-1}(w_k)\} + tr\{\hat{f}_j(w_k)\hat{f}_i^{-1}(w_k)\} - 2p], \text{ for } i \neq j$$

Then an agglomerative hierarchical cluster analysis is performed since it requires no prior information about the number of clusters. Linkage function is selected as Ward distance which means the distance between any two clusters is represented by the increase in sum-of-squares if these two clusters were merged. Due to space limitation, only selected estimated spectral densities (smoothed periodograms) for capacity and demand are listed in Appendix A1-3. By visually inspecting the outcome dendogram, see Figure 5, 2 major clusters have been revealed

(cluster 1 contains 3237 days, group 2 contains 341 days). Need to mention that based on the dendrogram, there is also another cluster which accounts for less than 2% of total sample. Thus, this "small" cluster should be treated as an "outlier", not a representative cluster. In Figure 6, representative profiles are characterized by their mean values with 90% confidence intervals as depicted by dashed lines.

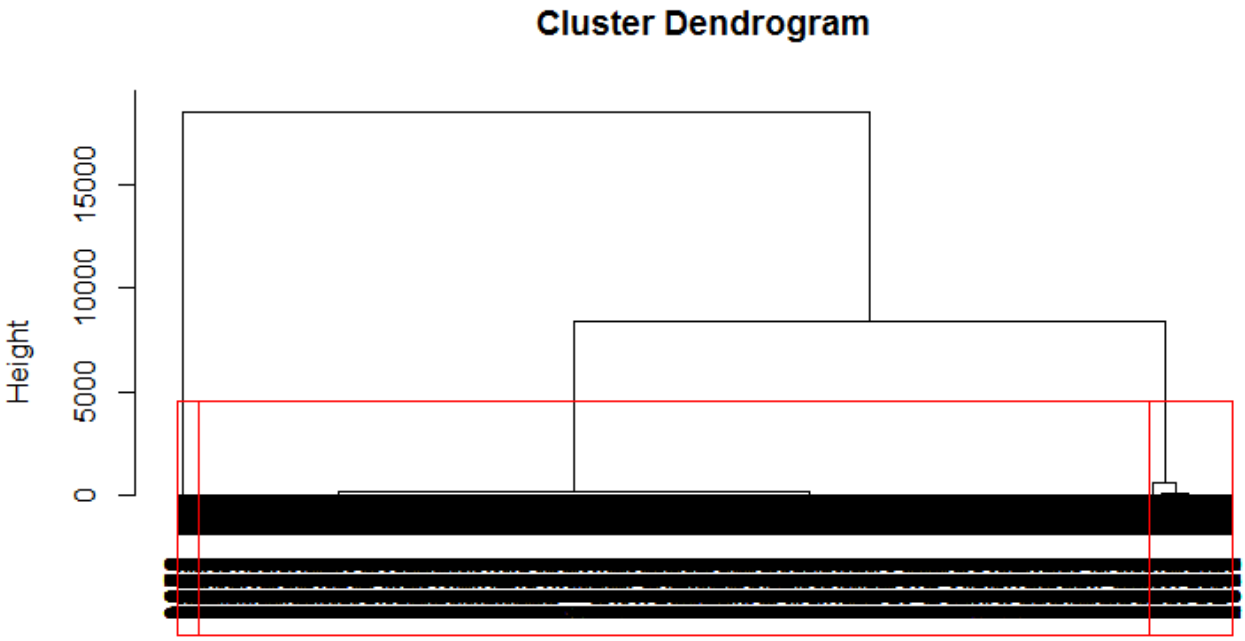


Figure 5 Clustering Result-Dendrogram

It can be seen that cluster 1 represents "good" weather days with relative high AARs and low intending to land rates. This result suggests during "good" weather days, quarter-hourly AARs are, on average, about 12 which are very close to upper bound 16. The corresponding intending to land rates has lower values which indicate less delay for traffic. However, during "bad" weather days, as indicated by cluster 2, airport capacity is generally low which is on average about 9. Due to low capacity, the corresponding intending to land rates would be influenced and more delays were incurred.

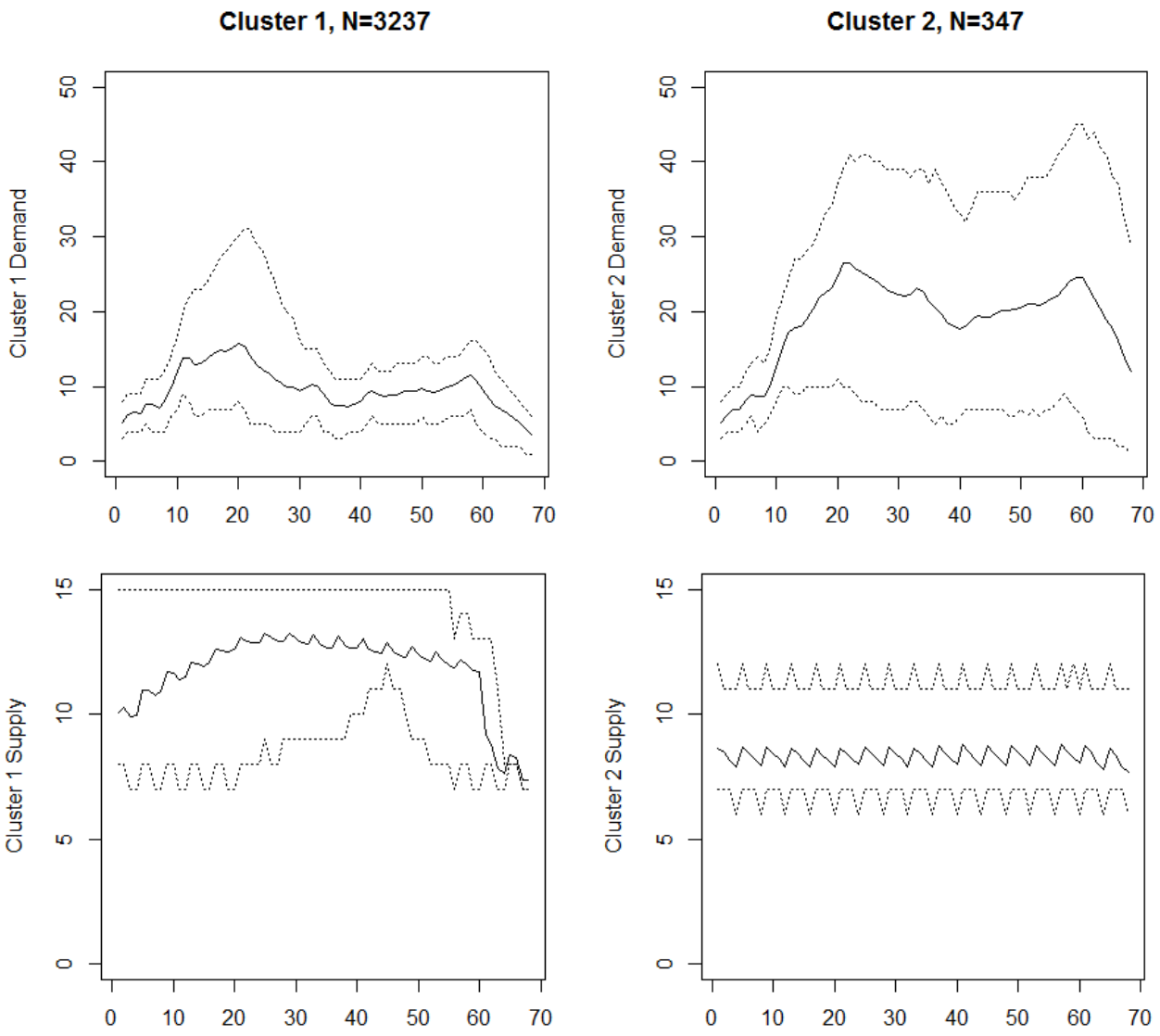


Figure 6 Representative Profiles

5. Discussion

Two interesting observations regarding Figure 6: Firstly, during "bad" weather days, capacity would generally be low and intending to land rates has higher variance than "good" days. This will generally create difficulty in air traffic management. Secondly, during "good" weather days, we can observe a "dip" at the end of the day for both capacity and demand. It makes sense for demand profile because we would expect little traffic around midnight during good weather days since no delay got spilled over to the next day. Based on the interview with FAA traffic

managers, we learnt that during non-busy period, airport controllers tend to enter AARs in an arbitrary way. Thus, the sharp drop in capacity profiles at the end of the day might simply pick up human reporting errors.

This study is a pilot project in charactering airport historical performance by using intending to land rates as demand and AARs as capacity. Bivariate time series clustering based on spectral density distance are employed. By examining hierarchical dendrogram, 2 representative clusters have been revealed and characterized by their mean values and 90% confidence intervals. One future direction is to conduct performance analysis regarding these 2 representative profiles including traffic delay, airport efficiency, airport predictability, etc.

One limitation of this study is that spill-over effect is not considered in clustering and an effective operation day is restricted to be in the same day since we don't have reliable capacity rates from midnight to 7:00am. But this assumption might be too restrictive because spill-over phenomenon sometimes interests traffic managers under the context of extreme event management. In spite of this limitation, by applying frequency domain distance measure and Ward linkage-based hierarchical clustering for historical capacity and demand profiles, my answer to the paper is that we can characterize and identify 2 major representative capacity-demand bundle profiles for SFO which can be used for future day-of-operation performance analysis.

