Application of Cluster Analysis and Multidimensional Scaling on Medical Schemes Data

by

Ian Roux

Thesis presented in partial fulfillment of the requirements for the degree of Master of Commerce

at

Stellenbosch University.

Department of Statistics and Actuarial Science
Faculty of Economic and Management Sciences

Supervisor: Prof. N.J. Le Roux

Co-supervisor: Prof. H. McLeod

Date: December 2008

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 10 October 2008

Summary

Cluster analysis and multidimensional scaling (MDS) methods can be used to explore the structure in multidimensional data and can be applied to various fields of study. In this study, clustering techniques and MDS methods are applied to a data set from the health insurance field. This data set contains information of the number of medical scheme beneficiaries, between ages 55 to 59, that are treated for certain combinations of chronic diseases. Clustering techniques and MDS methods will be used to describe the interrelations among these chronic diseases and to determine certain clusters of chronic diseases.

Similarity or dissimilarity measures between the chronic diseases are constructed before the application of MDS methods or clustering techniques, because the chronic diseases are binary variables in the data set. The calculation of dissimilarities between the chronic diseases is based on various dissimilarity coefficients, where a different dissimilarity coefficient will produce a different set of dissimilarities. One of the aims of this study is to compare different dissimilarity coefficients and it will be shown that the *Jaccard, Ochiai, Baroni-Urbani-Buser, Phi* and *Yule* dissimilarity coefficients are most suitable for use on this particular data set.

MDS methods are used to produce a lower dimensional display space where the chronic diseases are represented by points and distances between these points give some measurement of similarity between the chronic diseases. The classical scaling, metric least squares scaling and nonmetric MDS methods are used in this study and it will be shown that the nonmetric MDS method is the most suitable MDS method to use for this particular data set. The Scaling by Majorizing a Complicated Function (SMACOF) algorithm is used to minimise the loss functions in this study and it was found to perform well.

Clustering techniques are used to provide information about the clustering structure of the chronic diseases. Chronic diseases that are in the same cluster can be considered to be more similar, while chronic diseases in different clusters are more dissimilar. The robust clustering techniques: PAM, FANNY, AGNES and DIANA are applied to the data set. It was found that AGNES and DIANA performed very well on the data set, while PAM and FANNY performed only marginally well.

The results produced by the MDS methods and clustering techniques are used to describe the interrelations between the chronic diseases, especially focussing on chronic diseases mentioned in the

same body system rule (Council for Medical Schemes, 2006, p.6-7). The cardiovascular diseases: *Cardiomyopathy*, *Coronary Artery Disease*, *Dysrhythmias* and *Hypertension* are strongly related to each other and to *Asthma*, *Chronic Obstructive Pulmonary Disease*, *Diabetes Mellitus Type* 2, *Hypothyroidism* and *Hyperlipidaemia*. The gastro-intestinal conditions: *Crohn's Disease* and *Ulcerative Colitis* seem to be very strongly related. *Bipolar Mood Disorder*, *Epilepsy*, *Schizophrenia* and *Parkinson's Disease* also seem to be strongly related. The chronic diseases: *Addison's Disease* and *Diabetes Insipidus* also show a very strong relation to each other.

Opsomming

Trosontleding- en multidimensionele skalering (MDS) metodes kan gebruik word om die struktuur van multidimensionele data te ondersoek en kan in verskeie toepassingsgebiede aangewend word. Trosontleding- en MDS metodes gaan in hierdie studie op ’n datastel afkomstig uit die mediese skema bedryf toegepas word. Hierdie datastel bevat inligting oor die getal mediese skema lede, tussen die ouderdomme 55 en 59, wat behandeling ontvang vir sekere chroniese siektes. Die trosontleding- en MDS tegnieke gaan gebruik word om verwantskappe tussen die chroniese siektes te beskryf.

Ongelyksoortigheidsmaatstawwe tussen die chroniese siektes moet eers bepaal word voordat die trosontleding- en MDS tegnieke toegepas kan word, want die chroniese siektes is binêre veranderlikes in die datastel. Verskeie ongelyksoortigheidsmaatstawwe kan gebruik word in die praktyk. Verskillende ongelyksoortigheidsmaatstawwe mag egter verskillende resultate oplewer. Een van die doelwitte met hierdie studie is om verskeie ongelyksoortigheidsmaatstawwe te vergelyk en dit is bevind dat die *Jaccard*, *Ochiai*, *Baroni-Urbani-Buser*, *Phi* en *Yule* ongelyksoortigheidsmaatstawwe die mees toepaslike vir gebruik op hierdie datastel is.

MDS metodes word gebruik in hierdie studie om ’n laer dimensionele figuur te produseer, waar die chroniese siektes as punte in die figuur voorgestel word en afstande tussen die punte ’n aanduiding verskaf van die gelyksoortigheid tussen die chroniese siektes. Klassieke skalering, metriese kleinste kwadrate skalering en nie-metriese MDS gaan in hierdie studie toegepas word. Dit is bevind dat die nie-metriese MDS metode die mees gepaste MDS metode is vir toepassing op hierdie datastel. Die “*Scaling by Majorizing a Complicated Function*” (SMACOF) algoritme word gebruik in hierdie studie om sekere verliesfunksies te minimeer en hierdie algoritme het baie bevredigend vertoon.

Trosontledingsmetodes word gebruik in hierdie studie om inligting oor die trosvorming van die chroniese siektes te verskaf. Chroniese siektes wat in dieselfde tros aangetref word, is meer soortgelyk, terwyl chroniese siektes in verskillende trosse weer meer van mekaar verskil. Die trosontledingsalgoritmes: PAM, FANNY, AGNES en DIANA word in hierdie studie toegepas op die datastel. Dit is bevind dat AGNES en DIANA baie goed vaar op die datastel, terwyl PAM en FANNY bevredigend vaar.

Die resultate van die trosontleding- en MDS tegnieke word gebruik om die verwantskappe tussen die chroniese siektes te beskryf, met spesiale klem wat geplaas word op chroniese siektes wat in dieselfde "*body system rule*" voorkom (Council for Medical Schemes, 2006, p.6-7). Daar is bevind dat die hartsiektes: "*Cardiomyopathy*", "*Coronary Artery Disease*", "*Dysrhythmias*" en "*Hypertension*" sterk verwant is aan mekaar en aan "*Asthma*", "*Chronic Obstructive Pulmonary Disease*", "*Diabetes Mellitus Type* 2", "*Hypothyroidism*" en "*Hyperlipidaemia*". Dit lyk ook of "*Crohn's Disease*" en "*Ulcerative Colitis*" baie nou verwant is aan mekaar. "*Bipolar Mood Disorder*", "*Epilepsy*", "*Schizophrenia*" en "*Parkinson's Disease*" vertoon ook 'n sterk verwantskap. Verder lyk dit ook of "*Addison's Disease*" en "*Diabetes Insipidus*" baie nou verwant is aan mekaar.

Acknowledgements

- I would like to thank Prof. N.J. Le Roux, my supervisor, for his constant encouragement and invaluable guidance.

- I also would like to thank Prof. Heather McLeod, my co-supervisor, for her help and efforts to make the data set available.

- I want to thank the following members of the REF Study 2005 Consortium: André Deetlefs, Andre Meyer, Brett Mill and Pierre Robertson for their efforts to make the data available.

- I would like to thank Old Mutual Healthcare (Pty) Ltd., Medscheme (Pty) Ltd., Discovery Health (Pty) Ltd. and Metropolitan Health Group (Pty) Ltd. for providing the data.

- I would like to thank the Harry Crossley foundation for their funding.

- I want to thank all my friends, especially those living at 15 Karee Street, for their friendship and support.

- I want to thank the Gertenbach's on Belvedere farm for their hospitality, support and encouragement.

- I would like to thank Alechia for her love, friendship and constant support throughout my studies.

- Finally, but most significantly, I want to thank my parents for their encouragement, love and support throughout my studies and throughout my life.

Contents

Chapter 1

Introduction

1.1 Introduction

Consider a data set consisting of *p* variables measured for each of *n* objects. Well-known visual displays, like scatterplots for example, can be used to explore the data if *p* = 2 or 3. However, it becomes more difficult to display the relationships among the objects, the relationships among the variables or the variable-object relationships when *p* becomes larger than three. Multidimensional scaling (MDS) and cluster analysis are two statistical techniques that can be used to explore the structure of multidimensional data. Although both clustering techniques and MDS methods are used to describe how the objects or variables in a certain data set are related, the two approaches are different.

Borg & Groenen (2005) describe MDS as a “method that represents measurements of similarity (dissimilarity) between objects as distances between points in a lower dimensional space”. For example, points in this lower dimensional (usually two dimensional) space that lie close to one another indicate that the corresponding objects have a greater similarity. The objects are not classified into clusters, but distances between the points can be used to assess the relationships among the objects. However, the interpretation of the distances is rather subjective, as different observers might reach different conclusions with regard to possible clusters. Furthermore, it is impossible to capture all the variability of the multidimensional data in a lower dimensional display.

Clustering techniques, on the other hand, produce information about the clustering structure of the objects. Anderberg (1973) mentions that “the objective of the clustering techniques is either to group data objects or the variables into clusters, in such a way that the elements belonging to the same cluster resemble each other, whereas elements in different clusters are dissimilar” . The clustering structure found is based on actual similarity measurements between the objects. Clustering techniques are also well suited to high dimensional data. One of the possible drawbacks of clustering techniques is that it can be difficult to interpret the relationships among the objects. For example, many clustering methods only produce a list of clusters and their objects as output that might be difficult to interpret. However, certain visual displays have been developed to overcome this drawback, and are known as *clusplots* (Pison *et al.*, 1998). Clusplots use MDS to produce a configuration of points in a two dimensional display and use clustering techniques to find a clustering structure. The clusters are then represented by ellipses in the lower dimensional display. Clusplots illustrate that it is useful to supplement MDS configurations with information about the clustering structure of the objects. However, clusplots do not usually use the correct aspect ratio. This means that a unit change in the horizontal direction of the graphical display is not equal to a unit change in the vertical direction and distances between objects in the graphical display can therefore not be fully appreciated, which is vital for interpretation. Care must therefore be taken by the user when implementing clusplots to ensure that the correct aspect ratio is used. Another limitation of clusplots is that clusplots do not provide information about the variables when the objects are clustered. Biplots (Gower & Hand, 1996) can be used to overcome this limitation by providing a graphical display with information on both the objects and the variables simultaneously.

Cluster analysis and MDS methods are applied to various fields including medical research, ecology, economics, psychometrics, chemometrics, and many more (Kaufman & Rousseeuw, 1990). It will be shown that the clustering techniques and MDS methods can also be used on a South African data set in the health insurance field. This data set is known as the *REF Study 2005* raw data set and it was provided by the Risk Equalisation Fund (REF) Study 2005 consortium. The REF will equalise risk with regard to the age and disease profile of medical scheme beneficiaries in South Africa in relation to the Prescribed Minimum Benefit (PMB) conditions (Council for Medical Schemes, 2005, p.1). Risk equalisation is not unique to South Africa and has been introduced by many countries that include: Belgium, Finland, Germany, Czech Republic, Germany, Ireland, Netherlands, Norway, Russian Federation, Sweden, Switzerland, United Kingdom, Israel, Australia, New Zealand, Colombia, Canada and the United States of America (Parkin & McLeod, 2001).

The REF Study 2005 raw data set was collected from four major medical scheme administrators in South Africa in order to produce the REF Contribution Table 2007 (RETAP, 2007). These four administrators provided services to approximately 4.25 million beneficiaries, which represented approximately 63% of South Africa’s total medical scheme beneficiaries reported in 2005 (RETAP, 2007, p.16). The data set contains information about the costs and exposure for unique combinations of age, gender and disease, where the exposure is measured in beneficiary months. There are 19 different age bands and 25 REF chronic diseases. It is important to investigate how the chronic diseases are related to one another, where a strong relationship between two chronic diseases indicates that they tend to co-occur often. Results from such an investigation could confirm medical knowledge of relationships among certain chronic diseases. Also, the results could be used if the REF Entry and Verification Criteria, with regard to multiple chronic diseases, are changed in the future. The REF Entry and Verification Criteria will be discussed in Section 1.2.

Scatterplots and histograms cannot be used to display the relationships among all the chronic diseases simultaneously, because the REF Study 2005 raw data set is multidimensional. It is therefore necessary to use MDS methods or clustering techniques to display the relationships among the chronic diseases, and it will be shown that these methods are well suited to this purpose. Both MDS methods and clustering techniques will be used, as Gordon (1999) points out that "such combined analyses will reduce the element of subjectivity in assessing results. If similar conclusions would be reached from the separate analyses, one can have more confidence in the accuracy of the results."

1.2 The REF Entry and Verification Criteria and the Body System Rules

Medical schemes in South Africa have to submit monthly data to the REF on a quarterly basis. The data that medical schemes submit to the REF are in the form of *REF Grid Count* data (Council for Medical Schemes, 2006). A medical scheme will then either receive or make a payment to the REF, dependent on this REF Grid Count data. The payments will depend on the age, gender of the beneficiaries and the number of beneficiaries treated for a certain chronic disease. The type of chronic disease and the number of multiple chronic diseases will also influence the payments. It is therefore important that the REF enforces strict Entry and Verification Criteria of the data submitted, to ensure a fair system. Note however that the REF is not yet in full operation, and currently operates in a "shadow period" where medical schemes submit data but no money actually changes hands (RETAP, 2007).

There was a change in the REF Entry and Verification Criteria, applicable to all medical scheme administrators from 1 January 2007 (Council for Medical Schemes, 2006, p.6-7) and known as the REF Entry and Verification Criteria version 2. The change regards the following eight rules, which will be called *body system rules* in this study. These rules will affect the REF Grid Count data if beneficiaries are treated for certain multiple chronic diseases. The body system rules are the following:

1. For Count purposes, only one of the following chronic respiratory diseases can be assigned to the same beneficiary: Chronic Obstructive Pulmonary Disease, Asthma and Bronchiectasis (COP, AST or BCE).

2. For Count purposes, only one of the following cardiovascular diseases can be assigned to the same beneficiary: Cardiomyopathy, Coronary Artery Disease, Dysrhythmias and Hypertension (CMY, IHD, DYS or HYP).

3. For Count purposes, only one of Hypertension or Chronic Renal Failure can be assigned to the same beneficiary (HYP or CRF).

4. For Count purposes, only one of the following gastro-intestinal conditions can be assigned to the same beneficiary: Crohn’s disease and Ulcerative Colitis (CSD or IBD).

5. For Count purposes, only one of Bipolar Mood Disorder and Schizophrenia can be assigned to the same beneficiary (BMD or SCZ).

6. For Count purposes, only one of Multiple Sclerosis, Bipolar Mood Disorder and Epilepsy can be assigned to the same beneficiary (MSS, BMD or EPL).

7. For Count purposes, only one of Systemic Lupus Erythematosus and Rheumatoid Arthritis can be assigned to the same beneficiary (SLE or RHA).

8. Diabetes Mellitus Type 1 and Type 2 cannot co-occur.

Various chronic diseases mentioned in the same body system rule present clinical diagnostic challenges in the absence of convincing clinical evidence (RETAP, 2007, p.74). The body system rules will prevent over counting of these chronic diseases and will help to ensure a fair system.

An important distinction is made between whether a life (defined as a beneficiary month) is *treated* for a certain disease or *diagnosed* with a certain disease, with *treated* being the more strict criterion. A life is treated for a certain disease when the life meets the REF Entry and Verification criteria version 2 (Council for Medical Schemes, 2006, p.19-35). The REF Entry and Verification criteria version 2 require specific diagnosis-related information and proof of treatment for each of the REF chronic diseases and HIV, before the life can be classified as a "treated patient".

The *REF Study 2005* raw data set is based on the REF Entry and Verification criteria version 2 for CDL (chronic disease list) conditions (Council for Medical Schemes, 2006, p.19-35) and known as TREATED data (RETAP, 2007, p.18). The chronic disease list is a list of 25 chronic diseases as prescribed in the *Amendment to the General Regulations made in terms of the Medical Scheme Act of 1998*, (Act 131 of 1998, p.35).

It must be noted that the last body system rule was actually applied to the TREATED REF Study 2005 raw data set. The REF Entry and Verification criteria version 2 state that "for REF purposes, Type 1 and Type 2 diabetes cannot occur concurrently. Evidence of use of oral euglycaemic or hypoglycaemic medicines automatically leads to the classification of a diabetic case as Type 2." (Council for Medical Schemes, 2006, p.26). The Diabetes Mellitus Type 1 and Type 2 chronic diseases therefore do not show any co-occurrence in the TREATED REF Study 2005 raw data set. The other body system rules were not applied to the TREATED REF Study 2005 raw data set.

1.3 The Medical Scheme 55-59 Data Set

MDS methods and clustering techniques will be used to describe the relationships among the chronic diseases, which are binary variables in the TREATED REF Study 2005 raw data set. However, the application of the MDS methods and clustering techniques on the TREATED REF Study 2005 raw data set is not a straightforward task.

Firstly, measurements of similarity (dissimilarity) between the binary variables must be calculated and there are at least 43 possible similarity coefficients that could be used (Hubàlek, 1982). The choice of similarity coefficient is very important as it will heavily influence the graphical displays produced by the MDS methods and clustering techniques. Some of these similarity coefficients will be more appropriate to use than others, depending on the actual data set. Cox & Cox (2001) recommend using several different similarity coefficients in practice, hoping for robustness against a specific choice.

Secondly, several MDS methods can be applied to the data set and these methods will produce different configurations of points in the graphical displays. Several clustering techniques can also be used, and these techniques will also produce different clustering structures. Gordon (1999) mentions that it is rarely the case that only one method of analysis is particularly appropriate for a certain study. It therefore seems more appropriate to use several MDS methods and several clustering techniques, and not just a single method.

“Different age bands have different disease profiles, with multiple conditions becoming more apparent later in life” (Council for Medical Schemes, 2005, p.26). This means that MDS methods and clustering techniques are very likely to produce different results for the different age bands. Therefore, it is important to produce results according to particular age bands. This study will focus only on one of these age bands, as it is not practical to repeat the same methodology for all 19 age bands, remembering that the methodology will depend on several similarity coefficients, several MDS methods and several clustering techniques. However, other age bands can be analysed in a similar way in future. According to McLeod (2007), leader of the REF Study 2005 consortium, the age band of 55 to 59 years contains the most lives treated for chronic diseases, out of all 19 age bands, in the TREATED REF Study 2005 raw data set. This age band also shows the occurrence of every single chronic disease and many of these chronic diseases co-occur, which makes this a suitable age band to use when describing the relationships among the chronic diseases. The *Medical Scheme 55-59* data set is a subset of the TREATED REF Study 2005 raw data set, where only the age band of 55 to 59 years is considered. The results from the complete analysis of this particular age band can be used to determine which choices of various similarity coefficients, MDS methods and clustering techniques are more appropriate, and these results can then be used if similar analyses of the other age bands are carried out in future. Still, the use of only one age band in this study has the drawback that the important relationship between age and chronic disease cannot be explored.

The variables of the Medical Scheme 55-59 data set used in this study, along with the respective codes, are given in Table 1.1. The REF chronic diseases and HIV are binary variables where “1” indicates that the lives are treated for the corresponding disease and “0” indicates that the lives are not treated for the particular disease. The variable NON is also a binary variable where “1” indicates that the lives are not treated for any of the 25 REF chronic diseases or HIV and the lives are not claiming for maternity benefits. The variable *Nr.Lives* indicates the number of lives that are treated for the same combination of diseases. This means that each row of the Medical Scheme 55-59 data set represents multiple identical lines of individual lives that are treated for the same unique combination of diseases. The last three variables are continuous variables that measure the average

Table 1.1: *Variables of the Medical Scheme 55-59 data set that will be used in this study.*

Codes	Description
NON	Not treated for any chronic disease or HIV, and no maternity benefit was claimed.
ADS	Addison's Disease
AST	Asthma
BCE	Bronchiectasis
BMD	Bipolar Mood Disorder
CMY	Cardiomyopathy
COP	Chronic Obstructive Pulmonary Disease
CRF	Chronic Renal Failure
CSD	Crohn's disease
DBI	Diabetes Insipidus
DM1	Diabetes Mellitus Type 1
DM2	Diabetes Mellitus Type 2
DYS	Dysrhythmias
EPL	Epilepsy
GLC	Glaucoma
HAE	Haemophilia
HYL	Hyperlipidaemia
HYP	Hypertension
IBD	Ulcerative Colitis
IHD	Coronary Artery Disease
MSS	Multiple Sclerosis
PAR	Parkinson's Disease
RHA	Rheumatoid Arthritis
SCZ	Schizophrenia
SLE	Systemic Lupus Erythematosus
TDH	Hypothyroidism
HIV	HIV/AIDS
Nr.Lives	Number of lives that have the same gender and disease combination
Hospital	Average DTP Hospital costs
Medicine	Average CDL Medicine costs
Related	Average Total Related costs

cost for each combination of chronic diseases that were treated. The average cost is described in three categories: *DTP Hospital cost*, *CDL Medicine cost* and *Total Related cost*. The Total Related costs in the REF Study 2005 were defined to be the sum of DTP Related costs, CDL Related costs and Capitation costs (RETAP, 2007, p.28). Appendix B contains the first 30 rows of the Medical Scheme 55-59 data set.

The *Male MS (55-59) data set* and *Female MS (55-59) data set* are two subsets of the Medical Scheme 55-59 data set, reflecting male and female lives. These two data subsets will be used to describe differences between male and female lives, with regard to the occurrence of chronic diseases.

1.4 Aims of this Study

MDS methods and clustering techniques will be applied to the Medical Scheme 55-59 data set, where an attempt will be made to describe the relationships among the 25 REF chronic diseases, HIV and NON. Chronic diseases and HIV that co-occur more often in the Medical Scheme 55-59 data set are expected to show a stronger relationship in the MDS and clustering graphical displays. The results produced by these methods can then be used to investigate whether chronic diseases mentioned in the same body system rule are strongly related. It is also important to investigate which other chronic diseases, not mentioned in the body system rules, also have strong relationships.

Differences between male and female lives with regard to the occurrence of chronic diseases also need to be investigated. MDS methods can be used to produce separate displays for male and female lives where points in these displays represent the various chronic diseases. Clustering techniques can also be used to produce separate displays for male and female lives, where these displays will show the clustering structures of the chronic diseases. These graphical displays will be used to describe differences between male and female lives.

This study will only focus on the 55-59 year age band and several similarity coefficients, several MDS methods and several clustering techniques will be used. The results from the study, for this particular age band, can then be used to determine which choices of similarity coefficient, MDS method and clustering technique are more appropriate to use for other age bands, should a similar analysis of the other age bands be carried out in the future.

1.5 Methods used in this Study

All of the statistical techniques used in this study were carried out by using the programming and graphical environment of the computer package *R*, available from `http://www.cran.r-project.org`. Some of the standard *R* functions were used in this study. Also, several new functions were constructed for use in this study. Details of these functions are provided in Appendix A.

The Medical Scheme 55-59 data set will be discussed in detail in Chapter 2. Histograms and tables will be used to describe the cost aspects and the occurrence of the various chronic diseases.

The calculation of dissimilarities between binary variables, based on several similarity coefficients, will be discussed in Chapter 3. These dissimilarities will provide some measurement of how similar (dissimilar) chronic diseases are to each other. The dissimilarities will be used in Chapter 4 and Chapter 5, where an attempt will be made to display the different chronic diseases. Some metric and Euclidean properties of the various dissimilarity coefficients will also be discussed in Chapter 3. The metric and Euclidean properties of these dissimilarity coefficients are important to consider when the chronic diseases are displayed in Chapters 4 and 5 (Gower & Legendre, 1986). Some of these displays will only be appropriate to use if certain metric requirements are met.

Metric and nonmetric MDS methods will be discussed in Chapter 4. The main difference between metric and nonmetric MDS methods is that metric MDS methods require dissimilarity coefficients with metric properties and nonmetric MDS methods only use the ranking order of the dissimilarities and can therefore accept dissimilarity coefficients with or without metric properties. However, the display space for both metric and nonmetric MDS methods will be chosen to be Euclidean. This means that the displays will have a direct distance-like interpretation. Points representing chronic diseases that lie close to each other indicate that these chronic diseases tend to co-occur often. Nonmetric MDS methods produce graphical displays where the configuration of points is determined by minimising a certain loss function. Scaling by Majorizing a Complicated Function (SMACOF) is a technique that was developed to minimise complicated functions and will be used in this study to minimise loss functions. It will be shown how SMACOF can be used for nonmetric and metric MDS methods (Borg & Groenen, 2005). The performance of metric and nonmetric MDS methods, in the case of the Medical Scheme 55-59 data set, will also be compared. This will be discussed in Chapter 4.

Chapter 5 is concerned with the cluster analysis of the Medical Scheme 55-59 data set, where clustering structures of the chronic diseases will be produced. This study will focus on the robust clustering algorithms of Kaufman & Rousseeuw (1990), as these clustering algorithms are very well suited to analysing the Medical Scheme 55-59 data set. The graphical representations accompanying the Kaufman and Rousseeuw algorithms, showing the clustering structures of the chronic diseases, are of particular importance. These clustering structures group the REF chronic diseases, HIV and NON into clusters, in such a way that the diseases belonging to the same cluster resemble each other, whereas diseases in different clusters are dissimilar. The clusplot graphical display (Pison *et al.,* 1998) will also be discussed. The dissimilarities between chronic diseases calculated in Chapter 3, will be used as input structures for the clustering algorithms. It is important to consider the performance of the various similarity (dissimilarity) coefficients. The performance of the clustering algorithms of Kaufman & Rousseeuw will also be compared to the more traditional clustering methods.

1.6 Notation and Abbreviations

x	Scalar
X	Random variable
$\underline{x}$	Column vector
$\mathbf{X}^T$	The transpose of matrix $\mathbf{X}$
$\underline{x}^T$	The transpose of vector $\underline{x}$
$\mathbf{X}: n \times p$	A matrix consisting of n rows and p columns
$\mathbf{I}_n$	Identity matrix with n rows and n columns
s_{ij}	Similarity coefficient between objects i and j
d_{ij}	Dissimilarity coefficient between objects i and j
δ_{ij}	Distance between objects i and j in the lower dimensional display space
$s(i)$	Silhouette value of object i
L	The best fitting r-dimensional subspace
SVD	Singular value decomposition
MDS	Multidimensional scaling
PCA	Principal component analysis
SMACOF	Scaling by majorizing a complicated function
PAM	Partitioning around medoids
CLARA	Clustering large applications

FANNY	Fuzzy clustering
AGNES	Agglomerative nesting
DIANA	Divisive analysis
MONA	Monothetic analysis
AC	Agglomerative coefficient
DC	Divisive coefficient
REF	Risk Equalisation Fund

1.7 Definitions and Terminology used in this Study

CDL Chronic disease list conditions, which is a list of 25 chronic diseases as prescribed in the *Amendment to the General Regulations made in terms of the Medical Scheme Act of 1998,* (Act 131 of 1998, p.35).

Chronic disease A disease that is listed as a CDL.

Age Age last birthday on 1 January.

Life A single beneficiary month.

Treated patient Patient that is treated for a certain disease, which means that the patient meets the REF Entry and Verification criteria version 2 for CDL conditions (Council for Medical Schemes, 2006, p.19-35). The REF Entry and Verification criteria version 2 require specific diagnosis-related information and proof of treatment for each of the chronic diseases and HIV.

TREATED data Data based on the REF Entry and Verification criteria version 2 for CDL conditions.

Medical Scheme 55-59 data A subset of the TREATED REF Study 2005 raw data set provided by the REF Study 2005 consortium. This subset contains information regarding the average costs and the number of lives treated for unique combinations of gender and disease, where the lives are 55 to 59 years old.

Male MS (55-59) data set A subset of the Medical Scheme 55-59 data set, containing male lives.

Female MS (55-59) data set A subset of the Medical Scheme 55-59 data set, containing only female lives.

Chapter 2

Description of the Medical Scheme 55-59 Data Set

2.1 Introduction

The Medical Scheme 55-59 data set, introduced in Chapter 1, will be described further in this chapter by using univariate statistical techniques. Histograms and tables will be used to describe the occurrence of the chronic diseases and the costs of treatment.

2.2 Univariate Description of the Medical Scheme 55-59 Data Set

The Medical Scheme 55-59 data set contains 1 636 rows of unique combinations of gender and disease and represents 2 199 367 lives. The Male MS (55-59) data set contains 750 rows of unique combinations of disease and represents 1 056 917 male lives. The Female MS (55-59) data set contains 886 rows and represents 1 142 450 female lives.

Table 2.1 contains the number of male and female lives that were treated for certain chronic diseases. Most of the lives in the Medical Scheme 55-59 data set were not treated for any of the chronic diseases or HIV. Observe from Table 2.1 that the chronic diseases HAE, DBI, ADS, BCE and MSS did not occur often, while the chronic diseases AST, TDH, DM2, HYL and HYP occurred more often. A similar observation was made for the 2002 data where HYP, HYL and AST were found to be some of the most common chronic diseases (Council for Medical Schemes, 2005, p.37). Observe from Table 2.1 that there is a difference between the number of male and female lives treated for certain diseases, but a direct comparison cannot easily be made because the total number of male and female lives in the Medical Scheme 55-59 data set are different. It is therefore more convenient to use chronic disease rates to make a direct comparison between male and female lives. RETAP (2007) used a “chronic rate per 1,000 lives”. The chronic disease rate that will be used in this study is defined in Definition 2.1.

Definition 2.1 Chronic disease rate

- The number of lives per 1 000 that are treated for a certain chronic disease, where *lives* are defined as beneficiary months. (For example, a female treated for HAE for four months is counted as four female lives.)

Table 2.1: The *number of male and female lives treated for a certain chronic disease. Many lives are treated for multiple chronic diseases. The number of lives that are not treated for any chronic disease or HIV, coded as NON, is also provided.*

Variables	Male	Female	Total
HAE	0	4	4
DBI	21	53	74
ADS	116	91	207
MSS	19	215	234
BCE	156	326	482
CSD	362	329	691
SLE	140	631	771
SCZ	356	772	1 128
CRF	1 094	598	1 692
IBD	897	808	1 705
PAR	1 029	703	1 732
BMD	802	1 226	2 028
HIV	2 451	1 236	3 687
GLC	3 984	4 716	8 700
DYS	5 321	4 043	9 364
DM1	7 050	4 386	11 436
EPL	5 954	6 252	12 206
COP	7 215	6 396	13 611
RHA	4 401	11 423	15 824
CMY	13 120	11 229	24 349
IHD	28 615	13 887	42 502
AST	20 169	30 328	50 497
TDH	7 498	61 755	69 253
DM2	49 123	34 066	83 189
HYL	104 036	68 270	172 306
HYP	196 341	224 680	421 021
NON	763 387	813 432	1 576 819
Total	1 056 917	1 142 450	2 199 367

Table 2.2: *Chronic disease rates (see Definition 2.1) of the Medical Scheme 55-59 data set.*

Disease	Male lives	Female lives	All lives
HAE	0.0000	0.0035	0.0018
DBI	0.0199	0.0464	0.0336
ADS	0.1098	0.0797	0.0941
MSS	0.0180	0.1882	0.1064
BCE	0.1476	0.2854	0.2192
CSD	0.3425	0.2880	0.3142
SLE	0.1325	0.5523	0.3506
SCZ	0.3368	0.6757	0.5129
CRF	1.0351	0.5234	0.7693
IBD	0.8487	0.7073	0.7752
PAR	0.9736	0.6153	0.7875
BMD	0.7588	1.0731	0.9221
HIV	2.3190	1.0819	1.6764
GLC	3.7695	4.1280	3.9557
DYS	5.0345	3.5389	4.2576
DM1	6.6703	3.8391	5.1997
EPL	5.6334	5.4724	5.5498
COP	6.8265	5.5985	6.1886
RHA	4.1640	9.9987	7.1948
CMY	12.4135	9.8289	11.0709
IHD	27.0740	12.1555	19.3247
AST	19.0829	26.5465	22.9598
TDH	7.0942	54.0549	31.4877
DM2	46.4776	29.8184	37.8241
HYL	98.4335	59.7575	78.3435
HYP	185.7677	196.6651	191.4283

For example, the female HAE chronic disease rate is calculated as follows from Table 2.1:

$$\frac{\text{Number of female lives treated for HAE}}{\text{Total number of all female lives}} \times 1000 = \frac{4}{1\,142\,450} \times 1000 = 0.0035$$

Table 2.2 contains the chronic disease rates of male and female lives of the Medical Scheme 55-59 data set. The values in Table 2.2 show that male and female chronic disease rates differ dramatically for some chronic diseases. For example, the chronic disease MSS occurred 10 times more often in female lives. The chronic disease rates are graphically displayed in Figures 2.1 and 2.2. It can be observed from Figure 2.1 and Figure 2.2 that female lives have much higher chronic disease rates for DBI, MSS, BCE, SLE, SCZ, BMD, RHA and TDH, while the male lives have higher chronic disease rates for CRF, PAR, IHD and HYL. The chronic disease with the highest chronic disease rate is HYP.

