Internship Report

Niels Jongs
Student Number: S1168517

1st Supervisor: Prof. dr. Willem Heiser
2nd Supervisor: Dr. Wija Oortwijn

Mathematical Institute
Leiden University
Netherlands
Date: 31-03-2015
E-mail: n.jongs@me.com

Introduction

This report contains an overview of the activities during my two month intern period at Ecorys. In the first section a company profile will be given by briefly describing the main characteristics and history of Ecorys. The second section contains a detailed description of the activities during my intern period. The third section gives a comprehensive description of the intern assignment. the intern assignment consists of using multidimensional reduction techniques for analysing healthcare cost data. In the third section a introduction and the results of this assignment will be given. In the appendix of this document a manual is found for redoing the analyses that is related to the intern assignment. The final and fourth section of the report contains a reflection concerning the intern period, in this section I will describe my learning experience and the problems I encountered during this two month period. Additionally to this section I will describe non proficient skills that require development.

During the development of this document the emphasis was placed on producing a document that is accessible for several disciplines. This is due to fact that both supervisors have excellent skills in different disciplines and the goal is that the document is accessible for several different disciplines . Due to this technical mathematical notation of methods used is not provided in this document. I would thank Ecorys and especially Dr. Wija Oortwijn and her team for offering me chance to get an impression of the consultancy industry and also Prof. Dr. Willem Heiser for the valuable advise he provided during the process of my intern assignment.

Contents

1 Company Profile

Ecorys is a leading european research and consultancy firm, their aim is to deliver real benefit to society through the work they do. Ecorys is active in 120 countries, their headoffice is located in Rotterdam and have permanent offices in the United Kingdom, Spain, Russia, South Africa and much more. The history of Ecorys dates back to 1929 and was formerly known as the Netherlands Economic Institute (NEI). The NEI was founded by a group of businessmen from Rotterdam and their goal was to stimulate the collection and the analysis of economic data. In 1999 NEI merged with Kolpron Consultants and in 2000 the diversification and internationalisation was continued by merging with the British company ECOTEC Research and Consulting. The name Ecorys was chosen as new name for the Group.

Ecorys has clients in as well public, private and not-for-profit sectors across the globe. Their clients consists of:

- National government institutions departments
- Regional and local government institutions
- EU (EC Directorates-General)
- Health (care) institutions
- Transport and logistic companies and organisations

Ecorys is divided in three practices, Finance, Industry and Trade, Connectivity and Social and Regional economics, these practices are again divided in several sectors. During my intern period I was active within the practice Social and Regional Economics and in the sector health. The practices and their sectors are best displayed in figure 1.1. The clients within the sector health range from local and foreign governments to major multinational companies and is a well respected consultancy firm for the european parlement. With their expertise in the quality and efficiency of healthcare they

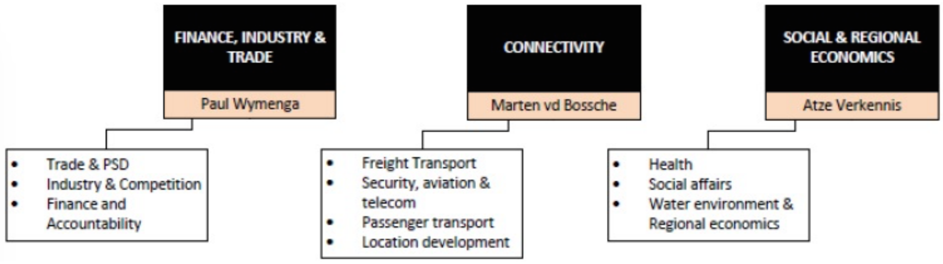


Figure 1.1: Overview practices and the belonging sectors

aim to deliver high quality advice for their clientele and thereby improving healthcare. The range of topics within this sector varies enormous but is best summarised by the following topics:

- The effect of policy changes on the quality of healthcare
- The effect of financial incentives on improving the efficiency of healthcare
- What are the costs and benefits of healthcare developments
- What are the consequence for patients, healthcare professionals, governments and insurance companies of a new policy

2 Activities

The activities during my intern period can be divided in three themes, visualisation of data, data analyses and tender writing. All these themes will be individual described in the following subsections. The emphasis will be on discussed the main tasks and results.