6. Reference

Agrawal R., Faloutsos C., Swami A., 1993. Efficient similarity search in sequence databases. In Foundations of Data Organization and Algorithms, volume 730 of Lecture Notes in Computer Science, 69-84. Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-57301-2.

Grabbe S., Sridhar B., Mukherjee A., Similar days in the NAS: an airport perspective, Aviation Technology, Integration, and Operations Conference, 2013.

Liao T.W., 2005. Clustering of time series data—a survey. Pattern Recognition 38, 1857-1874.

Liu Y., Seelhorst M., Pozdnukhov A., Hansen M., Ball M., 2014. Assessing Terminal Weather Forecast Similarity for Strategic Air Traffic Management.

Mori U., Mendiburu A., Lozano J., 2014. Distance measures for time series in R: the TSdist package. Available at: http://cran.r-project.org/web/packages/TSdist/vignettes/TSdist.pdf

Montero P., and Vilar J., 2014. TSclust: An R Package for Time Series Clustering. Journal of Statistical Software 62, 1-43.

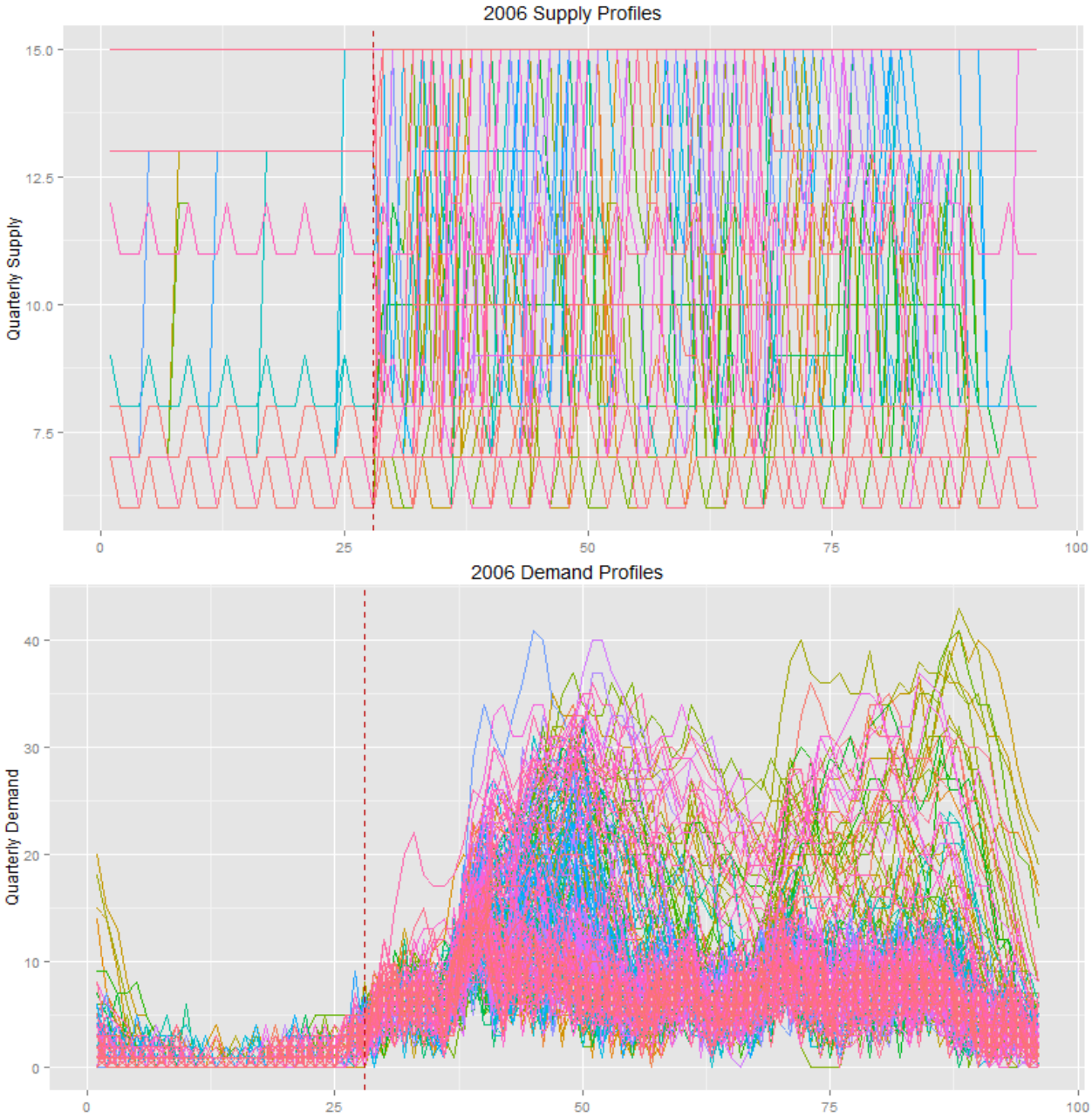
Shumway R., and Stoffer D., 2011. Time series analysis and its applications. Third edition. Springer New York Dordrecht Heidelberg London.

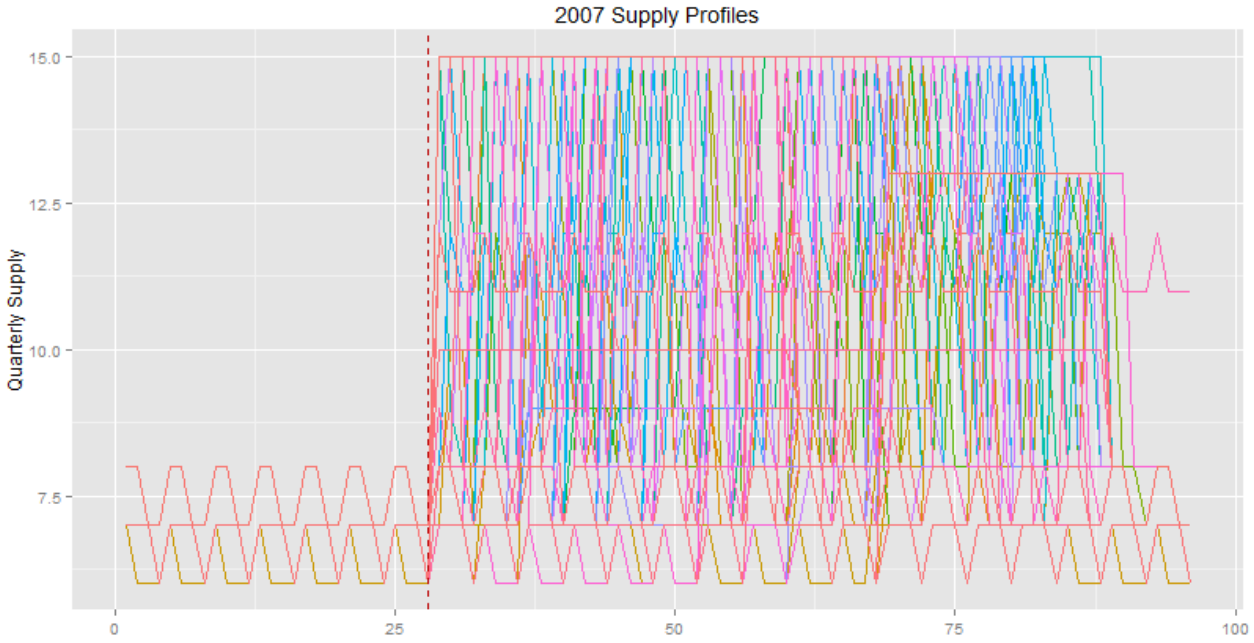
Wolfe S.R. and Rios J.L., A method for using historical ground delay programs to inform day-of-operation programs, AIAA Gidance, Navigation, and Control Conference, 2011.

