



Universitat
de les Illes Balears

MASTER'S THESIS

AIRRM SYSTEM, IMPLEMENTATION OF BUSINESS RULES UNDER CLUSTER ANALYSIS

Robert Fortuny Borysiewicz

**Master's Degree in Màster Universitari en Anàlisi de Dades Massives
en Economia i Empresa (MADM)**

Centre for Postgraduate Studies

Academic Year 2020-21



Universitat
de les Illes Balears

AIRRM SYSTEM, IMPLEMENTATION OF BUSINESS RULES UNDER CLUSTER ANALYSIS

Robert Fortuny Borysiewicz

Master's Thesis

Centre for Postgraduate Studies

University of the Balearic Islands

Academic Year 2020-21

Key words:

AirRM, revenue management, airline, clustering, price strategy

Thesis Supervisor's Name: Heiko Jürgen Rachinger



Abstract

Business Rules is one out of many tools AirRM system provides to deal with revenue management strategies, however, it is a recommended tool to set a solid pricing strategy by using historical data.

In this paper I will discuss, the first step in using AirRM system in an airline I currently work in as a revenue manager. The main topic is the implementation of AirRM's tool Business Rule, its value and the analysis of how many rules are necessary to optimize revenue strategy in a leg, using cluster analysis.

The results, through statistical evidence, show an efficient grouping of flights, according to its booking behaviour and total revenue; and through Business Rule assignment, a useful booking forecasting method for our department.

Key words: clustering, revenue management, forecasting, airline, systems

1. Introduction

1.1. Motivation and Structure

In April 2021 AirRM system arrived to our revenue department. AirRM is a revenue management system that provides many tools to optimize flight revenue, however, by recommendation of the system providers, Business Rules was the best starting point.

This paper resumes the process our Revenue Department has followed to reach to Business Rule's implementation.

First, I will describe some airline nomenclature and the data used in this paper. Second, I'll provide the cluster analysis. And last, through the combination of the extracted clusters and AirRM Business Rule tool, a useful method in order to predict fare class protections.

1.2. General Nomenclature

Numerous airline terminology will be used through subsequent points. Below are some definitions, majority defined by *Jeffrey S. Zickus(1998)*:

- **Leg:** Origin and destination of a flight
- **Load Factor or Occupation:** Percentage of seats, in a flight, booked on departure.

- **Fare Class:** Classification signifying fare/price paid by a passenger.
- **Protections:** Number of permitted bookings in every fare class.
- **Capacity:** Number of seats for sale in an airplane
- **Time Frame:** Time period between seat optimization.
- **Revenue Management (RM):** The allocation of protection levels on every flight based on expected demand.
- **Booking Curve:** General profile by which passengers of a specific type book their reservations.
- **Pickup:** Number of incremental bookings in a specific time frame
- **RAS:** Total revenue of a flight divided between its capacity.
- **AirRM:** Revenue management system
- **NDO:** Day or period of days before flight departure
- **Business Rule:** AirRM tool which distributes protection levels among the fair classes according to each NDO

1.3. Clustering applied to airline revenue management

The vast literature in RM for airline companies deal with much more complex methods which take into account all the Airlines network, dealing with connectivity through their HUBs, introducing the concept of bid price and booking demands combining more than one leg; see *Jeffrey S. Zickus(1998)* or *Bertsimas & de Boer (2003)*.

Some other theory related to the use of clustering techniques in airline RM, focus on grouping fare classes in relation to price levels and set maximum selling amount of these groups for each leg, see *Bertsimas & de Boer (2003)*.

However, I approached a clustering analysis on flights that have similar characteristics year over year, which comes handy mainly for two reasons. First it provides a summary of ticket reservation behaviour and revenue potential for each of the flight grouping; and second, the number of clusters obtained shows the number of pricing strategies needed to capture most of the revenue potential of a leg, optimizing a revenue manager's job.