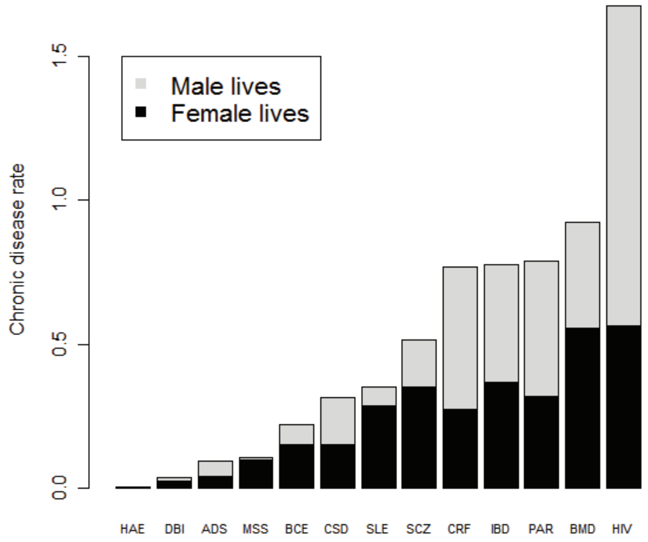


Figure 2.1: *Chronic disease rates of chronic diseases that occurred least often. The breakdown for male and female lives is also displayed. See Definition 2.1.*

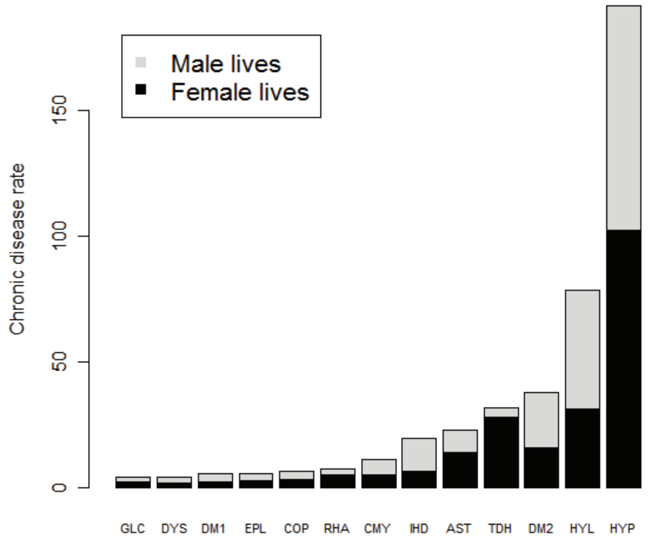


Figure 2.2: *Chronic disease rates of chronic diseases that occurred most often.*

Some lives are not treated for any chronic diseases, others for a single chronic disease, while others are treated for multiple chronic diseases. Table 2.3 shows the number of lives that are treated for a certain number of chronic diseases. It can be observed from Table 2.3 that approximately 72% of all lives are not treated for any chronic disease. Approximately 17% of all lives are only treated for one chronic disease.

Table 2.3: *Number and percentage of male and female lives that were treated for a certain number of chronic diseases.*

Number of chronic diseases (including HIV)	Male lives	Female lives	Male percentage	Female percentage
None	763 387	813 432	72.2277	71.2007
1	174 065	210 677	16.4691	18.4408
2	82 491	86 171	7.8049	7.5426
3	28 564	25 002	2.7026	2.1885
4	6 839	5 718	0.6471	0.5005
5	1 300	1 205	0.1230	0.1055
6	225	214	0.0213	0.0187
7	43	31	0.0041	0.0027
8	3	0	0.0003	0.0000

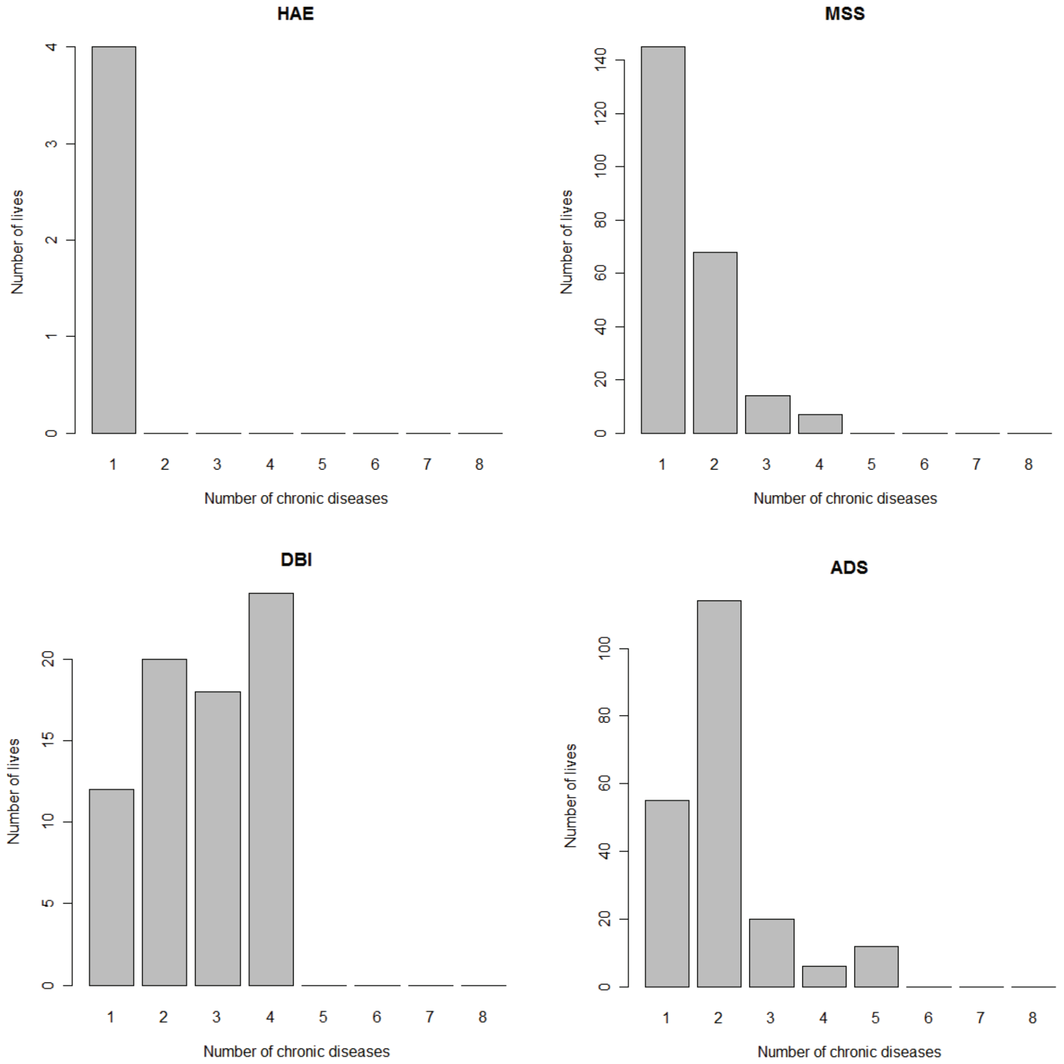


Figure 2.3: *Number of lives treated for a particular chronic disease and also treated for none, one, two, three, …, seven other chronic diseases. The number of lives corresponding to 1 is the number of lives treated only with one particular chronic disease. The number of lives corresponding to 2 is the number of lives that are treated for the particular chronic disease and only one other chronic disease.*

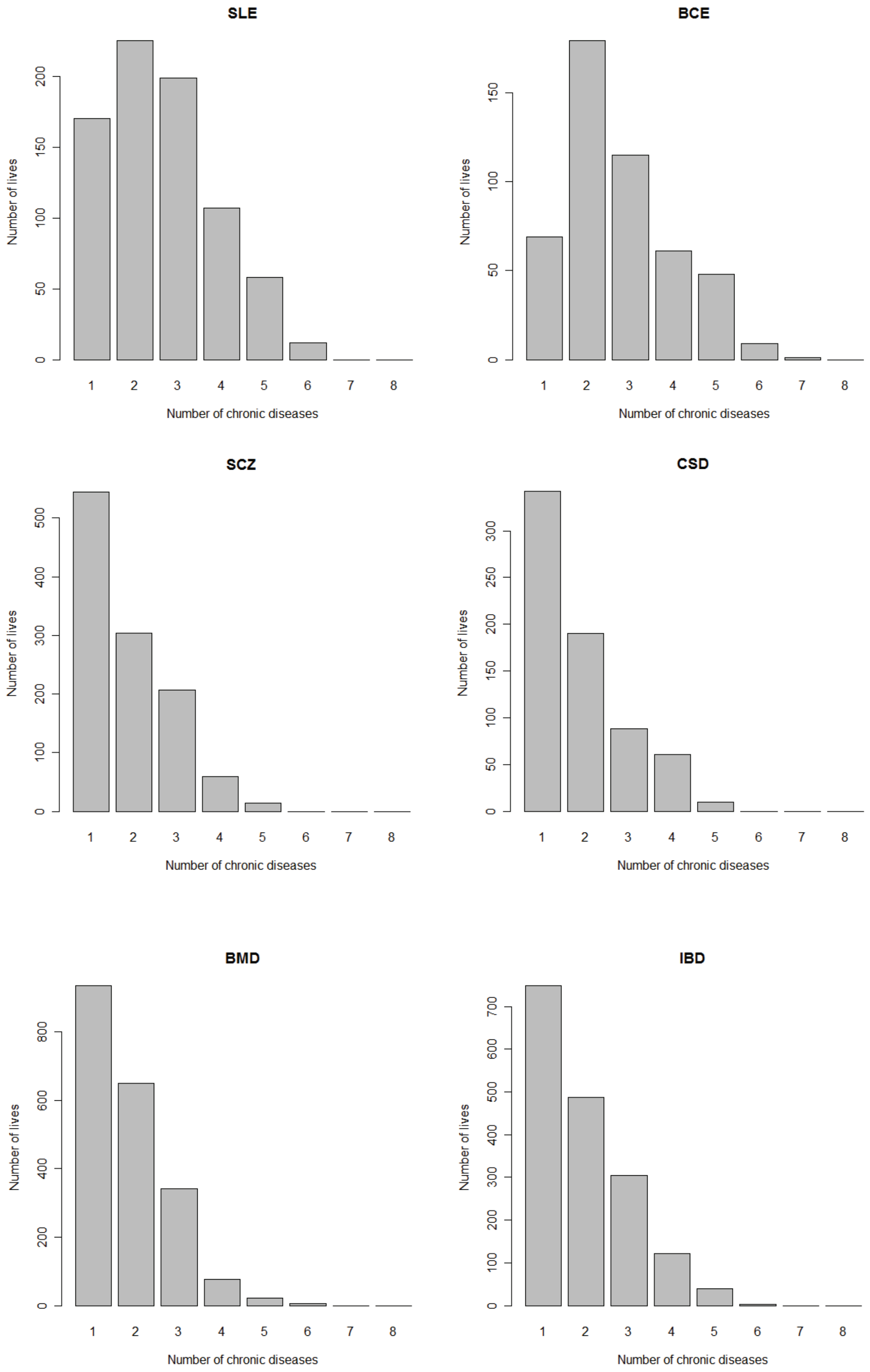


Figure 2.3: Continued.

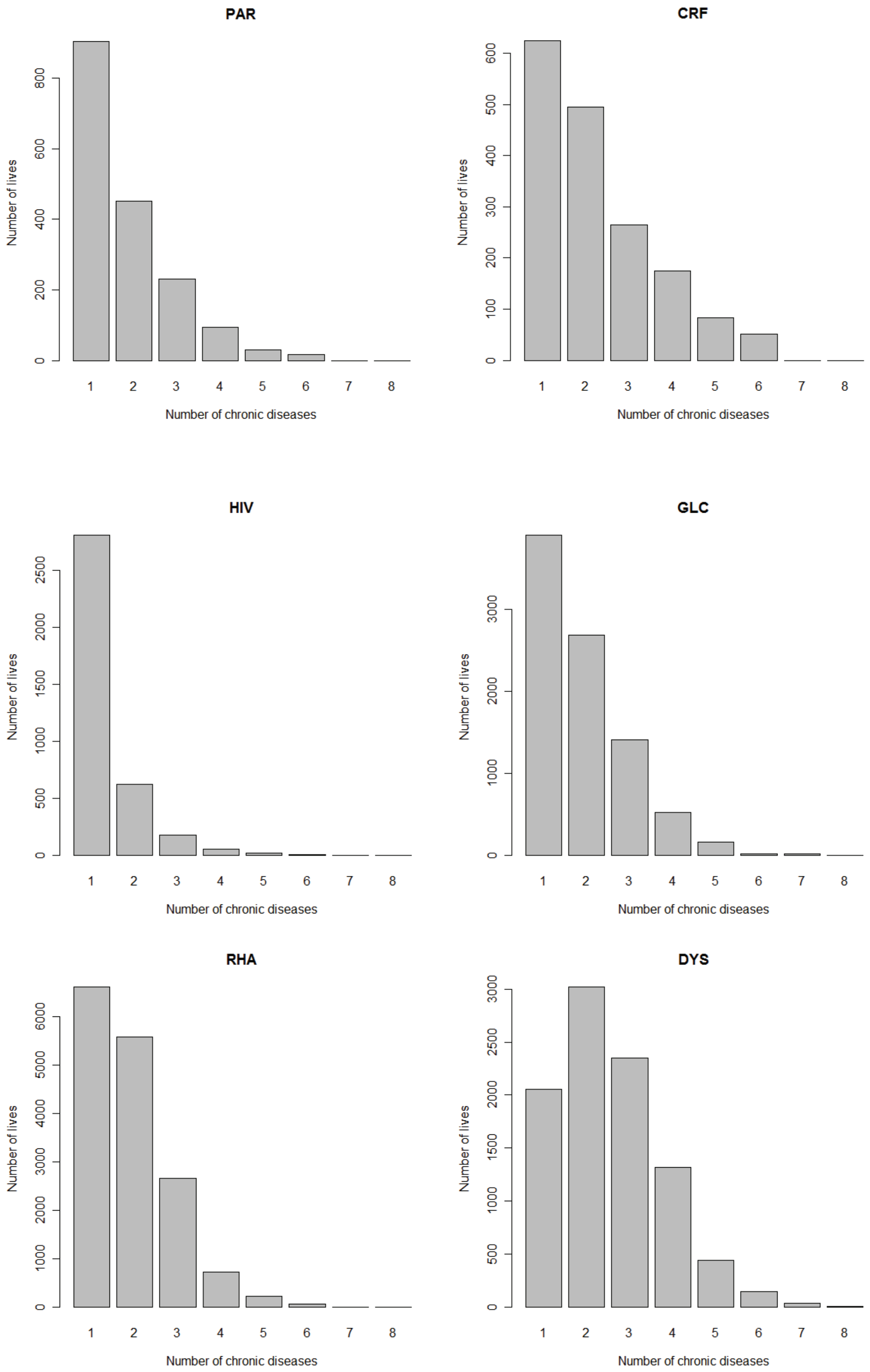


Figure 2.3: Continued.

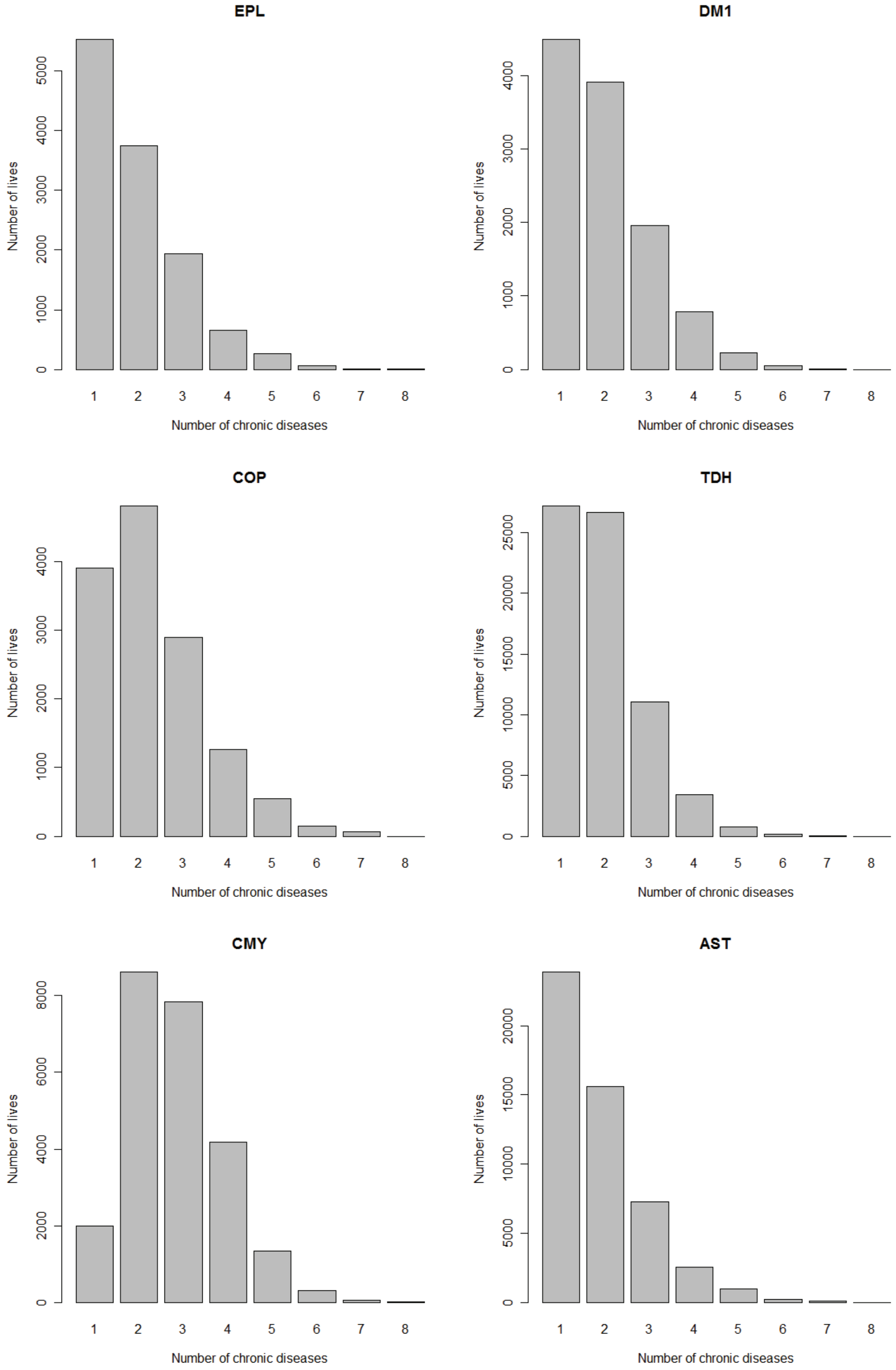


Figure 2.3: Continued.

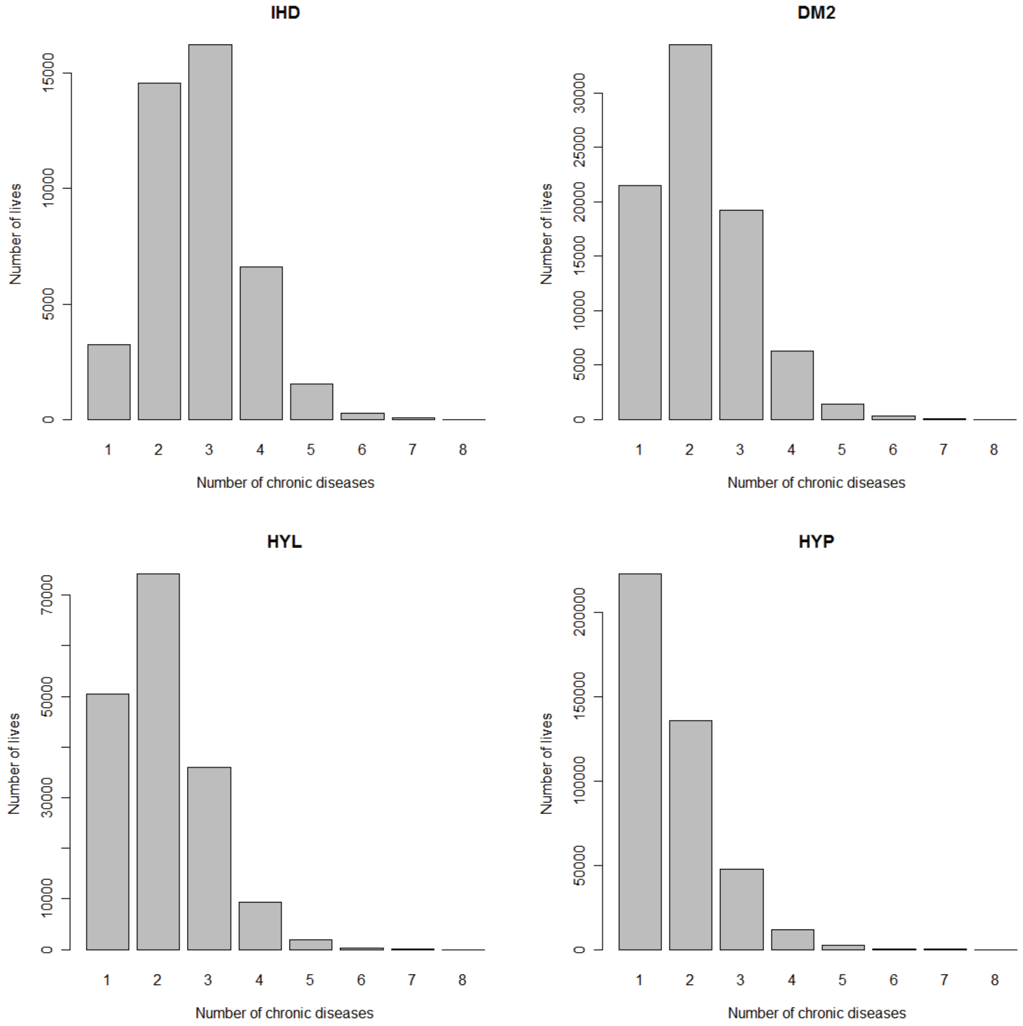


Figure 2.3: Continued.

Figure 2.3 gives a breakdown of the number of lives treated for 1, 2, 3, …, 8 chronic diseases, where the breakdown is done for each chronic disease. This will give information about the co-occurrences of certain diseases. It can be observed from Figure 2.3 that CMY, IHD, DM2, HYL, DYS, BCE, SLE, ADS and DBI tend to co-occur often with other chronic diseases, while the chronic diseases MSS, PAR, HIV and HAE show much less co-occurrence with other chronic diseases.

The histograms in Figure 2.3 provide information about the co-occurrences of certain chronic diseases, but the usefulness of these histograms is limited. The histograms show that some chronic diseases, like CMY, co-occur often with other chronic diseases, but with which diseases do CMY co-occur? The histograms do not show which of the chronic diseases are strongly related.

Table 2.4: *Number of co-occurrences between pairs of chronic diseases.*

	ADS	AST	BCE	BMD	CMY	COP	CRF	CSD	DBI	DM1	DM2	DYS	EPL
ADS	–	22	0	0	0	3	0	0	5	9	14	1	2
AST	22	–	238	61	2,110	5,812	21	74	9	343	3,580	457	588
BCE	0	238	–	6	82	151	0	0	0	0	27	14	0
BMD	0	61	6	–	35	15	0	0	0	11	96	12	270
CMY	0	2,110	82	35	–	1,399	201	0	0	686	4,144	1,864	337
COP	3	5,812	151	15	1,399	–	6	12	0	110	948	278	198
CRF	0	21	0	0	201	6	–	0	0	191	269	26	14
CSD	0	74	0	0	0	12	0	–	0	0	55	4	0
DBI	5	9	0	0	0	0	0	0	–	0	0	0	12
DM1	9	343	0	11	686	110	191	0	0	–	0	78	171
DM2	14	3,580	27	96	4,144	948	269	55	0	0	–	958	830
DYS	1	457	14	12	1,864	278	26	4	0	78	958	–	135
EPL	2	588	0	270	337	198	14	0	12	171	830	135	–
GLC	16	392	0	23	178	81	26	5	0	164	719	58	41
HAE	0	0	0	0	0	0	0	0	0	0	0	0	0
HYL	8	5,900	35	276	6,595	1,824	259	103	24	2,368	23,242	2,600	1,904
HYP	29	17,570	186	451	18,822	4,983	794	137	37	5,558	51,451	5,011	4,379
IBD	0	108	2	0	30	37	0	74	0	6	95	7	44
IHD	0	1,488	9	32	5,069	786	210	12	12	862	6,575	1,890	652
MSS	0	8	0	0	0	6	0	0	0	0	6	0	11
PAR	0	45	0	1	29	0	12	8	0	23	143	20	71
RHA	0	867	32	13	447	237	16	67	0	80	1,196	72	171
SCZ	0	23	0	77	7	11	0	0	0	17	79	0	117
SLE	0	87	0	12	43	33	0	0	0	8	78	5	30
TDH	111	3,057	41	284	2,047	705	81	38	29	624	4,333	901	1,038
HIV	0	82	12	0	66	32	10	0	0	65	194	16	25
TOTAL	220	42,942	835	1,675	44,191	17,667	2,136	589	128	11,374	99,032	14,407	11,040

Table 2.4: *Continued.*

	GLC	HAE	HYL	HYP	IBD	IHD	MSS	PAR	RHA	SCZ	SLE	TDH	HIV
ADS	16	0	8	29	0	0	0	0	0	0	0	111	0
AST	392	0	5,900	17,570	108	1,488	8	45	867	23	87	3,057	82
BCE	0	0	35	186	2	9	0	0	32	0	0	41	12
BMD	23	0	276	451	0	32	0	1	13	77	12	284	0
CMY	178	0	6,595	18,822	30	5,069	0	29	447	7	43	2,047	66
COP	81	0	1,824	4,983	37	786	6	0	237	11	33	705	32
CRF	26	0	259	794	0	210	0	12	16	0	0	81	10
CSD	5	0	103	137	74	12	0	8	67	0	0	38	0
DBI	0	0	24	37	0	12	0	0	0	0	0	29	0
DM1	164	0	2,368	5,558	6	862	0	23	80	17	8	624	65
DM2	719	0	23,242	51,451	95	6,575	6	143	1,196	79	78	4,333	194
DYS	58	0	2,600	5,011	7	1,890	0	20	72	0	5	901	16
EPL	41	0	1,904	4,379	44	652	11	71	171	117	30	1,038	25
GLC	–	0	1,519	3,434	12	313	2	3	137	12	26	682	2
HAE	0	–	0	0	0	0	0	0	0	0	0	0	0
HYL	1,519	0	–	96,806	300	24,333	17	292	1,839	126	120	13,767	98
HYP	3,434	0	96,806	–	547	29,343	48	525	6,609	281	393	30,946	556
IBD	12	0	300	547	–	99	0	0	149	11	0	114	0
IHD	313	0	24,333	29,343	99	–	0	99	499	13	57	2,356	44
MSS	2	0	17	48	0	0	–	0	0	0	0	19	0
PAR	3	0	292	525	0	99	0	–	5	10	2	123	0
RHA	137	0	1,839	6,609	149	499	0	5	–	21	201	1,506	7
SCZ	12	0	126	281	11	13	0	10	21	–	0	146	0
SLE	26	0	120	393	0	57	0	2	201	0	–	141	0
TDH	682	0	13,767	30,946	114	2,356	19	123	1,506	146	141	–	23
HIV	2	0	98	556	0	44	0	0	7	0	0	23	–
TOTAL	7,845	0	184,355	278,896	1,635	74,753	117	1,411	14,171	951	1,236	63,112	1,232

Table 2.4 provides the actual number of co-occurrences between pairs of chronic diseases and gives an indication of the relatedness between pairs of chronic diseases. For example, it can be observed from Table 2.4 that CMY co-occurs often with HYP, HYL, IHD, DM2 and DYS. It must be remembered though that this table should not just be interpreted in absolute terms, but in relative terms by comparing the number of co-occurrences in each cell in Table 2.4 to the total number of co-occurrences, which is given as “TOTAL”.

The cost aspects of the Medical Scheme 55-59 data set will also be described. Each line of the Medical Scheme 55-59 data set contains the average Hospital, average Medical and average Related costs associated with a certain unique combination of disease and gender (RETAP, 2007). The term *average* is used to indicate that each line of unique combination of disease and gender does not represent a single life, but a multiple number of lives. Only lives treated with a single chronic disease and lives not treated with any chronic disease will be considered. The reason for this is that the Medical Scheme 55-59 data set does not provide a breakdown of the average cost for each chronic disease when the diseases occur simultaneously. For example, lives simultaneously treated for HYP, CMY and DYS might show an average Hospital cost of R1000, but the apportioning of the R1000 among HYP, CMY and DYS will be unknown. The lives with multiple chronic diseases can therefore not be used when average cost differences of the chronic diseases are described. This means that only approximately 89% of lives can be considered, as seen from Table 2.3.

Table 2.5 provides information of the average treatment costs associated with female lives. The histograms provided in Figure 2.4 and Figure 2.5 display the breakdown of average costs for female lives. The chronic diseases CRF and MSS have very high average costs. The chronic diseases TDH and HYP have some of the lowest average costs. The female lives not treated for HIV or any chronic diseases, coded as NON, also display very low average costs. Various chronic diseases show a different breakdown of the different types of average costs. The chronic diseases DBI, IBD, SCZ, PAR, CSD and MSS show a higher proportion of average Medicine cost, while TDH, SLE, CMY, IHD, COP, ADS and BCE show a higher proportion of average Hospital cost.

Table 2.6 provides information of the average treatment costs associated with male lives. There were no male lives treated for HAE, as seen from Table 2.1. There were also no male lives treated only for DBI. Some male lives were treated for DBI, but these male lives were also simultaneously treated for other chronic diseases. That is why DBI and HAE are not listed in Table 2.6.

Table 2.5: *The breakdown of average cost related to female lives treated for only one chronic disease and female lives that were not treated for any chronic disease or HIV. That means only 89.64% of female lives are considered. The average cost is apportioned into three categories: Hospital, Medicine and Related.*

Disease	Hospital	Medicine	Related	Total
HAE	0.00	0.00	0.00	0.00
NON	160.33	19.30	39.43	219.06
TDH	184.07	74.59	52.36	311.02
DYS	133.17	168.99	47.51	349.67
HYP	185.62	153.75	60.09	399.46
IBD	93.87	340.44	35.39	469.70
RHA	177.92	238.36	69.44	485.72
HYL	234.16	204.40	69.27	507.83
AST	229.49	241.41	54.80	525.70
SLE	371.61	101.01	55.48	528.10
GLC	303.17	189.50	68.82	561.49
DBI	0.00	620.40	0.00	620.40
CMY	419.99	221.55	67.89	709.43
DM2	353.97	235.35	214.19	803.51
EPL	372.66	345.55	111.19	829.40
SCZ	130.17	525.00	175.67	830.84
PAR	172.83	750.28	125.00	1 048.11
HIV	384.17	589.07	89.98	1 063.22
CSD	385.35	632.00	59.55	1 076.90
BMD	187.63	552.23	339.52	1 079.38
IHD	931.07	200.39	54.54	1 186.00
COP	705.70	366.97	116.52	1 189.19
DM1	460.16	471.90	266.57	1 198.63
ADS	950.77	117.59	227.38	1 295.74
BCE	1 501.07	194.17	62.29	1 757.53
MSS	793.38	6 967.87	131.18	7 892.43
CRF	5 961.12	1 764.95	4 987.43	12 713.50

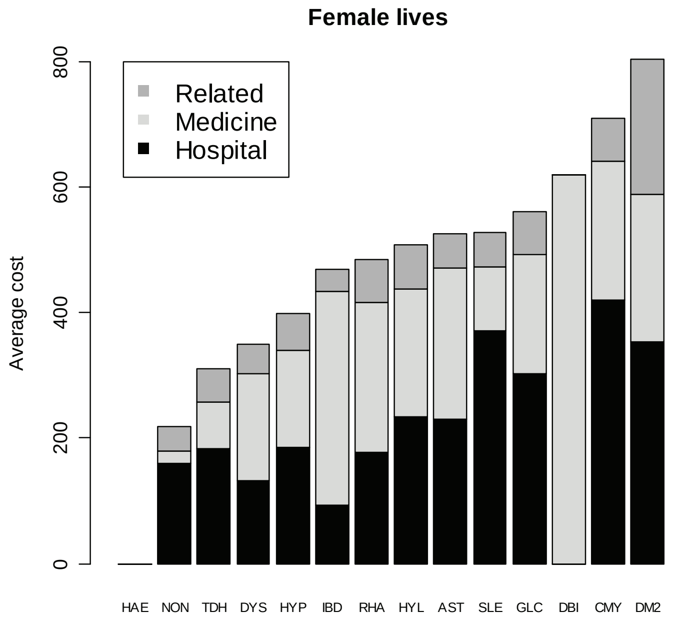


Figure 2.4: *The breakdown of average costs related to female lives treated for only one chronic disease and female lives that were not treated for any chronic disease or HIV.*

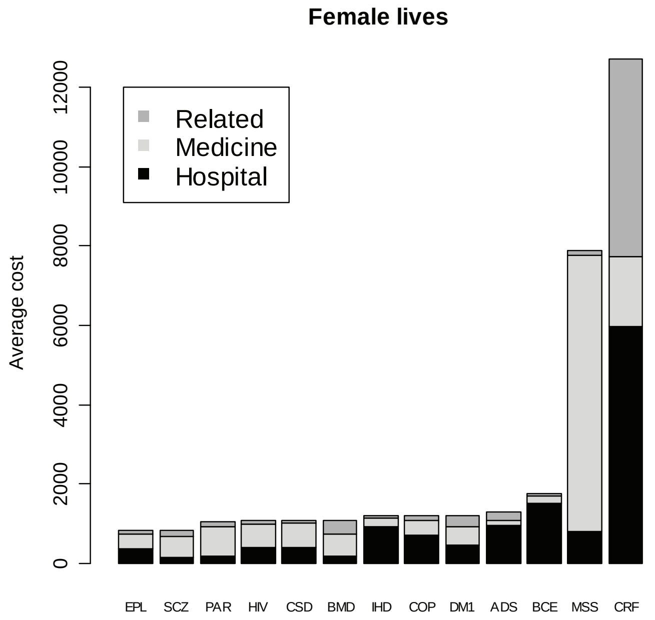


Figure 2.5: *The breakdown of the average costs related to female lives treated for only one chronic disease and female lives that were not treated for any chronic disease or HIV. These chronic diseases are associated with relatively high average costs.*

Table 2.6: *The breakdown of the average cost related to male lives treated for only one chronic disease and male lives that were not treated for any chronic disease or HIV. The average cost is split into three categories: Hospital, Medicine and Related.*

Disease	Hospital	Medicine	Related	Total
SLE	0.00	50.76	13.82	64.58
NON	214.47	22.02	34.49	270.98
TDH	257.12	83.42	72.60	413.14
ADS	226.82	216.74	14.14	457.70
HYP	261.52	167.19	62.70	491.41
HYL	261.04	213.25	48.09	522.38
IBD	122.78	396.76	28.98	548.52
BCE	0.00	297.64	267.72	565.36
BMD	88.99	426.97	62.82	578.78
AST	294.14	241.89	60.50	596.53
GLC	358.53	205.30	63.10	626.93
RHA	346.44	233.64	95.53	675.61
DYS	411.20	202.32	226.60	840.12
DM2	438.06	237.96	250.15	926.17
IHD	518.39	255.50	160.34	934.23
EPL	529.41	350.31	66.86	946.58
CSD	580.28	488.04	33.43	1 101.75
CMY	712.39	249.19	163.20	1 124.78
HIV	414.84	648.40	71.16	1 134.40
COP	698.92	337.30	124.63	1 160.85
SCZ	785.80	411.92	67.33	1 265.05
PAR	600.11	893.81	55.52	1 549.44
DM1	682.35	521.56	346.10	1 550.01
MSS	0.00	5 253.77	76.85	5 330.62
CRF	4 374.02	2 292.40	5 639.28	12 305.70

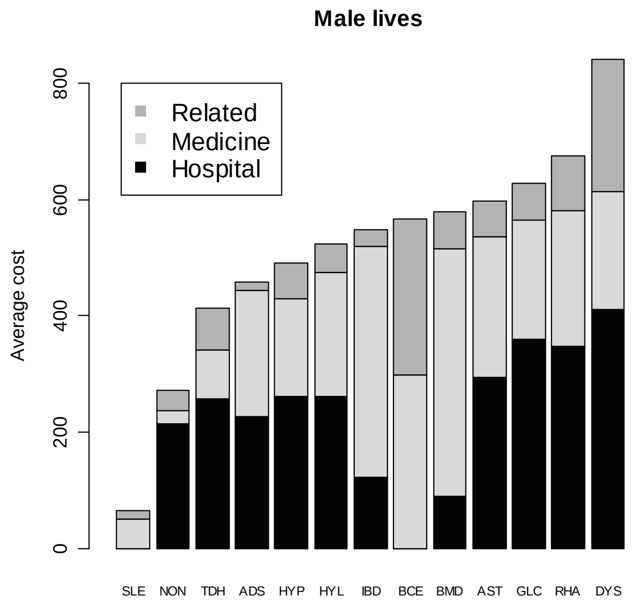


Figure 2.6: *The breakdown of the average costs related to male lives treated for only one chronic disease and male lives that were not treated for any chronic disease or HIV. These chronic diseases are associated with relatively low average costs.*

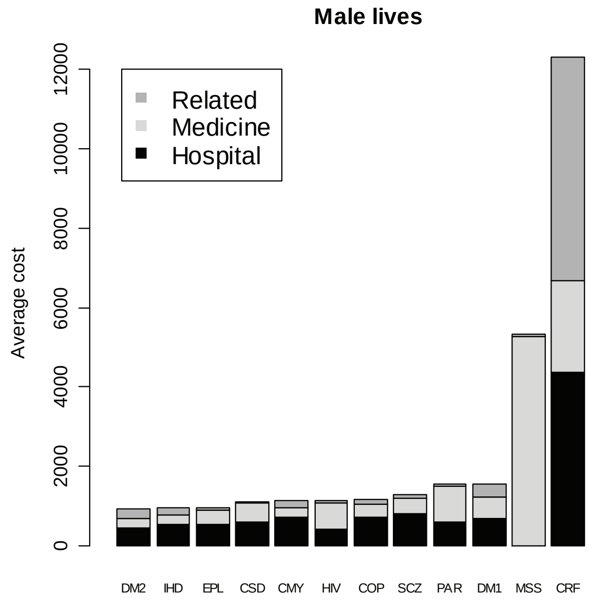


Figure 2.7: *The breakdown of the average costs related to male lives treated for only one chronic disease and male lives that were not treated for any chronic disease or HIV. These chronic diseases are associated with relatively high average costs.*

Figures 2.6 and 2.7 display the breakdown of average costs for male lives. The chronic diseases CRF and MSS have high average costs, which is also the case for female lives. The chronic diseases SLE, TDH and ADS have some of the lowest average costs. Various chronic diseases show a different breakdown of average costs. The chronic diseases SLE, IBD, PAR, BMD and MSS show a higher proportion of average Medicine cost, while TDH, CMY, IHD and COP show a higher proportion of average Hospital cost. The average Related costs tend to be less substantial. Only the chronic diseases CRF, BCE and DYS have a substantial proportion of average Related cost.