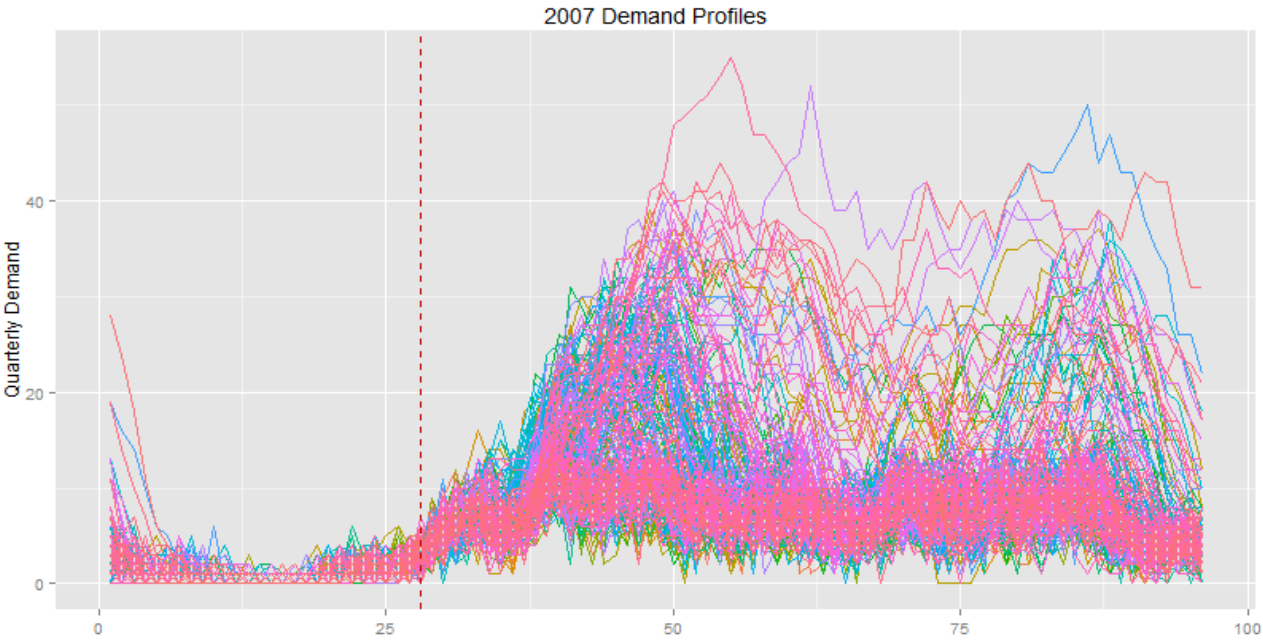
7. Appendix

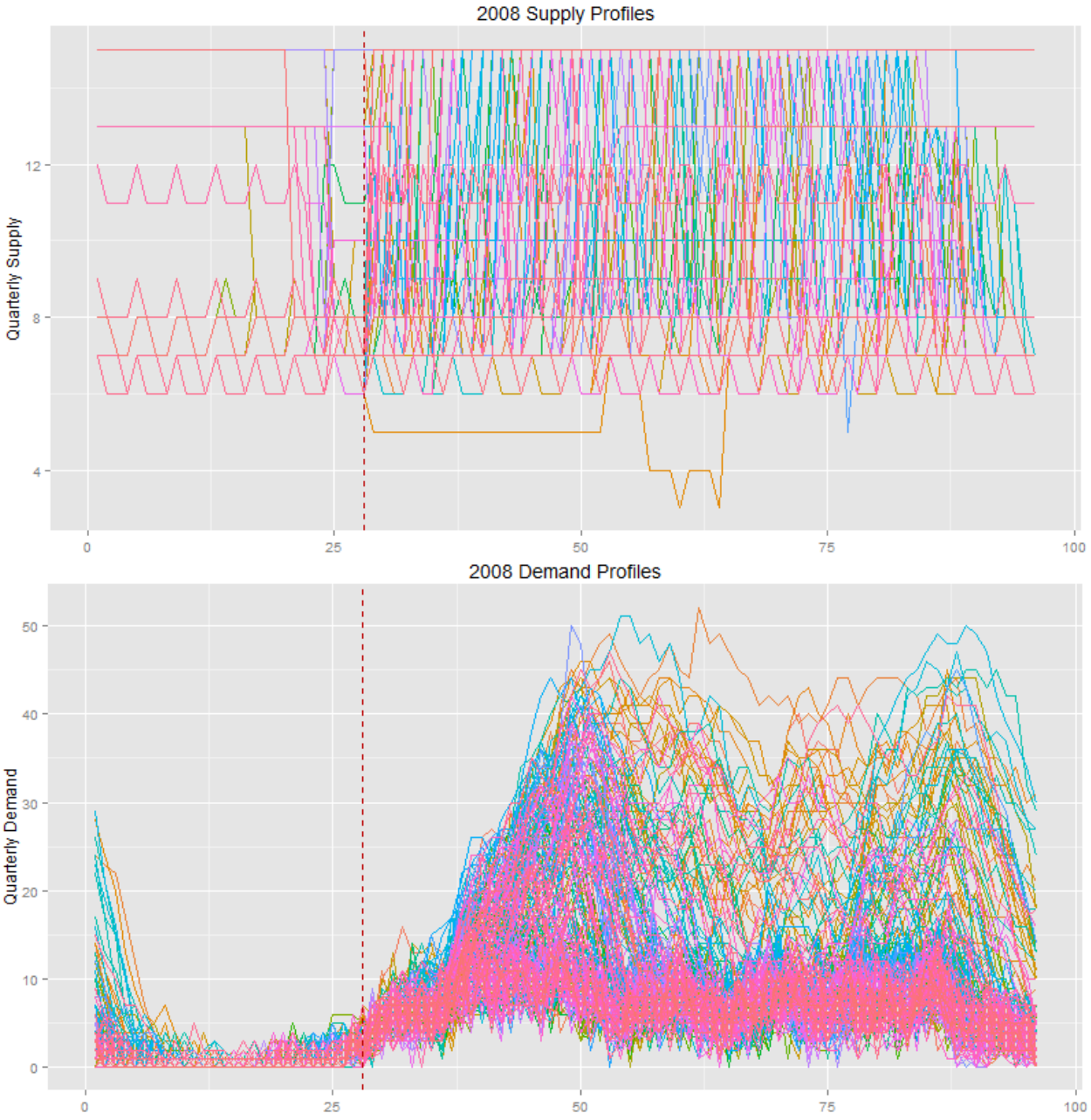
A1. Extra EDA Plots

A1-1. Demand profiles and Supply profiles for 2006 through 2014

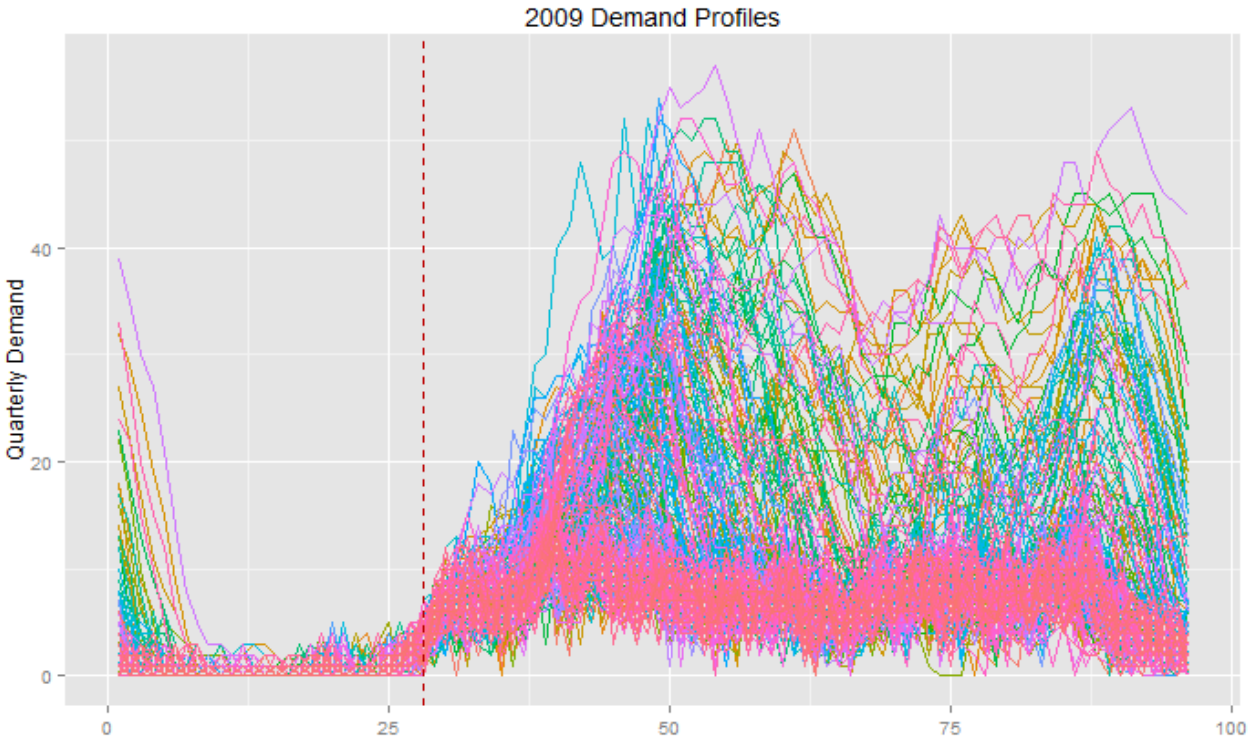


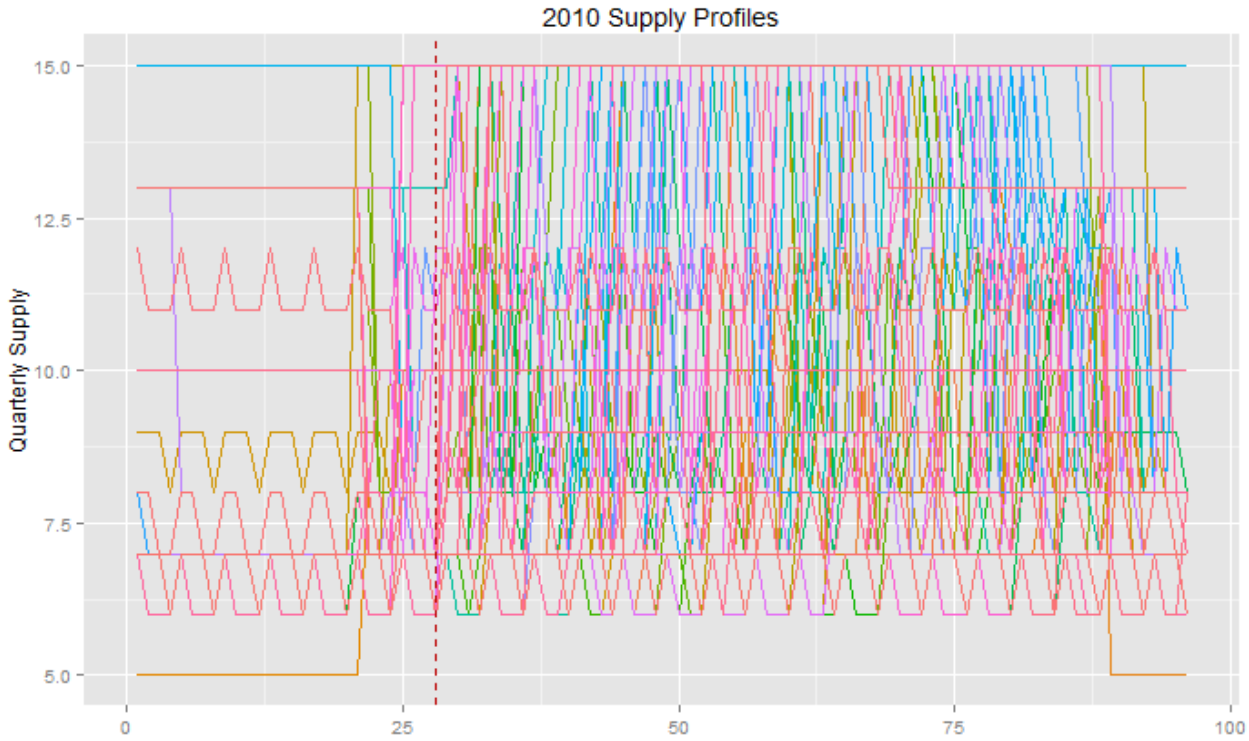

2007 Supply Profiles
Quarterly Supply
15.0