Hence, for each grouping of similar flights we will be able to set a suitable price strategy with the Business Rule tool. Then assign future flights, which meet with similar characteristics, to BRs and establish predictions; explained in further chapters.

2. Data and Clustering

2.1. Data

2.1.1. Input data

The data used is from all the flights flown in 2019 from a single origin and destination, which in total add up to 3882 flights, an average of 8 flights a day. And for each flight I extracted the following variables:

- Origin
- Destination
- Flight Number
- Total bookings
- Group bookings
- Tour operator bookings
- Booking curve
- RAS
- Total Revenue

For the analysis, the group and tour operator bookings have been extracted from the flights because their prices are fixed for the majority of cases and the amount are negotiated by the commercial department.

As revenue managers, in our case airline, we are in charge of all the public bookings which are more dynamic than the bookings mentioned above; and these are the bookings that I included in the analysis.

The fact that these last bookings are more dynamic, is the main reason why price strategies also have to adapt to their dynamic behaviour, and hence, require numerous price strategies.

Moreover, another variable had to be calculated; the area of the booking curve.

I realized, in previous analysis, that without this variable, flights were grouped only by the revenue results on departure, without taking into account how in advance the bookings were purchased. See Figure 1 as an example.

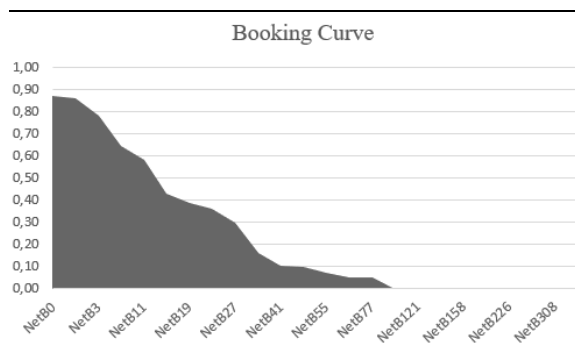


Figure 1: Example of accumulative load factor according to days prior. Ex: NetB27 stands for Net Load Factor 27 days before flight

This figure shows the evolution and accumulation of load factor (percentage of seats booked) in the previous days from departure.

The days previous to the flight, which the tickets are purchased on, is a key information to distinguish differences between flight behaviour.

Then the outliers were removed. The majority of outlier flights were on peak reservation days, which corresponded to national and local holidays.

And no more modifications were made to the data, except from removing bad data or missing values provided by the booking system used by the company.

2.1.2. Clustering in R.

Once I had all the data set ready and structured, I used R statistical programme to study a clustering analysis using k-means, introducing the Load Factor, flight RAS and Booking Curve Area as the three dimensions in order to classify flights. See R code in Figure 2 in Appendix.

K-means clustering (MacQueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centres and then iteratively refining them as follows:

1. Each instance is assigned to its closest cluster centre.

2. Each cluster centre is updated to be the mean of its constituent instances.

The final issue is how to choose k. Using the function “fviz_cluster” is a good option to find a k that minimizes the error. According to R documentation, “fviz_cluster” provides a ggplot2-based elegant visualization of partitioning methods including kmeans [stats package]; pam, clara and fanny [cluster package]; dbscan [fpc package]; Mclust [mclust package]; HCPC [FactoMineR]; hkmeans [factoextra]; observations are represented by points in the plot, using principal components if $\text{ncol}(\text{data}) > 2$. See Figure 3.