2.3 Summary

It is very important to note that the results produced in this chapter only apply to the age band of 55 to 59 years and thus should not be generalised to all age bands of the complete data set. Nonetheless, it was shown that approximately 72% of lives were not treated for any chronic disease and approximately 17% of lives were treated only for one chronic disease. This means that only approximately 11% of the lives were treated for multiple chronic diseases for this age band.

It was also shown that the chronic diseases HAE, DBI, ADS, BCE and MSS did not occur often. On the other hand, the chronic diseases AST, TDH, DM2, HYL and HYP occurred more often. Chronic disease rates were used to make a direct comparison between male and female lives, with regard to the occurrence of chronic diseases. It was found that female lives have higher chronic disease rates for DBI, MSS, BCE, SLE, SCZ, BMD, RHA and TDH, while the male lives have higher chronic disease rates for CRF, PAR, HIV, IHD and HYL.

Histograms were used to describe the average costs of lives treated for a single chronic disease and lives not treated for any chronic disease. This means that approximately 89% of lives were considered. The reason for this was that average costs could not correctly be allocated between chronic diseases when the chronic diseases co-occur. This is unfortunate because all information regarding average costs of lives treated for multiple chronic diseases had to be discarded, which may limit the relevance of the results. Still, it was found that the chronic diseases CRF, MSS and DM1 were associated with very high average costs for both male and female lives. The male and female lives displayed very low average costs for HYP, HYL and TDH. Lives not treated for any chronic disease or HIV also displayed very low average costs.

It was also found that CMY, IHD, DM2, HYL, DYS, BCE, SLE, ADS and DBI tend to co-occur often with other chronic diseases, while the chronic diseases MSS, PAR, HIV and HAE show much less co-occurrence with other chronic diseases. The usefulness of this information however is limited, as it cannot be used to describe the relationships among the chronic diseases. For example, the histograms showed that the chronic disease CMY co-occurred often with other chronic diseases, but the histograms did not show with which chronic diseases.

It will be shown in Chapters 4 and 5 how MDS methods and clustering methods can be used to overcome this limitation of histograms. These methods will show which chronic diseases are more related to one another and which chronic diseases are less related to one another. This can be done by calculating similarities (dissimilarities) between the chronic diseases. Therefore, the question of how to obtain these similarities (dissimilarities) must firstly be addressed. This is done in Chapter 3.

Chapter 3

Calculating Proximities between the Chronic Diseases

3.1 Introduction

There are two measures of proximity in a statistical context, similarity and dissimilarity, which are used to measure how similar and dissimilar objects are to each other. The construction of the similarities (dissimilarities) between the chronic diseases of the Medical Scheme 55-59 data set will be described in this chapter. The similarities (dissimilarities) are based on the number of lives treated for different combinations of chronic diseases. These dissimilarities will be used in Chapters 4 and 5, where the relationships between the chronic diseases will be graphically displayed.

The calculation of the dissimilarities is based on various dissimilarity coefficients. Various authors, for example Anderberg (1973), Hubàlek (1982), Gower & Legendre (1986), Cox & Cox (2001) have discussed several dissimilarity coefficients to be used with certain types of variables. However, the discussion in Section 3.2 will focus only on dissimilarity coefficients that are based on binary data, because the chronic diseases are binary variables in the Medical Scheme 55-59 data set.

Metric and Euclidean properties of the various dissimilarity coefficients, which are based on binary data, will be discussed in Section 3.3. The metric and Euclidean properties of these dissimilarity coefficients are important to consider when the chronic diseases are graphically displayed in Chapters 4 and 5. It will only be appropriate to use some of these displays if certain metric requirements are met.

The *R* function *Dissim.CDL* was developed to calculate the dissimilarities between the various chronic diseases. This will be discussed in Section 3.4. Details of the function *Dissim.CDL* are provided in Appendix A. Various different dissimilarity coefficients will be used to calculate dissimilarities between the chronic diseases. It might not be wrong to consider only one appropriate dissimilarity coefficient, but it might be better to consider various different appropriate dissimilarity coefficients, hoping for robustness against a specific choice (Cox & Cox, 2001, p.12). The techniques that will be discussed in Chapters 4 and 5 take dissimilarities as input, and the results will therefore greatly depend on the choice of dissimilarity coefficient. It is because of this that several dissimilarity coefficients, and not just one, will be used in this study.

3.2 Construction of Similarity and Dissimilarity Coefficients for Binary Data

Let s_{ij} represent the similarity coefficient between objects *i* and *j*. Most of the similarity coefficients have values in the range [0, 1]. Large values of s_{ij} indicate that the two objects are very similar. Dissimilarity coefficients of quantitative data can be obtained directly from the data, but dissimilarity coefficients based on binary, nominal and ordinal data are constructed by transforming the similarity coefficients. Let d_{ij} represent the dissimilarity coefficient between objects *i* and *j*. A value d_{ij} close to zero indicates that the two objects *i* and *j* are very similar. Conversely, large values of d_{ij} indicate that the two objects are very dissimilar.

Dissimilarities often only satisfy the first three of the following axioms of a metric (Kaufman & Rousseeuw, 1990, p. 13):

1. $d_{ii} = 0$
2. $d_{ij} \geq 0$
3. $d_{ij} = d_{\mathrm{ji}}$
4. $d_{ij} \leq d_{ih} + d_{hj}$

The fourth axiom does not have to be satisfied in general. However, if it is satisfied then the dissimilarities are actual measures of distance.

Table 3.1: *Measure of similarity (dissimilarity) between object r and object s.*

		Object *s*		
		1	**0**	
Object *r*	**1**	*a*	*b*	
	0	*c*	*d*	*p = a+b+c+d*

In Table 3.1, *a* indicates the number of variables out of *p* that score 1 for both objects *r* and *s*, *b* indicates the number of variables that score 1 for object *r* and 0 for object *s*, *c* indicates the number of variables that score 0 for object *r* and 1 for object *s* and *d* indicates the number of variables that both score 0 for object *r* and *s*.

Similarity coefficients are usually constructed when the variables are binary. These similarity coefficients are then transformed into dissimilarity coefficients, which are used to construct a dissimilarity matrix. The calculation of the similarity coefficient between objects *r* and *s* is based on Table 3.1.

Table 3.2: *Various similarity coefficients for binary data.*

Similarity coefficient	***Formula***
Jaccard	$s_{rs} = \frac{a}{a+b+c}$
Dice, Sorensen	$s_{rs} = \frac{2a}{2a+b+c}$
Kulczynski 1	$s_{rs} = \frac{a}{b+c} \quad \text{if } r \neq s$ $= 0 \quad \text{if } r = s$
Ochiai	$s_{rs} = \frac{a}{\left[(a+b)(a+c)\right]^{0.5}}$
Phi	$s_{rs} = \frac{ad-bc}{\left[(a+b)(a+c)(d+b)(d+c)\right]^{0.5}}$
Baroni-Urbani, Buser	$s_{rs} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
Kulczynski 2	$s_{rs} = \frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
Rao, Russell	$s_{rs} = \frac{a}{a+b+c+d}$
Simple matching coefficient	$s_{rs} = \frac{a+d}{a+b+c+d}$
Yule	$s_{rs} = \frac{ad-bc}{ad+bc}$
Sokal, Sneath, Anderberg	$s_{rs} = \frac{a}{a+2(b+c)}$

Several different similarity coefficients are used in practice. In fact, Hubàlek (1982) gives a very comprehensive list, containing 43 similarity coefficients that are used for binary data. Cox & Cox (2001) and Gower & Legendre (1986) also listed several similarity coefficients. Table 3.2 lists 11 similarity coefficients that are readily used in practice and are based on the suggestions made by Hubàlek (1982). Hubàlek (1982) found that the *Jaccard*, *Dice-Sorensen*, *Kulczynski 1* and *Ochiai* similarity coefficients generally work well and that the *Phi* and *Baroni-Urbani-Buser* similarity coefficients are reasonable. The results by Hubàlek (1982) are based on an ecological data set and these results are very relevant for the Medical Scheme 55-59 data set, because the ecological data set has similar characteristics to the Medical Scheme 55-59 data set. This will be discussed in detail in Section 3.4.

The similarity coefficients have to be transformed into dissimilarity coefficients. There are many possible transformations, but Gower & Legendre (1986) suggest the following transformations:

$$d_{ij} = 1 - s_{ij} \tag{3-1}$$

$$d_{ij} = (1 - s_{ij})^{0.5} \tag{3-2}$$

Kaufman & Rousseeuw (1990) prefer to use the transformation in (3-1), because it tends to lead to a clearer clustering structure. The transformation in (3-2) makes the difference between large similarities more important, but makes it more difficult to obtain small dissimilarities. However, the transformation in (3-2) leads to more dissimilarity coefficients with metric and Euclidean properties. This will be discussed in Section 3.3.

3.3 Metric and Euclidean Properties of Dissimilarity Coefficients for Binary Data

Gower & Legendre (1986) discuss metric and Euclidean properties of many dissimilarity coefficients, for both binary and quantitative data. However, only the Euclidean and metric properties of dissimilarity coefficients based on binary data will be discussed in this section, because the chronic diseases are binary variables in the Medical Scheme 55-59 data set.

Let all dissimilarities d_{ij} be placed in the dissimilarity matrix **D**, where $[\mathbf{D}]_{\mathrm{ij}} = d_{ij}$. Gower & Legendre (1986) define the terms *metric* and *Euclidean* in the context of dissimilarity coefficients as follows:

Definition 3.1 Metric property

- **D** is said to be *metric* if the metric (triangle) inequality $d_{ij}+d_{ik} \geq d_{jk}$ holds for all triplets (i, j, k) and d_{ii}=0 for all i.

Definition 3.2 Euclidean property

- **D** is said to be *Euclidean* if n points P_i (i=1,2,...,n) can be embedded in a Euclidean space such that the Euclidean distance between points P_i and P_j is d_{ij}. This implies that d_{ij} must be non-negative.

If **D** is Euclidean, it is also a metric. However, if **D** is a metric it is not necessarily Euclidean. Gower & Legendre (1986) established various results to determine when dissimilarity coefficients display metric and Euclidean properties. For binary variables, define:

$$S_\theta = \frac{a+d}{a+d+\theta(b+c)} \qquad\qquad T_\theta = \frac{a}{a+\theta(b+c)}$$

where *a*, *b*, *c* and *d* are defined as in Table 3.1. Most of the similarity coefficients displayed in Table 3.2 can be obtained by using the appropriate choice of *θ*. Dissimilarities can then be formed by transforming these similarity coefficients by a particular transformation function. The metric and Euclidean properties of these formed dissimilarities will be dependent on which transformation function was used. Gower & Legendre (1986) consider the following two transformations:

$$D_\theta = 1 - S_\theta \qquad\qquad D_\theta = 1 - T_\theta$$

$$D_\theta = \sqrt{1 - S_\theta} \qquad\qquad D_\theta = \sqrt{1 - T_\theta}$$

Gower & Legendre (1986) obtain the following results:

The dissimilarity $1-S_\theta$ is metric for $\theta \geq 1$ and the dissimilarity $\sqrt{1-S_\theta}$ is metric for $\theta \geq 1/3$. The dissimilarity $1-S_\theta$ may be non-metric for $\theta < 1$ and the dissimilarity $\sqrt{1-S_\theta}$ may be non-metric for $\theta < 1/3$. There are similar results when S_θ is replaced by T_θ.

If $\sqrt{1-S_\theta}$ is Euclidean, then so is $\sqrt{1-S_\phi}$ for all $\phi \geq \theta$. A similar result holds when S_θ is replaced by T_θ.

The dissimilarity $\sqrt{1-S_\theta}$ is Euclidean for $\theta \geq 1$ and the dissimilarity $\sqrt{1-T_\theta}$ is Euclidean for $\theta \geq 1/2$. However, $1-S_\theta$ and $1-T_\theta$ may be non-Euclidean.

Gower & Legendre (1986) use these results to investigate the metric and Euclidean properties of among others the dissimilarity coefficients derived from the similarity coefficients listed in Table 3.2. These results are given in Table 3.3.

Table 3.3: *Metric and Euclidean properties of various dissimilarity coefficients for binary data.*

Coefficients	$D_\theta = (1 - T_\theta)$		$D_\theta = \sqrt{1 - T_\theta}$	
	Metric	**Euclidean**	**Metric**	**Euclidean**
Jaccard	YES	NO	YES	YES
Dice, Sorensen	NO	NO	YES	YES
Kulczynski 1				
Ochiai	NO	NO	YES	YES
Phi	NO	NO	YES	YES
Baroni-Urbani, Buser				
Kulczynski 2	NO	NO	NO	NO
Rao, Russell	YES	NO	YES	YES
Simple matching coefficient	YES	NO	YES	YES
Yule	NO	NO	NO	NO
Sokal, Sneath, Anderberg	YES	NO	YES	YES

Note that the Euclidean and metric properties of the *Kulczynski 1* and *Baroni-Urbani-Buser* dissimilarity coefficients are not displayed in Table 3.3. The *Kulczynski 1* dissimilarity coefficient can take negative values, so its metric and Euclidean properties are irrelevant (Gower & Legendre, 1986). The *Baroni-Urbani-Buser* similarity coefficient cannot be written in the form S_θ or T_θ and was not considered by Gower & Legendre (1986). The Euclidean and metric properties of the *Baroni-Urbani-Buser* dissimilarity coefficient can therefore not be listed. The transformation $d_{ij} = (1 - s_{ij})^{0.5}$ may be preferred to the transformation $d_{ij} = 1 - s_{ij}$, because more of the dissimilarities will have metric or Euclidean properties. However, the choice of transformation will also depend on the problem at hand. Table 3.3 will be used in Chapter 4 where it is important to consider metric and Euclidean properties.

3.4 Calculation of Dissimilarities between the Chronic Diseases

As described in Chapter 2, the Medical Scheme 55-59 data set consists of binary variables where "1" indicates that the lives are treated for the respective disease and where a "0" indicates that the lives are not treated for the disease. It is important to investigate how these chronic diseases are related to one another. Dissimilarities between the chronic diseases need to be constructed and will be based on the number of lives treated for different combinations of chronic diseases. These dissimilarities will then be used in Chapters 4 and 5 where the relationships between the chronic diseases will be graphically displayed.

A dissimilarity matrix of the objects, and not the variables, of a data set will usually be constructed for most applications. However, in the case of the Medical Scheme 55-59 data set, a dissimilarity matrix of the respective binary variables needs to be constructed, remembering that the chronic diseases are binary variables. This can be done by following the approach of Johnson & Wichern (2002, p.677). The approach uses Table 3.4, which is very similar to Table 3.1, except that the roles of the binary variables and objects are interchanged. The substitution of *n* (the total number of lives) for *p* (the number of binary variables) is also required. The similarity coefficients listed in Table 3.2 can then be used in the usual manner to construct the similarities between the chronic diseases. These similarities can then be transformed into dissimilarities.

Table 3.4: *Measure of similarity between chronic diseases r and s, where a indicates the number of lives out of n (total number of lives) that are treated for both chronic diseases r and s, b indicates the number of lives treated for chronic disease r, but not treated for chronic disease s, c indicates the number of lives treated for chronic disease s, but not treated for chronic disease r and d indicates the number of lives that are not treated for chronic diseases r and s.*

		Chronic disease *s*		
		1 (YES)	**0 (NO)**	
Chronic disease *r*	**1 (YES)**	*a*	*b*	
	0 (NO)	*c*	*d*	*n = a+b+c+d*

Hubàlek (1982) shows that it is important to consider whether the binary variables are symmetric or asymmetric, as this will greatly influence the choice of similarity coefficient to be used. Symmetric binary variables are binary variables where code “0” and “1” are equally important. Asymmetric binary variables do not attach equal importance to codes “0” and “1”. The most meaningful outcome is usually coded as “1” and the less meaningful outcome as “0”, where “1” refers to the presence of an attribute and “0” to its absence (Kaufman & Rousseeuw, 1990, p.26). The chronic diseases of the Medical Scheme 55-59 data set should therefore be regarded as asymmetric binary variables, because the presence of a chronic disease is more meaningful than its absence (Kaufman & Rousseeuw, 1990, p.26). It is typical for asymmetric binary variables to have large *d* values in Table

3.4, because of more 0-0 matches (Hubàlek, 1982). It was found that the average *d* value for the Medical Scheme 55-59 data set is 432 times greater than the average *a* value, which suggests that the chronic diseases are indeed asymmetric binary variables. Various authors suggest omitting *d* values in the denominator of similarity coefficients when 0-0 matches do not have great importance or if the *d* values are very large in comparison with the *a* values (Hubàlek, 1982; Gower, 1985; Sibson *et al.*, 1981; Gower and Legendre, 1986; Kaufman & Rousseeuw, 1990). It can be seen from Table 3.2 that the *Simple matching* coefficient and the *Rao-Russell* similarity coefficient include *d* values in the denominator, and are therefore unlikely to be suitable for application to the Medical Scheme 55-59 data set. So which similarity coefficients should be used?

Hubàlek (1982) did a comprehensive study of similarity coefficients based on an ecology data set. The ecology data set is very similar to the Medical Scheme 55-59 data set, in the sense that the ecology data set has large *d* values and consists of asymmetric binary variables. Hubàlek (1982) compared the different similarity coefficients based on certain criteria and admissibility conditions using the ecology data set. Hubàlek (1982) found that the *Jaccard*, *Dice-Sorensen*, *Kulczynski 1* and *Ochiai* similarity coefficients performed satisfactory according to his specified criteria, while the *Phi* and *Baroni-Urbani-Buser* similarity coefficients also performed satisfactory, but to a lesser extent than the first-mentioned four similarity coefficients. Various authors also recommend using the *Jaccard* similarity coefficient when using asymmetric binary variables (Sibson *et al.*, 1981; Kaufman & Rousseeuw, 1990).

The *R* function *Dissim.CDL* was developed to calculate dissimilarities between all the chronic diseases. Details of *Dissim.CDL* are provided in Appendix A. The variable *Nr.Lives* in the Medical Scheme 55-59 data set contains the number of lives that have the same gender and diseases combination. This variable will therefore be used to determine the values of *a*, *b*, *c* and *d* for all pairs of chronic diseases, as described in Table 3.4. These values are then used to construct the various similarity coefficients s_{ij} between all pairs of chronic diseases. The *Dissim.CDL* function constructs all similarities $\{s_{ij}\}$, which are then transformed to dissimilarities $\{d_{ij}\}$ by using one of the following transformations: $d_{ij} = 1-s_{ij}$ or $d_{ij} = (1-s_{ij})^{0.5}$. All dissimilarities $\{d_{ij}\}$ are then placed in a dissimilarity matrix **D,** where $[\mathbf{D}]_{ij} = d_{ij}$. Various dissimilarity matrices will be constructed, where each dissimilarity matrix is based on a different similarity coefficient listed in Table 3.2. These dissimilarity matrices will be used as input structures for the MDS methods and clustering techniques, which will be discussed in Chapters 4 and 5.

3.5 Summary

The construction of dissimilarities between the chronic diseases of the Medical Scheme 55-59 data set was described in this chapter. Various authors suggested many different similarity coefficients that could be used to construct dissimilarities between binary data. Some of these coefficients are more appropriate to use than others, depending on the actual data set. It is also very important to consider whether the binary variables are symmetric or asymmetric when similarities (dissimilarities) are constructed, as this will greatly influence the choice of similarity coefficient to be used. The chronic diseases are asymmetric binary variables of the Medical Scheme 55-59 data set, where the presence (coded as "1") of the chronic disease is more important than its absence (coded as "0"). Various authors suggest omitting *d* values, which is the number of 0-0 matches, in the denominator of similarity coefficients when the data consist of asymmetric binary variables. The *Simple matching* coefficient and the *Rao-Russell* similarity coefficient include *d* values in the denominator, and are therefore unlikely to be suitable when applied to the Medical Scheme 55-59 data set. Many authors recommend using the *Jaccard* similarity coefficient when the data consist of asymmetric binary variables. The *Dice-Sorensen*, *Kulczynski 1*, *Ochiai*, *Phi* and *Baroni-Urbani-Buser* similarity coefficients are also recommended when the data consist of asymmetric binary variables.

The metric and Euclidean properties of the different similarity coefficients were also considered. Dissimilarities were formed by transforming the similarities by an appropriate transformation function. Two of these transformation functions were considered in this chapter. The first transformation: $d_{ij} = 1-s_{ij}$ is usually preferred for the clustering techniques discussed in Chapter 5, as it tends to lead to a clearer clustering structure. The second transformation: $d_{ij} = (1-s_{ij})^{0.5}$ leads to more dissimilarities with metric and Euclidean properties, which are important when certain MDS methods are applied to the Medical Scheme 55-59 data set in Chapter 4.

Chapter 4

Displaying the Relationships between Chronic Diseases using MDS Techniques

4.1 Introduction

Cox & Cox (2001) give a narrow definition of MDS as the "search for a lower dimensional space, usually Euclidean, in which points in the space represent the objects, one point representing one object, and such that the distances between the points in the space, $\{\delta_{ij}\}$, match, as well as possible, the original dissimilarities $\{d_{ij}\}$". A wider definition of MDS can include any multivariate data analysis technique which produces a graphical representation of objects from multivariate data. The techniques used for the search of the lower dimensional space and the associated configuration of points form nonmetric and metric MDS (Cox & Cox, 2001, p.1-3).

The main difference between metric and nonmetric MDS is that metric MDS requires dissimilarity coefficients with metric properties and nonmetric MDS can accept dissimilarity coefficients with or without metric properties. Classical scaling and metric least squares scaling are well-known metric MDS methods. Classical scaling treats dissimilarities directly as distances and metric least squares scaling matches distances to transformed dissimilarities with the transformation $\delta_{ij}=f(d_{ij})$, where f is a continuous parametric monotonic function. Nonmetric multidimensional scaling methods abandon the metric nature of the transformation f, but the transformation must preserve the rank order of the dissimilarities. Nonmetric MDS has the advantage of making minimal assumptions about how distances and dissimilarities are related (Gordon, 1999).

Metric least squares scaling and the nonmetric MDS method find a suitable configuration of points by minimising a certain loss function. Borg & Groenen (2005) describe the Scaling by Majorizing a Complicated Function (SMACOF) algorithm, which will be used to minimise the loss functions in this study. Classical scaling uses spectral decomposition on a doubly centered matrix of dissimilarities to find a lower dimensional display space (Gower & Hand, 1996, p.233-234). Nonmetric MDS makes the heaviest demands on computing resources and classical scaling makes the least demands on computing resources. Given this, classical scaling might be preferred when the data set contains many objects. Another disadvantage of metric least squares scaling and nonmetric

MDS is that the minimization algorithm provides no guarantee that a global minimum solution will be obtained (Borg & Groenen, 2005).

Classical scaling, metric least squares scaling and nonmetric MDS will be used to display the relationships between the chronic diseases of the Medical Scheme 55-59 data set. The dissimilarities, which were described in Chapter 3, will be used as input for these MDS techniques. Graphical displays will be produced, where the chronic diseases will be displayed in a two-dimensional representation. The display space, for both metric and nonmetric MDS, will be chosen to be Euclidean. This means that the displays will have a direct distance-like interpretation. Chronic diseases that tend to co-occur often are expected to be situated close to each other and chronic diseases that do not co-occur often are expected to be situated further apart. This will be discussed in Section 4.3. Several *R* functions, which are provided in Appendix A, were also developed to implement the metric and nonmetric MDS methods. Nonmetric MDS will also be used to produce separate configurations for the Male MS (55-59) data set and the Female MS (55-59) data set, and will be discussed in Section 4.4.

4.2 Multidimensional Scaling Methods used in this Study

4.2.1 Metric multidimensional scaling

The two main metric MDS methods, classical scaling and metric least squares scaling, will be discussed in this section.

4.2.1.1 Classical scaling

The classical scaling process of finding the configuration of points in the lower dimensional space for a set of dissimilarities $\{d_{ij}\}$ will be described in this section.

Suppose there are *n* objects with dissimilarities $\{d_{ij}\}$ measured between all pairs of objects. Define matrix **A** as $[\mathbf{A}]_{ij} = a_{ij}$, where $a_{ij} = -\frac{1}{2}[d_{ij}]^2$. Define matrix **B** as $[\mathbf{B}]_{ij} = b_{ij}$, where

$$b_{ij} = a_{ij} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet}\text{, and}$$

$$a_{i\bullet} = \tfrac{1}{n}\sum_{j=1}^{n} a_{ij}\ ,\ a_{j\bullet} = \tfrac{1}{n}\sum_{i=1}^{n} a_{ij}\ ,\ a_{\bullet\bullet} = \tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\ .$$

The matrix **B** can be expressed as a doubly centered matrix of dissimilarities

$$\mathbf{B} = \mathbf{HAH}$$

where **H** is the centering matrix,

$$\mathbf{H} = \mathbf{I}_n - \tfrac{1}{n}\underline{1}\,\underline{1}^{\mathrm{T}},$$

with $\underline{1} = (1,1,\ldots,1)^{\mathrm{T}}$, a vector of n ones.

The configuration of points can be found by expressing **B** in terms of its spectral decomposition (Gower & Hand, 1996, p.233-234) as

$$\mathbf{B} = \mathbf{V\Lambda V}^{\mathrm{T}},$$

where $\mathbf{\Lambda}$=diag(λ_1, λ_2, . . . , λ_n), the diagonal matrix of eigenvalues {λ_i} of **B** with $\lambda_1 \geq \lambda_2 \geq . . . \geq \lambda_n$. The matrix of corresponding eigenvectors is **V**=[$\underline{v}_1$, $\underline{v}_2$, . . ., $\underline{v}_n$] where the eigenvectors are normalised such that $\underline{v}_i^{\mathrm{T}}\underline{v}_i = 1$ for all i=1, 2, . . . , n. The configurations of the points in a r-dimensional display space can then be represented by the coordinate matrix **X**: $n \times r$ given by

$$\mathbf{X} = \mathbf{V}_r \mathbf{\Lambda}_r^{1/2},$$

where the columns of matrix $\mathbf{V}_r$: $n \times r$, consists of the first r eigenvectors of **B** that correspond to the r largest eigenvalues of **B**, and the matrix $\Lambda_r^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_r^{1/2})$. The coordinate matrix **X** will be used to display the points which represent the objects. It must be remembered that the arbitrary sign of the eigenvectors {$\underline{v}_i$} leads to invariance of the configurations with respect to reflection in the origin. The display space will not necessarily be Euclidean. Cox & Cox (2001) points out that if **B** is positive semi-definite of rank r, then a configuration in r dimensional Euclidean space can be found, so that the associated distances between the points $\{\delta_{ij}\}$ are such that $\delta_{ij} = d_{ij}$ for all i, j. Dissimilarity coefficients with Euclidean properties will always lead to an Euclidean display space as their dissimilarity matrices are Euclidean embeddable (Gower & Legendre, 1986). Dissimilarity coefficients with no metric or Euclidean properties may give rise to a matrix **B** which is not positive semi-definite.

How many dimensions should be used in the display space? It is easily shown that **B** has at least one zero eigenvalue, since $\mathbf{B}\underline{1} = \mathbf{HAH}\underline{1} = \underline{0}$ where $\underline{0}$ represents a vector of n zeroes. A configuration of points in an $r = n-1$ dimensional Euclidean space can therefore always be found. The configuration obtained could be rotated to its principal axes in the principal component sense (Cox & Cox, 2001, p.36). The principal axes are orthogonal to each other. Only the first r ($r \leq n-1$) principal axes are chosen for representing the objects, as this will explain the maximum variation in r dimensions. It turns out that **X** already has the points referred to their principal axes, since

$$\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{X} &= \left(\mathbf{V}_r\mathbf{\Lambda}_r^{1/2}\right)^{\mathrm{T}}\left(\mathbf{V}_r\mathbf{\Lambda}_r^{1/2}\right)\\ &= \mathbf{\Lambda}_r^{1/2}\mathbf{V}_r^{\mathrm{T}}\mathbf{V}_r\mathbf{\Lambda}_r^{1/2} = \mathbf{\Lambda}\,,\end{aligned}$$

where **Λ** is a diagonal matrix. The distances between the points in the full *n*−1 dimensional Euclidean space are given by

$$\delta_{ij}^2 = \sum_{s=1}^{n-1} \lambda_s (x_{is} - x_{js})^2\,,$$

and hence relatively small eigenvalues contribute far less to the squared distance δ_{ij}^2. If only *r* eigenvalues of **B** are retained as being significantly large, then the *r* dimensional Euclidean space spanned by the first *r* eigenvectors of **B** can be used to represent the objects.

A measure of the proportion of variation explained using only *r* dimensions is

$$\frac{\sum_{s=1}^{r}\lambda_s}{\sum_{s=1}^{n-1}\lambda_s}\,.$$

Cox & Cox (200, p.38) suggest a modified measure when **B** is not positive semi-definite:

$$\frac{\sum_{s=1}^{r}\lambda_s}{\sum_{s=1}^{n-1}|\lambda_s|} \text{ or } \frac{\sum_{s=1}^{r}\lambda_s}{\sum(\text{positive eigenvalues})}\,.$$

Choice of *r* can then be assessed with this measure, but for practical purposes, *r* will usually be chosen to be 2 or 3.

4.2.1.2 Metric least squares scaling

Metric least squares scaling is a metric MDS method that finds a configuration **X** : *n* ×*m* by matching δ_{ij} to d_{ij} by minimising a certain loss function (Cox, 2001, p.49). The term δ_{ij} refers to the distance between points *i* and *j* in this *m*-dimensional space **X** : *n* ×*m*, and Euclidean distance is usually used. The points in the display represent the original objects. A two-dimensional space (*m* = 2) is usually used.