12.5
10.0
7.5
0
25
50
75
100

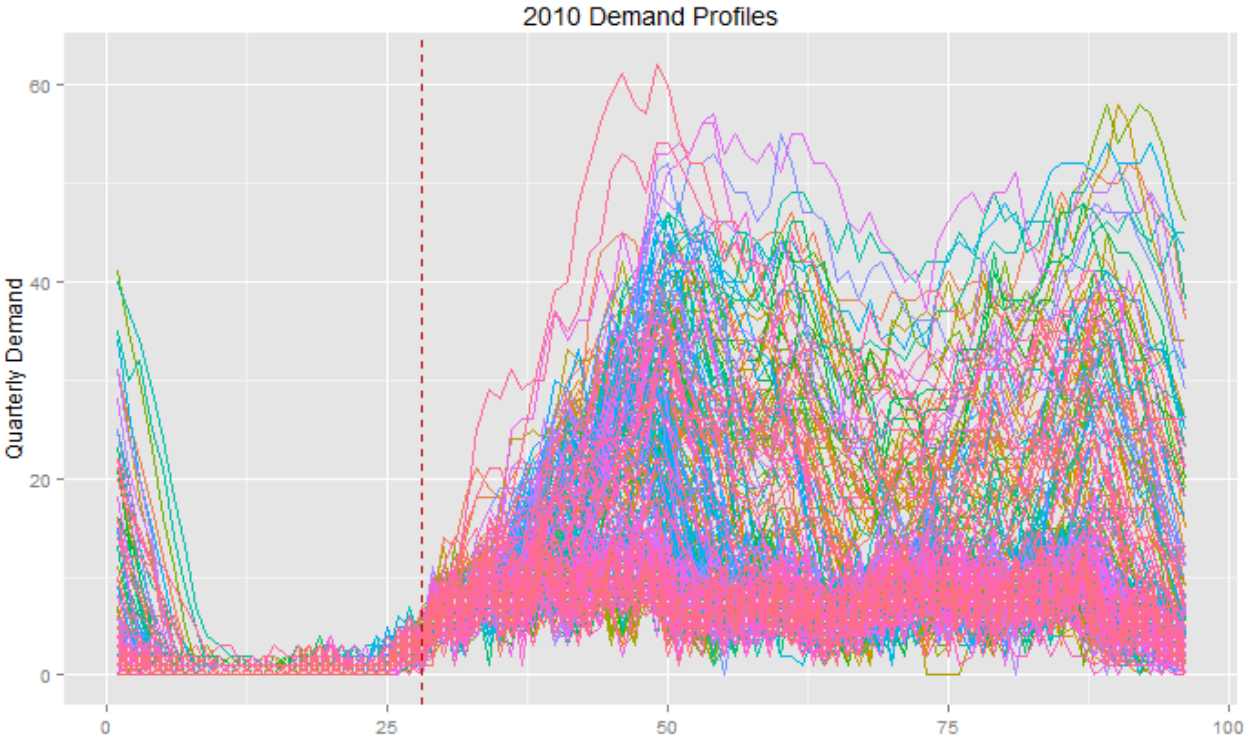

2007 Demand Profiles
Quarterly Demand
40
20
0
0
25
50
75
100

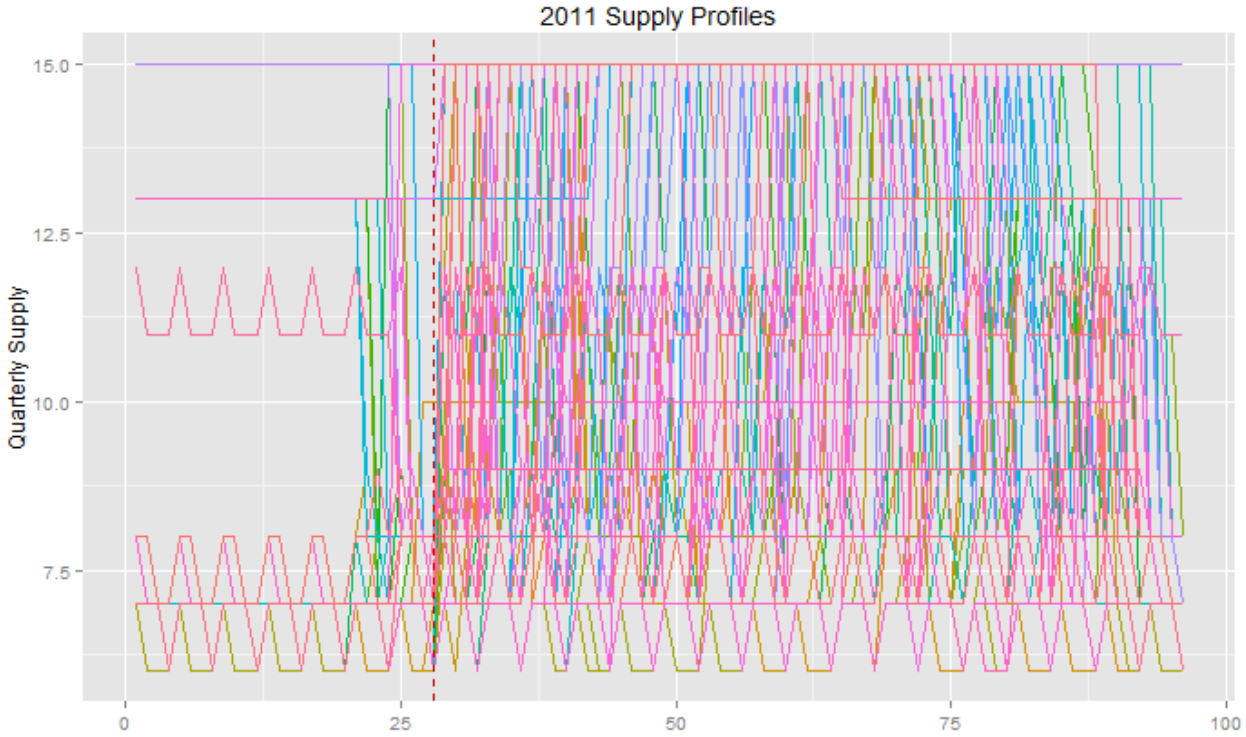

2008 Supply Profiles
Quarterly Supply
12
8
4
0
25
50
75
100
2008 Demand Profiles
Quarterly Demand
50
40
30
20
10
0
0
25
50
75
100