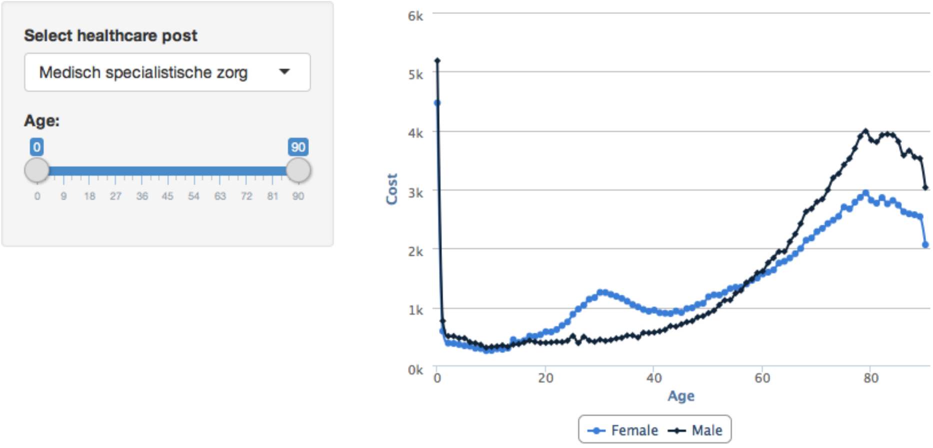
2.1 Data Visualisation

The sector health within Ecorys has a high demand for new and more modern visualisations of data. During my intern period I experimented with several data-sets to develop interactive visualisations of data. All the visualisations developed during this period are produced by using R[1] and supplementary packages such as Rcharts[2], Plotly[3] and Shiny[4]. During the development of these visualisations I simultaneously wondered how employees of Ecorys could reproduce these kind of visualisations and thereby quickly realised that the degree of inadequacy in working with these packages prohibits employees of reproducing the visualisations. This degree of inadequacy is due to two factors:

1. For the use these packages within R advanced coding skills are required. The majority of the employees within the sector health has some basic coding skills outside the scope of R but is lacking knowledge of R
2. The most of these visualisations require data engineering skills. For producing interactive visualisations complex data structures are needed and automatic adjustment of data is required for the interactive components

If the need for these modern visualisation is substantial I recommend Ecorys to attract new employees with advanced skills in developing visualisations of data. It is also possible to outsource the development of visualisations, although the downside is the high expense of outsourcing. In contrast to the high expense of outsourcing is the use of R and the available packages, R has a open-source construction and thereby the expense of its use is zero.

I primarily developed visualisations based on healthcare cost data. The used data is publicly available at www.vektis.nl, a detailed description of the data is given in the section intern assignment since the data is also used for the intern assignment. The data is engineered in such a way that in the interactive menu of the visualisation a healthcare section is selected and represented in a line graph over age and separated by gender. Also it is possible to adjust the range of age. Due to the interactive structure it is not possible to display the results in this report, the interactive application is available at jongs.shinyapps.io and is 25 hours active at request. In figure 2.1 a screenshot of the application is displayed. This figure clearly reveals the functionality of the application in exploring the healthcare expenses per healthcare section over age en grouped by gender. Besides developing the application in shiny I also produced some other less complex interactive visualisations. One of these visualisation concerns the comparison between european countries and the minimum sentence for several offences. For this I developed a bar-chart in which the countries of interest and the type of offence can be selected.



Select healthcare post

Medisch specialistische zorg

Age: 0 – 90

Figure 2.1: Overview practices and the belonging sectors

2.2 Data Analyses

During my intern period I performed some very simple data analysis on hearing aid data. The analysis primary consisted of calculating means and frequencies of occurrence. A major side activity during the analysis was problematic data. The institution responsible for delivering the data kept changing their data and resending it, every time a new data set was delivered the institution had a different reason for changing the data set, the main reasons are summarised below:

- Inconstancy in the notation of missing values, The data partly consisted of numeric values ranging from 0 to 50000 and the institution changed the missing values to 0 or 999 without good communication and thereby these values are included in the calculations.
- Non workable data formats
- Missing categories in the data what suggested that certain categories were not observed. Somehow these categories were observed after changing the data
- Duplicated row entries what led to identical observations and an increases in observations

Eventually a final data set was delivered after a meeting with the institution in context, nevertheless, we decided to inspect the data ourselves. I checked the data on inconsistencies, double observations and unusual categories. Personally I strongly believe that if a third party is responsible for delivering the data it should always be checked on unusual observations before performing any calculations. By doing this time is saved by performing calculations on dirty data.

2.3 Tender Writing

Zorginstituut Nederland(ZiNL) petitioned a tender for several requests concerning the visualisation and analysis of healthcare data. This request was related to a specific program within ZiNL, Zinnige Zorg. The program Zinnige Zorg aims to facilitate access to good and sensible healthcare, no more than necessary and no less than necessary. Thus, both the quality and affordability of care is studied. For this tender I worked on two inquiries, these inquiries concern state-of-the-art visualisation and quantitative analysis in the context of the program within ZiNL. For both inquiries I have writen a small part about how these desires could be fulfilled. For the visualisation part I summarised the capabilities of two software packages, R and SAS, and how these capabilities could be applied within the program Zinnige Zorg. For this tender I also produced a Treemap as an example of the capabilities of R. For this Treemap I used healthcare data and the results are displayed in figure 2.2, the size of the blocks are determined by the amount healthcare activities for each individual hospital. The Treemap is divided in three healthcare activities for breast-cancer patients and the colour of the blocks is determined by the average of healthcare activities per patient for each hospital in 2012. It is clearly visible that the amount of healthcare activities for hospital, and patient varies enormously for between hospital. These kind of visualisations should help ZiNL to quickly identify hospital that deviate from the standard amount of healthcare activities per patient. For example, in figure 2.2 we clearly see that hospital 78 has an average of 3.5 radiological thorax examinations per patient, this obviously exceeds the standard of one examination for each patient per year. For the quantitative analysis part I also summarised the capabilities of R and SAS and discussed how their capabilities could be used for improving healthcare in the context of the program Zinnige Zorg. In this part I also wrote a small proposal concerning the use of publicly available health insurance data to increase the efficiency in healthcare