Various loss functions have been suggested in the past, see for example Sammon (1969). Minimising different loss functions will produce different optimal configurations $\mathbf{X} : n \times m$. Borg & Groenen (2005, p.187) use a general loss function, which will be referred to as *Raw Stress*:

$$\text{Raw Stress} = \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij})^2 \tag{4-1}$$

where δ_{ij} is the Euclidean distance between points i and j in the graphical display, d_{ij} is the dissimilarity between objects i and j and $\{w_{ij}\}$ are weights which must contain non-negative values. In practical research, some dissimilarity values d_{ij} may be missing. The weights can then be used to overcome this problem by simply letting w_{ij} = 0 if d_{ij} is missing, and w_{ij} = 1 if d_{ij} is not missing. Other values of w_{ij} are also allowed and different choices of w_{ij} lead to different loss functions (Borg & Groenen, 2005, p.255). More generally, let $w_{ij} = d_{ij}^q$. Different choices of q can be used to emphasize the representation of small or large dissimilarities. Large negative values of q may lead to a better representation of small dissimilarities, but not large dissimilarities. Conversely, large positive values of q lead to a better representation of large dissimilarities, but not small dissimilarities. For a relative presentation of both small and large dissimilarities, choose $q = -2$. If the dissimilarities have some clustering, then choosing a large value of q may reveal a clearer clustering structure (Borg & Groenen, 2005, p.256).

The Raw Stress value in (4-1) is a badness-of-fit measure, but it is not very informative. A large Raw Stress value does not necessarily indicate a bad fit, as it depends on the scale of distances in the configuration $\mathbf{X} : n \times m$. Instead, *Normalised Stress* should be used to avoid scale dependency, where

$$\text{Normalised Stress} = \frac{\sum_{i<j} w_{ij}(\delta_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} d_{ij}^2}. \tag{4-2}$$

Normalised Stress values in (4-2) depend on many factors. The higher n, representing the number of points, the higher the Normalised Stress in general. The higher m, the dimensionality of the display space, the lower the Normalised Stress values. The larger the squared errors $(\delta_{ij} - d_{ij})^2$, the higher the Normalised Stress value (Borg & Groenen, 2005, p.54).

Loss functions, such as Normalised Stress in (4-2), are indices that assess the mismatch between the dissimilarities and corresponding distances. The residual plot and the bubble plot can be used to describe this mismatch in more detail . The residual plot is a scatter diagram of the distances and the dissimilarities. A bisector is drawn from the lower left corner to the upper right corner, and the dissimilarities are drawn on this bisector. The sizes of the dissimilarities can therefore be noted immediately. The corresponding distances, that should match the original dissimilarities as well as possible, are also drawn in this residual plot. The vertical distance between the dissimilarities and the corresponding distances is a measure of the corresponding error $e_{ij} = (\delta_{ij} - d_{ij})$. The error gives an indication of the size of mismatch between of the dissimilarities and the corresponding distances. Larger errors will cause a higher Normalised Stress value in (4-2). The residual plot gives an indication of which dissimilarities are better represented in the metric MDS display, but the residual plot does not give an indication of how well the original objects are represented by points in the display. However, the bubble plot can be used to assess the fit of each point. The bubble plot uses the *Stress per point* measure, which is defined by Borg & Groenen (2005, p.45) as follows:

Definition 4.1 Stress per point

- The average of the squared errors between the current object and all other objects.

The bubble plot still uses the same configuration of points as the metric MDS plot to display the objects. The only difference is that the bubble plot uses bubbles to represent the objects, where bubbles with a larger radius indicate points with better fit. The viewer can therefore immediately see which objects are better represented in the display.

In practice, a two-dimensional display is mostly used to display the final configuration $\mathbf{X} : n \times m$ with $m = 2$. It is also possible to display the final configuration $\mathbf{X} : n \times m$ in a three-dimensional graph, with $m = 3$. A three-dimensional display will have a lower final Normalised Stress (4-2) and Raw Stress (4-1) value than a two-dimensional display, but it is easier to describe a two-dimensional display for the purpose of this study. The Normalised Stress (4-2) value in a two-dimensional display can be assessed by considering the upper and lower bounds of the Normalised Stress (4-2) value. The Normalised Stress value in (4-2) has the following lower and upper bounds in a two-dimensional display: [0, 0.4352] (Borg & Groenen, 2005, p.275).

Various methods can be used to minimise Normalised Stress (4-2) or Raw Stress (4-1). The aim of these methods is to find an optimal configuration $\mathbf{X} : n \times m$ of points, from which distances $\{d_{ij}\}$ can be calculated. The optimal configuration will be the configuration that produces distances that best match the dissimilarities $\{\delta_{ij}\}$, in the sense that a minimum stress value is reached. These methods usually operate in an iterative manner by changing the configuration of points in each step, until either a minimum stress value or a specified maximum number of iterations is reached. These minimising methods will usually require an initial configuration of points. It is common practice to use the configuration produced by classical scaling as the initial configuration (Borg & Groenen, 2005, p.277). Random initial configurations, where points are randomly produced using a uniform distribution, can also be used. Sammon (1969) uses a steepest descent method. Loss functions can also be minimised by the SMACOF algorithm. Borg & Groenen (2005, p.187) mention that the SMACOF theory is "simple and more powerful than the steepest descent methods, because it guarantees monotone convergence of the loss function" .

The SMACOF algorithm used for metric MDS methods operates in an iterative manner by changing the configuration of points in each step of the algorithm. The SMACOF algorithm used to perform metric MDS is explained in detail in Borg & Groenen (2005, p.185-194). The reader should note however that the notation used by Borg & Groenen (2005) is different to the notation used in this study. Borg & Groenen (2005, p.270) suggest using Normalised Stress (4-2) rather than Raw Stress (4-1) as loss function, because using the latter may lead to *degenerate* solutions. These degenerate solutions are configurations that were obtained by making the loss function very small, irrespective of the relationship between distances and the data. The SMACOF algorithm ensures that the Normalised Stress value in (4-2) reaches a local minimum, but the local minimum may not be a global minimum. The steepest descent methods can also not guarantee that the local minimum found is indeed the global minimum. Borg & Groenen (2005, p.276) point out that local minima in MDS are not necessarily bad. A final configuration with a slightly worse fit is acceptable if it has a clearer interpretation than a configuration with a better fit. The problem of whether the local minimum is indeed the global minimum can be overcome in several ways. One possibility is to use multiple starting configurations where the whole SMACOF algorithm is repeated for each starting configuration and a minimum Normalised Stress value in (4-2) is noted. The final chosen configuration will be the overall configuration of all the configurations, produced from each starting configuration, which leads to the lowest Normalised Stress value in (4-2). Another possibility is to use the *tunneling method* (Borg & Groenen, 2005, p.278).

The SMACOF algorithm can also be used for nonmetric MDS to minimise loss functions. Nonmetric MDS will be discussed in the next section.

4.2.2 Nonmetric multidimensional scaling

Nonmetric multidimensional scaling is also known as ordinal multidimensional scaling. As mentioned earlier in Section 4.1, the nonmetric MDS method abandons the metric nature of the transformation $\delta_{ij}=f(d_{ij})$, where f can now be arbitrary. The only requirement for nonmetric MDS is that the transformation must preserve the rank order of the dissimilarities. The aim with Nonmetric MDS is to find an optimal configuration $\mathbf{X} : n \times m$ by matching the *disparities* $\{\hat{d}_{ij}\}$ to $\{d_{ij}\}$ by minimising a certain loss function. This is similar to the metric least squares scaling method, the difference being that the dissimilarities $\{d_{ij}\}$ in the loss functions are now replaced by disparities, $\{\hat{d}_{ij}\}$. The actual dissimilarity values are only used to determine the rank-order of the disparities, $\{\hat{d}_{ij}\}$. This means

$$d_{ij} < d_{kl} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{kl}\,.$$

The disparities are also sometimes called *pseudo distances*. These disparities are chosen in an optimal manner and will be discussed later.

The loss function used by Borg & Groenen (2005) for nonmetric MDS is very similar to the loss function used for metric MDS. This loss function will also be referred to as *Raw Stress,* with

$$\text{Raw Stress} = \sum_{i<j} w_{ij}(\delta_{ij} - \hat{d}_{ij})^2 \tag{4-3}$$

where δ_{ij} is the Euclidean distance between points i and j, $\hat{d}_{ij}$ is the disparity between objects i and j and the weights $\{w_{ij}\}$ must contain non-negative values. The weights are usually chosen as

$$w_{ij} = 0 \text{ if } i = j \text{ and } w_{ij} = 1 \text{ otherwise.}$$

Other values of w_{ij} are also allowed and different choices of w_{ij} lead to different loss functions (Borg & Groenen, 2005, p.255).

The Raw Stress value in (4-3) is a badness-of-fit measure, but it is not very informative. A large value does not necessarily indicate a bad fit, as it depends on the scale of distances in the configuration **X**: $n \times m$. *Normalised Stress* can be used to remove the scale dependency where

$$\text{Normalised Stress} = \frac{\sum_{i<j} w_{ij}(\delta_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} w_{ij}\hat{d}_{ij}^2}. \tag{4-4}$$

The aim of the nonmetric MDS method is to find an optimal configuration **X** : $n \times m$ of points, from which distances { δ_{ij} } and disparities { $\hat{d}_{ij}$ } can be calculated, that will minimise the Normalised Stress (4-4) or Raw Stress (4-3) loss functions. However, the minimising of the Normalised Stress (4-4) function is not an easy task. The minimising is usually done by an iterative process. The difference between the iterative process of the metric least squares scaling method and this iterative process is that the disparities { $\hat{d}_{ij}$ } also need to be optimally chosen, which depends on the distances {δ_{ij}}. The distances {δ_{ij}}, in turn, depend on the configuration **X** : $n \times m$, which changes during each iteration. Therefore, the disparities { $\hat{d}_{ij}$ } and distances {δ_{ij}} need to be optimally chosen during each step of the iteration. The SMACOF algorithm can again be used to minimise the Normalised Stress (4-4) or Raw Stress (4-3) loss functions.

The SMACOF algorithm used for the nonmetric MDS method is described in detail by Borg & Groenen (2005, p.200-209). The algorithm starts off with an initial configuration **X** : $n \times m$. It is common practice to use the configuration produced by classical scaling as the initial configuration. Random initial configurations, where points are randomly produced using the uniform distribution, can also be used. Distances {δ_{ij}} (usually Euclidean distances) will then be calculated from the initial configuration **X** : $n \times m$ of points. The disparities { $\hat{d}_{ij}$ } will then be optimally computed for the distances {δ_{ij}}, with the constraint that the disparities must have the same rank-order as the dissimilarities {d_{ij}}. Monotone (isotonic) regression (Borg & Groenen, 2005, p.205) or the Up-and-Down-Blocks algorithm (Borg & Groenen, 2005, p.206-208) can be used to choose optimal disparities { $\hat{d}_{ij}$ }. The optimal disparities, produced by the Monotone (isotonic) regression or the Up-and-Down-Blocks algorithm, must also be normalised to avoid a degenerate solution. Borg & Groenen (2005) normalise the disparities in such a way that

$$\sum_{i<j} w_{ij}\hat{d}_{ij}^{2} = \tfrac{n(n-1)}{2} .$$

The Normalised Stress (4-4) or Raw Stress (4-3) value can then be calculated, using the distances and disparities that were calculated during this step. The configuration $\mathbf{X}$: $n \times m$, will then be transformed to an updated configuration $\mathbf{X}^{[\mathbf{k}]}$: $n \times m$. This updated configuration will then be used in the next iteration. New distances will be calculated from $\mathbf{X}^{[\mathbf{k}]}$: $n \times m$, and new disparities will then be calculated, which will depend on these new distances. The SMACOF procedure guarantees that the updated configurations will never lead to a higher Normalised Stress value in (4-4). The iteration continues until either a minimum Normalised Stress value in (4-4) or the specified maximum number of iterations is reached.

Borg & Groenen (2005, p.270) suggest using Normalised Stress (4-4) rather than Raw Stress (4-3) as loss function, because using the latter may lead to degenerate solutions. Heiser (1991) also points out that negative disparities could lead to degenerate solutions. The SMACOF algorithm ensures that the Normalised Stress value in (4-4) reaches a local minimum, but the local minimum may not be a global minimum. As mentioned earlier, the problem of whether the local minimum is indeed the global minimum can be overcome using multiple initial configurations or by using the tunneling method.

The Shepard diagram and bubble plot can be used to describe the badness-of-fit. The Shepard diagram plots the disparities {$\hat{d}_{ij}$} and distances {δ_{ij}} on the same graph, which gives an indication of how well the disparities are fitted to the distances. The (d_{ij}, $\hat{d}_{ij}$) pairs are plotted and these pairs all lie on a monotonically increasing regression line. The (d_{ij}, δ_{ij}) pairs are also plotted. The vertical distance between these points gives a measure of the corresponding error e_{ij} = ($\delta_{ij} - \hat{d}_{ij}$). The error gives an indication of the size of mismatch between the distances and the corresponding disparities. Larger errors will cause a higher Normalised Stress (4-4) and Raw Stress (4-3) value. The bubble plot can be used to assess the fit of each point in the nonmetric MDS display and uses the *Stress per point* measure, which is the average of the squared errors $e_{ij}^{2} = (\delta_{ij} - \hat{d}_{ij})^{2}$ between the current object and all other objects (Borg & Groenen, 2005, p.54).

4.3 Application of the MDS Methods to the Medical Scheme 55-59 Data Set

The MDS techniques previously discussed will be used to display the chronic diseases of the Medical Scheme 55-59 data set. The chronic diseases will be displayed in several two-dimensional displays where each point in the display represents one of the chronic diseases. Chronic diseases that tend to co-occur are expected to be situated close to each other and chronic diseases that do not co-occur often are expected to be further apart. This can be used to assess whether chronic diseases mentioned in the same body system rule are related.

The dissimilarities formed by the transformation $d_{ij} = (1-s_{ij})^{0.5}$ which were calculated in Chapter 3, will be used as input for these MDS techniques. This transformation leads to more dissimilarity coefficients with metric and Euclidean properties, which is preferable for the metric MDS methods.

4.3.1 Metric multidimensional scaling

4.3.1.1 Classical scaling

The *R* function named *classical.scaling* was developed to implement the classical scaling method. Cox & Cox (2001, p.138) provide a practical algorithm of the classical scaling method, and it was this algorithm which was implemented in the *classical.scaling* function. The details of the function *classical.scaling* are provided in Appendix A.

The dissimilarities $\{d_{ij}\}$ measured between all pairs of chronic diseases will be used as input for this function. It was shown in Chapter 3 how these dissimilarities based on binary data could be constructed and these dissimilarities were calculated with the *R* function *Dissim.CDL*. Only dissimilarity coefficients with metric or Euclidean properties will be considered in this section.

The classical scaling method can only be performed when **B** is a positive semi-definite matrix of suitable rank, where **B** is derived from the dissimilarities. Dissimilarity coefficients, without metric or Euclidean properties, may sometimes lead to a positive semi-definite matrix **B**. But this depends on the actual data set. It is quite possible that these dissimilarity coefficients without metric properties may not lead to a positive semi-definite matrix **B**, and the classical scaling method will fail in these cases. However, dissimilarity coefficients with Euclidean properties will always lead to a positive semi-definite matrix **B** of suitable rank. This is one reason why only dissimilarity coefficients with Euclidean properties are considered when classical scaling is used in this section. Another reason is that the classical scaling method uses dissimilarities directly as distances. It will therefore be inappropriate to use dissimilarity coefficients that do not have metric properties, even if these dissimilarities lead to a positive semi-definite matrix **B** of suitable rank.

It was mentioned in Chapter 3 that the *Simple matching* and the *Rao-Russell* similarity coefficients are unlikely to be suitable for use on the Medical Scheme 55-59 data set, because the chronic diseases are asymmetric binary variables. The *Simple matching* and the *Rao-Russell* similarity coefficients are usually used when a data set consists of symmetric binary variables. These coefficients however will also be used in this section to provide some evidence that these two similarity coefficients are indeed unsuitable, even though they have Euclidean properties.

The *classical.scaling* function displays the chronic diseases in a two-dimensional plot where a unit change in the horizontal direction is equal to a unit change in the vertical direction. Otherwise distances cannot be appreciated, which is vital for interpreting the graphical displays.

It is also apparent that most of these graphical displays explain a small proportion of the point variability, because the full dimensional space consists of 26 ($n-1$) dimensions and the display space is only two-dimensional. This two-dimensional space however is still the optimal display space, as it explains most of the point variability when compared to any other two-dimensional space. These graphical displays tend to show some over-crowding of chronic diseases in some areas and the low proportion of variability explained by these graphical displays might be a cause for this over-crowding. The reader will sometimes find it difficult to read the labels of the points that represent the chronic diseases located in the over-crowded areas. It was therefore necessary to provide an enlargement of these over-crowded areas.

The classical scaling display based on the *Jaccard* dissimilarity coefficient is provided in Figure 4.1. The following instructions in *R* were used to construct Figure 4.1:

```
All.Dissim<-Dissim.CDL(x=final.dat[c(1:1636),c(1:28)],transformation=2)
classical.scaling(D=All.Dissim$Jaccard,cex=0.55)
```

The first line is used to construct all dissimilarity matrices based on the different similarity coefficients. The transformation $d_{ij} = (1-s_{ij})^{0.5}$ was used to construct the dissimilarities. Figures 4.2

to 4.7 represent classical scaling plots, based on the other dissimilarity coefficients, and was constructed in a similar manner.

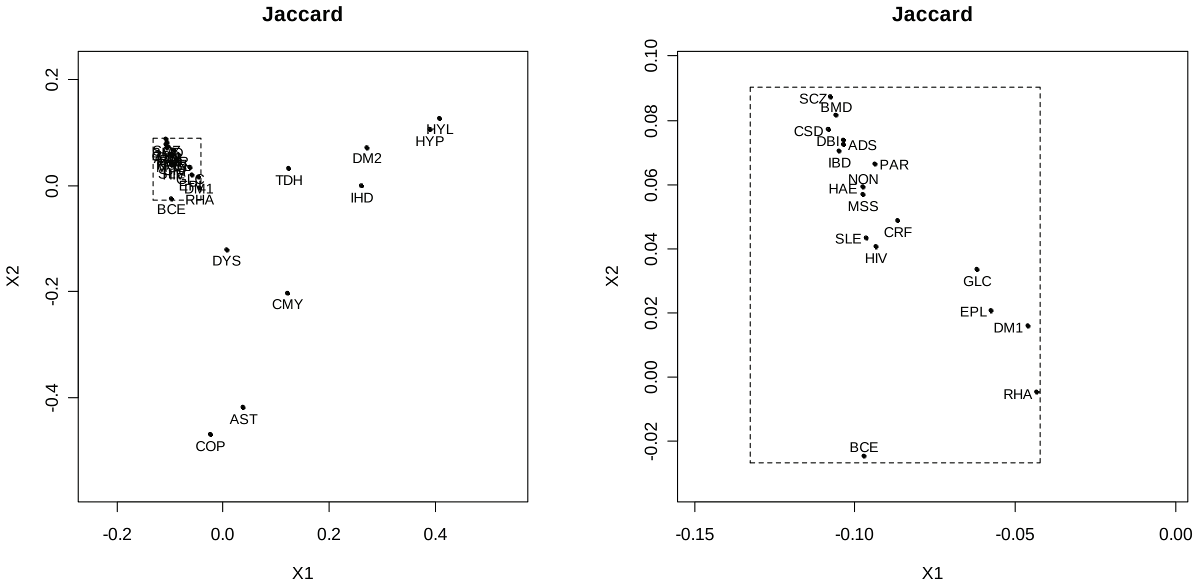


Figure 4.1: *Classical scaling of the Medical Scheme 55-59 data set using the Jaccard dissimilarity coefficient. The points in the display represent the various chronic diseases (The two dimensions explain 9.2% of the point variability). The panel on the right represents an enlargement of an area that is over-crowded in the left panel.*

The classical scaling plots based on various dissimilarity coefficients with Euclidean properties are different. Notice how the *Rao-Russell* and the *Simple matching* dissimilarity coefficients display totally different plots when compared to the plots based on the other dissimilarity coefficients. A possible reason for this is that the *Rao-Russell* and the *Simple matching* dissimilarity coefficients are usually used when the data consist of symmetric binary variables, which is not the case with the Medical Scheme 55-59 data set. The chronic diseases are ranked in Chapter 2 in accordance with the chronic disease rates that they display. It is shown that HYP, HYL, DM2, TDH, AST, IHD, CMY and RHA display the highest chronic disease rates (in decreasing order). Notice how these chronic diseases tend to be displayed in a similar order in Figures 4.6 and 4.7. It is also evident that the points representing chronic diseases with lower chronic disease rates are situated very close to each other. The *Rao-Russell* and the *Simple matching* dissimilarity coefficients are therefore unsuitable because the chronic diseases that do not occur often form strong similarities, even though these chronic diseases do not actually co-occur that often. The other dissimilarity coefficients used in this study are based on asymmetric binary variables and are more suitable for use on the Medical Scheme 55-59 data set.

The classical scaling plots based on the *Jaccard*, *Dice-Sorensen*; and *Sokal-Sneath-Anderberg* dissimilarity coefficients are very similar. Anderberg (1973, p.90) showed that the *Jaccard*, *Dice-Sorensen* and *Sokal-Sneath-Anderberg* similarity coefficients are all monotonic to each other. That is the reason why the three plots are so similar.

Figures 4.1 to 4.5 will be used to investigate whether the chronic diseases mentioned in the same body system rule are related. It is also important to investigate which other chronic diseases, not mentioned in the body system rules, also show a strong relation.

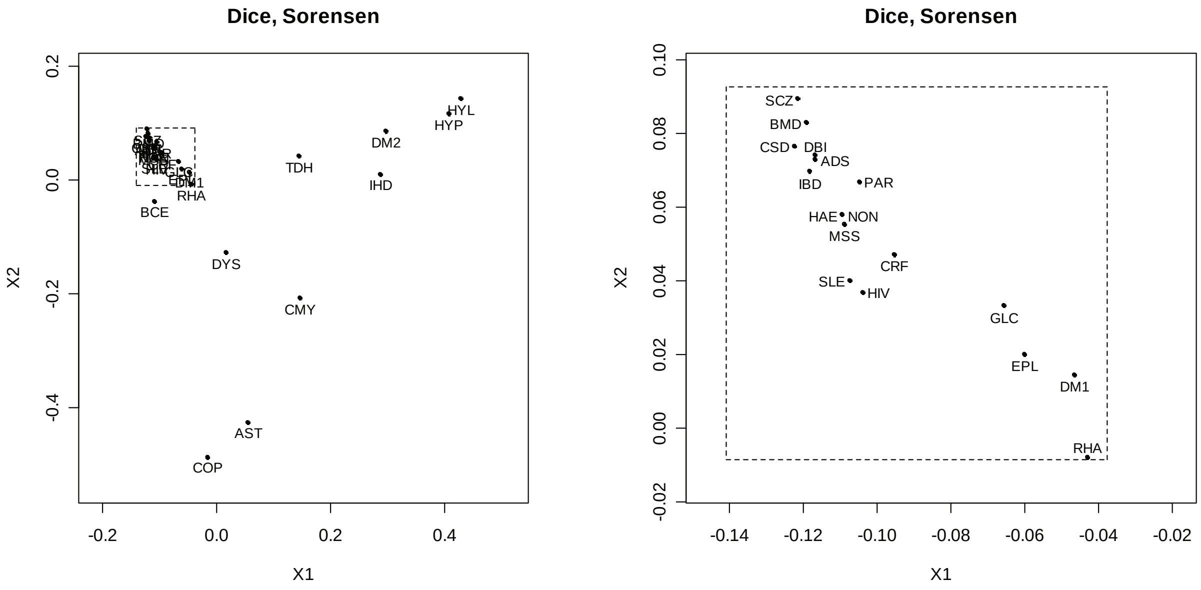


Figure 4.2: *Classical scaling of the Medical Scheme 55-59 data set using the Dice-Sorensen dissimilarity coefficient. The panel on the right represents an enlargement of an area that is over-crowded in the left panel. (The two dimensions explain 10.4% of the point variability).*

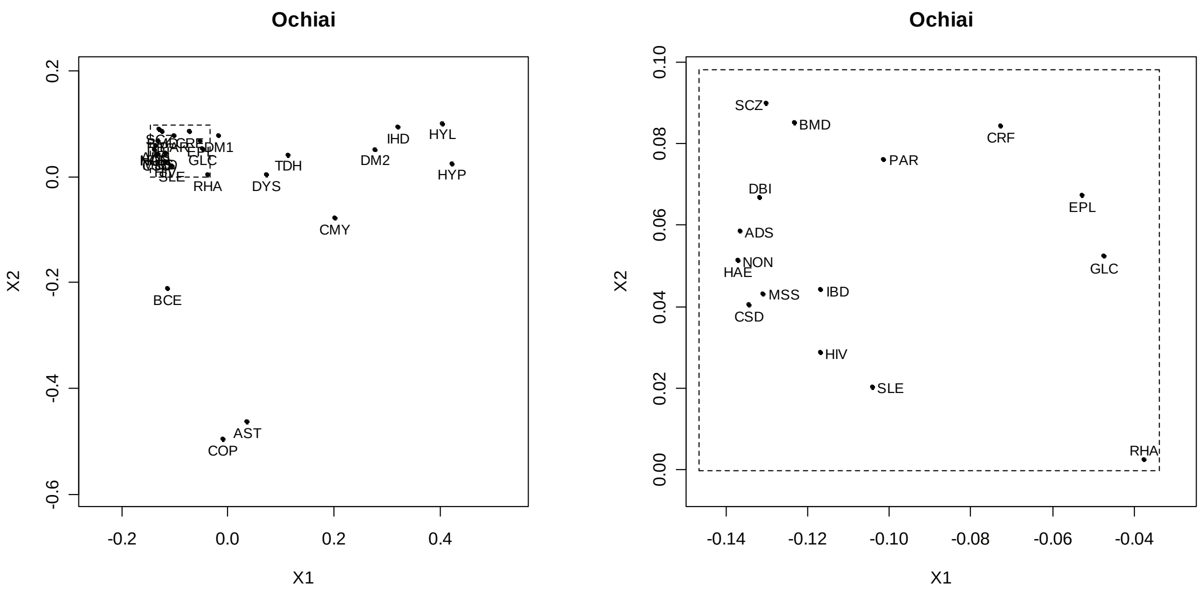


Figure 4.3: *Classical scaling of the Medical Scheme 55-59 data set using the Ochiai dissimilarity coefficient. The panel on the right represents an enlargement of an area that is over-crowded in the left panel. (The two dimensions explain 10.92% of the point variability).*

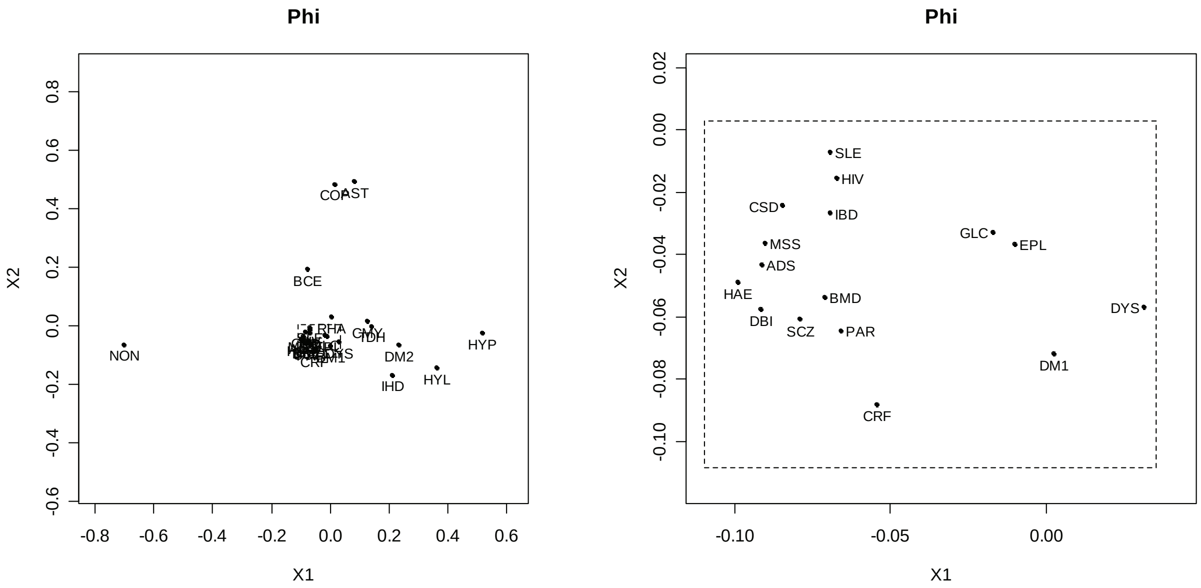


Figure 4.4: *Classical scaling of the Medical Scheme 55-59 data set using the Phi dissimilarity coefficient. The panel on the right represents an enlargement of an over-crowded area in the left panel. (The two dimensions explain 13.35% of the point variability).*

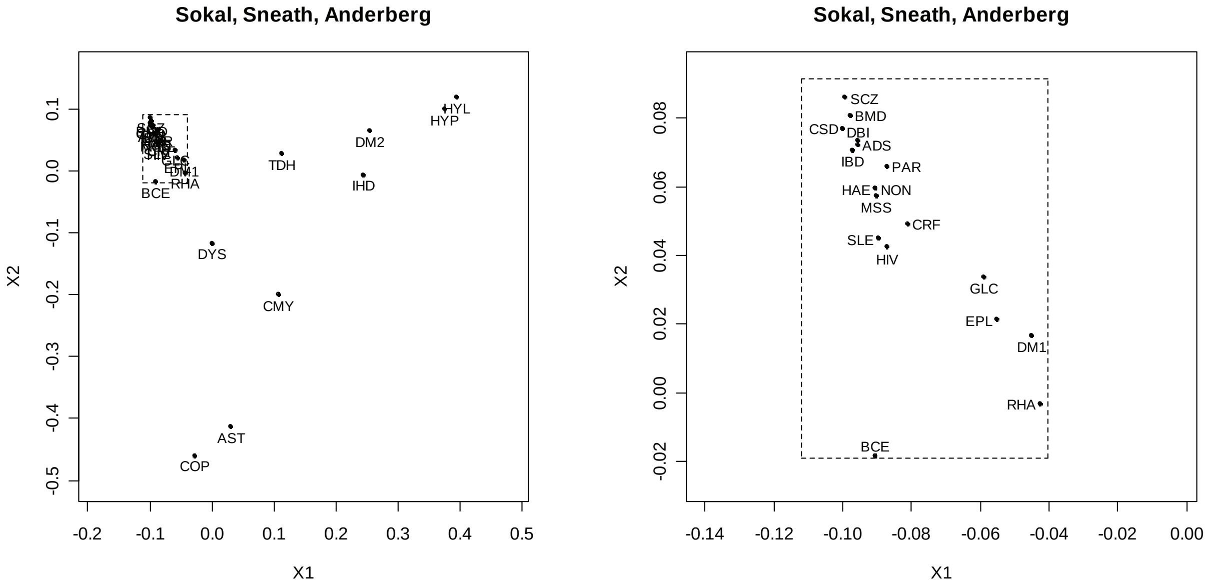


Figure 4.5: *Classical scaling of the Medical Scheme 55-59 data set using the Sokal-Sneath-Anderberg dissimilarity coefficient. The panel on the right represents an enlargement of an over-crowded area in the left panel. (The two dimensions explain 8.49% of the point variability).*

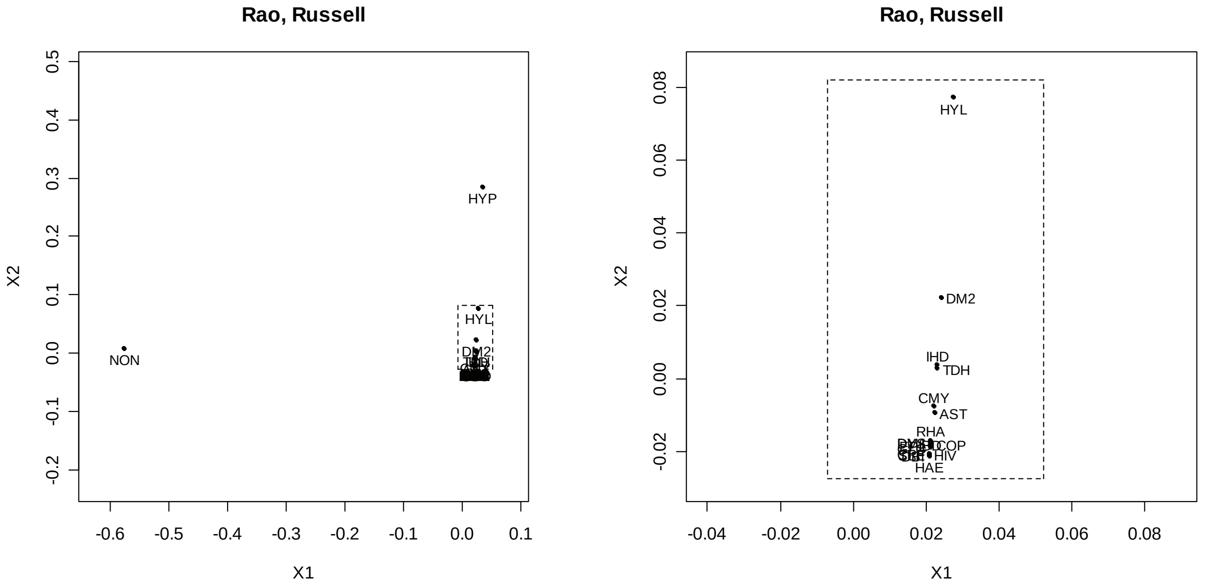


Figure 4.6: *Classical scaling of the Medical Scheme 55-59 data set using the Rao-Russell dissimilarity coefficient. (The two dimensions explain 80.82% of the point variability). The panel on the right represents an enlargement of an over-crowded area in the left panel. Unfortunately, the enlargement is also over-crowded.*

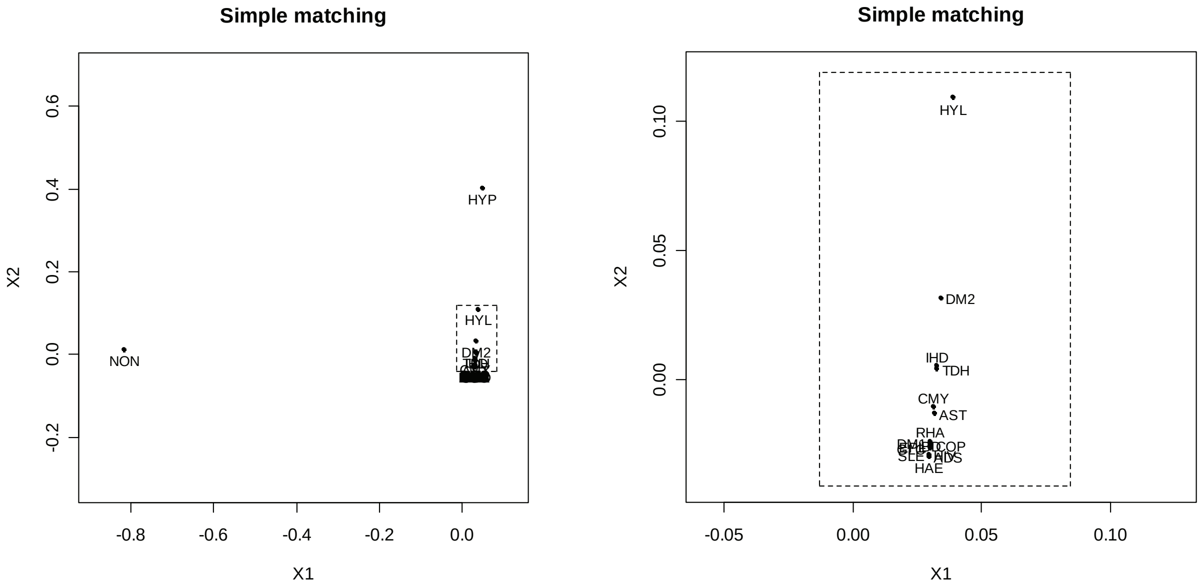


Figure 4.7: *Classical scaling of the Medical Scheme 55-59 data set using the Simple matching dissimilarity coefficient. The panel on the right represents an enlargement of an over-crowded area. (The two dimensions explain 80.82% of the point variability).*

The first body system rule involves the respiratory diseases COP, AST and BCE. Many of the classical scaling plots appear to show that these chronic diseases are strongly related. The classical scaling plots based on the *Phi* and *Ochiai* coefficients show that the points representing these chronic diseases are not far apart. It seems that COP and AST are more strongly related, because the point representing BCE is usually further apart. The BCE point is also close to the other chronic diseases not mentioned in the first body system rule.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. The classical scaling plots based on the suitable dissimilarity coefficients show that these diseases are strongly related. These cardiovascular diseases also seem to co-occur often with chronic diseases HYL, DM2, TDH.