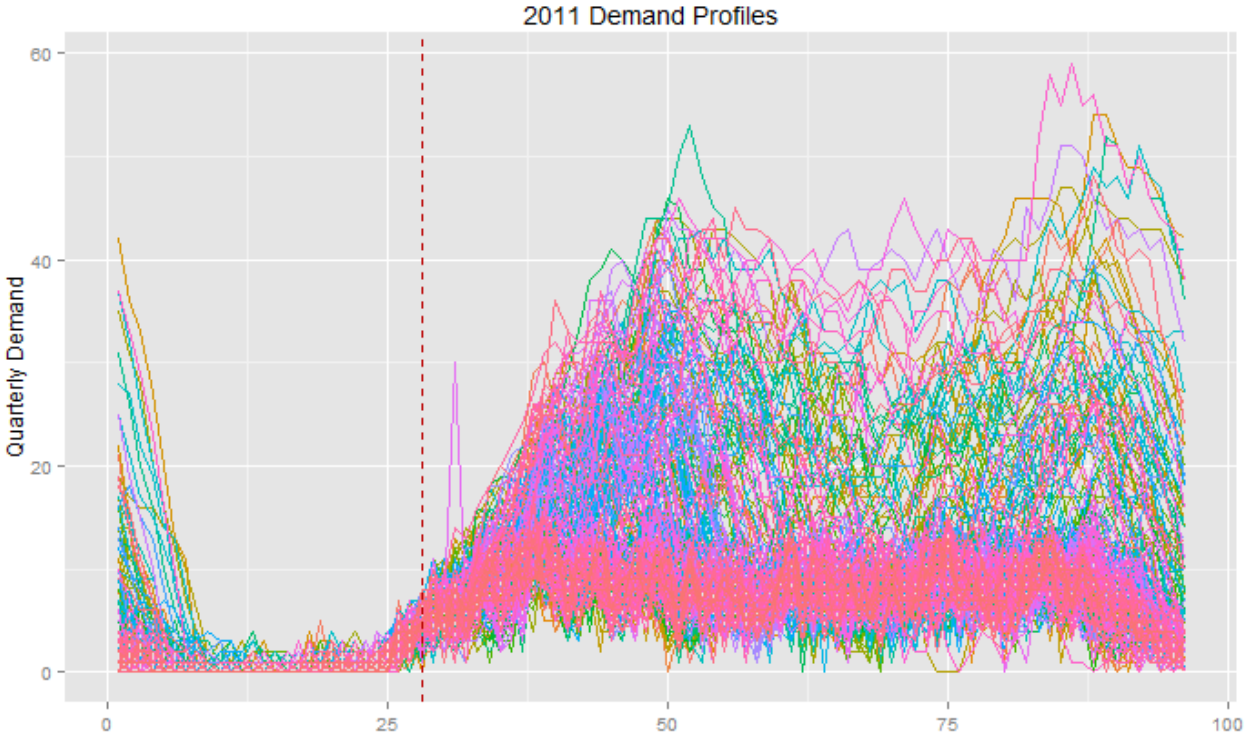

2009 Supply Profiles
Quarterly Supply
15.0
12.5
10.0
7.5
0
25
50
75
100

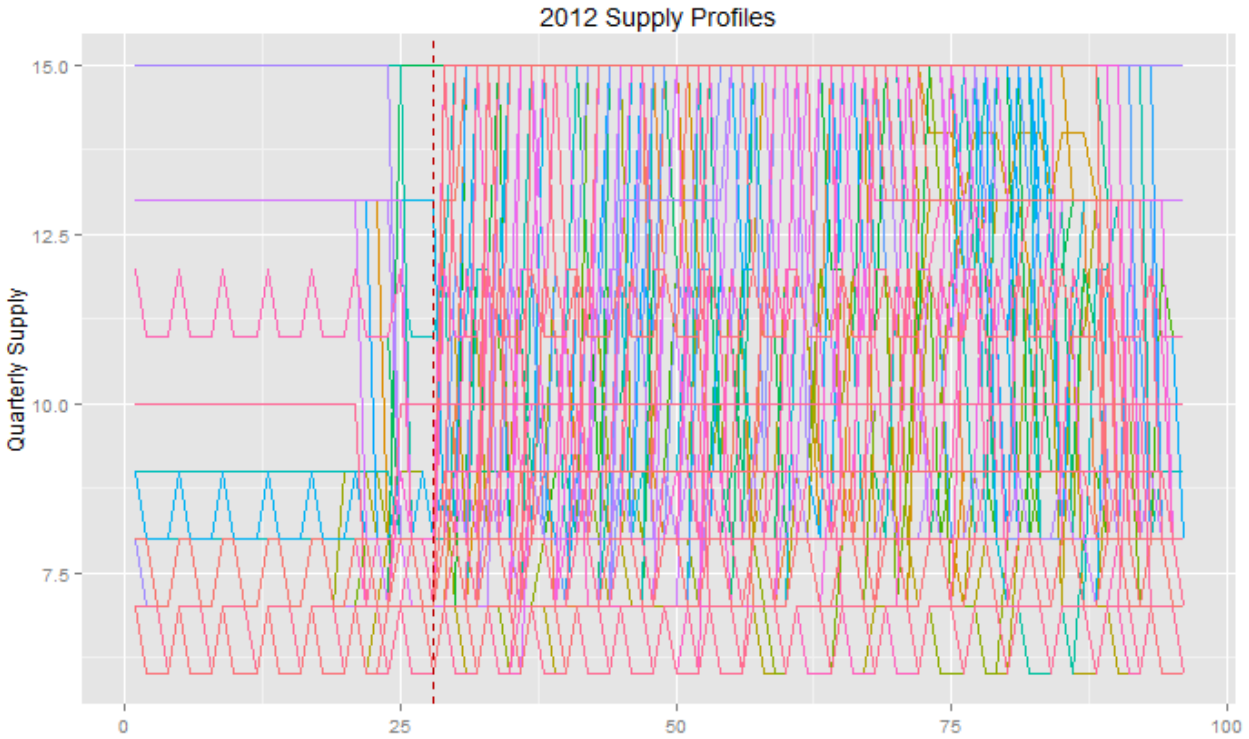

2009 Demand Profiles
Quarterly Demand
40
20
0
0
25
50
75
100

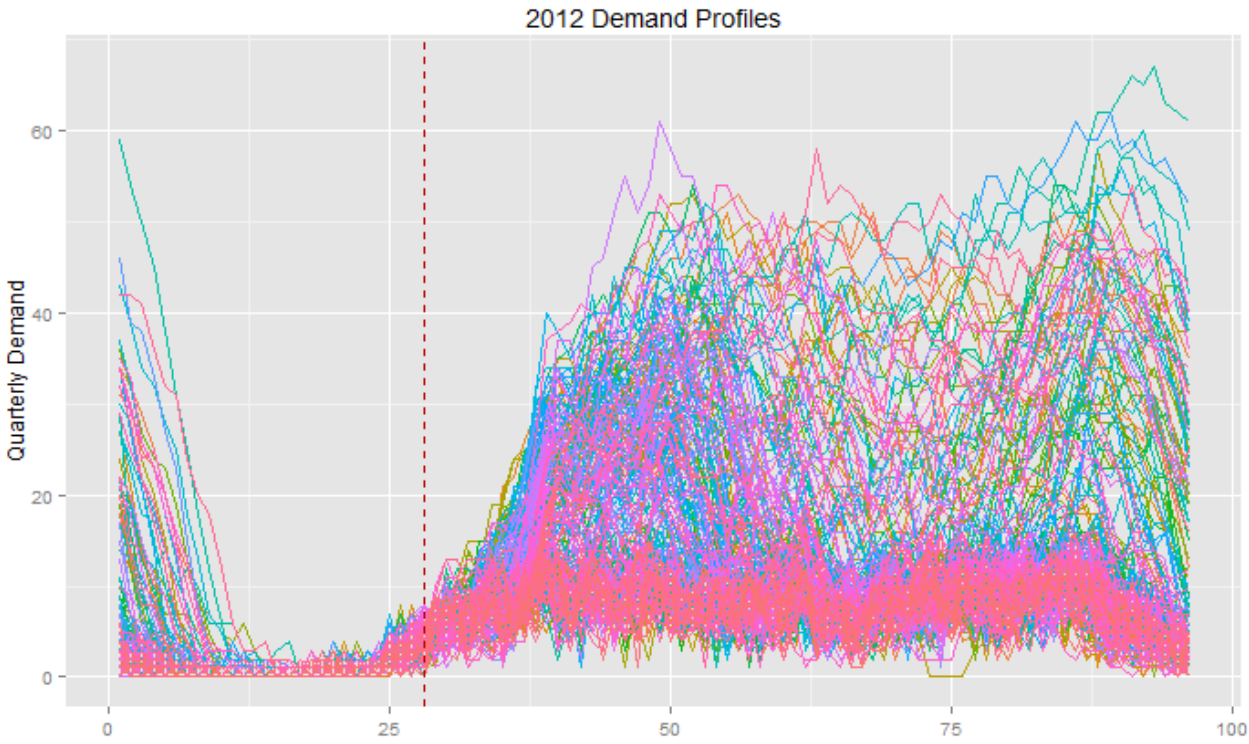

2010 Supply Profiles
Quarterly Supply
15.0
12.5
10.0
7.5
5.0
0
25
50
75
100

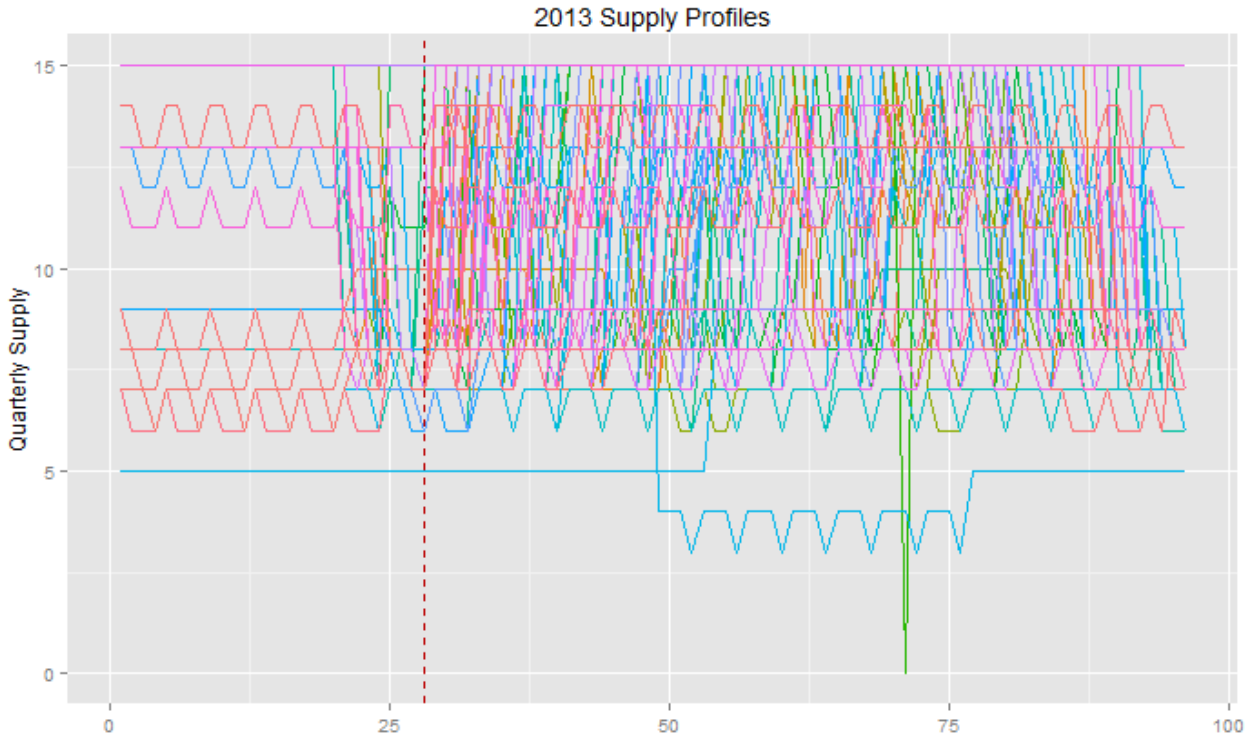

2010 Demand Profiles
Quarterly Demand
60
40
20
0
0
25
50
75
100

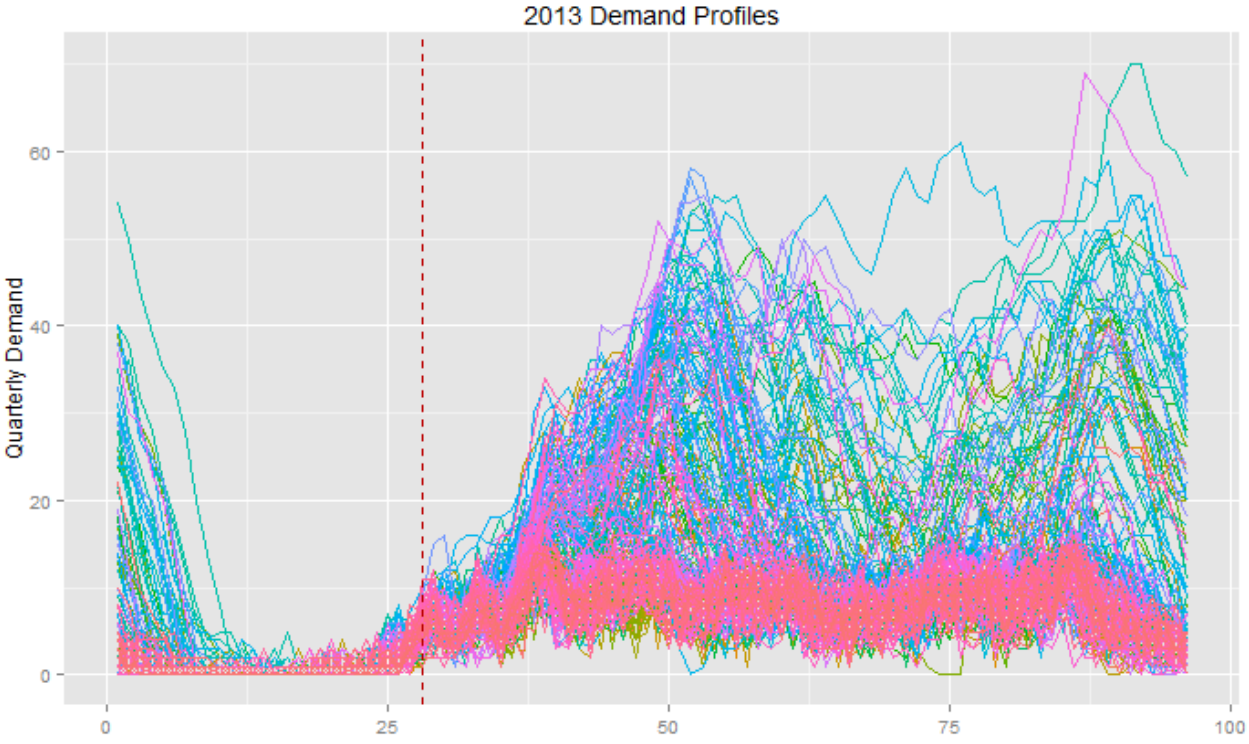

2011 Supply Profiles
Quarterly Supply
15.0
12.5
10.0
7.5
0
25
50
75
100

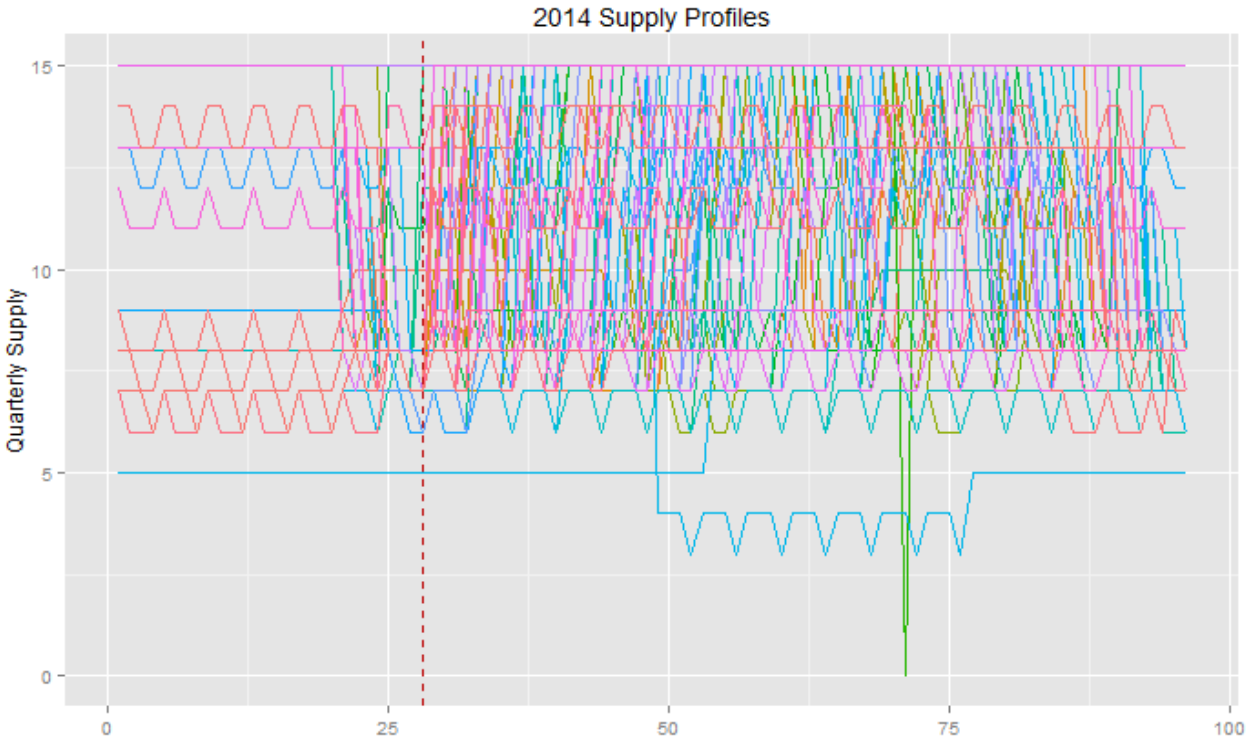

2011 Demand Profiles
Quarterly Demand
60
40
20
0
0
25
50
75
100

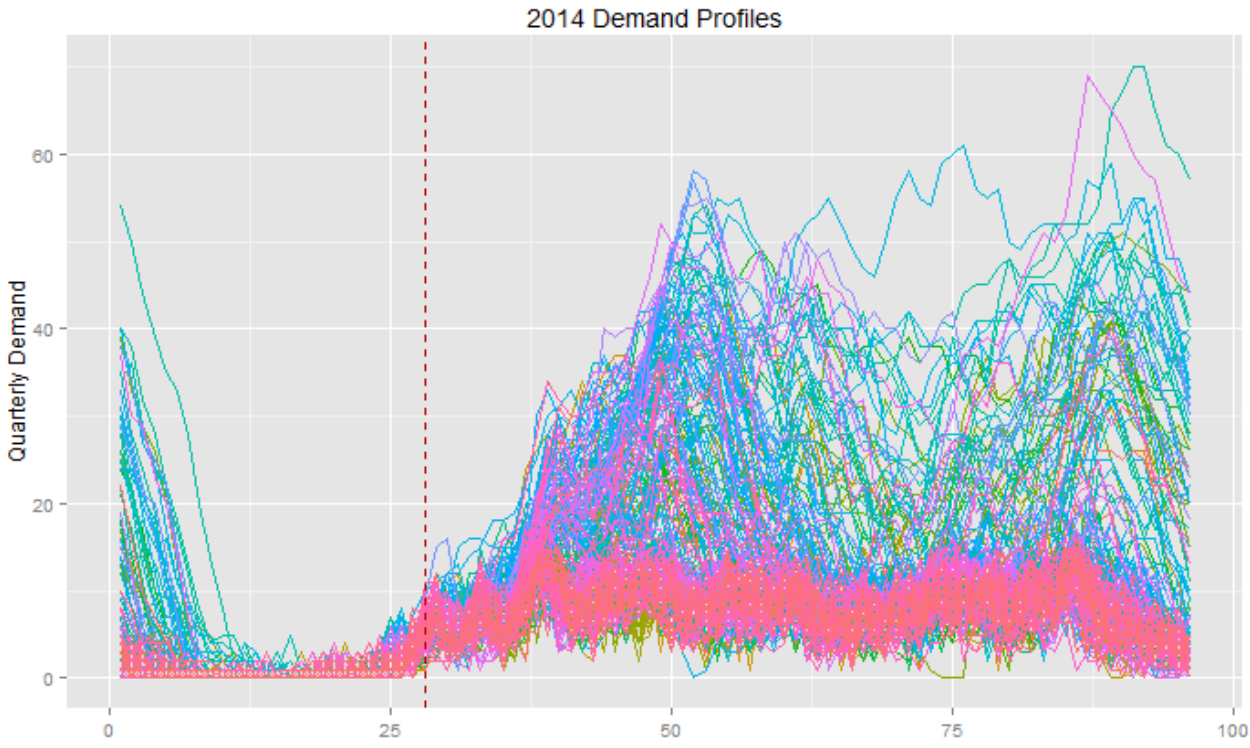

2012 Supply Profiles
Quarterly Supply
15.0
12.5
10.0
7.5
0
25
50
75
100


2012 Demand Profiles
Quarterly Demand
60
40
20
0
0
25
50
75
100

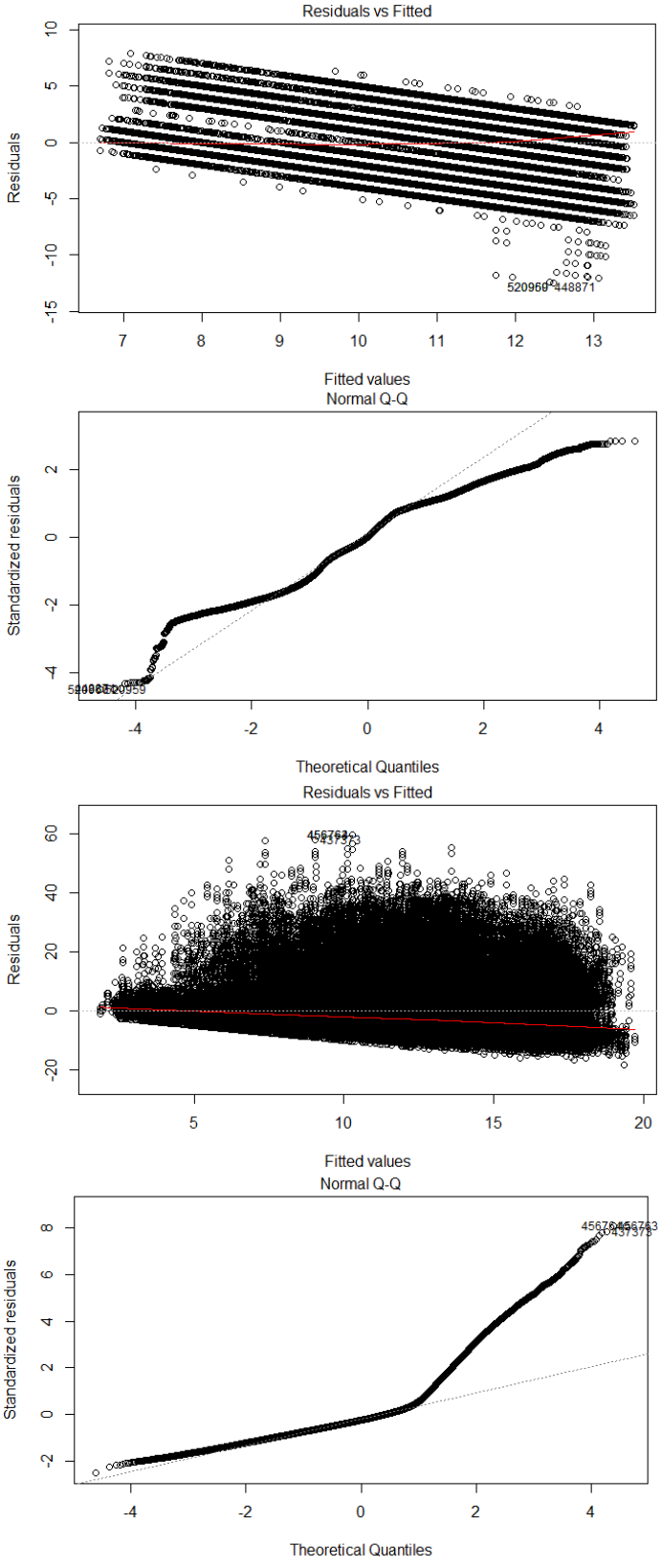

2013 Supply Profiles
Quarterly Supply
15
10
5
0
0
25
50
75
100


2013 Demand Profiles
Quarterly Demand
60
40
20
0
0
25
50
75
100

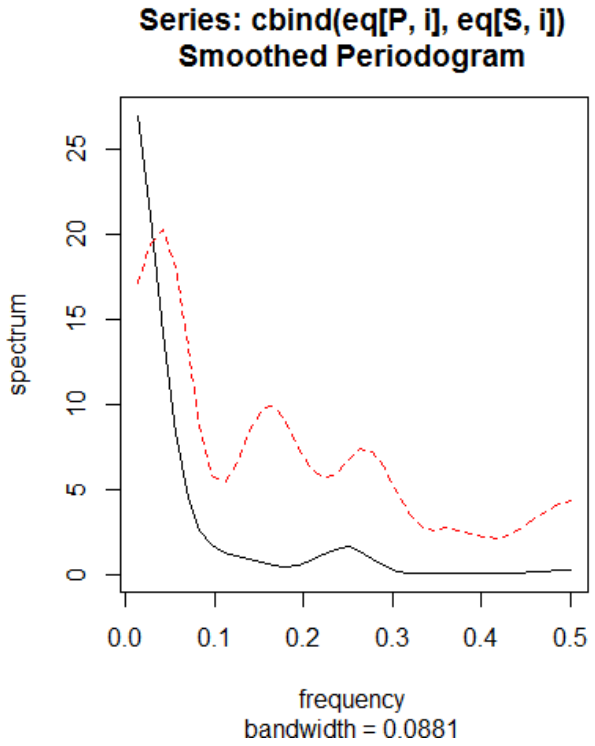

2014 Supply Profiles
Quarterly Supply
15
10
5
0
0
25
50
75
100

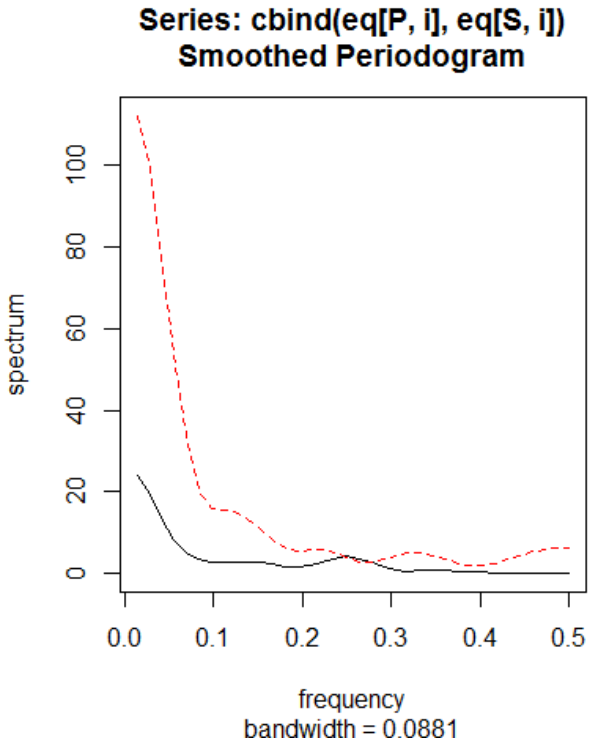

2014 Demand Profiles
Quarterly Demand
60
40
20
0
0
25
50
75
100

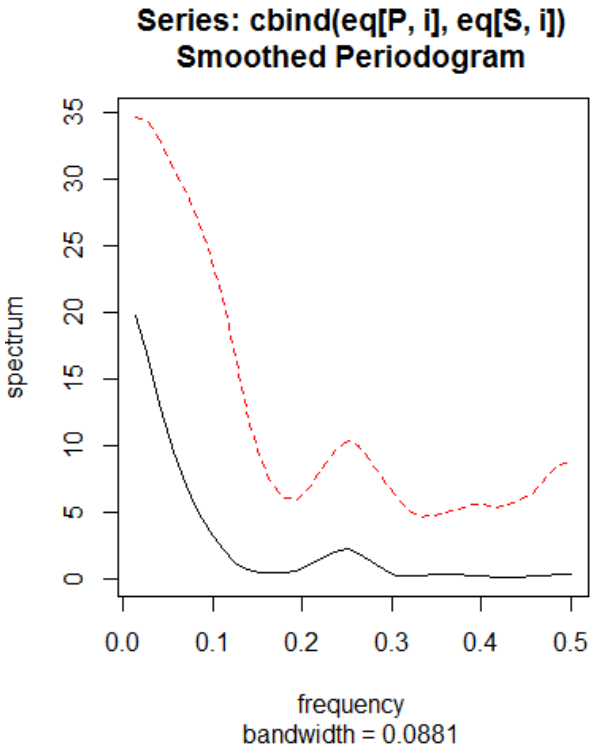
A1-2. Regression analysis for supply and demand

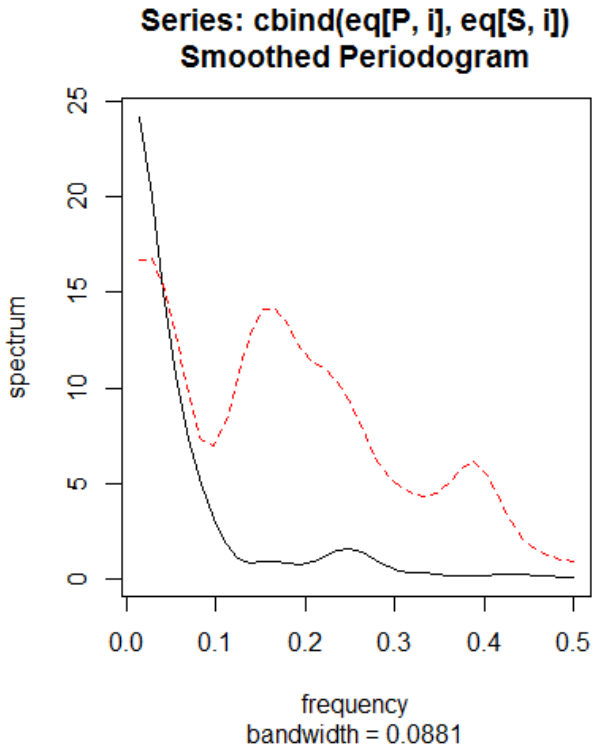