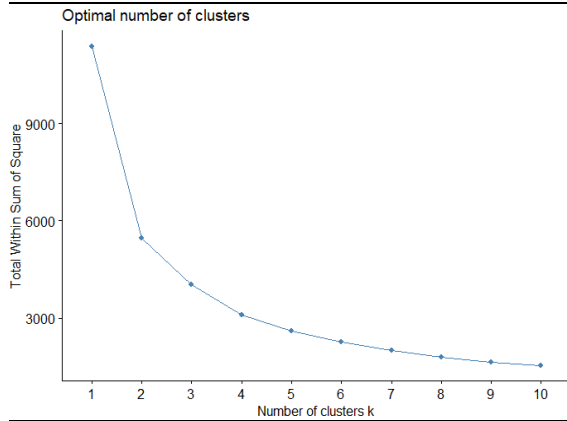


Figure 3: Output of fviz_cluster function. Method = "wss"

As we observe in figure3, clusters over 8 minimize significantly the distances between the cluster centre and the data points, so I defined $k = 8, 9$ and 10 as possible options. Then I used NbClust function (see RDocumentation), which provides 30 indexes for determining the best number of clusters. I set the range of k between 5 and 10. See figure 4:

```
*****
* Among all indices:
* 8 proposed 5 as the best number of clusters
* 5 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 6 proposed 8 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 2 proposed 10 as the best number of clusters
*****
```

Figure 4: Output of NbClust function.

The most proposed cluster is $k = 5$, however in our department we were keener on having a higher number of groupings under the idea that with more clusters, the easier will be to adapt to changes in flight behaviours; also, AirRM allows to deal with many price strategies comfortably.

Hence, out of the upper bound number of k , the major proposed cluster is $k = 8$, which is finally the number of clusters we will move forward with.

Also see the spatial plot of the clustering in Figure 5 in appendix, for $k = 8$.

2.1.3. Output data

The output extracted from R is a document that has the same data as before but, in addition, a column, which classifies each flight in a certain cluster. In the following table we observe the number of flights grouped in each cluster, and ordered by its average RAS, which gives us a hint of the revenue potential of each group.

Cluster	Num. Flights	Ave. RAS
2	246	87
4	424	59
5	155	50
3	416	45
8	675	29
6	375	24
7	935	16
1	562	9

Table 1: Cluster information

We observe in the table a slightly higher concentration of flights in the lower bound average RAS. There are groupings, up to five times, more than another cluster. This does not disturb or look like an inconvenience, as simply it is a feature of the revenue potential of the market in which this leg operates.

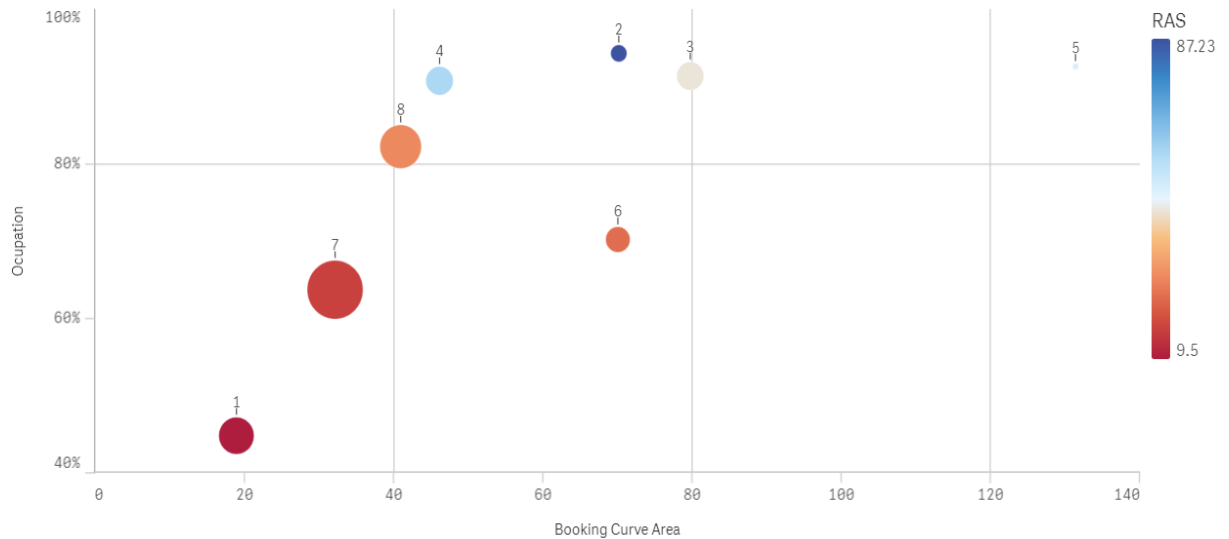


Figure 5: Distribution among Occupation and Booking Curve Area. Size represents the number of flights in the cluster. Represented on Qlik app

Moreover, in Figure 5 I represented the relation between booking anticipation, that can be interpreted from the booking curve area, and occupation.