The third body system rule involves HYP and CRF. It seems that HYP tends to co-occur more often with HYL than with CRF. It must be remembered though that these plots explain a very small proportion of the point variability, and this makes the results less convincing.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the classical scaling plots show that the points that represent these two diseases are always fairly close to each other. Notice also that these points are always situated in the over-crowded area. It therefore appears that these two chronic diseases co-occur often.

The fifth body system rule involves the chronic diseases BMD and SCZ. It seems that these two chronic diseases are also strongly related. Notice that these the points representing these two diseases are also situated in the over-crowded area.

The sixth body system rule involves MSS, BMD and EPL. The points that represent these chronic diseases always seem to be fairly close to each other and are always situated in the over-crowded area. Therefore, it appears that these chronic diseases might be related.

The seventh body system rule involves the chronic diseases SLE and RHA. The points that represent these chronic diseases always seem to be fairly close to each other which suggests that these chronic diseases co-occur often.

It is problematic that the classical scaling plots explain a very small proportion of the point variability, and this makes the results less convincing and leads to certain areas being over-crowded. These over-crowded areas seem to suggest that the chronic diseases found in over-crowded areas co-occur often. This however is not always the case. Figure 2.3 showed that the chronic disease HAE did not co-occur with any other chronic disease. The lives that are not treated for any chronic diseases or HIV are classified as NON. This implies that NON cannot co-occur with any of the chronic diseases. It should therefore be expected that the points representing NON and HAE should be relatively far away from the points representing the other chronic diseases, but this is not the case. Notice that the points representing HAE and NON are usually found in this over-crowded area, suggesting that HAE and NON co-occur often with the other chronic diseases that are located in this crowded area. The reason for this apparent error is the low point variability explained by the two-dimensional display space. It is likely that the points representing NON and HAE will be much further apart from other points if all dimensions could be visually displayed. Unfortunately, it is only possible to display these points visually in a one, two or three dimensional plot.

4.3.1.2 Metric least squares scaling

The chronic diseases are displayed in a two-dimensional plot where the configuration of points is derived using the SMACOF algorithm to minimise Normalised Stress in (4-2). The *R* function named *SMACOF.metric* was developed to implement the SMACOF algorithm for metric MDS. Borg & Groenen (2005, p.192) provide a practical algorithm to minimise loss functions associated with metric MDS, and it was this algorithm which was implemented in the *SMACOF.metric* function. The *SMACOF.metric* function displays the chronic diseases in a two-dimensional plot where a unit change in the horizontal direction is equal to a unit change in the vertical direction. Otherwise distances cannot be appreciated, which is vital for interpreting the graphical displays. The function also produces a residual plot and a bubble plot. The details of this function are provided in Appendix A.

The dissimilarities $\{d_{ij}\}$ measured between all pairs of chronic diseases will be used as input for this function. It was shown in Chapter 3 how these dissimilarities based on binary data could be constructed and these dissimilarities were calculated with the *R* function *Dissim.CDL*. Only dissimilarity coefficients with metric or Euclidean properties will be considered in this section. The metric and Euclidean properties of the various dissimilarity coefficients are shown in Table 3.3. It was shown in Section 4.3.1.1 that the *Simple matching* and the *Rao-Russell* similarity coefficients are not suitable, because the chronic diseases are not symmetric binary variables. Only the *Jaccard*, *Dice-Sorensen*, *Sokal-Sneath-Anderberg*, *Phi* and *Ochiai* dissimilarity coefficients will be used in this section. The SMACOF algorithm will also require an initial configuration of points. The configurations produced by classical scaling will be used as initial configurations.

Various graphical displays based on the different dissimilarity coefficients with metric and Euclidean properties will be shown in this section. Points in the metric least squares scaling plots and bubble plots will represent the various chronic diseases. These plots can be used to assess whether chronic diseases mentioned in the same body system rule are related. The bubble plot is used to assess the fit of the various chronic diseases. The residual plot also provides some information regarding the dissimilarities and the badness-of-fit.

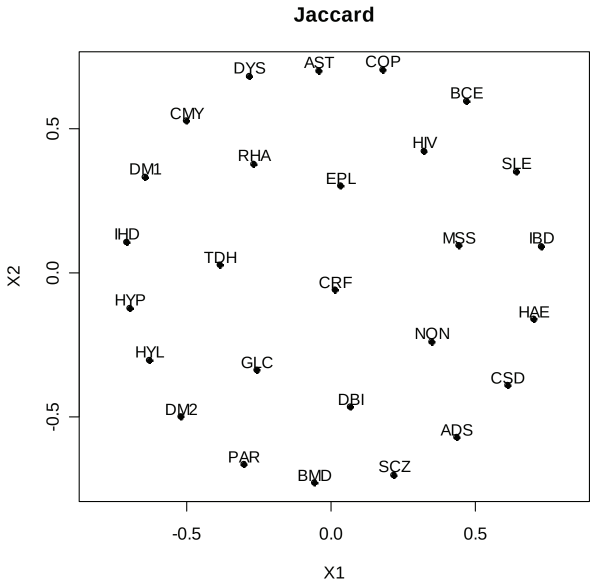


Figure 4.8: *Metric least squares scaling plot of the Medical Scheme 55-59 data set using the Jaccard dissimilarity coefficient. (Normalised Stress in (4-2) = 0.149).*

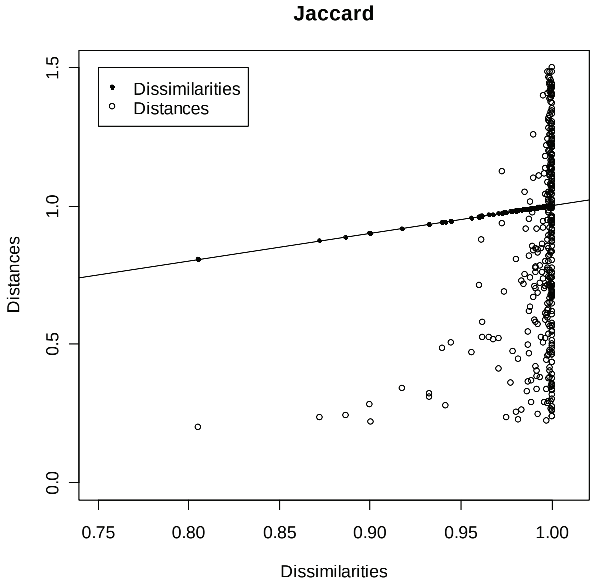


Figure 4.9: *Residual plot, which gives an indication of the badness-of-fit. The filled circles represent the dissimilarities and are located on the bisector. The open circles represent the corresponding distances. The vertical distance between these open and closed circles is the corresponding error of fit. Large vertical distances indicate large errors, and hence poor fitting of distances.*

Figure 4.8 provides the metric least squares scaling plot of the chronic diseases using the *Jaccard* dissimilarity coefficient. Points representing the chronic diseases that are close to each other indicate that the corresponding chronic diseases are strongly related.

The following instruction in *R* was used to construct Figures 4.8, 4.9 and 4.10:

```
SMACOF.metric(delta=All.Dissim$Jaccard,Bubble.plot=TRUE,Main="Jaccard").
```

The chronic diseases in Figure 4.8 do not appear to form any distinct clusters. It seems that the chronic diseases are equally spaced on circles and Borg & Groenen (2005, p.274) mention that this could be caused by equal dissimilarities. Figure 4.9 provides information on the dissimilarities and distances. Notice how all the dissimilarities have indeed very similar values. Figure 4.10 is a bubble plot that provides information on how well the chronic distances are fitted. Larger bubbles indicate chronic diseases that have better fit. For example, notice that HYP has a much better fit than CRF. The bubble plot has the same configuration of points than metric least squares scaling plot, but provides additional information and can therefore be used in place of the metric least squares scaling plot.

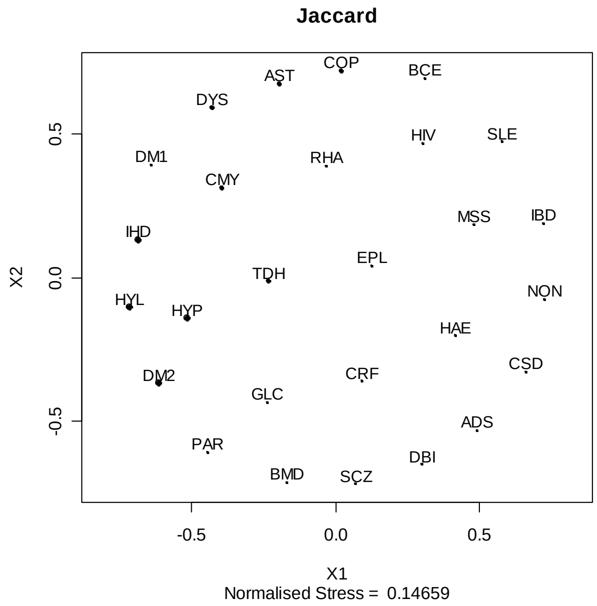


Figure 4.10: *Bubble plot of the Medical Scheme 55-59 data set using the Jaccard dissimilarity coefficient. Larger bubbles indicate chronic diseases that have better fit, i.e. chronic diseases that have a low Stress per point. See definition 4.1. (Normalised Stress in (4-2)= 0.149).*

The Normalised Stress value in (4-2) has the following lower and upper bounds in a two-dimensional display: [0, 0.4352], where values closer to zero indicate better fit. The Normalised Stress value of 0.14889 is reasonable which indicates that the distances in the metric least squares scaling are a reasonable approximation of the dissimilarities.

The weights that were used in Figures 4.8 to 4.10 are given as

$$w_{ij} = 1 \text{if } i \neq j \text{ and } w_{ij} = 0 \text{ if } i = j.$$

Different choices of weights can also be used. More generally, let $w_{ij} = \delta_{ij}^{q}$. Different choices of *q* can be used to emphasize the representation of small or large dissimilarities. Setting q = 0 will result in all $w_{ij} = 1$. This choice however, as seen in Figure 4.8, does not lead to a metric least squares scaling plot with a clear clustering structure. A possible reason for this is that most of the dissimilarities have very similar values. Large positive values of *q* may lead to a better representation of larger dissimilarities, but not smaller dissimilarities. If the dissimilarities have some clustering, then choosing a large value of *q* may reveal a clearer clustering structure (Borg & Groenen, 2005, p.256). This is confirmed by Figure 4.11 where metric least squares scaling plots and residual plots for different values of *q* are shown. The metric least squares scaling plot with weights $w_{ij} = \delta_{ij}^{100}$ has a clearer clustering structure. The corresponding residual plot shows that larger dissimilarities are better represented than the smaller dissimilarities, because the larger errors are associated with the smaller dissimilarities. Hence, smaller distances should be interpreted with care and larger distances can be interpreted in the usual manner.

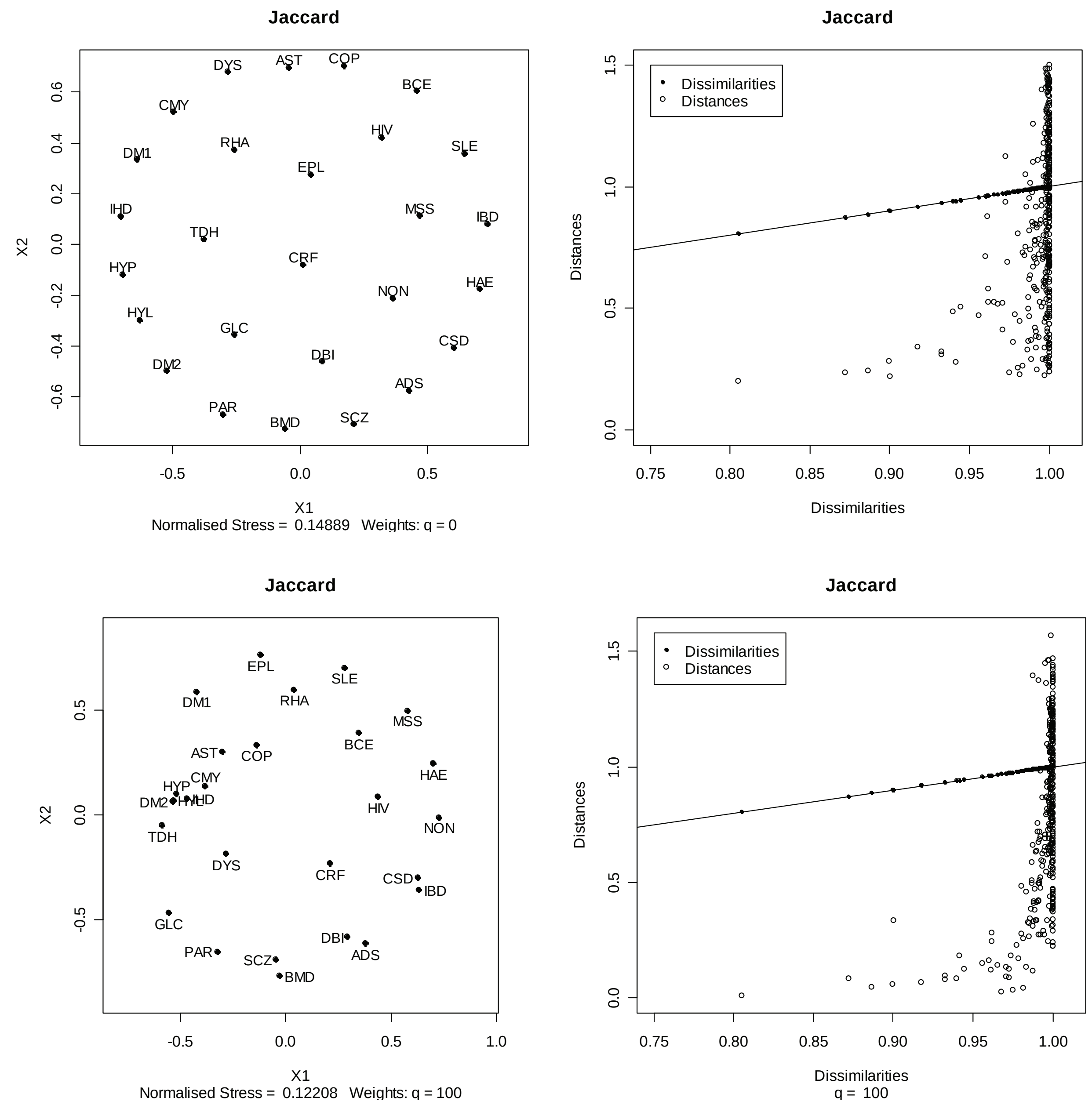


Figure 4.11: *Metric least squares scaling plots (left panels) and the corresponding residual plots (right panels) of the chronic diseases, where different weights* $w_{ij} = \delta_{ij}^{q}$ *are used. Larger positive values of q lead to a metric least squares scaling plot with a clearer clustering structure.*

It appears that large positive *q* values should be used when metric least squares scaling plots are constructed, as this leads to a clearer clustering structure. The other dissimilarity coefficients will also be used to display the chronic diseases. Figure 4.12 provides metric least squares scaling plots based on weights with large positive *q* values. Figures 4.11 and 4.12 can be used to describe how the different chronic diseases seem to be related. Metric least squares scaling plots with clearer clustering structures should be preferred.

The first body system rule involves the respiratory diseases COP, AST and BCE. Many of the metric least squares scaling plots seem to show that these diseases are strongly related, because the points representing these chronic diseases are usually not far apart. It seems that COP and AST are more strongly related, because the points representing BCE are usually further apart from AST and COP.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. The metric least squares scaling plots, with clearer clustering structures, show that these chronic diseases are strongly related. These cardiovascular diseases also seem to co-occur with chronic diseases HYL, DM2, TDH.

The third body system rule involves HYP and CRF. It appears that HYP co-occurs more often with HYL than with CRF. The chronic disease CRF, on the other hand, appears to co-occur often with DM1.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. The points that represent these two diseases are always fairly close to each other. It therefore appears that these two chronic diseases co-occur often and are strongly related.

The fifth body system rule involves the chronic diseases BMD and SCZ. These chronic diseases also appear to be strongly related, because these chronic diseases always tend to be located close to each other in the metric least squares scaling plots.

The sixth body system rule involves MSS, BMD and EPL. The points that represent the chronic diseases BMD and EPL tend to be fairly close to each other, but the point representing MSS tend to be further away. It is therefore not clear if MSS is strongly related to BMD and EPL.

The seventh body system rule involves the chronic diseases SLE and RHA. Some of the metric least squares scaling plots with clearer clustering structures do show that the points that represent these chronic diseases are situated fairly close to each other. It therefore seems that these chronic diseases tend to co-occur.

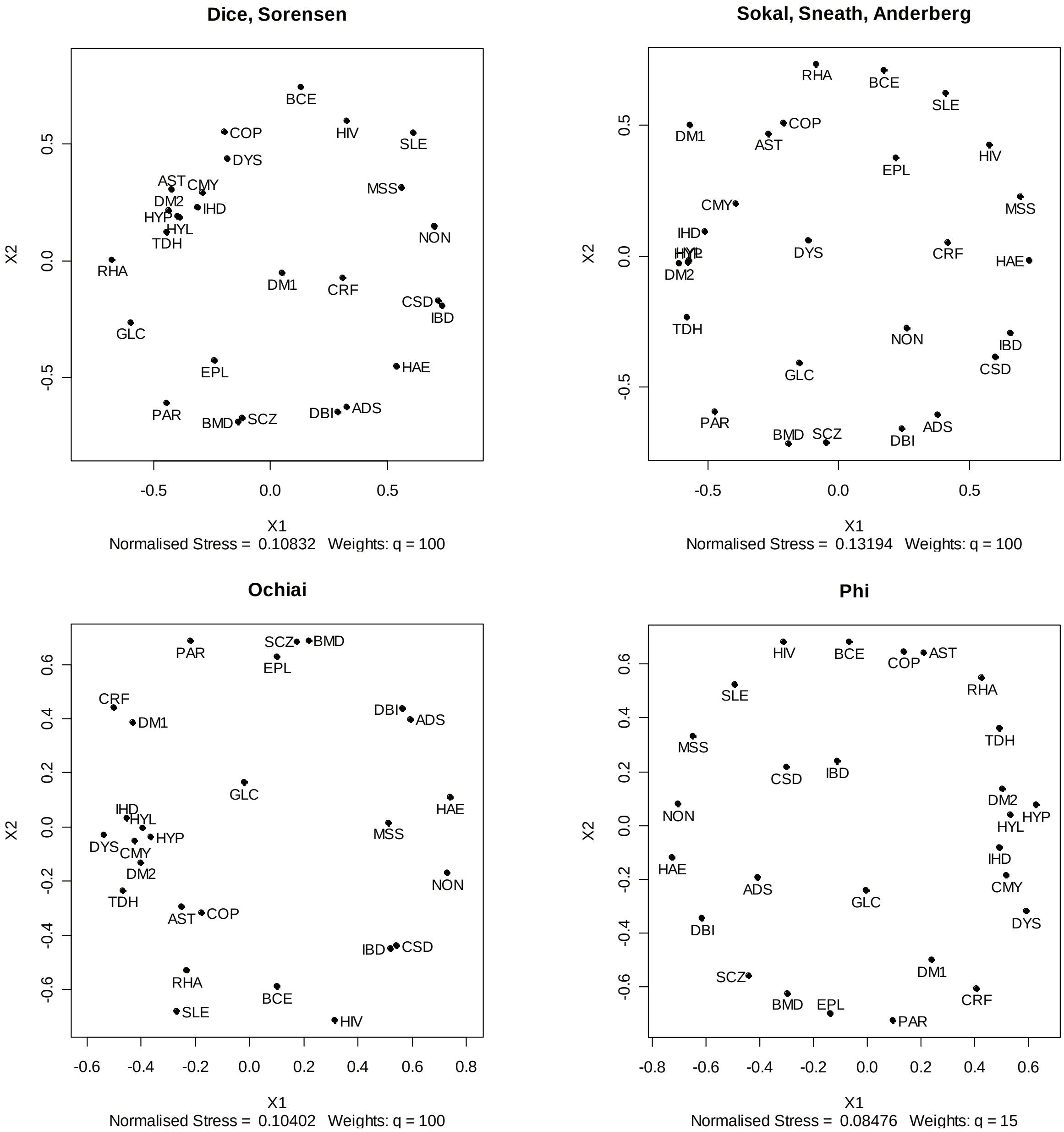


Figure 4.12: *Metric least squares scaling plots of the chronic diseases using different dissimilarity coefficients. Different weights* $w_{ij} = \delta_{ij}^{q}$ *are used and the values of q are displayed for each of the metric least squares scaling plots.*

The last system body rule involves DM1 and DM2. It seems that DM1 tends to co-occur with CRF, and that DM2 tends to co-occur with the cardiovascular diseases CMY, IHD, HYP and DYS.

It also seems that the chronic diseases ADS and DBI are strongly related, as the points representing these two diseases are situated close to each other in Figure 4.12. These chronic diseases however are not mentioned in any of the body system rules. It also seems that the respiratory chronic diseases AST and COP are related to the cardiovascular related chronic diseases and chronic diseases HYL, TDH and DM2, as these points tend to be situated relatively close to each other in Figure 4.12.

The metric least squares scaling plots with a clearer clustering structure are preferred to the classical scaling plots, because the classical scaling plots explain a very small proportion of the point variability. This causes an over-crowding of points in the classical scaling plot which makes interpretation more difficult. The metric least squares scaling plots with the clearer clustering structure are easier to interpret. The disadvantage of the metric least squares scaling plots is the subjective choice of loss function, weights and minimising methods. A different choice of these factors will lead to a different metric least squares scaling plot, which may lead to a different interpretation. The classical scaling method is far less subjective and produces the same configuration of points, up to a rotation of the points. This rotation does not distort the relative distances between the points. Furthermore, the classical scaling method takes less computation time than the metric least squares scaling method, because the iterative procedure to minimise the loss function may take very long in some cases. This however was not a problem with the Medical Scheme 55-59 data set as only 27 objects are presented in the metric least squares scaling plots.

The classical scaling and metric least squares scaling methods only use dissimilarity coefficients with metric or Euclidean properties. The nonmetric MDS method in Section 4.3 however also uses dissimilarity coefficients without metric properties.

4.3.2 Nonmetric multidimensional scaling

Nonmetric scaling will be applied to the Medical Scheme 55-59 data set in this section. The SMACOF algorithm for nonmetric MDS will be used to find a two-dimensional display that produces the lowest Normalised Stress value in (4-4). The *R* function named *SMACOF.Nonmetric* was developed to implement the SMACOF algorithm for nonmetric MDS. The details of this function are provided in Appendix A. Borg & Groenen (2005, p.205) provide a practical algorithm to minimise loss functions associated with nonmetric MDS, and it was this algorithm which was implemented in the *SMACOF.Nonmetric* function. The *SMACOF.Nonmetric* function displays the chronic diseases in a two-dimensional plot where a unit change in the horizontal direction is equal to a unit change in the vertical direction. The function also produces a Shepard diagram and a bubble plot.

The dissimilarities $\{d_{ij}\}$ measured between all pairs of chronic diseases will also be used as input for this function. It was shown in Section 4.3.1.1 that the *Simple matching* and the *Rao-Russell* similarity coefficients are not suitable when applied on the Medical Scheme 55-59 data set. These two similarity coefficients will therefore not be used in this section. The SMACOF algorithm will also require an initial configuration of points. The configurations produced by classical scaling can be used as initial configurations, but random initial configurations will also be considered in order to try to avoid local minima. The weights that will be used in this section are given as

$$w_{ij} = 1 \text{if } i \neq j \text{ and } w_{ij} = 0 \text{ if } i = j.$$

Various graphical displays based on the different dissimilarity coefficients will be shown in this section. Points in the nonmetric MDS plot and bubble plot will represent the various chronic diseases. The bubble plot is very similar to the nonmetric MDS plot. The bubble plot provides additional information of the goodness-of-fit of the chronic diseases and can therefore be used in place of the nonmetric MDS plot. Larger bubbles in the bubble plot indicate chronic diseases that have better fit. Chronic diseases that tend to co-occur will be situated close to each other in the bubble plot and chronic diseases that do not co-occur often will be further apart. The bubble plot can therefore be used to assess whether chronic diseases mentioned in the same body system rule are related. The Shepard diagram will also provide some information regarding the dissimilarities, fitted disparities, distances between points in the configuration and the badness-of-fit.

The following *R* instruction was used to construct the bubble plot and Shepard diagram in Figure 4.13:

```
SMACOF.Nonmetric(display.plots=TRUE,Main="Jaccard",
delta=All.Dissim$Jaccard,maxit =400, epsil =0.0000001)
```

Figures 4.14 to 4.21 were constructed in a similar manner. The bubble plots and Shepard diagrams based on the various dissimilarity coefficients are displayed in these figures. The configurations produced by classical scaling were used as initial configurations. Many MDS computer programs use the configuration produced by classical scaling as initial configuration (Borg & Groenen, 2005, p.277).

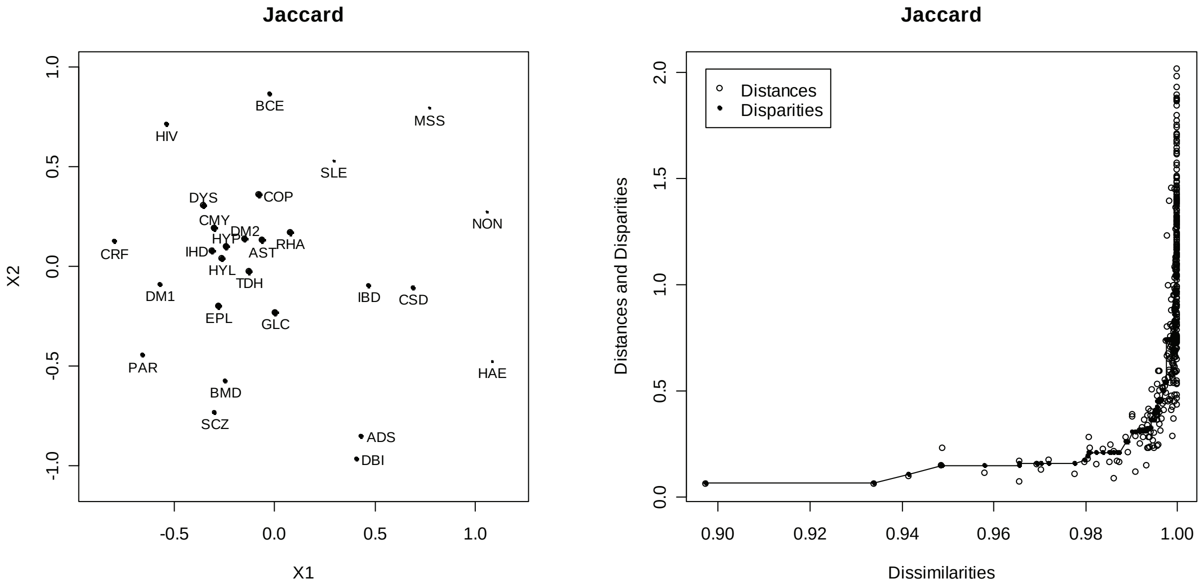


Figure 4.13: *Bubble plot (left panel) and Shepard diagram (right panel) produced using nonmetric MDS. The Jaccard dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.067). Large bubbles in the bubble plot indicate points with better fit. The vertical distance between disparities and distances in the Shepard diagram represent the errors of fit. A monotonic regression line of disparities is also provided.*

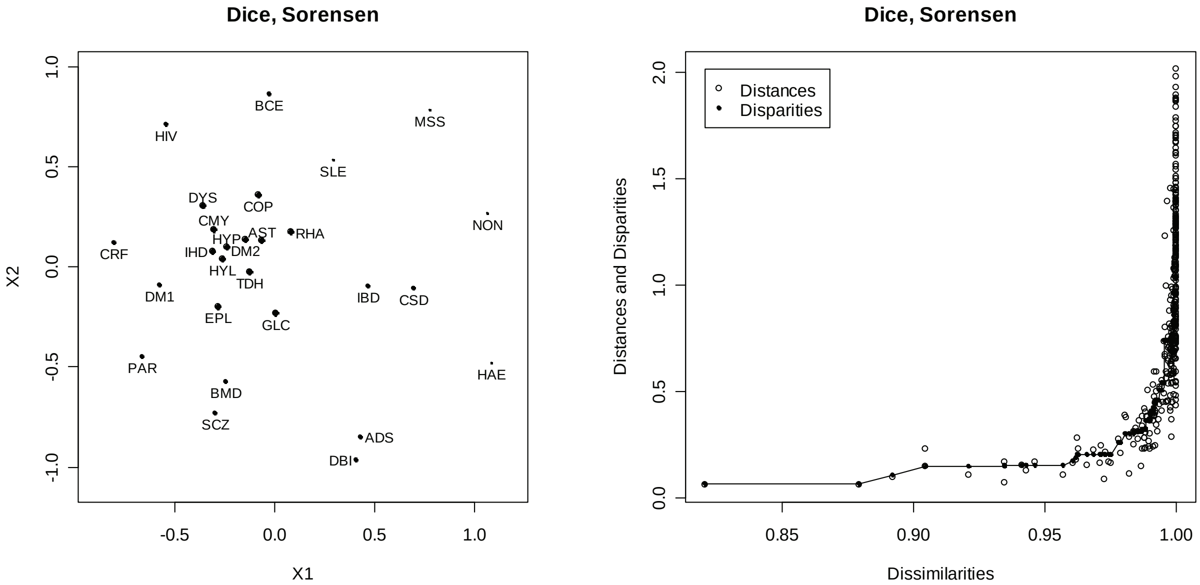


Figure 4.14: *Bubble plot (left panel) and Shepard diagram (right panel) produced using nonmetric MDS. The Dice-Sorensen dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.067).*