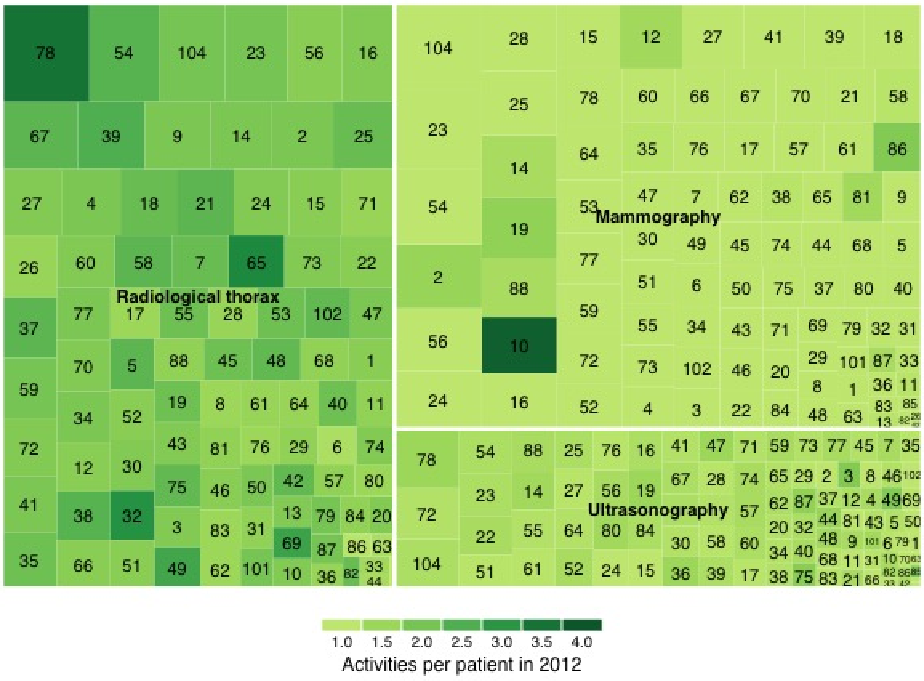


Figure 2.2: Treemap,the blocks are determinted by the amount of healthcare activities and the colour of the blocks by the average of healthcare activities per patient

2.4 R course

Within the sector health within Ecorys the knowledge of R is minimal and therefore my supervisor invited me to give a small R course. This crash course R consisted of a interactive 4 hour during lecture in which the basic functionally of R was discussed. I started with informing the benefits of using software like R-studio to improve the efficiency of R. Next I started with explaining methods for importing data and showed some examples of importing a csv and sav file. Also the use of library was discussed and how to evaluate the quality of the libraries. This was followed by a introduction in making sub-selections of data. As last I discussed how to produce some simple plots and showed some examples of constructing a plot in R.

2.5 XML data

In the last week of my intern a colleague requested my help in merging a large amount of single XML files into comma separated values(.csv) file. The structure of an XML file is best described by root, nodes and trees. Every root contains several nodes that contain information, comparable to a variable with a single value, and a tree is best defined by several variables in a single node. Transforming a single XML input directly into a csv file is problematic since the occurrence trees will produce a nasty solution with a large amount of missing values. To tackle this problem I extracted the data separately for each XML file by using a R script and column binding all the single variables after this I added the new file to data frame containing other xml files inputs.

3 Intern Assignment

For my intern assignment I investigated the use of multidimensional scaling(MDS) techniques on healthcare cost data. My first intention was to use the PROXSCAL procedure in SPSS with Torgerson[5] as initial configuration but due to an error in SPSS it was not possible to use more than 560 entries. By using the Torgerson initial configuration the convergence algorithm starts at a classical multidimensional scaling[6] solution. Due to this initial configuration the solution of the convergence algorithm does not get stuck at a local minima. Eventually I switched to the categorical principal component analysis(CATPCA) procedure since the solution is equivalent to using MDS but it uses different algorithms to obtain the solution, an advantage of CATPCA is the calculation time of the algorithm. In the following subsection a brief description of the data will be given and followed by a description of the CATPCA procedure. The last subsection discusses the results of the CATPCA, a step-by-step manual for reproducing the results is found in the appendix of this document.

3.1 Data

The data is publicly available at www.vektis.nl and contains data concerning healthcare declarations of basic health insurances over the year 2012. These healthcare declarations are divided into 18 different types of care that are distinguished within healthcare. These 18 different healthcare types are:

specialist medical care, pharmacy, secondary mental healthcare, GP registration fee, GP consultation, GP remainder, medical aid devices, dentist, physiotherapy, paramedical care, patient transport sitting, patient transport lying, maternity care, obstetric care, primary psychological care, foreign care, primary support and remaining expenses. A more formal description of the healthcare types is found on the website: www.vektis.nl .Every row entry is profiled by three individual variables, gender, age ranging from 0 to 90 and the first three digits of the postal code. For every first three digits of all postal codes in the Netherlands we have $2 * 90$ observations, two for gender and 0 to 90 for the age range. For each of these profiles the total healthcare expense per type is displayed per row. For the analysis we calculated the average healthcare expense per type for each profile by dividing the total healthcare expenses by the number of insured individuals per profile weighted according to the registration time in 2012. The data file has a total of 136324 observations, to reduce calculation time for CATPCA the data is restructured. This restructuring is done by merging postal codes into a collection of the first two digits and binning the age in 5 categories (0-17, 18-24, 25-44, 45-64 and 65+), this resulted in a data set of 900 observations. The data does not contain missing values, although it contains zero's these values indicate that these profiles did not had healthcare expenses for that specific type of healthcare. The data contained a postal code category of zero, this postal code is a residual category and therefore is deleted from the data-set. The R script for merging the data is found in the appendix.