A1-3 Selected estimated spectral density output









A2. R Code

```
###read in data
data<-read.csv("SFO.csv", colClasses =
c("character",rep("integer",8),rep("NULL",55),rep("integer",7),rep("NULL",7),rep("integer",4),r
ep("numeric",3)))
data<-data[data$YYYYMM<201500 & data$YYYYMM>=200500,] ##10 years

###Code to generate Figure 1
library(ggplot2)
data2005<-data[data$Year==2005,]
ggplot(data2005,aes(index,ARR_RATE,group=id,colour=id))+ggtitle("2005 Supply
Profiles")+geom_line()+xlab("")+ylab("Quarterly
Supply")+geom_vline(aes(xintercept=28),colour="#BB0000",linetype="dashed")
ggplot(data2005,aes(index,ARR_DEMAND,group=id,colour=id))+ggtitle("2005 Demand
Profiles")+geom_line()+xlab("")+ylab("Quarterly
Demand")+geom_vline(aes(xintercept=28),colour="#BB0000",linetype="dashed")

###Code to generate Figure 2
ggplot(data2005[1:1360,],aes(index3,ARR_RATE,colour=id))+ggtitle("First 20 Days' Supply
Profiles")+geom_line()+xlab("")+ylab("Quarterly Supply")
ggplot(data2005[1:1360,],aes(index3,ARR_DEMAND,colour=id))+ggtitle("First 20 Days'
Demand Profiles")+geom_line()+xlab("")+ylab("Quarterly Demand")

###Code to generate Figure 3
data<-data[data$HR_LOCAL>=7,] ##local hour 7am to 12am
sfoaar <- ts(data$ARR_RATE) ##supply series
sfodemand <- ts(data$ARR_DEMAND) ##demand series
acf(sfoaar,68*10,main="SFO Arrival Capacity",ylab="autocorrelation")
abline(v=68,col="Red",lty=3)