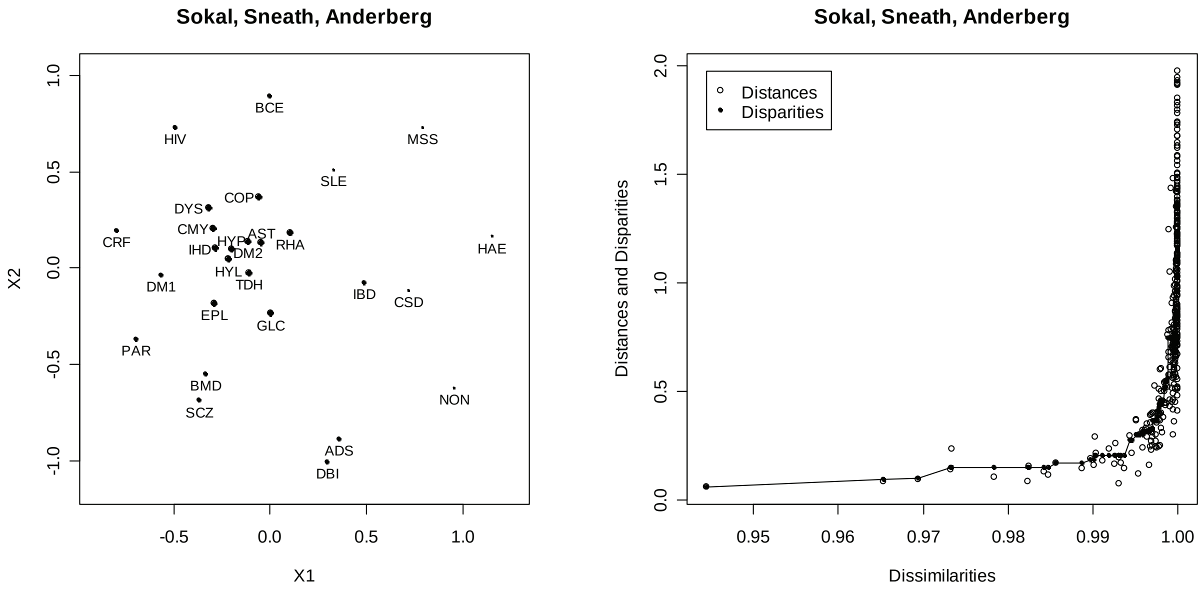


Figure 4.15: *Bubble plot (left panel) and Shepard diagram (right panel) produced using nonmetric MDS. The Sokal-Sneath-Anderberg dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.067).*

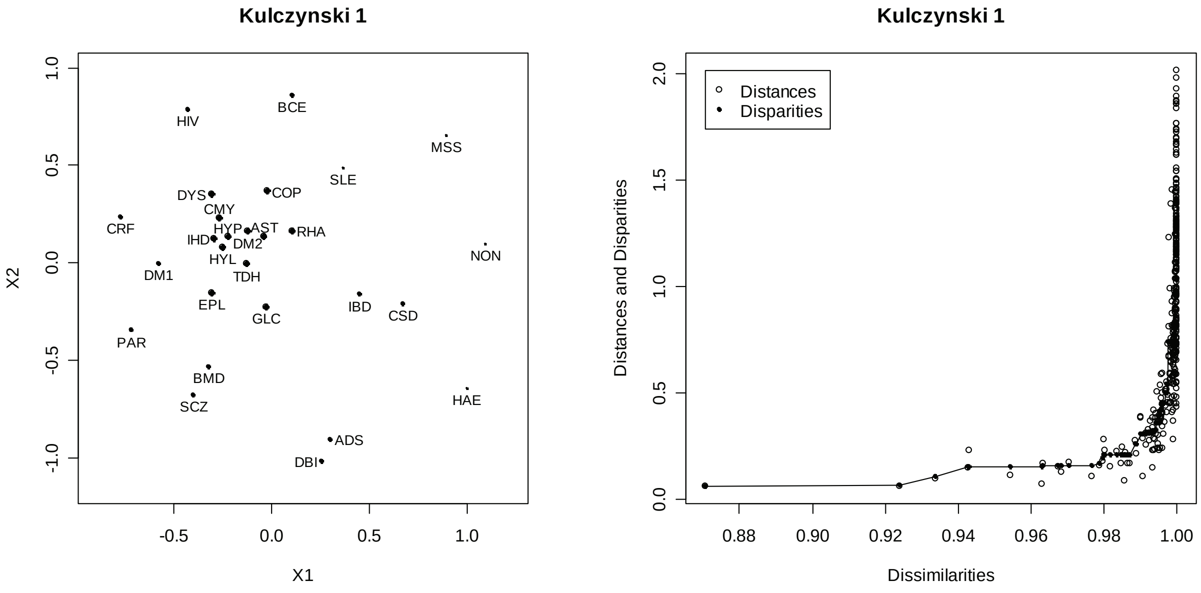


Figure 4.16: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Kulczynski 1 dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.067).*

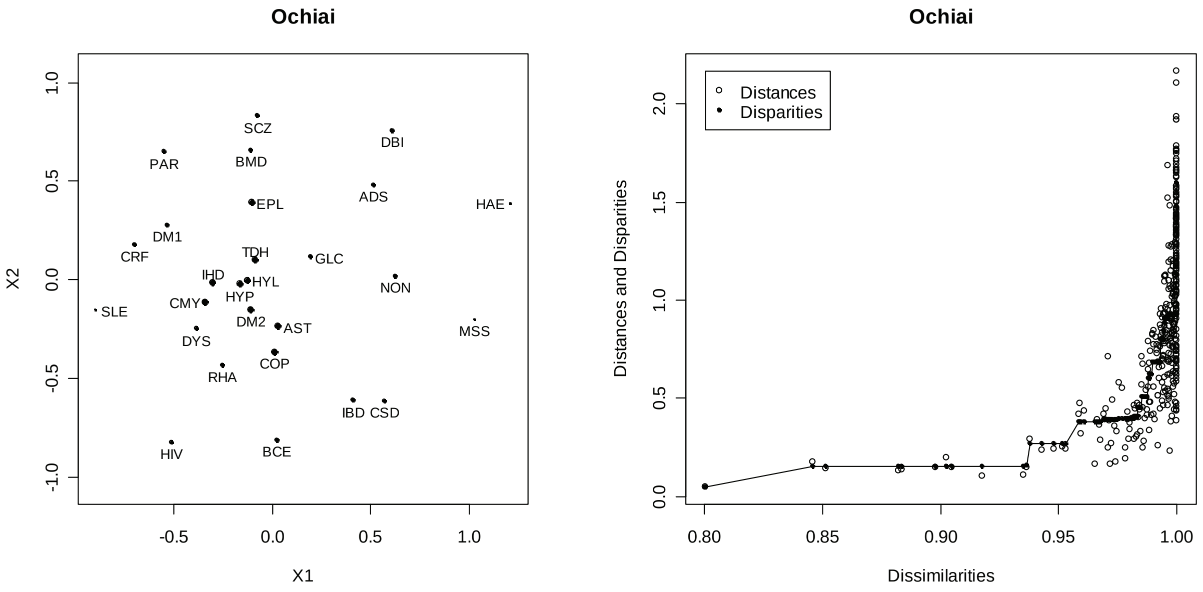


Figure 4.17: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Ochiai dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.068).*

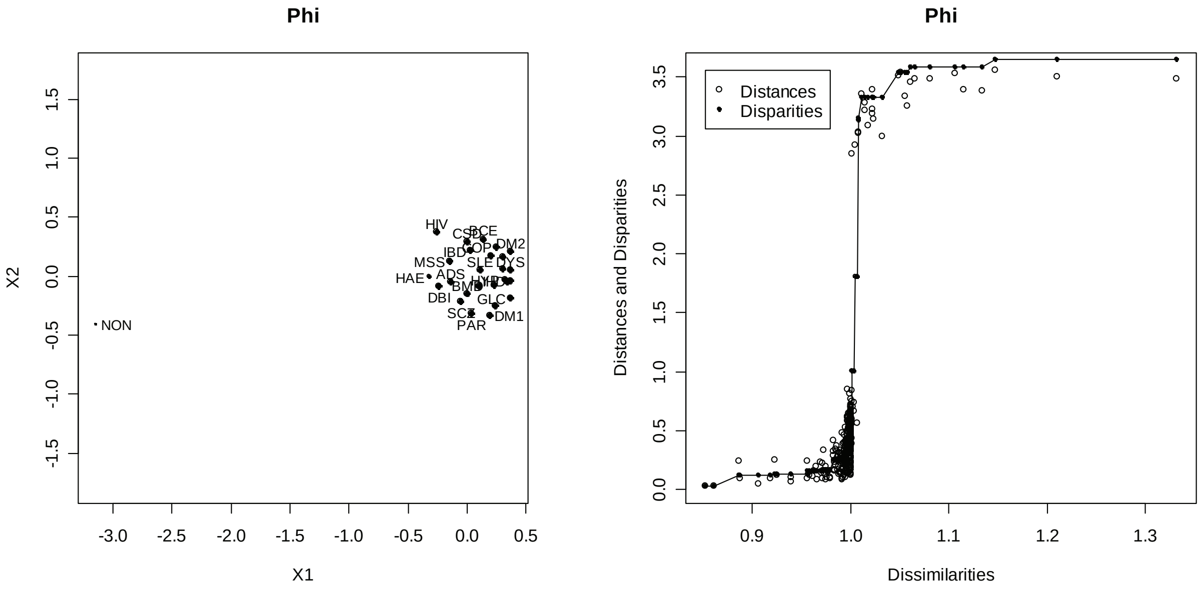


Figure 4.18: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Phi dissimilarity coefficient is used. Notice the large disparities and distances. (Normalised Stress in (4-4) = 0.036).*

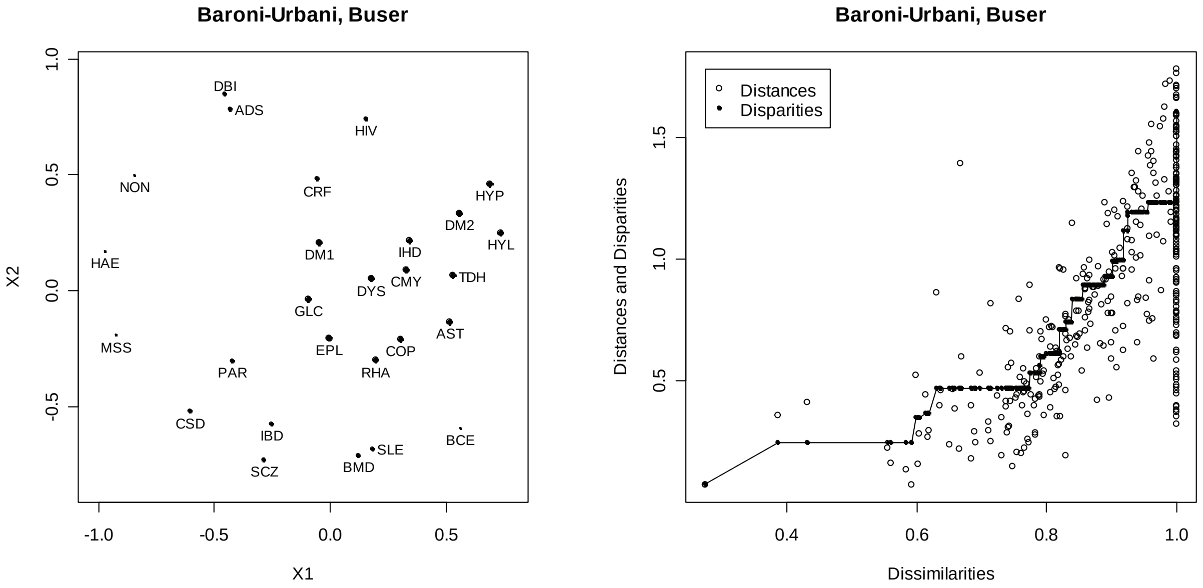


Figure 4.19: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Baroni-Urbani-Buser dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.081).*

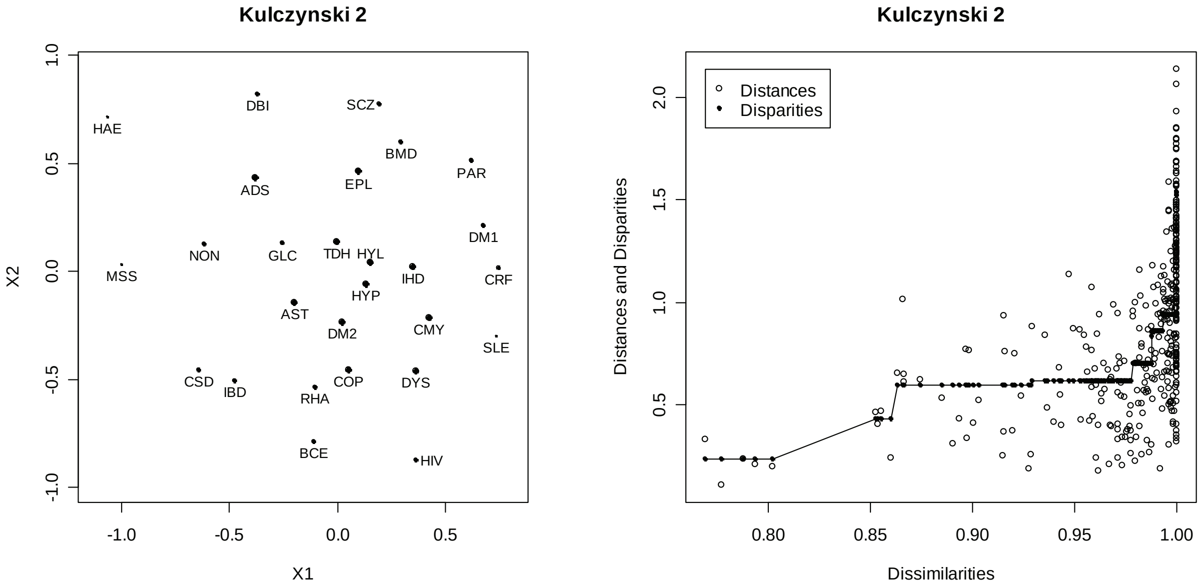


Figure 4.20: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Kulczynski 2 dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.084).*

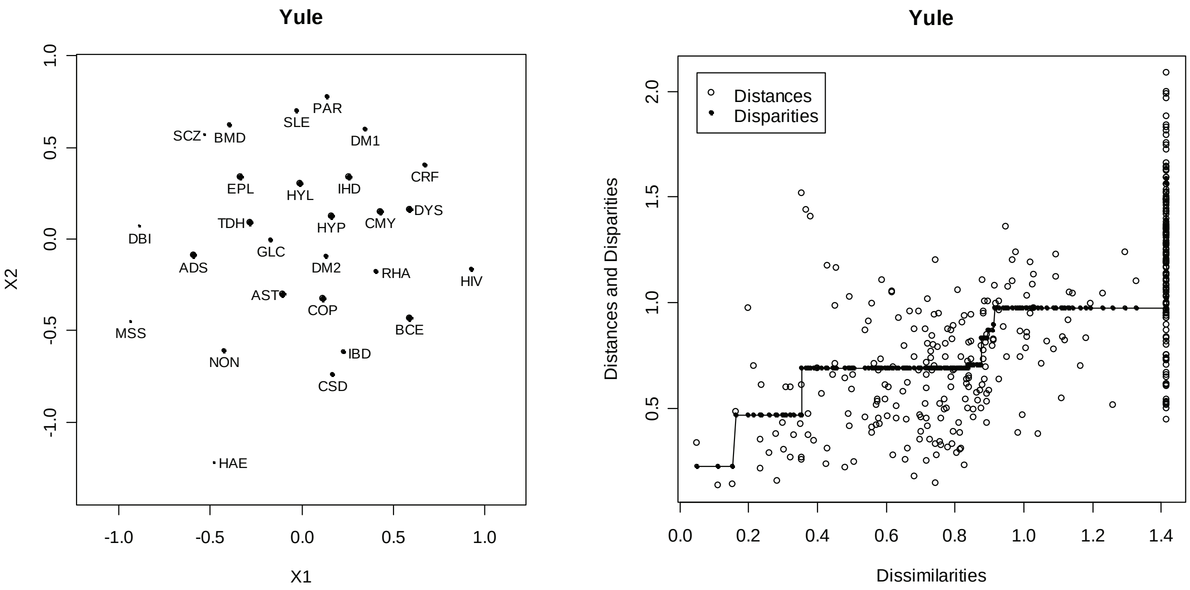


Figure 4.21: *Bubble plot and Shepard diagram produced using nonmetric MDS. The Yule dissimilarity coefficient is used. (Normalised Stress in (4-4) = 0.082).*

The bubble plots can be used to assess how well the points fit. It seems that the chronic diseases COP, AST, DYS, HYP, DM2, IHD, HYL, HYP TDH, and DYS are some of the best-represented chronic diseases. In contrast, the chronic diseases MSS, HAE and NON are usually poorly represented.

The Shepard diagrams based on the various dissimilarity coefficients provide information on how well the disparities are fitted to the distances. Monotone (isotonic) regression was used to fit the disparities, remembering that the rank-order of the disparities must be the same as the rank-order of the dissimilarities. It is for this reason that the isotonic regression line must be non-decreasing when dissimilarities are used. The isotonic regression line is displayed in the Shepard diagram and the disparities are located on this line. The vertical distance between the distances and corresponding disparities gives an indication of the error of fit. The *Phi* Shepard diagram looks different to the other Shepard diagrams. Notice the very large disparities and large distances. The distances also seem to be grouped into two separate ranges. The associated errors do not seem to be substantial and this is confirmed by the relatively low Normalised Stress (4-4) value of 0.03586. Notice in the *Phi* bubble plot that NON is located far away from the chronic diseases, and NON is poorly represented. This means that NON has higher error values. The *Jaccard, Dice-Sorensen, Sokal-Sneath-*

Anderberg, *Kulczynski 1*, *Kulczynski 2* and *Ochiai* Shepard diagrams seem to be reasonably similar. The *Baroni-Urbani-Buser* and *Yule* Shepard diagrams show a wider range of dissimilarities than the other Shepard diagrams. The other Shepard diagrams show dissimilarities of values close to 1 and these dissimilarities occur in a narrow range. Dissimilarities based on the *Jaccard*, *Dice-Sorensen* and *Sokal-Sneath-Anderberg* similarity coefficients have the narrowest range of values.

The various bubble plots can also be used to describe how the chronic diseases are related. It seems that most of the chronic diseases mentioned in the same body system rule are related, as seen in Figures 4.13 to 4.21. It was mentioned earlier that the configuration produced by classical scaling was used as the initial configuration for these figures. The final configuration shown in these bubble plots will be dependent on this chosen initial configuration. The final configuration might produce a local minimum Normalised Stress value in (4-4), but there is no guarantee that this local minimum will also be the global minimum. Further, the use of classical scaling to produce the initial configuration for dissimilarity coefficients without Euclidean or metric properties may not be appropriate. It is because of these reasons that random configurations, generated out of a uniform (0, 1) distribution, were also considered (Borg & Groenen, 2005, p.277). The *R* function *SMACOF.Nonmetric.random.starts* (see Appendix A) was developed to use random initial configurations. This function also uses the configuration generated by classical scaling as one of the possible initial configurations. The function calculates the minimum Normalised Stress value in (4-4), where each of these possible initial configurations produces a final configuration by an iterative procedure. The final configuration that produces the overall minimum Normalised Stress value in (4-4) will then be displayed.

The following *R* instruction was used to construct the *Jaccard* bubble plot in Figure 4.22:

```
SMACOF.Nonmetric.random.starts(D=All.Dissim$Jaccard,r=200,
Bubble=TRUE,main="Jaccard")
```

The bubble plots displayed in Figures 4.23 and 4.24 were constructed in a similar manner. A classical scaling configuration and 200 random configurations were used as initial configurations. The final configurations displayed in these figures are more likely to produce the global minimum Normalised Stress value in (4-4).

Notice that the Normalised Stress (4-4) values indicated in Figures 4.22, 4.23 and 4.24 are relatively low, keeping in mind that a value of 0.4352 is the highest possible value. These low Normalised Stress values indicate an overall good fit of the points. The *Phi* bubble plot produces the lowest Normalised Stress value. This does not mean that the *Phi* dissimilarity coefficient is the best, it just means that the points in the *Phi* bubble plot show better fit given the *Phi* dissimilarities. Some of the other bubble plots show a clear clustering structure, even though higher Normalised Stress values are displayed. Furthermore, notice that most of the Normalised Stress values are lower when random initial configurations are used. This means that using only the classical scaling as initial configuration did usually not produce a final configuration with the global minimum Normalised Stress value.

The bubble plots displayed in Figures 4.22, 4.23 and 4.24 can be used to describe how the chronic diseases are related.

The first body system rule involves the respiratory diseases COP, AST and BCE. Many of the bubble plots seem to show that these diseases are strongly related, because the points representing these chronic diseases are usually not far apart. The *Ochiai* and *Phi* bubble plots show that these three chronic diseases are strongly related. The other bubble plots show that COP and AST are more strongly related, because the point representing BCE is usually further apart from AST and COP. Overall, it seems that there is enough evidence based on the data to indicate that COP, AST and BCE are strongly related. The bubble plots also show that AST and COP are better represented than BCE, as indicated by the larger bubbles. This means that BCE has larger errors on average.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. These chronic diseases are fairly well represented in the bubble plots. Most of the bubble plots show that the points representing these chronic diseases are located close to each other, which suggests that these chronic diseases are strongly related.

The third body system rule involves HYP and CRF. It seems that HYP tends to co-occur more often with HYL, and CRF tends to co-occur more often with DM1. Many of the bubble plots show that the points representing HYP and CRF are not very close, but these points are not very far from each other either. There appears to be some relation between these chronic diseases, but the relation is not very strong.

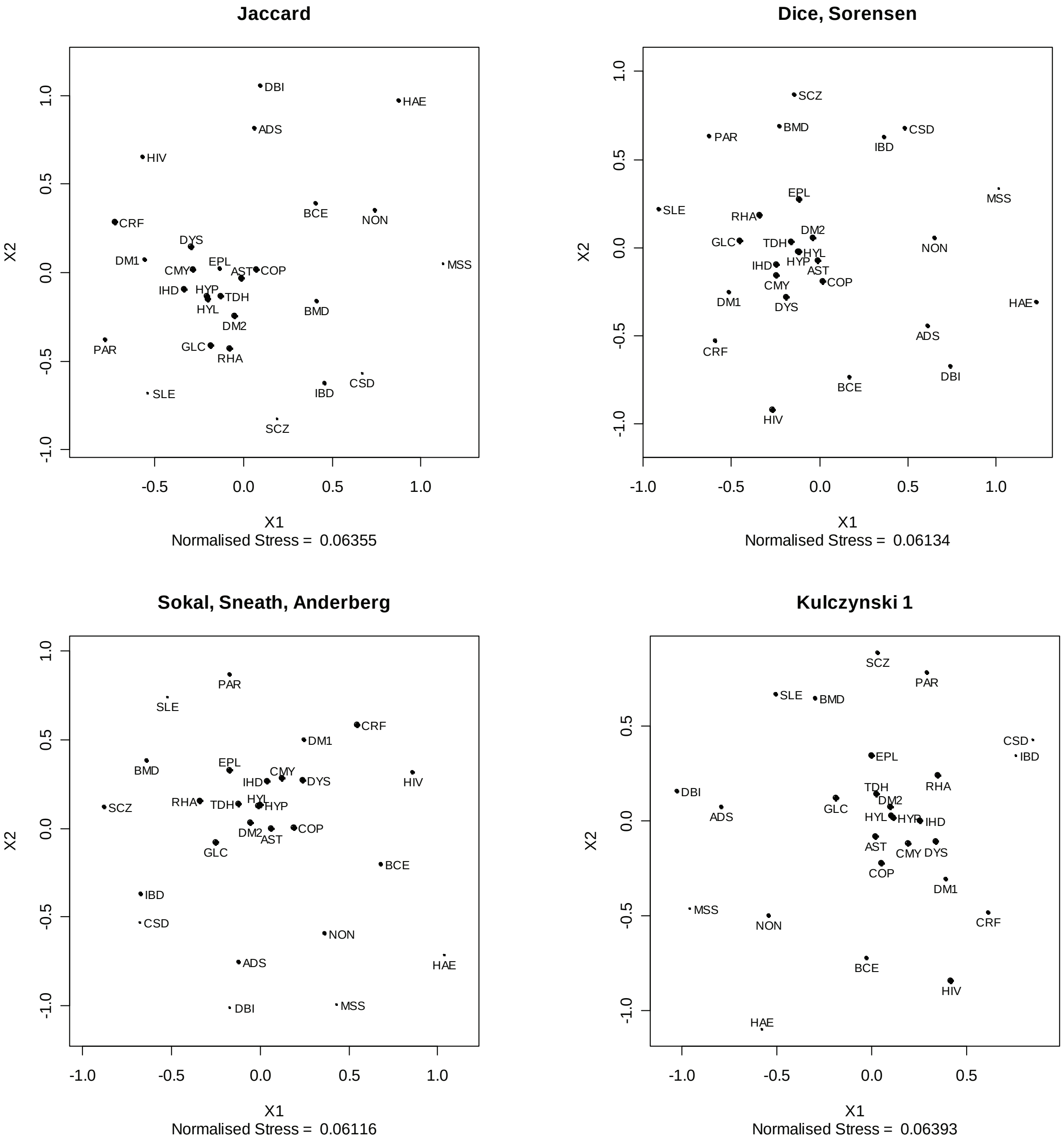


Figure 4.22: *Bubble plots produced using nonmetric MDS. Various different dissimilarity coefficients are used. A classical scaling configuration and 200 random configurations were used as initial configurations. The final configuration that produced the overall minimum Normalised Stress value in (4-4) is displayed in these bubble plots.*

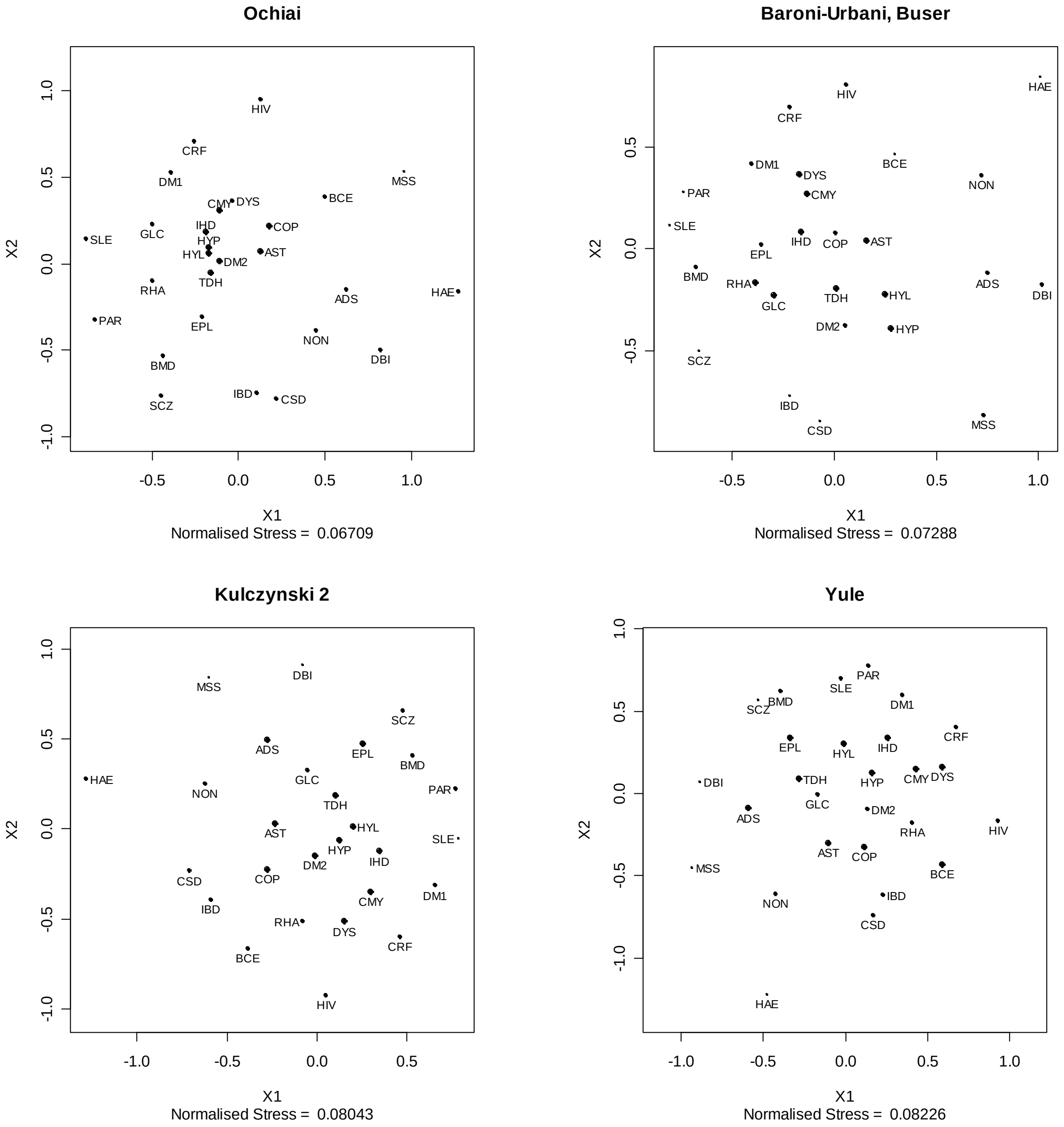


Figure 4.23: *Bubble plots produced using nonmetric MDS. Various different dissimilarity coefficients are used. A classical scaling configuration and 200 random configurations were used as initial configurations.*

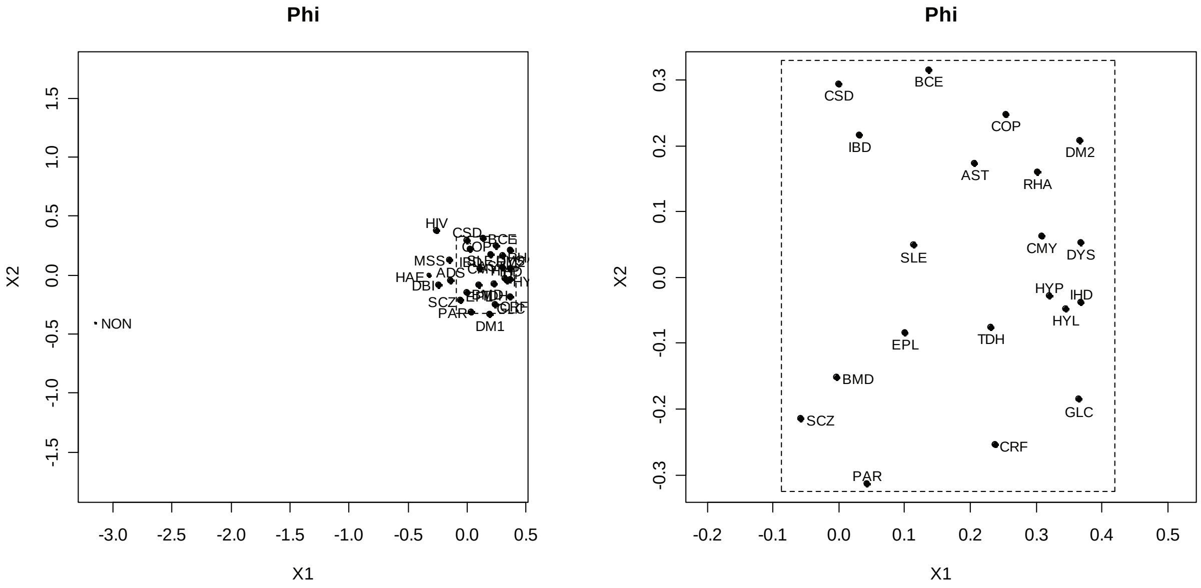


Figure 4.24: *Bubble plot produced using nonmetric MDS based on the Phi dissimilarity coefficient. A classical scaling configuration and 200 random configurations were used as initial configurations. The final configuration that produced the overall minimum Normalised Stress value is displayed in this bubble plot. (Normalised Stress in (4-4) = 0.0358). The panel on the right represents an enlargement of an over-crowded area in the left panel.*

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. The points that represent these two diseases are fairly close to each other. It appears that these two chronic diseases co-occur often.

The fifth body system rule involves the chronic diseases BMD and SCZ. These chronic diseases also seem to be strongly related, as these chronic diseases are fairly close to each other.

The sixth body system rule involves MSS, BMD and EPL. The points that represent the chronic diseases BMD and EPL tend to be fairly close to each other, but the point representing MSS ia always further away. It is therefore not clear if MSS is strongly related to BMD and EPL. The chronic disease MSS is poorly represented in most bubble plots. This means that distances involving MSS should be interpreted with more care.

The seventh body system rule involves the chronic diseases SLE and RHA. The chronic disease RHA are better represented in the bubble plots, because RHA has larger bubbles than SLE. This means RHA has smaller errors on average. Most of the bubble plots show that the points representing these two chronic diseases are not close together, but the points are not far away either. These bubble plots suggest that these two chronic diseases might have some weak relation.

The last system body rule involves DM1 and DM2. It seems that DM1 tends to co-occur with CRF, and that DM2 tends to co-occur with the cardiovascular diseases CMY, IHD, HYP and DYS.

It appears that the chronic diseases ADS and DBI are strongly related, as the points representing these two diseases are situated close to each other in most bubble plots. These chronic diseases however are not mentioned in any of the body system rules. It also seem that the respiratory chronic diseases AST and COP are related to the cardiovascular chronic diseases and chronic diseases HYL, TDH and DM2, as these points tend to be situated relatively close to each other in many bubble plots.

The nonmetric MDS method produces relatively clear clustering structures when compared to the metric MDS methods. The nonmetric MDS method might therefore be preferred over the classical scaling method for the Medical Scheme 55-59 data set, because the classical scaling plots explain a very small proportion of the point variability. This causes an over-crowding of points in the classical scaling plot, which makes interpretation very difficult. The nonmetric MDS method is also less subjective than the metric least squares scaling method as no different weights have to be used to produce reasonably clear clustering structures. However, the nonmetric MDS method takes more computation time than metric least squares scaling as optimal disparities and distances have to be computed during each iteration in the SMACOF algorithm. The classical scaling method is least computer intensive. However, the increased computation time of the nonmetric MDS method is not a problem with the Medical Scheme 55-59 data set as only 27 objects are presented in the plots. The computation times (measured in seconds) for the classical scaling, metric least squares scaling and nonmetric MDS methods were 0.4, 7 and 10 respectively when applied to the Medical Scheme 55-59 data set. The nonmetric MDS method will be more computer intensive when random initial configurations are used. The computation time for the nonmetric MDS method, using 200 random initial configurations, was approximately 29 minutes when applied to the Medical Scheme 55-59 data set.

Most of the dissimilarity coefficients used for the nonmetric MDS method appear to be appropriate. Many of these dissimilarity coefficients produce very similar configurations. However, the *Phi* dissimilarity coefficient does have a very different configuration.