3.2 Briefly Formalizing CATPCA

Categorical principal component analysis (CATPCA) also known as non-linear PCA is just like PCA a technique that reduces the number of dimensions of a data set into a smaller set of uncorrelated dimensions that represents the data as closely as possible. In contrast to CATPCA PCA is constraint by two limitations, it assumes that the relation between variables is linear and it can only handle numerical variables. CATPCA is suitable for ordinal, nominal and numerical variables that are non-linear related and was first described by Gutmann (1941). CATPCA achieves the same goals as PCA by using a optimal scaling process for the categorical variables. This optimal scaling process is a iterative procedure and converts categorical numbers into numeric values. This is necessary since a variance measure is needed for a PCA procedure and the concept of variance only applies to numeric variables. These categorical quantifications are assigned in such a way that as much as possible of the variance in the quantified variables is accounted for. Noteworthy is that numerical values can also be scaled by an optimal scaling procedure. Even if all the variables are numerical, the relation ship between these variables can be non-linear and therefore CATPCA is most suitable. However, if the relation between variables is linear

	Dimension		
Type	1	2	3
Medical specialist	,918	,243	-,031
Pharmacy	,940	,175	-,054
2nd mental care	-,510	-,064	-,410
GP registration fee	,917	,175	,031
GP consultation	,853	,445	-,076
GP remainder	,778	,274	,164
Medical aid devices	,969	,118	,004
Dentist	,596	-,408	,109
Physiotherapy	,886	-,096	-,181
Paramedical care	,715	-,322	-,193
Patient transport sitting	,696	,150	,132
Patient transport lying	,938	,136	,079
Maternity care	-,424	,871	-,052
Obstetric care	-,431	,864	-,046
1st mental care	-,642	,452	-,229
Foreign care	,297	,038	-,437
Primary support	-,030	,034	,847
Remaining expenses	,914	,067	-,066
Postal code	,011	,021	,430
Gender	,008	-,556	,088
Age	,985	,038	,025

Table 3.1: Factor loadings for all types of healthcare for each dimension

the solution of CATPCA is exactly equal to traditional PCA[8].

For the intern assignment dimension reduction is not the primary goal but the output of CATPCA is used to evaluate clustering in a 2 or 3 dimensional space. The amount of variance explained by each individual dimension is a scale for evaluating the performance of the solution.

3.3 Results

Interpretation of the results is done by inspecting the factor loadings for each dimension, the variance accounted for (VAF) by each dimension and the dimensional representation of the objects. The factor loadings indicate how the variables are represented in the dimensional space of the solution. The final solutions incorporates three dimensions and accounts for 75.27% of the variance in the data, the first dimension 54.55%, second dimension 13.61% and the third dimension accounts for 7.11% of the variance. The variance accounted for by each dimension is confirmed by the factor loadings displayed in table 3.1. The factor loadings clearly reveal that most of the healthcare types are strongly represented in the first dimension. Only maternity and obstetric care are stronger represented in the second dimension with respect to the first dimension. The third dimension is representative for the primary support healthcare, Also first and secondary mental healthcare are fairly represented in the third dimension. These factor loadings partly determine the final solution for the objects in a three dimension space. The three dimensional representation of the objects is found in figure 3.1.

This figure reveals a clear clustering in the first and second dimension

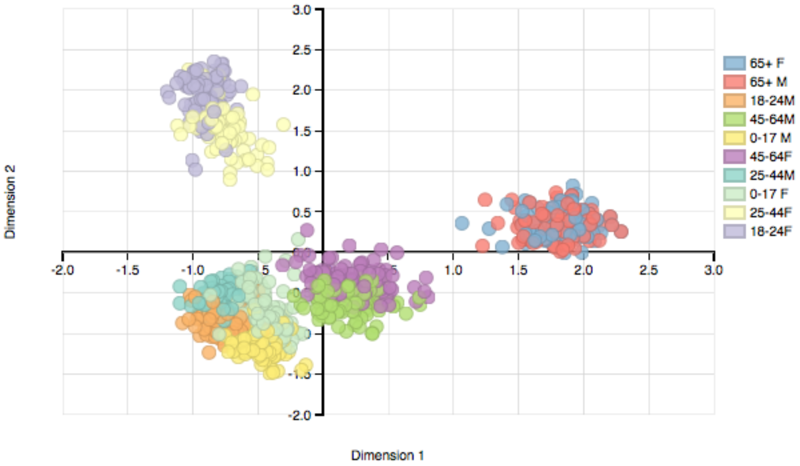


Figure 3.1: CATPCA output: objects plotted in a two dimension space, dimension 1 plotted against dimension 2

(figure 3.1) for age and gender. Individuals in the age category 65+ are clearly higher located in the first dimension, this indicates that their healthcare expenses for medical specialists, pharmacy, GP registration fee, Medical aid devices and patient transport are much higher with respect to the remaining age categories and gender. This effect is also confirmed by the factor loading of age in table 3.1.