For describing the type of flight I grouped, this figure can be very useful. For instance, in Cluster 1 and 7 we can find all the flights that have had a very bad results in terms of bookings, due to low demand. However, cluster 6 and 8 show similar results in revenue terms but the difference relies on booking anticipation behaviour. And for the rest clusters of flights with better results in occupation and RAS have also different booking curves, see Figure 6 in appendix.

These figures, 5 and 6, show the consistency in the classification by potential revenue and booking behaviour. They provide the differences in average RAS of each grouping, and in Figure 6, in appendix, provides an accurate classification of flights with similar booking curves of each grouping.

3. Forecasting raw demand models

3.1. Year over year passenger flow correlation

Correlation between traveling flows year over year, under own experience, is one the key statements that revenue managers, in an airline, take into account in order to predict bookings. Summer, Christmas and Easter have similar behaviours every year. There also are weekly trends of corporate travellers and weekend trips, that give insights of potential demand according to day of week and time of departure. Hence, we have as-

sumed annuity replication relation in passenger demands.

However, data from 2019, was chosen, as 2020 data is compromised by lockdowns and restriction policies due to COVID-19. Today, as restrictions are eventually disappearing, we assume that further traveling flows will continue moving to 2019 behaviour. Plus, we will use 2019 bookings as benchmark of further forecasting demand method.

In addition, I have chosen a week in October 2021 to approach to forecast its protection levels (bookings allowed for each price) to test this method.

I have assigned every date in 2021 to the same dates in the same week of the year from 2019; represented in next table. This comparison is available in this route, as every year the number of flights remain constant; on contrary, the analysis would be more complex.

Dates 2019	Dates 2021
21/10/2019	18/10/2021
22/10/2019	19/10/2021
23/10/2019	20/10/2021
24/10/2019	21/10/2021
25/10/2019	22/10/2021
26/10/2019	23/10/2021
27/10/2019	24/10/2021

Once we have clustered all the flights from year 2019, and assigned the flights of 2021 to the corresponding one in 2019; I have used our new rev-



enue management system AirRM to create the Business Rules with the clustering information and assign them to the future flights. This last procedure will be explained in the next point.

3.2. Forecasting raw demand with AirRM programme. Business Rules

AirRM is a continually developing airline management system with intelligent tools that integrate a variety of key information from multiple sources. This enables, currently up to 80 airlines, to make better and profitable decisions.

In our case, the application we will focus on is the use of Business Rules (BR). These are matrixes of Fare Classes (which are classified by letters), in which the rows represent every 2% of the flight's capacity, and each column represent the repartition of protection levels among the fare classes between every NDO noted on the header of the matrix. See Figure 6 in appendix, which represents Cluster 3.

For example, looking at Figure 6, I will explain the meaning of column 17. Each letter corresponds to Fare Class (price), and they are ordered by lower prices at the bottom and higher at the top. As we said, each row represents a 2% of the flight's capacity. Then, the Protections in each Fare Class will be distributed as follows: "Z" 4%, "N" 10%, "A" 18%, P 14%, "Q" 14%, "T" 6%, "U" 6%, "R" 6%, "S" 6%, "H" 4%, "E" 4%, "L" 4%, "B" 4%.

The next column to the left of column 17 is column 14, this states that this previous protection distribution will be implemented during 17 to 14 days prior to the flight; in other words, it will only apply during three days, and then it will change to the distribution from column 14 and so on.