4.4 Application of the Nonmetric MDS Method to the Male MS (55-59) and Female MS (55-59) Data Sets

The Medical Scheme 55-59 data set contains male and female lives that might display different configurations of chronic diseases. It was mentioned in Section 4.3.2 that the nonmetric MDS method seems to produce clearer clustering structures than the metric MDS methods. It is for this reason that only nonmetric MDS will be considered in this section. The *R* function *SMACOF.Nonmetric.random.*starts (see Appendix A) will again be used, where separate male and female dissimilarity matrices will be used as input. The *R* function *Dissim.CDL* will be used to generate these dissimilarity matrices based on separate Male MS (55-59) and Female MS (55-59) data sets. The following *R* instructions were used to construct all the male and female dissimilarity matrices, based on the various similarity coefficients listed in Table 3.2:

```
F.All.Dissim<-
Dissim.CDL(x=final.dat[c(1:886),c(1:28)],transformation=2)
```

```
M.All.Dissim<-
Dissim.CDL(x=final.dat[c(887:1636),c(1:28)],transformation=2)
```

The first instruction was used to construct all dissimilarity matrices based only on the Female MS (55-59) data set. The second instruction was used to construct all dissimilarity matrices based only on the Male MS (55-59) data set. The first 886 lines of the Medical Scheme 55-59 data set, named *final.dat* in *R*, consisted of female lives and the remaining lines consisted of male lives. The transformation $d_{ij} = (1-s_{ij})^{0.5}$ was used to construct the dissimilarities.

The initial configurations used in this section will consist of 200 random configurations and a configuration produced by the classical scaling method. The bubble plots displayed in this section will represent the final configuration that produced the overall minimum Normalised Stress value in (4-4). It is important to investigate if the separate male and female configurations support the body system rules. Most of the dissimilarity coefficients listed in Table 3.2 will be used.

The following *R* instructions were used to construct the *Female: Jaccard* bubble plot and the *Male: Jaccard* bubble plot displayed in Figure 4.25:

```
SMACOF.Nonmetric.random.starts(D=F.All.Dissim$Jaccard,r=200,
Bubble=TRUE,main="Female: Jaccard")

SMACOF.Nonmetric.random.starts(D=M.All.Dissim$Jaccard,r=200,
Bubble=TRUE,main="Male: Jaccard")
```

The other male and female bubble plots in Figures 4.25, 4.26 and 4.27 were constructed in a similar manner.

It was mentioned in Chapter 2 that none of the male lives were treated for the chronic diseases HAE. This means that HAE cannot be displayed in the male bubble plots. This can be seen in the bubble plots displayed in Figures 4.25, 4.26 and 4.27.

The first body system rule involves the respiratory diseases COP, AST and BCE. Many of the bubble plots seem to show that these diseases are strongly related, because the points representing these chronic diseases are usually not too far apart. This seems to be the case for both the Male MS (55-59) and Female MS (55-59) data sets. The *Male: Ochiai*, *Female: Kulczynski 2* and *Female: Yule* bubble plots especially show that the points representing these chronic diseases are located close to each other. Some bubble plots, like the *Female: Ochiai*, *Male: Kulczynski 2* and *Male: Yule* bubble plots, show that these chronic diseases are not located close to each other, which suggests that these chronic diseases are not very strongly related. However, most of the bubble plots show that COP and AST are more strongly related, because the point representing BCE is usually further apart from AST and COP. The bubble plots also show that AST and COP are better represented than BCE, as indicated by the larger bubbles.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. These chronic diseases are very well represented in the bubble plots of the Male MS (55-59) and Female MS (55-59) data sets. Most of the bubble plots, for both Male MS (55-59) and Female MS (55-59) data sets, show that the points representing these chronic diseases are located very close to each other. This suggests that these cardiovascular related diseases seem to be very strongly related.

These cardiovascular diseases also seem to be strongly related to the chronic diseases AST, COP, HYL, DM2 and TDH.

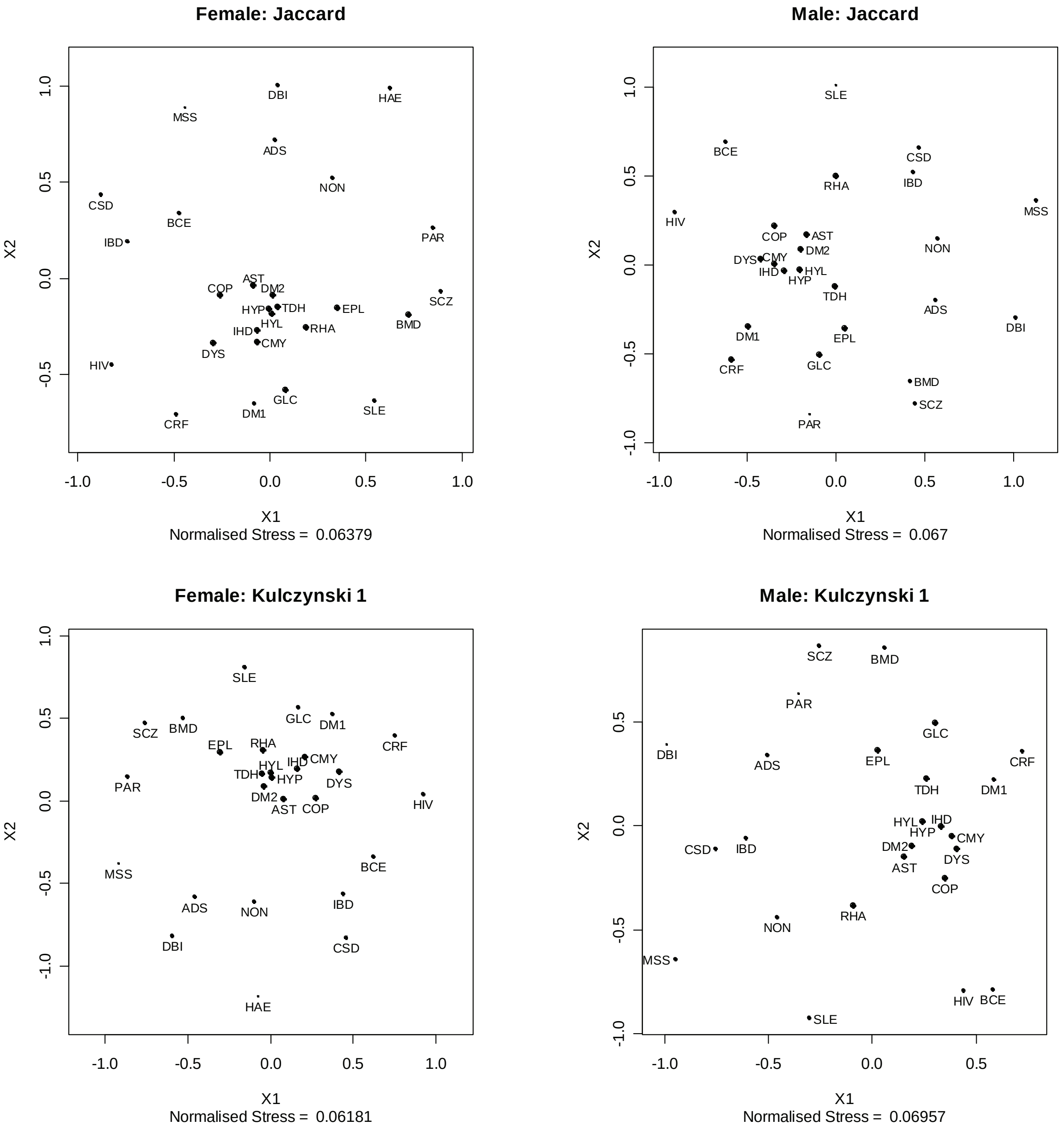


Figure 4.25: *Bubble plots of the Male MS (55-59) and Female MS (55-59) data sets, produced using nonmetric MDS and the Jaccard and Kulczynski 1 dissimilarity coefficients. A classical scaling configuration and 200 random configurations were used as initial configurations. The final configuration that produced the overall minimum Normalised Stress value in (4-4) is displayed in these bubble plots.*

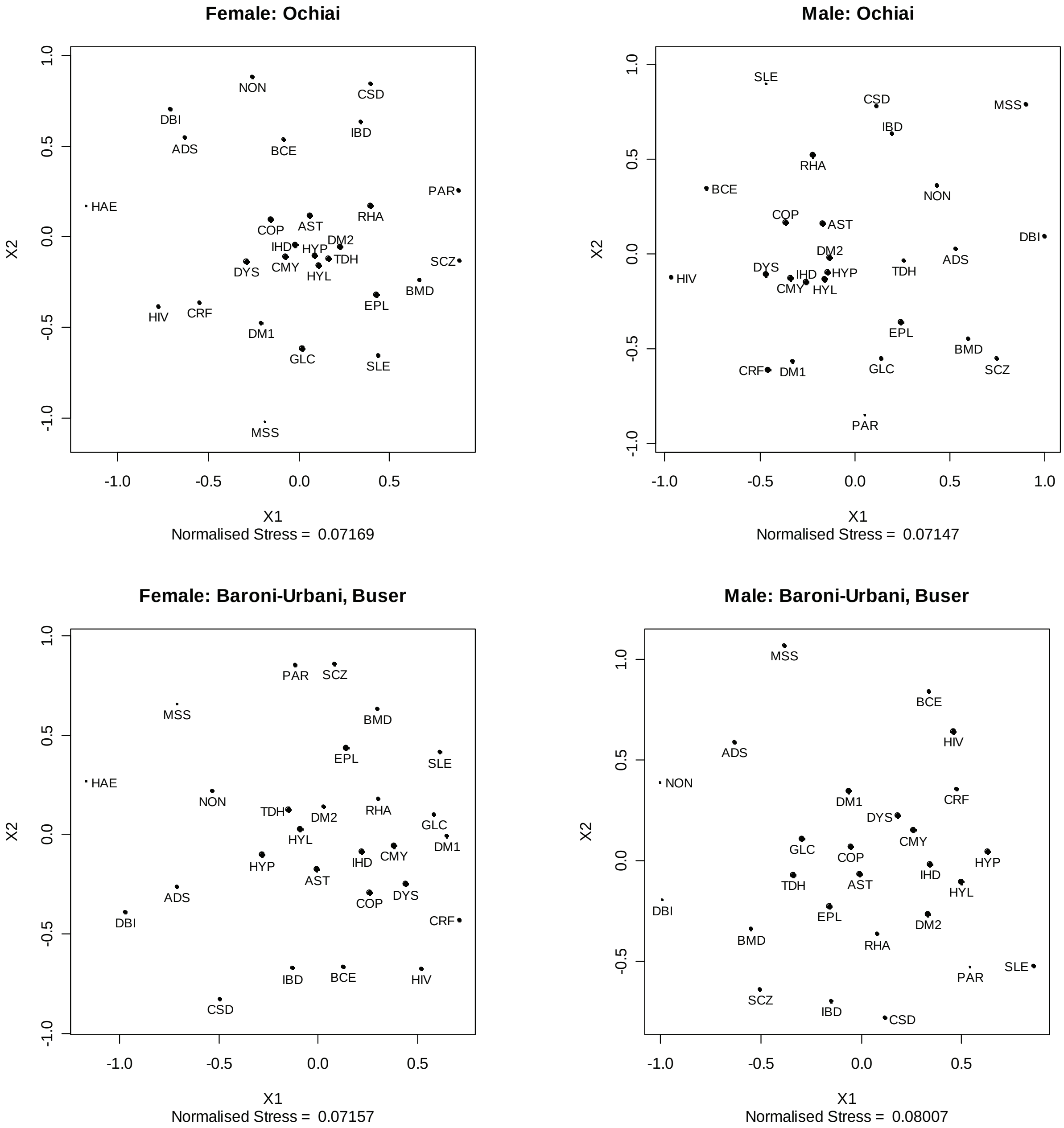


Figure 4.26: *Bubble plots of the Male MS (55-59) and Female MS (55-59) data sets, produced using nonmetric MDS and the Ochiai and Baroni-Urbani-Buser dissimilarity coefficients. A classical scaling configuration and 200 random configurations were used as initial configurations.*

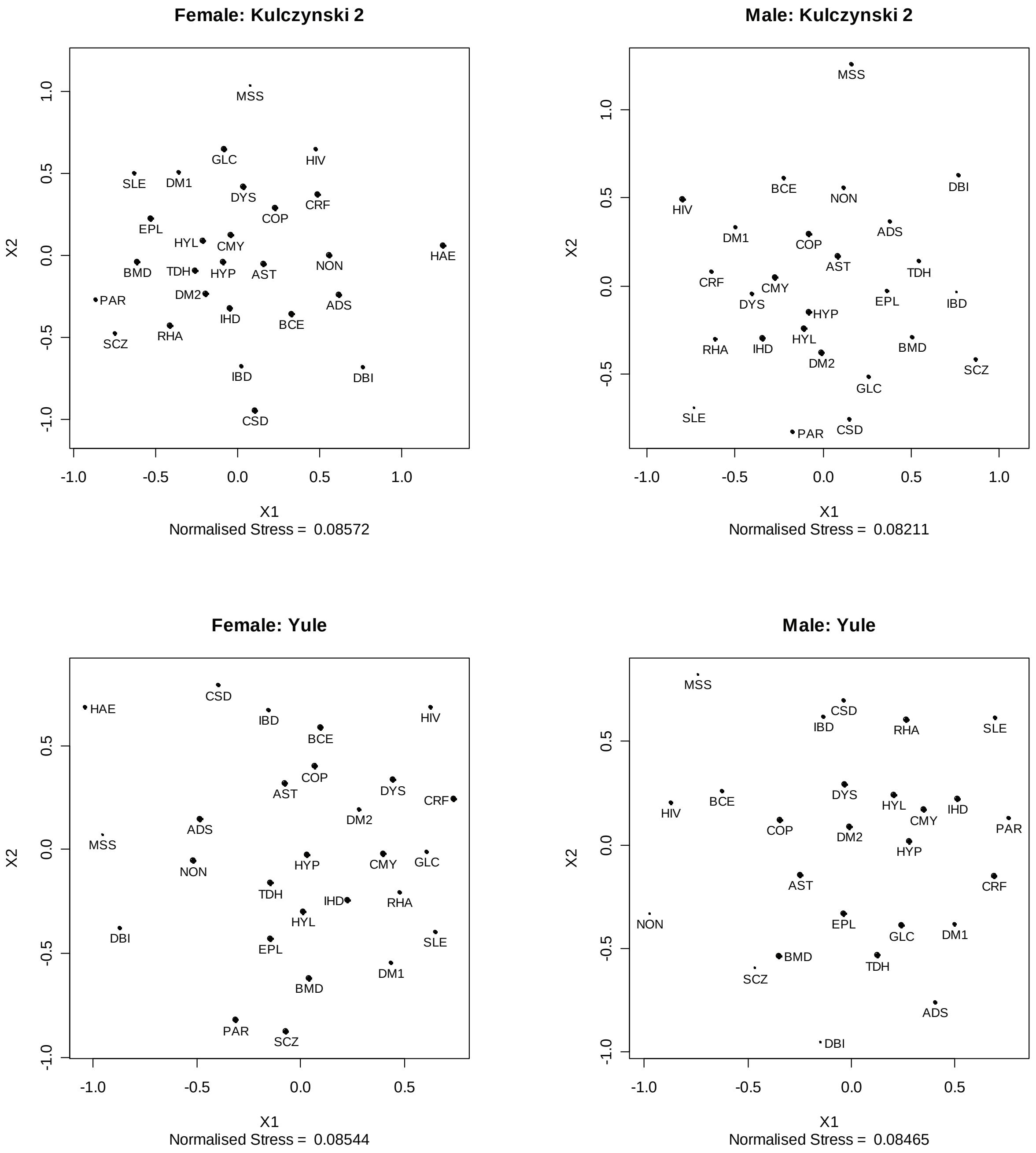


Figure 4.27: *Bubble plots of the Male MS (55-59) and Female MS (55-59) data sets, produced using nonmetric MDS and the Kulczynski 2 and Yule dissimilarity coefficients. A classical scaling configuration and 200 random configurations were used as initial configurations.*

The third body system rule involves HYP and CRF. Many of the bubble plots show that the points representing HYP and CRF are not located very close to each other, but these points are not very far from each other either. There seems to be some relation between these chronic diseases, but the relation is not very strong. This seems to be the case for both male and female lives.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. The points that represent these two diseases always seem to be close to each other. Only the *Female: Baroni-Urbani-Buser, Male: Baroni-Urbani-Buser* and *Male: Kulczynski 2* bubble plots show that the points are not close together. Still, it seems that these two chronic diseases co-occur often and are therefore related.

The fifth body system rule involves the chronic diseases BMD and SCZ. These chronic diseases appear to be strongly related, because these chronic diseases are usually located close to each other. This is the case for male and female bubble plots. It also seems that BMD, SCZ, PAR and EPL are strongly related, because the points representing these chronic diseases are often close to each other.

The sixth body system rule involves MSS, BMD and EPL. The points that represent the chronic diseases BMD and EPL tend to be fairly close to each other, but the point representing MSS is always further away. It is therefore not clear if MSS is strongly related to BMD and EPL. The chronic disease MSS is poorly represented in most of the male and female bubble plots. This means that distances involving MSS should be interpreted with more care.

The seventh body system rule involves the chronic diseases SLE and RHA. The chronic disease RHA appears to be better represented in the bubble plots, because RHA has larger bubbles than SLE. Most of the bubble plots show that the points representing these two chronic diseases are not close together, nor are they far away. These bubble plots suggest that these two chronic diseases might have some weak relation. This seems to be the case for both of the Male MS (55-59) and Female MS (55-59) data sets.

The chronic diseases ADS and DBI also seem to be strongly related, as the points representing these two diseases are situated close to each other in most bubble plots. These chronic diseases however are not mentioned in any of the body system rules. The male bubble plots show that these chronic diseases are usually further apart when compared to the female bubble plots. The relation between these chronic diseases therefore seems to be stronger for female lives.

4.5 Summary

The purpose of this chapter was to display the various chronic diseases in such a way that the relationships between the chronic diseases can be described. Several multidimensional scaling techniques were used to construct two-dimensional displays, where each point in the display represents a single chronic disease. These displays were used to investigate whether the chronic diseases mentioned in the same body system rules are related. This was done for the Medical Scheme 55-59 data set and for the separate Male MS (55-59) and Female MS (55-59) data sets.

The classical scaling, metric least squares scaling and nonmetric MDS methods were used to display the chronic diseases, but it seems that the nonmetric MDS method was most useful. It seems that the nonmetric MDS method should be preferred over the classical scaling method for the Medical Scheme 55-59 data set, because the classical scaling plots explained a very small proportion of the point variability. This caused an over-crowding of points in the classical scaling plot which made interpretation more difficult. The nonmetric MDS method was also less subjective than the metric least squares scaling method as no different weights had to be used to produce reasonably clear clustering structures. However, the nonmetric MDS method takes more computation time than metric least squares scaling as optimal disparities and distances had to be computed during each iteration of the SMACOF algorithm. The classical scaling method is least computer intensive. Fortunately, the computation time was not a major concern with the Medical Scheme 55-59 data set as only 27 objects are displayed in the plots. Nonmetric MDS also has the advantage of making minimal assumptions about how distances and dissimilarities are related. The metric MDS methods should only consider dissimilarity coefficients with metric or Euclidean properties, because it is not appropriate to use dissimilarities directly as distances when the dissimilarities do not have metric properties. Nonmetric MDS does not treat dissimilarities as distances, and can therefore also use dissimilarities without metric properties.

The dissimilarities formed by the transformation $d_{ij} = (1-s_{ij})^{0.5}$ which were described and calculated in Chapter 3, were used as input for these MDS techniques. This transformation was used because it leads to more dissimilarity coefficients with metric and Euclidean properties. Several dissimilarity coefficients were used in this chapter, hoping for some robustness against a specific choice. But not all dissimilarity coefficients listed in Table 3.2 were appropriate to use. It was shown that the *Rao-Russell* and the *Simple matching* dissimilarity coefficients are unsuitable to use with the Medical Scheme 55-59 data set. The plots of these dissimilarity coefficients showed that chronic diseases with similar chronic disease rates were located close to each other. This is undesirable, as chronic diseases should be located close to each other when they tend to co-occur often and not because the chronic diseases have similar chronic disease rates. The Medical Scheme 55-59 data set consists of asymmetric binary variables. The *Rao-Russell* and the *Simple matching* dissimilarity coefficients should only be used when the data consists of symmetric binary variables. The other dissimilarity coefficients were more suitable to use. It was also found that the *Jaccard*, *Dice-Sorensen*, *Kulczynski 1* and *Sokal-Sneath-Anderberg* dissimilarity coefficients produced very similar graphical displays.

The plots produced by the various MDS methods were compared to the various body system rules. There were some differences between the displays produced by these MDS methods. This is not of serious concern as it was expected that different displays would be produced, but it makes overall interpretation more difficult. The chronic diseases mentioned in the second (CMY, IHD, DYS and HYP), fourth (CSD and IBD) and the fifth (BMD and SCZ) body system rules seem to be strongly related in most MDS plots. The chronic diseases mentioned in the first (COP, AST and BCE) and sixth (MSS, BMD and EPL) body system rules seem to be reasonably related. Points representing COP and AST tend to be located close to each other, but the point representing BCE was usually located further away. It seems therefore that COP and AST have a strong relation with each other, but not with BCE. A similar observation was made with regard to the sixth body system rule. It seemed that BMD and EPL have a reasonable strong relation with each other, but MSS did not show a strong relation with BMD and EPL. The chronic diseases mentioned in the seventh (SLE and RHA) and the third (HYP and CRF) body system rules do seem to be weakly related. Some evidence suggested that there might be some relation between SLE and RHA and some relation between HYP and CRF, but these relations do not seem to be strong. The last body system rule, involving the Diabetes Mellitus diseases DM1 and DM2, was the only rule that was actually applied to the Medical Scheme 55-59 data set used in this study. The Medical Scheme 55-59 data set shows no co-occurrence between these two diseases. It is therefore expected that these two diseases should not be located very close to each other in the MDS plots. This was indeed the case. It was found that the points representing DM1 and CRF tend to be located close to each other. This means that there seems to be a strong similarity between DM1 and CRF.

It was interesting to note that some other chronic diseases, not mentioned in the same body system rule, also seem to be strongly related. The cardiovascular chronic diseases {HYP, DYS, IHD, CMY} did show a strong similarity with chronic diseases AST, COP, DM2, TDH and HYL. It also seems that the chronic diseases BMD, EPL, SCZ and PAR are strongly related. The chronic diseases

ADS and DBI also showed a strong similarity, even though they are not mentioned in any of the body system rules. It therefore seems that ADS and DBI tend to co-occur often.

The MDS plots based on separate Male MS (55-59) and Female MS (55-59) data were also compared to the various body system rules. The relations between chronic diseases mentioned in the same body system rule were also very similar to the results described previously for the Medical Scheme 55-59 data set.

It is important to note that these MDS methods were applied only to lives between the ages 55 to 59. The other age bands might show similar results, but this is not known for certain. However, the same MDS methods could also be used for the other age bands, and the results could be interpreted in a similar manner. It seems though that the nonmetric MDS method should be the preferred choice.

It was shown how the MDS methods could be applied to the Medical Scheme 55-59 data set and how the results could be interpreted. However, the interpretation of these MDS configurations is rather subjective, as different observers might reach different conclusions. It is also difficult to capture all the variability of the data in a low dimensional display. It will therefore be useful to supplement these MDS displays with information about the clustering structure of the data. This is why a cluster analysis of the chronic diseases will be considered in Chapter 5.

Chapter 5

Clustering of the Chronic Diseases using the Algorithms of Kaufman & Rousseeuw

5.1 Introduction

Cluster analysis encompasses many different techniques for discovering structure within a data set. In general, the clustering methods are of two types: Partitioning methods or hierarchical methods. Partitioning methods are based on specifying an initial number of clusters, and then reallocating observations between clusters in an iterative manner until some form of equilibrium is reached. Hierarchical methods do not really compete with partitioning methods, because they do not pursue the same goal. The goal of partitioning methods is to try to select the *best* clustering structure with *k* groups. A hierarchical algorithm describes a method yielding an entire hierarchy of clusterings for the given data set, without specifying an initial number of clusters (Kaufman & Rousseeuw, 1990; Gordon, 1999).

One of the best known traditional clustering methods is the *K-means* partitioning method (Struyf *et al*., 1997). This method determines group membership by calculating the centroid (mean) for each of *k* groups and assigning each observation to the group with the closest centroid. There are also model-based clustering techniques that are based on the assumption that the underlying data are generated from a mixture of probability distributions (Banfield & Raftery, 1993; Dasgupta & Raftery, 1998; Fraley & Raftery, 2002; Raftery & Dean, 2006). Unfortunately, the model-based methods and most of the traditional clustering methods, like *K-means,* cannot be used for the Medical Scheme 55-59 data set. The reason why these methods cannot be used on the Medical Scheme 55-59 data set will be discussed in Section 5.4.1.

The robust clustering algorithms of Kaufman & Rousseeuw (1990) were designed to accept a dissimilarity matrix **D**, and not just a data matrix **X**: $n \times p$, as an input structure. It is because of this that the clustering techniques of Kaufman & Rousseeuw (1990) will be the focus in this chapter. Kaufman & Rousseeuw (1990) describe several different hierarchical and partitioning clustering methods, which they believe cover most applications. These methods are not based on an underlying statistical model but their construction was based on theoretical considerations, such as logical consistency and robustness, and the past experience of the two authors.

The clustering algorithms of Kaufman & Rousseeuw (1990) and their input structures will be described in Section 5.2 and Section 5.3. It will be shown that these clustering algorithms are well suited for analysing the Medical Scheme 55-59 data set. Of particular importance are the graphical representations accompanying the Kaufman and Rousseeuw (1990) algorithms, showing the clustering structures of the chronic diseases. The clustering structures will be used to investigate whether the chronic diseases mentioned in the same body system rules are strongly related. This will be done in Section 5.4. Furthermore, the results from the cluster analysis can then be compared to those obtained from applying the MDS methods, to evaluate their support of the findings discussed in Chapter 4. The clustering structures of male and female lives will also be compared to the body system rules and will be discussed in Section 5.5.

The dissimilarities calculated in Chapter 3, with the *R* function *Dissim.CDL*, will be used as input structure to the various clustering methods in Section 5.4 and Section 5.5.

5.2 Input Structures for Cluster Analysis

Clustering algorithms will usually operate on either of two structures. The first structure is in the form of a data matrix **X**: $n \times p$ with

$$\mathbf{X}=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

The rows of this matrix represent the *n* objects and the columns represent the *p* variables.

The second structure is a **D**: $n \times n$ dissimilarity matrix:

$$\mathbf{D}=\begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d_{n1} & d_{n2} & \cdots & \cdots & 0 \end{bmatrix}$$

where $d_{ij} = d_{ji}$ measures the dissimilarity between objects *i* and *j*.

The dissimilarity structures were discussed in Chapter 3. The dissimilarity matrix contains all the pair-wise dissimilarities between the objects. Dissimilarities are used to measure how different objects are from one another. A value d_{ij} close to zero indicate that the two objects i and j are very similar. Conversely, large values of d_{ij} indicate that the two objects are very different.

5.3 The Clustering Algorithms of Kaufman & Rousseeuw

5.3.1 Partitioning methods

Partitioning methods are based on specifying an initial number of clusters, and then reallocating observations between clusters in an iterative manner until some form of equilibrium is reached. The partitioning around medoids (PAM), clustering large applications (CLARA) and fuzzy clustering (FANNY) methods will be discussed in this section. CLARA employs PAM in its algorithm, but was specifically adjusted to be more effective for larger data sets than PAM. CLARA only takes the $\mathbf{X} : n \times p$ data matrix as input data structure, but PAM and FANNY can take both the $\mathbf{X} : n \times p$ data matrix and the dissimilarity matrix **D** as input structure. The number of clusters, k, must be specified by the user for all of these methods (Kaufman & Rousseeuw, 1990).

5.3.1.1 Partitioning around medoids (PAM)

“The main objective of partitioning objects into k clusters is to find clusters in such a way that objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible.” (Kaufman & Rousseeuw, 1990). PAM is a partitioning algorithm that tries to achieve this objective. An $\mathbf{X}$: $n \times p$ data matrix or a dissimilarity matrix **D** can be used as input structure. The algorithm PAM forms k clusters and assigns each one of the n objects to one of these k clusters. The algorithm also computes k representative objects of the clusters, which are called *medoids*. The medoids are not the average values of each cluster, but an actual object of the data set. The number of clusters, k, must be specified by the user.

The PAM algorithm proceeds in two steps. The build step is the first of these steps. This step selects k centrally located objects to be used as initial medoids. Each object is then assigned to a cluster that corresponds to the nearest medoid. This means that object i is classified into cluster v_i when medoid m_{v_i} is nearer than any other medoid m_w:

$$d_{i,m_{v_i}} \leq d_{i,m_w} \text{ for all } w = 1, \ldots, k.$$

A value for the objective function is then determined during the build step. The objective function is the sum of the dissimilarities of all the objects to their nearest medoid:

$$\text{objective function}=\sum_{i=1}^{n} d_{i,m_{v_i}} \qquad (5\text{-}1)$$

The swap step is the last step in the algorithm. If the objective function in (5-1) can be reduced by swapping a selected object (a medoid) with an unselected object, then the swap is carried out. This is continued in an iterative manner until the objective function cannot be further reduced. This means that PAM selects *k* medoids that minimise the sum of dissimilarities, as defined in (5-1).

Graphical display: The silhouette plot

The silhouette plot shows how well each object is classified into a certain cluster. This means the quality of all the different clusters can be compared. Silhouette values can be calculated as follows (see also Kaufman & Rousseeuw, 1990, p.83-88):

For each object *i*. Let the following be defined:

A: Cluster to which object *i* belongs.

$a(i)$: Average dissimilarity of object *i* to all other objects contained in cluster A.

C_j: Any other cluster different from A with $j = 1, 2, 3,..., k - 1$.

$d(i, C_j)$: Average dissimilarity of object *i* to all other objects contained in cluster C_j.

$b(i) = \min_{j=1,2,3,...,k-1} d(i, C_j)$

B: Cluster where minimum is obtained, hence $b(i) = d(i, B)$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The silhouette value *s(i)* always lies between −1 and 1. A value of *s(i)* close to −1 can be interpreted as that object *i* is badly classified. A value of *s(i)* close to 1 means that the object *i* is well classified and a value close to 0 means that object *i* lies between two clusters. The silhouette of a cluster is a plot of the *s(i)* values of all the objects *i* in a certain cluster. The silhouette plot graphically displays the silhouettes of all the clusters next to each other. The quality of the classification of the clusters can then be compared. The overall average silhouette width of the silhouette plot is the average of the silhouette values of all the objects in the data set. The value of *k* can be appropriately chosen by considering the overall average silhouette width. Kaufman & Rousseeuw (1990) suggest that the user should run PAM with different values of *k* and the overall average silhouette width value should be noted for each run. The most appropriate *k* value can then be chosen as the one that provides the highest overall average silhouette width. The overall average silhouette width gives an indication of the clustering structure found in the data. Kaufman & Rousseeuw (1990, p.88) make a subjective interpretation of the highest average silhouette width values based on past experience. A value of above 0.7 indicates very clear clustering structure; a value between 0.5 and 0.7 shows reasonable clustering structure; a value between 0.25 and 0.5 shows weak clustering structure and a value below 0.25 indicates that no substantial structure has been found.

Graphical display: The clusplot

The clusplot was developed by Pison *et al.*, (1998) and provides information of the clustering structure in a two-dimensional display. The observations are represented by points in the plot and an ellipse is drawn around each cluster of observations, which provides an indication of cluster membership. The configuration of points in the two-dimensional space is determined using principal components analysis (PCA) or classical scaling. The clusplot will therefore produce a very similar display to the classical scaling technique discussed in Chapter 4 when a dissimilarity matrix **D** is used as input structure. The only difference, in this case, is that the clusplot provides additional information of the clustering structure by drawing ellipses to indicate certain clusters. This is a good example of how MDS displays can be supplemented with information about the clustering structure of the data.

PCA will be used when the data matrix $\mathbf{X}$: $n \times p$ is used as input structure to the clustering algorithm. The method of PCA can be seen as a dimension reduction technique (Gower & Hand, 1996, p.9). The objective is to find a lower dimensional display space (usually a two dimensional space) that can be visualised and where the choice of lower dimensional space is optimally chosen. Let L represent the r-dimensional ($r \leq p$) space which provides the best fit in a least squares sense. The best fitting r-dimensional subspace L is spanned by the columns of the matrix $\mathbf{V}_{r:}$ $n \times r$, consisting of the first r eigenvectors of $\mathbf{X}^T\mathbf{X}$ that correspond to the r largest eigenvalues. This matrix V_r can be obtained from the singular value decomposition (SVD) of the centered data matrix $\mathbf{X}$: $n \times p$ (Gower & Hand, 1996, p.239-241). These r eigenvectors define a natural set of orthogonal coordinate axes for the r-dimensional subspace L. The clusplot algorithm will choose $r = 2$ and hence provide a best fitting two-dimensional display.

Classical scaling will be used when the dissimilarity matrix **D** is used as input structure. Classical scaling was described in Chapter 4. It was mentioned in Chapter 4 that it is inappropriate to use dissimilarity coefficients without metric properties when the classical scaling method is used, because classical scaling treats dissimilarities directly as distances. The same applies to these clusplots.

The ellipses that indicate certain clusters are based on the average and the covariance matrix of each cluster. These ellipses are known as spanning ellipses, i.e. the smallest ellipse that covers all objects within that cluster. Clusplots might be extended by replacing the ellipse with the convex hull of all points in the cluster (Pison *et al.,* 1998) or possibly using alpha bags (Gardner, 2001, p.325).