The second dimension in figure 3.1 indicates an increase in expenses for maternity and obstetric care for females within the age category of 18-24 and 25-44. These effects are reasonable to expect since females within the age categories of 18-24 and 25-44 generally give birth in this age. Also it is commonly known that increased healthcare costs in elderly people, category 65+, are due to a deterioration in health. The factor loadings in combination with figure 3.1.a also reveal that most of the expenses for secondary mental healthcare are due to individuals in the age category of 0-17 and 18-24 and mainly for boys, this reflects the trend in the increasing amount of attention deficit hyperactivity disorder(ADHD) diagnoses in young boys. Primary mental healthcare expenses are mainly due to females.

In figure 3.1.b the first dimension is plotted against the third dimension, this figure reveals some clustering but not as clearly as in figure 3.1. The third dimension mainly represents postal codes and primary support healthcare. This indicates increased healthcare expenses in primary support healthcare for some sub-selection of postal codes. Inspection of the data revealed that the clustering in the third dimension is due to geo-spatial separation of the Netherlands, this separation is best described by a clustering for the province Friesland and the rest of the Netherlands. This indicates that

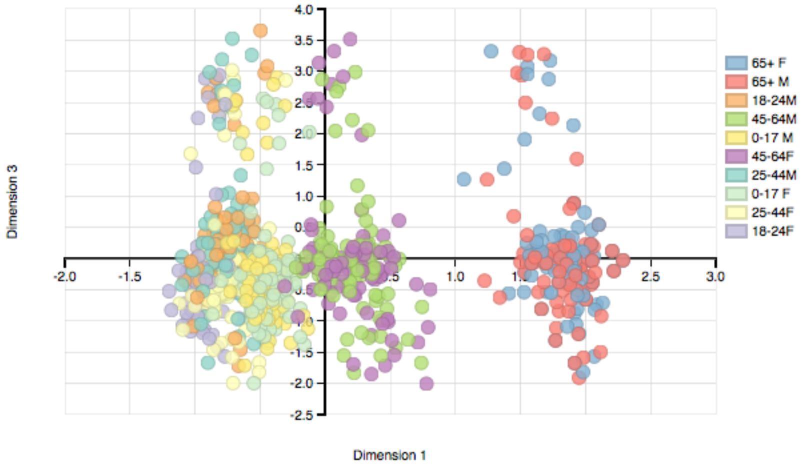


Figure 3.2: CATPCA output: objects plotted in a two dimension space, dimension 1 plotted against dimension 3

the province Friesland has increased healthcare expenses on primary support healthcare with respect to the rest of the Netherlands. The factor loadings for the third dimension also indicates that the rest of the Netherlands has higher healthcare expenses in foreign and second mental healthcare. This also indicates that the trend in diagnosing ADHD is less in the province Friesland, although I was not capable of finding concrete evidence supporting this suggestion but the results in the third dimension requires extra research to find out why there is a difference in primary support healthcare, foreign care and second mental healthcare between the Netherlands and the province Friesland.

4 Reflection

4.1 Communication

In the last two years I primary communicated with fellow students and teachers about mathematical statistics and never experienced problems when discussing certain models and their output. During my intern period I quickly realised that a more general form of communication about statistics is needed to explain how certain techniques function and how to interpret the results. Due to this I concluded that some terms require explanation in a non-technical context.

So during meetings with colleagues I tried to use non-technical communication about the results of my activities, however, this new form of communication was harder then I expected. Eventually a substantial amount of time was invested in considering how to explain my results without using technical verbalisation. Although, I must conclude that during this eight week period I became more acquainted in communicating without using technical terminology. Nevertheless, sometimes it is necessary to explain some technical details, in that case I learned it is important to use simple language and clear explanation. for example, during my intern I first tried to conduct a multidimensional scaling technique in SPSS (PROXSCAL) and use Torgerson as initial configuration. Unfortunately PROXSCAL in combination with the Torgerson configuration did not function with more then 561 observations. In my final presentation I needed to explain to colleagues why the use of Torgerson is of great importance, so I explained that Torgerson starts with a classical multidimensional scaling solution that does not require a iterative algorithm and guarantees that the solution is not stuck in a local minima. I explained this local minimal by using a drawing of how the optimal solution is found. Due to this intern period I became more capable in communicating about statistics in a general way and I realised that the relevance of this is enormous since it allows me to communication my results to different disciplines.

4.2 Organisation

Although a schedule was constructed at the start of my intern I realised that two major characteristics of organising requires development.

Problems with SPSS and PROXSCAL quickly became eminent and much time was invested in solving this problem while maintaining a workable solution for colleagues. Due to these problems the constructed schedule became worthless and fell several weeks behind. Eventually a solution was found but I did not catch up with the schedule again. I think when making a schedule it is useful to account for problems during the development of the results. A more practical solution in my case is to construct a new schedule since a large amount of time was lost.