These BR have been created through the following steps. First, we introduced in AirRM a matrix that assigned each flight with its cluster number, and then using the Business Rule Builder tool, AirRM calculates the average protection levels that each grouping of flights have had in each of their NDOs.

Once the 8 BRs are built, via date assignment we have previously done in AirRM, future flights are assigned to a BR that matches with the cluster of the flight assigned from 2019.

3.2.1. Testing with flights from 18th to 24th October 2021.

By means of October's week from 18th to the 24th I will test the results of the efficiency of this

method in forecasting fare class protection levels. In total the analysis is made over 56 flights.

Also, I will include the current prediction that AirRM provides for these flights, which is a default strategy that is manually modified by a revenue manager with no forecasting analysis.

I set as benchmark the final bookings from 2019 week, as we assume in the close future bookings will return to 2019 levels.

Hence the first approach is a general analysis of the sum of all week's protection levels according to price ticket intervals. See Table 2:

TOTAL WEEK 18 th TO 24 th	FARE 5-20€	FARE 20-60€	FARE 60-120€	FARE >120€
CURRENT PREDICTION	8301	1717	639	195
2019 BOOK- INGS	5031	3519	693	490
BR PREDIC- TION	4647	3471	1847	674

Table 2: Bookings and predictions

We observe in the table that in general numbers Business Rules achieve a very good estimate, taking into account all the flights. However, I also randomly chose 4 flights to show some individual results. See Table:

RANDOM FLIGHTS		FARE 5-20€	FARE 20- 60€	FARE 60- 120€	FARE >120€
F1	CURRENT.P	115	66	10	1
	2019 BOOKINS	112	66	6	2
	BR P.	105	66	21	6
F2	CURRENT.P	35	75	60	12
	2019 BOOKINS	23	156	7	0
	BR P.	0	74	68	48
F3	CURRENT.P	167	14	9	5
	2019 BOOKINGS	128	49	8	0
	BR P.	118	46	27	0
F4	CURRENT.P	115	70	6	1
	2019 BOOKINGS	34	126	26	1
	BR P.	61	84	32	13

Table 3: Ranfom flights forecast

The default prediction that nowadays is being implemented has a 0.87 level of correlation with 2019 bookings, and on the other hand, Business Rules predictions show a higher correlation rate, 0.97.

Except for flight F1 and F2, the rest of cases show a better performance under BR strategy. BRs provide a more proportional repartition



among Fare Classes and also matching closely with 2019 bookings. What seems clear, is that BRs perform better estimates than the default strategy provided by AirRM.

However, I still consider this estimate as a raw prediction, but it is a good start in order to set a base demand using BRs, and move forward to its implementation and improvement during use.

4. Conclusion

This paper has provided the analysis performed before an implementation of Business Rules in AirRM. Cluster analysis has been a useful approach in order to find the necessary number of Business Rules a leg required, and through these BRs an improvement in current forecasting protection levels, setting as benchmark 2019 bookings.

Therefore, next step is to move forward to its implementation in our revenue management department and continue improving the forecast analysis under the result