abline(v=68*2,col="Red",lty=3)
```

```
abline(v=68*3,col="Red",lty=3)
acf(sfodemand,680,main="SFO Arrival Demand",ylab="autocorrelation")
abline(v=68,col="Red",lty=3)
abline(v=68*2,col="Red",lty=3)
abline(v=68*3,col="Red",lty=3)
```

###Code to generate Figure 4

```
a<-spec.pgram (sfoaar,spans=40,detrend=T,demean=T,taper=0.1,log='no',main="SFO Arrival Capacity-Smoothed Periodogram")
b<-spec.pgram (sfodemand,spans=40,detrend=T,demean=T,taper=0.1,log='no',main="SFO Demand Capacity-Smoothed Periodogram")
###Code for regression analysis
fit1<-
lm(ARR_RATE~as.factor(Year)+as.factor(Month)+as.factor(DAYNUM)+as.factor(HR_LOCAL)+as.factor(QTR),data=data)
plot(fit1)
fit2<-
lm(ARR_DEMAND~as.factor(Year)+as.factor(Month)+as.factor(DAYNUM)+as.factor(HR_LOCAL),data=data)
plot(fit2)
```

###Code for J-divergence distance measure

```
library(astsa)
library(MASS)
library(cluster)
P = 1:68; S = P+68; p.dim = 2; n =68
trasdata=matrix(0, 136, 3652) ##136=68*2 and 3652 days
for (i in 1:3652){
  aar<-data.frame(data[data$id==dateid[i],]$ARR_RATE)
  colnames(aar)<-c("value")
  demand<-data.frame(data[data$id==dateid[i],]$ARR_DEMAND)
```

```
  colnames(demand)<-c("value")
  trasdata[,i]<-as.vector(unlist(rbind(aar,demand)))
}
b<-trasdata
eq = as.ts(b[, 1:3652])
L = c(4, 4) #for smoothing
f = array(dim=c(3652, 2, 2, 36))
for (i in 1:3652){ # compute spectral matrices