I also realised that I am insufficiently in keep track of the steps in the process of developing a solution. During one of the meetings with my supervisor it became clear that I was incapable of explaining why I made certain decisions in the process of developing a solution. After this meeting I invested more time in keeping track of my decisions so that I am capable of supporting these decisions. Also this is useful in the communication with others since it allows others to give feedback on decisions made during the process of developing a solution. A simple solution for this is to use notes in the syntax of SPSS or in the scripts of R.

Overall the most important learning experience is to what extend my mathematical background is of value in the consultancy industry. I expected that the consultancy industry was more dependent on the use of statistics, however, during my intern I realised that statistics is a small part of the industry. The competitive structure of winning a contract is what surprised me the most, therefore a substantial amount of time is spend in producing tenders that represent the company as best as possible. This competition is a bit like a mating dance, show your best moves and feathers and your chance of winning is optimal.

References

[1] R Core Team (2014). *R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/*

[2] Ramnath Vaidyanathan (2013). *rCharts: Interactive Charts using Polycharts.js. R package version 0.4.2*

[3] Chris Parmer, Scott Chamberlain, Karthik Ram, Toby Hocking, Marianne Corvellec, Pedro Despouy and Carson Sievert (). *plotly: Interactive,publication-quality graphs online. R package version 0.5.25. https://github.com/ropensci/plotly*

[4] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2015). *shiny: Web Application Framework for R. R package version 0.11.1. http://CRAN.R-project.org/package=shiny*

[5] Torgerson, W. S. (1958). *Theory & Methods of Scaling. New York: Wiley*

[6] Torgerson, W. S. (1952). *Multidimensional scaling: I. Theory and method. Psy- chometrika, 17, 401-419*

[7] Guttman, L. (1941). The quantification of a class of attributes: A theory and a method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 319–348). New York: Social Science Research Council.

[8] Linting, M., Meulman, J. J., Groenen, J. F., van der Kooij, A. J. (2007). Nonlinear Principal Components Analysis: Introduction and Application. *Psychological Methods, 12,* 336–358. DOI: 10.1037/1082-989X.12.3.336.

[9] Meulman, J. J. & Heiser, W. J. (2012). IBM SPSS Categrories [PDF file]. Available from http://www.sussex.ac.uk/its/pdfs/SPSS_Categories_21.pdf.

Appendix

R-Script

```r
setwd("~/Stat /Intern/MDS") # setworking directory
# Load data:
dat <- read.csv2("data1.csv",header=T)
dat$ZIP <- as.factor(dat$ZIP) # ZIP as categorical variable

levels(dat$AGE) <- c(0:90)
dat$AGE <- as.numeric(dat$AGE) # Make age a numeric variable

# Calculate average healthcare expenses per profile
dat[6:23] <- apply(dat[6:23],2,function(x){
  x <- as.numeric(x)/dat$COUNT_Y
})

# Merge data into smaller age categories
out <- matrix(NA,1,21)
for(i in 1:nlevels(dat$ZIP)){
  print(c(i,nlevels(dat$ZIP)))
  k <- levels(dat$ZIP)[i]
  test <- dat[dat$ZIP==k,]
  Fem <- test[test$SEX=="V",]
  Male <- test[test$SEX=="M",]
  out <- rbind(out,
  rbind(age0.17M <- as.integer(c(k,1,0,
          apply(Male[Male$AGE >= 0 & Male$AGE < 18,][,6:23],2,mean))),
         age0.17F <- as.integer(c(k,0,0,
           apply(Fem[Fem$AGE >= 0 & Fem$AGE < 18,][,6:23],2,mean))),
         age18.24M <- as.integer(c(k,1,1,
            apply(Male[Male$AGE >= 18 & Male$AGE < 25,][,6:23],2,mean))),
         age18.24F <- as.integer(c(k,0,1,
            apply(Fem[Fem$AGE >= 18 & Fem$AGE < 25,][,6:23],2,mean))),
         age25.44M <- as.integer(c(k,1,2,
            apply(Male[Male$AGE >= 25 & Male$AGE < 45,][,6:23],2,mean))),
         age25.44F <- as.integer(c(k,0,2,
            apply(Fem[Fem$AGE >= 25 & Fem$AGE < 45,][,6:23],2,mean))),
          age45.64M <- as.integer(c(k,1,3,
             apply(Male[Male$AGE >= 45 & Male$AGE < 65,][,6:23],2,mean))),
          age45.64F <- as.integer(c(k,0,3,
             apply(Fem[Fem$AGE >= 45 & Fem$AGE < 65,][,6:23],2,mean))),
           age65M <- as.integer(c(k,1,4,
              apply(Fem[Male$AGE >= 65,][,6:23],2,mean))),
        age65F <- as.integer(c(k,0,4,
           apply(Fem[Fem$AGE >= 65,][,6:23],2,mean)))))
}

out[is.nan(out)] <- NA
out <- (na.omit(out))
out <- as.data.frame(out) # make dataframe instead of matrix
out$V2 <- as.factor(out$V2)
levels(out$V2) <- c(”F”,”M”) # give correct level to sex variable
out$V3 <- as.factor(out$V3)
levels(out$V3) <- c(”0-17”,”18-24”,”25-44”,”45-64”,”65+”)
out$V1 <- as.factor(out$V1)

# create new collum names
colnames(out)[1:21] <- c(”ZIP”,”SEX”,”AGE”,”COST_SPEC”,
                         ”COST_FARM”,”COST_2GGZ”,
                         ”COST_GP”,”COST_GPCON”,
                         ”COST_GPLFT”,”COST_AID”,”COST_DEN”,
                         ”COST_PARAF”,”COST_PARALFT”,”COST_TRANS”,
                         ”COST_TRANL”,”COST_BABY”,”COST_BABY2”,
                         ”COST_1GGZ”,”COST_FOR”,”COST_1SUP”,”COST_O”)

out$ZIP <- as.numeric(as.character(out$ZIP))
out <- out[out$ZIP >= 100 & out$ZIP <= 999,] # delete postal codes 0,

# merging postal codes for 3 numbers to 2 numbers
MZIP <- NULL
indx.D <- 90
indx.U <- 100
while(indx.U <= 991){
  indx.D <- indx.D + 10
  indx.U <- indx.U + 10
  temp <- out[out$ZIP >= indx.D & out$ZIP < indx.U,]
  list <- split(temp, list(temp$SEX,temp$AGE))
  inzip <- NULL
  for(j in 1:length(list)){
    inzip <- rbind(inzip,cbind(data.frame(ZIP = indx.D,
    SEX = list[[j]][1,2],
    AGE = list[[j]][1,3]),t(apply(list[[j]][,4:21],2,mean))))
  }
  MZIP <- rbind(MZIP,inzip)
}
levels(MZIP[,1]) <- factor(10:99)
# make .csv file to load in spss
```

```
write.csv(MZIP,"SPSS.csv") # make CSV file
```