5. Appendix

Figure 2: R Code

```
setwd("c:/Users/robfo/Desktop/TFM")
datos <- read.csv2("datos_r.csv", sep = ";", header = TRUE)
plot(datos$AREA, datos$RAS)

#remove bad data
datos <- datos[!(datos$TotBk==0 | datos$TotBk==0),]
plot(datos$AREA, datos$RAS)

#PATA BCNPMI

Q1 = quantile(datos$RAS, 0.25)
Q3 = quantile(datos$RAS, 0.75)
IQR = Q3 - Q1
boxplot(datos$RAS)$out

#Interquartile Range (IQR) method
#IQR= Q3 - Q1 => And an outlier would be a point below [Q1 - (1.5)IQR] or above
#[Q3 + (1.5)IQR]

#limit_under = Q1 - 1.5*IQR no lo sacamos porque no hay negativo
limit_up = Q3 + 2*IQR
outliers = datos[datos$RAS > limit_up,]
new_data = datos[!(datos$RAS > limit_up),]
newdata <- subset(new_data, select = -outlier)
df = data.frame(newdata$AREA,
                newdata$RAS,
                newdata$LF)
names(df) = c("Area", "RAS", "LF")
df = scale(df)
head(df, 5)
names(df) = c("Area", "RAS", "LF")
#después de sacar los datos ponemos el clustering

library(ggplot2)
library(factoextra)
fviz_nbclust(df, kmeans, method = "wss")

k5 <- kmeans(df, centers = 5, nstart = 25)
k8 <- kmeans(df, centers = 8, nstart = 25)
k10 <- kmeans(df, centers = 10, nstart = 25)
str(k2)
fviz_cluster(k2, data = df)

p1 <- fviz_cluster(k5, geom = "point", data = df) + ggtitle("k = 5")
p2 <- fviz_cluster(k8, geom = "point", data = df) + ggtitle("k = 8")
p3 <- fviz_cluster(k10, geom = "point", data = df) + ggtitle("k = 10")

library(gridExtra)
grid.arrange(p1, p2, p3, nrow = 3)

#Estudio de clustering finalizado

#Sacamos nuevamente los datos con sus clusters
z = data.frame(newdata$FlightLine, newdata$Leg,
              newdata$Dept.Date,
              newdata$Season,
              k8$cluster, newdata$AREA, newdata$RAS, newdata$LF,
              newdata$Cap, newdata$Tot.Rev, newdata$NetRev, newdata$ID)

names(z) = c("FlightLine", "Leg", "Dept.Date", "Season", "cluster",
            "Area", "Rask", "LF", "Cap", "Rev", "Bkgs", "ID")

write.csv(z, "output")
```