  f[i,,,] = mvspec(cbind(eq[P,i], eq[S,i]),L,demean=T,detrend=T,log='no')$fxx
}
JD = matrix(0, 3652, 3652)
for (i in 1:3651){
   for (j in (i+1):3652){
     for (k in 1:36) { # use all freqs
       tr1 = Re(sum(diag(f[i,,,k]%*%ginv(f[j,,,k]))))
       tr2 = Re(sum(diag(f[j,,,k]%*%ginv(f[i,,,k]))))
       JD[i,j] = JD[i,j] + (tr1 + tr2 - 2*p.dim)
     }
   }
}
JD = (JD + t(JD))/n
colnames(JD) = c(1:3652)
rownames(JD) = colnames(JD)
```

##Code for hierarchical clustering

```
JD1<-as.dist(JD)
clus<-hclust(JD1, method="ward.D2") ##freq domain
plot(clus)
```

###Code for Cluster representation

```
par(mfrow=c(1,2))
```

```
g1<-meg[meg$groups==2,]
g1supply<-g1[,2:69]
g1demand<-g1[,70:137]
g1demand_mean<-colMeans(g1demand)
g1supply_mean<-colMeans(g1supply)
g1demand_up<-sapply(1:68,function (i) quantile(g1demand[,i],c(0.90)))
g1demand_low<-sapply(1:68,function (i) quantile(g1demand[,i],c(0.10)))

g1supply_up<-sapply(1:68,function (i) quantile(g1supply[,i],c(0.90)))
g1supply_low<-sapply(1:68,function (i) quantile(g1supply[,i],c(0.10)))
plot(c(1:68),g1demand_mean,type="l",ylim=c(0,50),ylab="Cluster 2 Demand",main="Cluster 2,
N=347")
lines(c(1:68),g1demand_up,lty=3)
lines(c(1:68),g1demand_low,lty=3)
plot(c(1:68),g1supply_mean,type="l",ylim=c(0,15),ylab="Cluster 2 Supply")
lines(c(1:68),g1supply_up,lty=3)
lines(c(1:68),g1supply_low,lty=3)
```