Manual CATPCA

Introduction

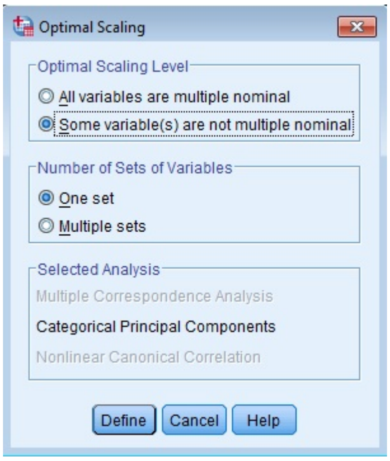
This manual contains a step by step procedure for performing the categorical principal component analysis (CATPCA) in SPSS as preformed in the intern assignment. For more detailed information and extra examples I refer to the official document of IBM SPSS categories 21 manual by Meulman and Heiser(2012). This official IBM SPSS document will be provided to my supervisor at Ecroys. This document contains information about extra function that are not used in the solution for the intern assignment. These function include discretize and missing in the CATPCA menu. The discretization option allows the user to select a recoding for the variables with approximately a normal distribution. The missing option allows the user to select a method for dealing with missing values in CATPCA.

The data used in this document is the merged data as discussed in my inter report. This data is available on request and is found on the server of Ecroys with the following path:

W:\2310 Health\Projecten\NL2310-30175 Data visualisatie\Data\data_SHORT.SAV

Performing CATPCA

1. When the data set is open in SPSS choose the following in the menu to produce CATPCA output:
   Analyze → Dimension Reduction → Optimal Scaling



Optimal Scaling

Optimal Scaling Level

○ All variables are multiple nominal

● Some variable(s) are not multiple nominal

Number of Sets of Variables

● One set

○ Multiple sets

Selected Analysis

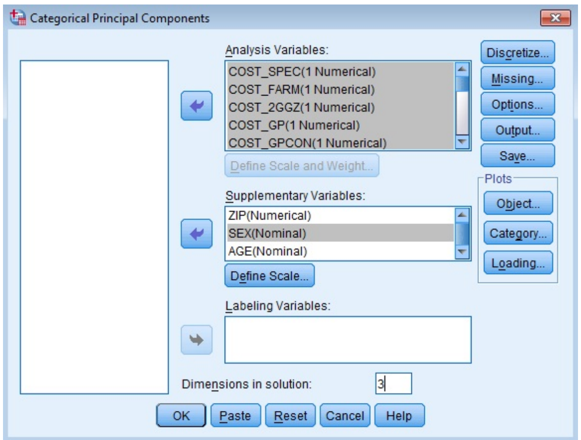
Multiple Correspondence Analysis

Categorical Principal Components

Nonlinear Canonical Correlation

   - In the optimal scaling menu select: *Some variables(s) are not multiple nominal* and click on *define*

2. Select all numerical variables as *analysis variables* and the three nominal variables as *supplementary variables* (ZIP, SEX, AGE).