Figure 5: Plot of Clustering with $K = 8$

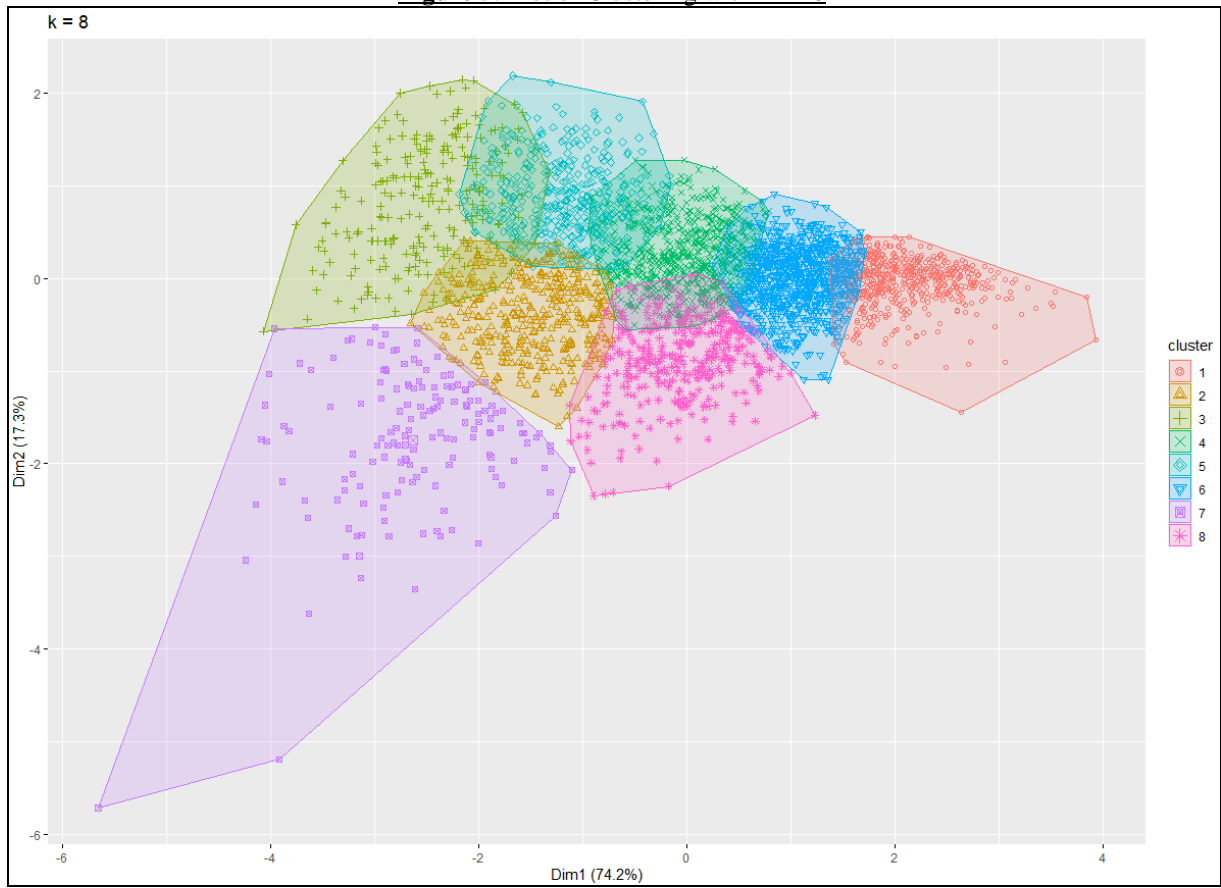




Figure 6: Plot of booking curves randomly picked from each cluster

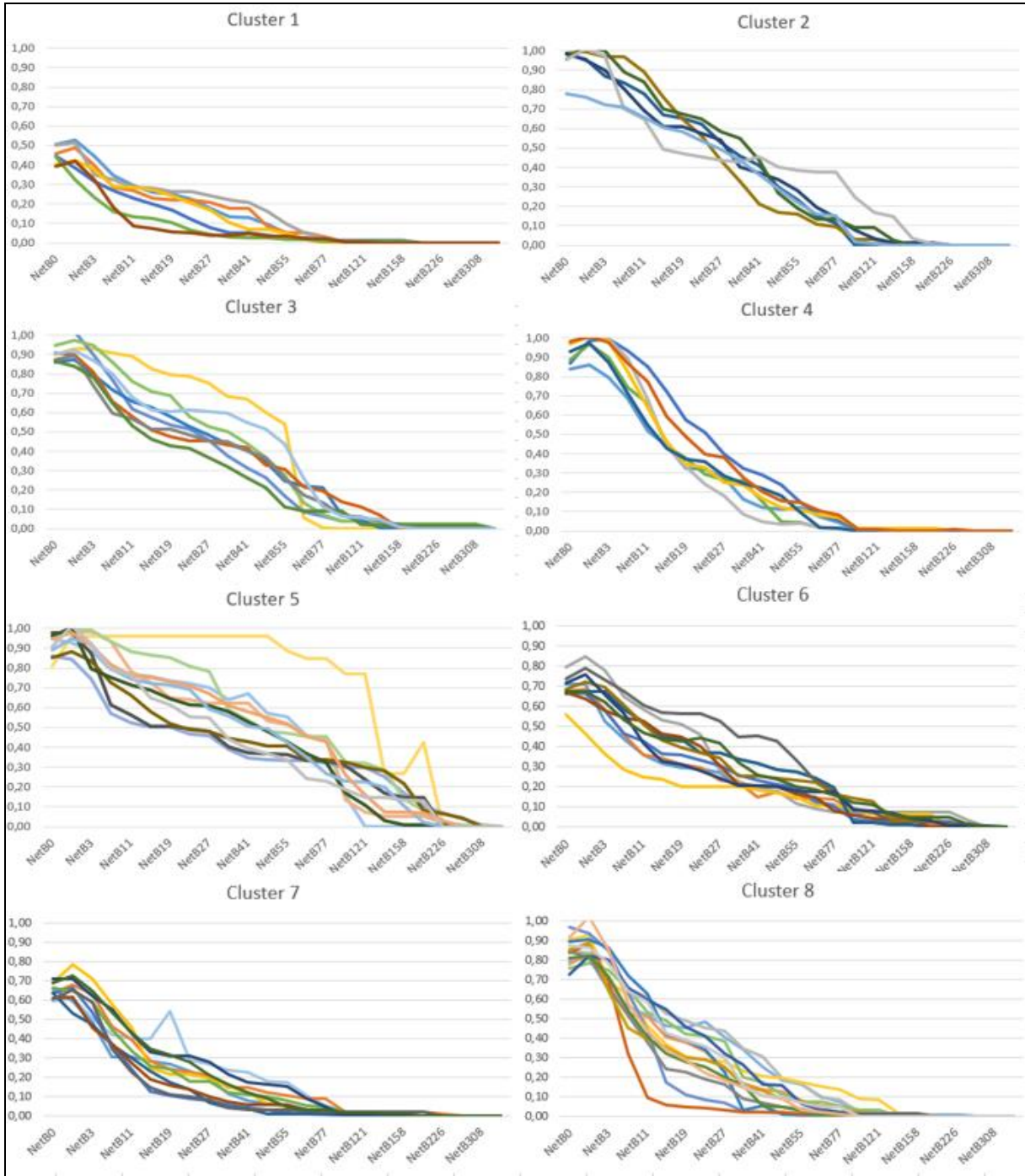




Figure 7: Example of Busines Rule in AirRM

0	2	4	6	8	10	12	14	17	20	28	32	36	40	45	50	60	80	100	120	180	260	365	
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
H	H	H	H	H	H	H	H	H	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
H	H	H	H	H	H	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
H	H	H	H	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
S	S	S	S	S	R	R	R	R	R	R	S	S	S	S	S	S	S	S	S	S	S	S	S
U	U	U	U	U	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
U	U	U	U	U	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
U	U	U	U	U	U	U	U	U	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
U	U	T	T	T	U	U	U	U	U	U	U	U	R	R	R	R	R	U	R	R	R	R	R
U	U	T	T	T	U	U	U	U	U	U	U	U	R	R	R	U	U	R	U	U	U	U	U
T	T	Q	Q	Q	T	T	T	T	T	T	T	T	T	U	U	U	T	U	U	U	U	U	U
T	Q	Q	Q	Q	T	T	T	T	T	T	T	T	T	U	U	T	T	T	T	T	T	T	T
T	Q	Q	Q	Q	Q	Q	Q	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	T	T	T	T	T	T	T	T	T	T
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	T	T	T	T	T	T	T	T	T	T
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	T	T	T	T	Q	T	T	T	T	T
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	T	T	T	Q	Q	Q	Q	Q	Q	Q
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
Q	Q	P	P	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
Q	P	P	P	P	P	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	P	Q	Q	Q	Q	Q
P	P	P	P	P	P	P	P	P	Q	Q	Q	Q	Q	Q	Q	Q	P	P	P	P	P	P	P
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	A	A	A
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	A	A	A	A	A	A
P	P	A	A	A	A	A	A	P	P	P	P	P	A	A	A	P	A	A	A	A	A	A	A



6. References

- MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
- Zickus, J. S. (1998). *Forecasting for airline network revenue management: Revenue and competitive impacts* (Doctoral dissertation, Massachusetts Institute of Technology).
- Bertsimas, D. J., & Boer, S. V. D. (2003). *Advances in airline revenue management and pricing* (Doctoral dissertation, Massachusetts Institute of Technology).
- NbClust function Rdocumentation:
<https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust>
- Fviz_cluster function Rdocumentation:
https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_cluster