Categorical Principal Components

Analysis Variables:

COST_SPEC(1 Numerical)
COST_FARM(1 Numerical)
COST_2GGZ(1 Numerical)
COST_GP(1 Numerical)
COST_GPCON(1 Numerical)

Supplementary Variables:

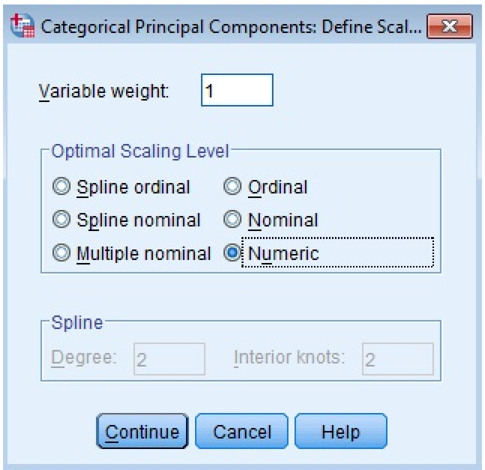
ZIP(Numerical)
SEX(Nominal)
AGE(Nominal)

Labeling Variables:

Dimensions in solution: 3

   - Set *dimensions in solution* to three since we are interested in a three dimensional solution.

3. Select all variables in *analysis variables* en click on *define scale and weight.*



Categorical Principal Components: Define Scal...

Variable weight: 1

Optimal Scaling Level

○ Spline ordinal ○ Ordinal

○ Spline nominal ○ Nominal

○ Multiple nominal ● Numeric

Spline

Degree: 2 Interior knots: 2

   - Select *numeric* in the field *optimal scaling* level and click on *continue*

4. Now select the variable ZIP in *supplementary variables* and click on *define scale*

   - In this dialog box select *numeric* in the field *optimal scaling level* and click on *continue*

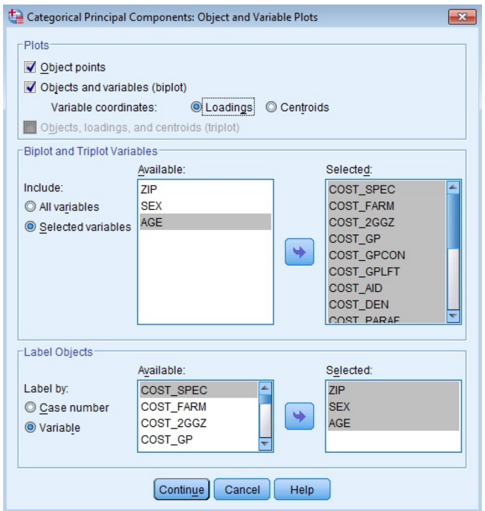
5. Now select the variables SEX and AGE in *supplementary variables* and click on *define scale*

   - In this dialog box select *nominal* in the field *optimal scaling level* and click on *continue*

6. Now select *output* and in the field *table* select *object scores* and *variance accounted for* and click on *continue.* The object scores are the coordinates for every row entry in a three dimensional space

   - If the user would like to export the object scores select *save* and in the field *object scores* select save to the *active data set* to add 3 columns of coordinates to the active data set and click on continue

7. In the following step the output of CATPCA will be adjusted to produce relevant plots. Click in the field *plots* on *object*



Categorical Principal Components: Object and Variable Plots

Plots

☑ Object points

☑ Objects and variables (biplot)

Variable coordinates: ● Loadings ○ Centroids

□ Objects, loadings, and centroids (triplot)

Biplot and Triplot Variables

Include: ○ All variables ● Selected variables

Available: ZIP, SEX, AGE

Selected: COST_SPEC, COST_FARM, COST_2GGZ, COST_GP, COST_GPCON, COST_GPLFT, COST_AID, COST_DEN, COST_PARAF

Label Objects

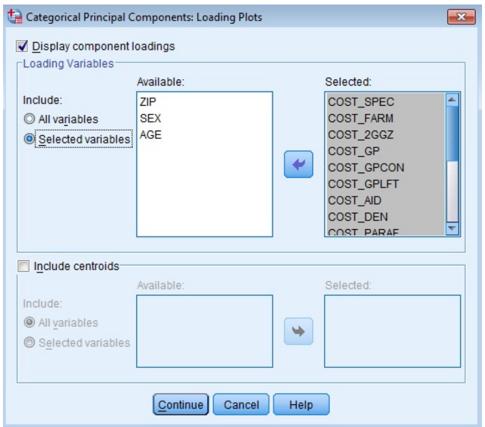
Label by: ○ Case number ● Variable

Available: COST_SPEC, COST_FARM, COST_2GGZ, COST_GP

Selected: ZIP, SEX, AGE

- In the field *plots* select *object points and object and variables(biplot)*, under object and variables(biplot) select *loadings*
- In the field *biplot and triplot variables* select *selected variables* and add all the variables except ZIP, SEX and AGE to the box *selected*
- In the field *label objects* select *variable* and add the variables ZIP, AGE and SEX to the box *selected* and click on *continue*

8. Click on *loadings* in the field Plots



Categorical Principal Components: Loading Plots

☑ Display component loadings

Loading Variables

Include: ○ All variables ● Selected variables

Available: ZIP, SEX, AGE

Selected: COST_SPEC, COST_FARM, COST_2GGZ, COST_GP, COST_GPCON, COST_GPLFT, COST_AID, COST_DEN, COST_PARAF

□ Include centroids

Available:

Selected:

Include: ● All variables ○ Selected variables

- select *display component loadings* and in the field *loading variables* select *selected variables.* select all the variables except ZIP, AGE, SEX and add the variables in the box *selected* and click on *continue*

9. Click on *paste* to get the syntax and run the syntax for the output.