One of the drawbacks of clusplots is that clusplots do not usually use the correct aspect ratio. This means that a unit change in the horizontal direction of the graphical display is not equal to a unit change in the vertical direction and distances between objects in the graphical display can therefore not be fully appreciated, which is vital for interpretation. Care must therefore be taken by the user when implementing clusplots to ensure that the correct aspect ratio is used. Another possible limitation of clusplots is that clusplots do not provide information about the variables when the objects are clustered. Biplots (Gower & Hand, 1996) can be used to overcome this limitation since they provide information simultaneously on both the objects and the variables in the same graphical display. Furthermore, biplots do use the correct aspect ratio, which is vital for correct interpretation. Various types of biplots are described by Gower & Hand (1996) and the different biplots can be used in different practical situations. The PCA biplot (Gower & Hand, 1996) is based on the PCA methodology (described earlier in this section) and is used on quantitative data. The PCA biplot should not be applied to binary data and hence cannot be applied to the Medical Scheme 55-59 data set. The MCA biplot (Gower & Hand, 1996) however can be used on binary and categorical data and it provides information simultaneously on both the objects and the binary (or categorical) variables in the same graphical display. The MCA biplot could therefore be applied to the Medical Scheme 55-59 data set.

5.3.1.2 Clustering large applications (CLARA)

The PAM method performs reasonably well in many clustering applications, but its large memory requirements is a major drawback. The PAM method constructs dissimilarities between all pairs of observations and the memory requirements are quadratic in the number of observations, while the computation time goes up accordingly. CLARA uses a less memory intensive algorithm. This means that CLARA can deal with much larger data sets than PAM. Internally, this is achieved by considering data subsets of fixed size, so that the overall time and storage requirements are linear in the number of observations, rather than quadratic (Struyf *et al*., 1997).

CLARA only accepts input of the **X**: $n \times p$ data matrix. The CLARA algorithm takes a subset of the whole data set and then applies the PAM algorithm to this subset. This causes the data subset to be divided into k clusters. The remaining objects of the larger data set are then assigned to the nearest medoid. The objective function in (5-1) is then calculated for the whole data set.

This procedure is repeated for several subsets of the large data set. The clustering that gave the lowest objective function value is retained. This means the CLARA algorithm also minimises a sum of dissimilarities. This procedure only needs to compute and store the dissimilarity matrix of one data subset at a time, which causes a linear increase in time and storage requirements when n, the number of objects in the data set, increases.

Graphical display: The silhouette plot and the clusplot

The silhouette plots provided by CLARA only represent the clustering structure related to the best-chosen subset. The best-chosen subset is the subset with the lowest value of the objective function in (5-1). This silhouette plot should be interpreted in the same way as the silhouette plot provided by PAM. The clusplot can also use the clustering data provided by CLARA. The clustering structure related to the best-chosen subset will be used to construct the clusplot.

5.3.1.3 Fuzzy Analysis (FANNY)

The PAM and CLARA methods are crisp clustering methods, because each observation can only belong to a single cluster and cannot belong to multiple clusters simultaneously. Fuzzy clustering gives each observation fractional membership in multiple clusters. The main advantage of fuzzy clustering over hard clustering is that it provides more detailed information of the clustering structure. However, more information is also more difficult to interpret. Other disadvantages of FANNY include that it does not assign representative objects for each cluster and the considerable computation time (Kaufman & Rousseeuw, p.165). FANNY however can either accept a data matrix **X**: $n \times p$ or a dissimilarity matrix **D**.

The FANNY algorithm assigns each object i to each cluster v with membership coefficient u_{iv} that indicates how strongly object i belongs to cluster v. The membership coefficients u_{iv} are unknown, but are defined through minimization of the following objective function (Kaufman & Rousseeuw, p.171):

$$\text{objective function} = \sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d_{ij}}{2\sum_{j=1}^{n} u_{jv}^2}. \tag{5-2}$$

The dissimilarities d_{ij} are known, unlike the membership coefficients u_{iv}. The minimization is carried out in an iterative manner until the objective function in (5-2) reaches a minimum value. The membership coefficient u_{iv} must obey the following conditions when the objective function in (5-2) is calculated:

1. $u_{iv} \geq 0$ for all $i = 1, \dots, n$ and all $v = 1, \dots, k$.
2. $\sum_{v=1}^{k} u_{iv} = 1$ for all $i = 1, \dots, n$.

Dunn's partition coefficient (Kaufman & Rousseeuw, p.171) will be computed after the objective function in (5-2) has reached its minimum value. Dunn's partition coefficient is computed as:

$$F_k = \sum_{i=1}^{n} \sum_{v=1}^{k} \frac{u_{iv}^2}{n}. \tag{5-3}$$

Dunn's partition coefficient in (5-3) lies between $1/k$ and 1. This coefficient gives an idea of the fuzziness of the resulting clustering. A value close to $1/k$ indicates very fuzzy clustering, while a value close to 1 indicates crisp clustering. The *normalised* version of Dunn's partition coefficient (Kaufman & Rousseeuw, p.171) is given by:

$$F_k^{'} = \frac{kF_k - 1}{k - 1} \tag{5-4}$$

The normalised version of Dunn's partition coefficient in (5-4) always has a value in the range [0, 1], whatever the value of k.

Graphical representations: The silhouette plot and the clusplot

The nearest crisp clustering is considered for graphical output (Struyf *et al.*, 1997). It assigns each object i to the cluster v in which it has the highest membership coefficient u_{iv}. The nearest crisp clustering is then represented graphically by means of a silhouette plot. The clusplot can also be constructed using the resulting nearest crisp clustering.

5.3.2 Hierarchical methods

A hierarchical algorithm describes a method yielding an entire hierarchy of clusterings for the given data set. In general, the hierarchical clustering methods are of two types: Agglomerative methods or divisive methods. Agglomerative methods start with each object in the data set representing its own small cluster, and then successively merging clusters until the whole data set is merged into one single large cluster. Divisive methods work in the opposite direction, as the divisive algorithms start off with the whole data set merged into one large cluster, and then splits up clusters until each object is separate (*S-PLUS ® 8 Guide to Statistics, Volume 2*, 2007, Chapter 23). Agglomerative nesting (AGNES) is one of the agglomerative methods. Divisive analysis (DIANA) and Monothetic analysis **(**MONA) are divisive methods.

Hierarchical and partitioning clustering algorithms are different. The goal of partitioning methods is to try to select the *best* clustering with *k* groups. The drawback of partitioning methods is that the final clustering structure will depend on the value of *k*, and the clustering structure might not necessarily be very good (Kaufman & Rousseeuw, 1990). Hierarchical methods do not really compete with partitioning methods, because they do not pursue the same goal. Hierarchical methods try to describe the data in a totally different way. A hierarchical method can never repair what was done in previous steps, where agglomerative methods cannot separate two objects once they have been joined in any previous steps, and divisive algorithms cannot merge two objects once they have been separated. This rigidity of hierarchical methods has the advantage that it leads to small computation times. However, the disadvantage is the inability to correct wrong decisions in retrospect (Kaufman & Rousseeuw, 1990, p.44-45).

5.3.2.1 Agglomerative nesting (AGNES)

The algorithm AGNES accepts either a dissimilarity matrix **D** or a data matrix **X**: $n \times p$ as input structure. If a data matrix is used as input, AGNES will first compute a dissimilarity matrix and use this dissimilarity matrix in its algorithm. The user also does not have to specify the number of clusters *k*, which is different from the partitioning methods.

AGNES is an agglomerative hierarchical clustering method and produces a sequence of clusterings. The algorithm starts with each object representing a different cluster. It then merges the two objects with the smallest dissimilarity between each other, into a single cluster. This cluster represents the two most similar objects, out of all the objects. The dissimilarities between the new cluster and the remaining clusters are then calculated. The process is then repeated until all the objects are merged into one single large cluster. The between-cluster dissimilarity can be defined in various ways such as the group average (average linkage) method, nearest neighbour (single linkage) method and the furthest neighbour (complete linkage) method. Kaufman & Rousseeuw (1990, p.243) preferred the group average linkage method to the other linkage methods, because they stated that "it is suitable for many situations and possess a certain robustness with respect to slight distortions". Several simulation studies were carried out to compare these different methods and the authors found that the group average linkage method was the most effective method in most cases (Kaufman & Rousseeuw, 1990, p.243). This is why the average linkage method will be used in this study.

The algorithm AGNES also provides the agglomerative coefficient, which gives an indication of the quality of the clustering structure. For each object *i, d(i)* denotes its dissimilarity to the first cluster it is merged with, divided by the dissimilarity of the merger in the last step of the algorithm (Struyf *et al*., 1997). The agglomerative coefficient (AC) is then defined as follows:

Definition 5.1 Agglomerative coefficient (AC)

- The average value of all {1−*d(i)*} values.

The value of the AC tends to increase with the number of objects, unlike the overall average silhouette width of the partitioning methods (Kaufman & Rousseeuw, 1990). This means that the AC values of data sets of different sizes should not be directly compared. This is a drawback of the AC value.

The clustering hierarchy obtained from AGNES can be graphically displayed by means of a clustering tree or by a banner plot.

Graphical display: The clustering tree

The agglomerative clustering tree is a graphical display that looks almost like a tree, where each leaf represents one of the objects from the data set (Struyf *et al*., 1997). The leaves are connected by small branches and the leaves belonging to the same small branch represent a cluster. These small branches are also joined by a larger branch until all objects are merged into a single large branch, representing one large cluster. The vertical coordinate where two branches join equals the dissimilarity between the corresponding clusters (Struyf *et al*., 1997). A drawback of clustering trees is that it becomes difficult to display many objects simultaneously (Kaufman & Rousseeuw, 1990).

Graphical display: The banner plot

The banner plot is a graphical display that shows the successive mergers from left to right, with objects listed from top to bottom. The mergers are represented by horizontal bars of a certain length, where the lengths represent the between cluster dissimilarities. The banner thus contains the same information as the clustering tree. The AC value is also provided by the banner plot and can also be defined as the average width of the banner plot (Struyf *et al.*, 1997).

5.3.2.2 Divisive analysis (DIANA)

The algorithm DIANA accepts either a dissimilarity matrix **D** or a data matrix **X** : $n \times p$. If the data matrix is used as input, DIANA will first compute a dissimilarity matrix and use this dissimilarity matrix in its algorithm. The user also does not have to specify the number of clusters k. DIANA is a divisive hierarchical method and the algorithm starts with all objects belonging to a single large cluster, and then splits up clusters until each object is separate.

The algorithm DIANA has been programmed in the *R* package *cluster* and computes a divisive hierarchy. This makes DIANA probably unique as most software for hierarchical clustering uses agglomerative methods (Struyf *et al.*, 1997). The main reason for this is the computational problems associated with divisive methods. In the first step of agglomerative methods, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ possible ways to merge two clusters. However, in the first step of divisive methods, there are $2^{n-1}-1$ possible ways to split the data set into two clusters. In practice, the latter number is much larger than the first due to larger data sets (*S-PLUS ® 8 Guide to Statistics, Volume 2*, 2007, Chapter 23). Kaufman & Rousseeuw (1990) consider a proposal of Macnaughton-Smith *et al.* (1964) to construct a divisive algorithm that uses less computational time.

DIANA divides the data set in the following iterative procedure. The first step is to find the object with the highest average dissimilarity to the other objects. This object initiates the *splinter group (S)*. The next step is to compute:

$$V_i = average_{j \notin S} d_{ij} - average_{j \in S} d_{ij}$$

for each object i. The object h for which this difference is largest must then be found. If $V_h > 0$, then object h is on average more similar to the splinter group (S) than the remaining objects, so object h is added to the splinter group (S). The previous step is repeated until all remaining differences V_h are negative. The data set is then split into two clusters. The next step is to select the cluster with the largest diameter, where the cluster diameter is the largest dissimilarity between any of the objects within the cluster. This cluster with the larger diameter is then divided in the same process mentioned above. The whole process is repeated until all clusters only contain a single object.

The *divisive coefficient* (DC) can be used to measure the clustering structure of the data set. For each object *i, d(i)* denotes the diameter of the last cluster to which it belonged, just before being split off as a single object, divided by the diameter of the whole data set (Struyf *et al*., 1997). The DC is defined as follows:

Definition 5.2 Divisive coefficient (DC)

- The average value of all {*d(i)*} values.

The DC value also increases with the number of objects in the data set. This means that the DC values of data sets of different sizes should not be directly compared.

The clustering hierarchy obtained from DIANA can be graphically displayed in two ways, by means of a clustering tree or by a banner plot. These are similar to the graphical displays provided by AGNES.

Graphical display: The clustering tree and the banner plot

Here the stem of the clustering tree represents the entire data set. The vertical coordinate where the branch splits into two smaller clusters equals the diameter of that cluster before splitting. The diameter of a cluster is the maximum dissimilarity between any two objects contained in the cluster. The resulting clustering tree provided by DIANA and the clustering tree provided by AGNES may be different, because the divisive algorithm is not the exact counterpart of the agglomerative algorithm (Struyf *et al*., 1997).

The banner plot shows the successive splits from left to right, with objects listed from top to bottom. The objects are put together by horizontal bars of a certain length, where the length represents the diameter of the cluster being split. The banner plot thus contains the same information as the clustering tree (Struyf *et al*., 1997). The divisive coefficient (DC) value can also be defined as the average width of the banner plot.

5.3.2.3 Monothetic analysis (MONA)

MONA is also a divisive hierarchical method. MONA can only operate on a data matrix **X** : $n \times p$ and cannot operate on a dissimilarity matrix **D**. Another restriction of MONA is that the data matrix **X** : $n \times p$ may only consist of binary variables. MONA uses a single well-chosen variable to perform each split in the data set. The MONA algorithm starts by choosing a variable to separate the data into two subsets. The objects with value of 1 for this binary variable are clustered into one group and the remaining objects make up the second cluster. Each one of these clusters is then separated into smaller clusters by choosing a suitable remaining variable. This process is continued until each cluster consists of objects having identical values for all variables. A final cluster is thus a single object or an indivisible cluster (*S-PLUS ® 8 Guide to Statistics, Volume 2*, 2007, Chapter 23).

The choice of the variable that is used to separate each cluster is an important part of the algorithm. The idea is to select the most centrally located variable, i.e. a variable that has the greatest sum of similarities to all other variables. There are several measures of association that may be used to define similarity between binary variables, as seen in Table 3.2. Kaufman & Rousseeuw (1990) choose to use a reasonably simple measure of association (similarity). This simple measure uses the number of objects for each combination of values that the two variables can take, similar to Table 3.4. The measure of association used by Kaufman & Rousseeuw (1990) between variables *i* and *j* is given below:

$$A_{ij} = |ad - bc|$$

where *a* indicates the number of objects out of *n* that score 1 for both variables *i* and *j*, *b* indicates the number of objects that score 1 for variable *i* and 0 for variable *j*, *c* indicates the number of objects that score 0 for variable *i* and 1 for variable *j* and *d* indicates the number of objects that both score 0 for variables *i* and *j*. This association measure is not a real similarity coefficient, in the strict sense, because it can take values larger than 1. The criterion also assumes that the binary variables are symmetric (Kaufman & Rousseeuw, 1990, p.300). Symmetric binary variables are binary variables where codes 0 and 1 are equally important. Asymmetric binary variables do not attach equal importance to codes 0 and 1. It is therefore less appropriate to use asymmetric binary variables in the algorithm MONA, as explained by Hubàlek (1982). This can be seen as a drawback of MONA.

Graphical display: The banner plot

The clustering hierarchy constructed by MONA can be represented by means of a banner plot. This banner plot is again a divisive banner, similar to the banner plot provided by MONA. However, the length of the bar is now given by the number of divisive steps needed to make a particular split. The variable responsible for this split is also listed in the inside of the bar.

5.4 Application of the Clustering Techniques on the Medical Scheme 55-59 Data Set

5.4.1 Choice of clustering techniques

The chronic diseases are binary variables of the Medical Scheme 55-59 data set. Clustering of the chronic diseases means that clustering of the binary variables, and not the objects, needs to be performed. The algorithms of Kaufman & Rousseeuw (1990) are very well suited for the clustering of the chronic diseases. The reason for this is that most of the clustering algorithms of Kaufman & Rousseeuw (1990) can take a dissimilarity matrix **D**, which contains the dissimilarities between the chronic diseases, as input structure. These methods also allow the $\mathbf{X} : n \times p$ data matrix as input structure, but the Medical Scheme 55-59 data set is not in the usual $\mathbf{X} : n \times p$ data matrix form. The reason is that each row of the Medical Scheme 55-59 data set does not represent a single life, but represents multiple lives that have the same unique combination of chronic diseases. It will therefore be wrong to use the Medical Scheme 55-59 data set directly as input structure. One possibility to overcome this problem is to expand the multiple lines into individual lines and then proceed accordingly, but this also leads to problems. The expanded matrix will then contain 2 199 367 rows, which is impractical to use. Another possibility is to assign a weight to each row of the Medical Scheme 55-59 data set, where the weight is based on the particular number of multiple lives in that row. A vector of weights can then be determined and used along with the Medical Scheme 55-59 data set as input structure. This approach however is also limited, because many of the clustering algorithms do not allow a vector of weights and the $\mathbf{X}: n \times p$ data matrix as input structure. Furthermore, it must be remembered that the Medical Scheme 55-59 data set consists of binary data, which means that only clustering algorithms suited for binary data and clustering algorithms that accept a dissimilarity matrix **D** as input structure, can be used on the Medical Scheme 55-59 data set. The clustering algorithms PAM, FANNY, AGNES and DIANA can therefore be used, because these methods can take a dissimilarity matrix **D** which contains the dissimilarities between the chronic diseases, as input structure. Unfortunately, the method CLARA cannot be used, as it cannot be used on binary data and it does not allow dissimilarity matrices as input structure. The method MONA does not take a dissimilarity matrix as input structure, but it can be used on binary data. The only practical requirement for the use of MONA on the Medical Scheme 55-59 data set is that the multiple lines must be expanded into individual lines and then this expanded data matrix used as input structure. However, the expanded matrix will then contain 2 199 367 rows, which is impractical to use. Another problem is that the association measure used in MONA assumes that the data set consists of symmetric binary variables (Kaufman & Rousseeuw, 1990, p.300), but the Medical Scheme 55-59 data set consists of asymmetric binary variables and it is therefore not really appropriate to use MONA, as explained by Hubàlek (1982). The *K-means* partitioning method can be implemented by the *R* function *kmeans* found in the package *cluster*, available from `http://www.cran.r-project.org`. The *R* function *kmeans* however does not accept a dissimilarity matrix as input structure and cannot be used on binary data and therefore cannot be applied to the Medical Scheme 55-59 data set (*S-PLUS ® 8 Guide to Statistics, Volume 2*, 2007). The model-based clustering algorithms can also not be used on the Medical Scheme 55-59 data set, because they do not take a dissimilarity matrix as input structure and the model-based clustering algorithms cannot be used on binary data (*S-PLUS ® 8 Guide to Statistics, Volume 2*, 2007).

The calculation of dissimilarity matrices was discussed in Chapter 3. The *R* function *Dissim.CDL* was developed to calculate dissimilarities between all chronic diseases. This function provides several dissimilarity matrices as output, which can be used as input structure for the clustering algorithms: PAM, FANNY, AGNES and DIANA.

5.4.2 Application of AGNES

This agglomerative hierarchical clustering method can be implemented by the *R* function *agnes* found in the package *cluster*, available from `http://www.cran.r-project.org`. Many different similarity coefficients were used to construct dissimilarity matrices, although the *Jaccard* coefficient is one of the most widely used coefficients (Gordon, 1999, p.18; Kaufman & Rousseeuw, 1990, p.26). The similarities were transformed to dissimilarities by the function $d_{ij} = 1 - s_{ij}$. The average linkage method was used to define the between-cluster dissimilarities, because Kaufman & Rousseeuw (1990) believed it was the most appropriate method to use.

Clustering trees are constructed in Figures 5.1 to 5.3 using different similarity coefficients. Various banner plots are also constructed in Figure 5.4. These graphical representations can be used to describe the clusters of chronic diseases.

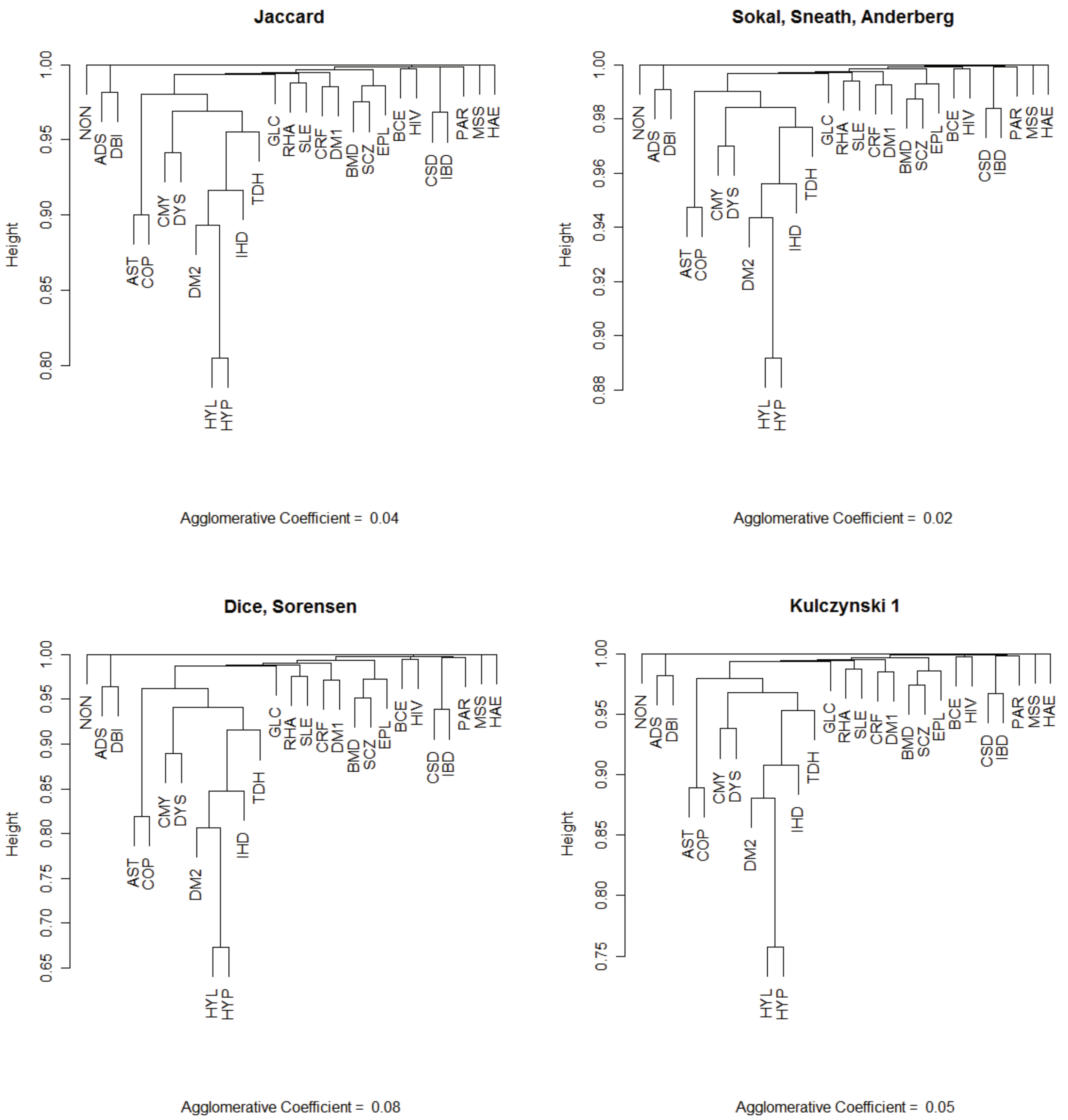


Figure 5.1: *Clustering trees of the chronic diseases of the Medical Scheme 55-59 data set using different similarity coefficients and AGNES. Notice how the clustering trees show almost identical clustering structures.*

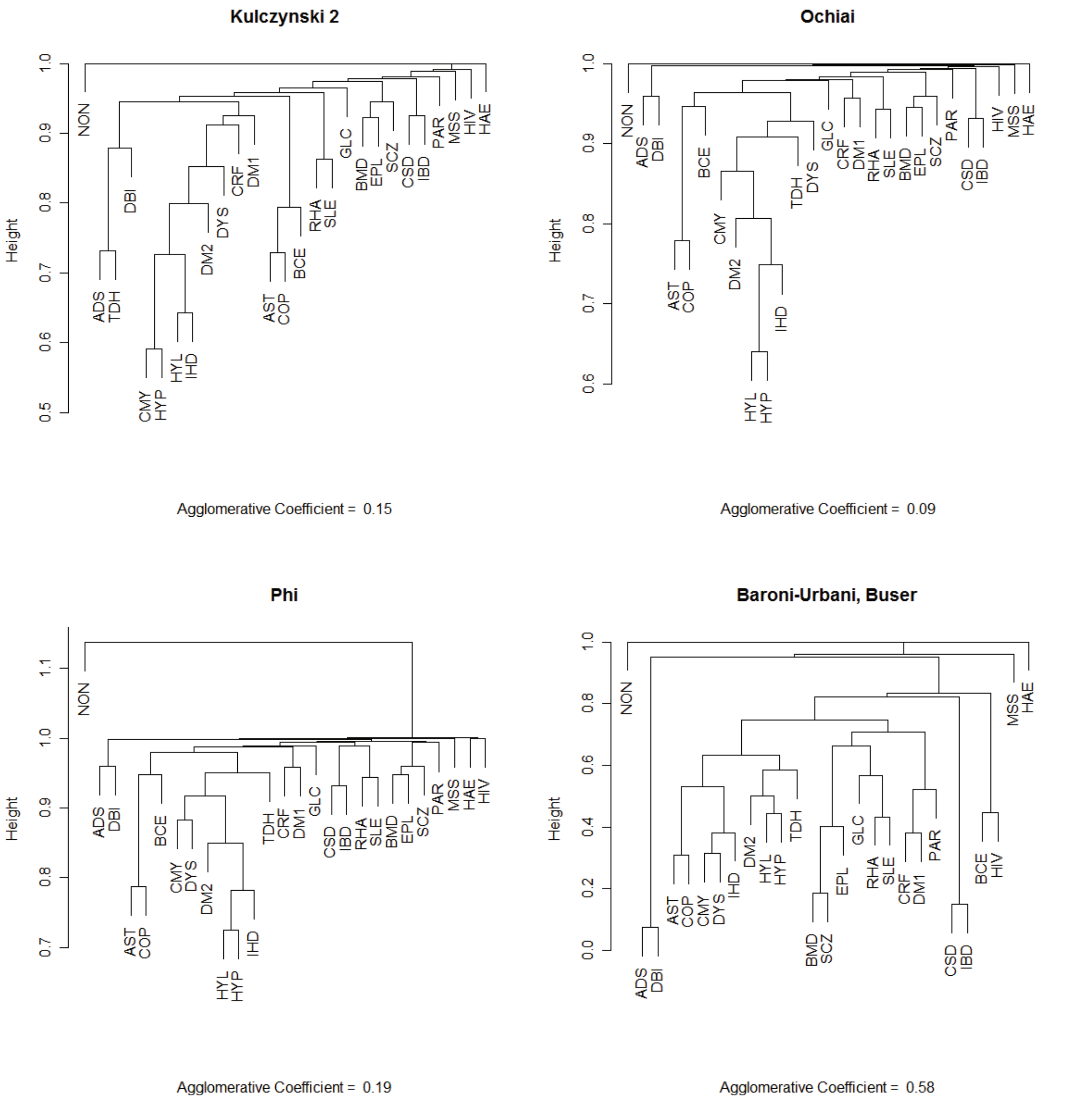


Figure 5.2: *Clustering trees of the chronic diseases produced by AGNES using more similarity coefficients. The Baroni-Urbani, Buser similarity coefficient leads to a reasonable clear clustering structure, because it has the highest AC value.*

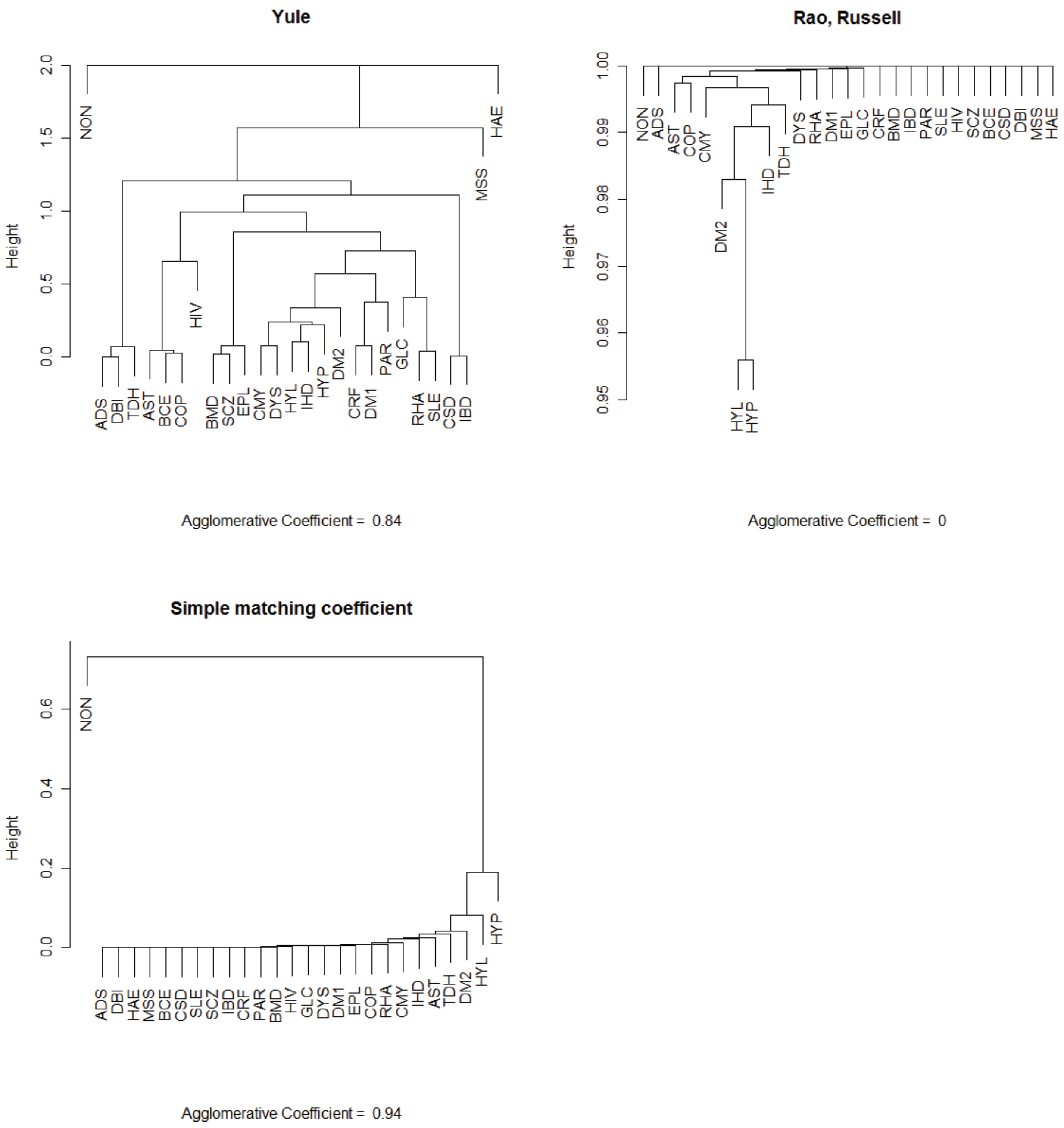


Figure 5.3: *Clustering trees of the chronic diseases produced by AGNES using more similarity coefficients. The Rao-Russell and the Simple matching coefficients show an unsuitable clustering structure.*

Figure 5.1 shows different clustering trees that have almost identical clustering structures. This is in agreement with Kaufman & Rousseeuw (1990). It was mentioned by Kaufman & Rousseeuw (1990) that the *Jaccard*, *Dice-Sorensen* and *Sokal-Sneath-Anderberg* similarity coefficients are all monotonic to each other and will show very similar clustering structures.

It also seems that the *Simple matching* and the *Rao-Russell* similarity coefficients are not appropriate to use with asymmetric binary variables. The chronic diseases were ranked in Chapter 2 in accordance with the chronic disease rates that they displayed. It was shown that HYP, HYL, DM2, TDH, AST, IHD, CMY and RHA displayed the highest chronic disease rates, in decreasing order. It seems that the chronic diseases are grouped in the *Simple matching* and the *Rao-Russell* clustering trees according to their chronic disease rate and not according to whether the chronic diseases actually co-occur. This is undesirable, as chronic diseases should be clustered together when they tend to co-occur often and not because the chronic diseases have similar chronic disease rates. It was also found in Chapter 4 that the *Simple matching* and the *Rao-Russell* similarity coefficients were unsuitable for use on the Medical Scheme 55-59 data set.

The *Yule* and *Baroni-Urbani-Buser* similarity coefficients have very clear clustering structures. It is interesting to note that the *Yule* and *Baroni-Urbani-Buser* similarity coefficients produced the widest range of dissimilarity values, as seen in Section 4.3.2. Most of the other similarity coefficients do not produce clear clustering structures. Many of the clustering trees displayed in Figures 5.1 to 5.3 show similar clustering structures. These clustering structures can be compared to the body system rules to see which chronic diseases are related.

The first body system rule involves the respiratory diseases COP, AST and BCE. Many of the clustering trees, like *Yule*, do show that these diseases form a cluster. It seems that COP and AST have a stronger relation with each other, than with BCE. The clustering tree based on the *Jaccard* coefficient however shows that COP and AST form a cluster, and BCE forms a cluster with HIV.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. Most of the clustering trees show that these diseases are located in the same cluster. These cardiovascular diseases also seem to cluster with chronic diseases HYL, DM2, TDH, AST and COP. The chronic diseases HYL and HYP seem to co-occur regularly. It seems that there is enough evidence based on the data to suggest that the chronic diseases mentioned in the second body system rule are strongly related.

The third body system rule involves HYP and CRF. There was very little evidence to indicate that these two chronic diseases are strongly related. Only the *Kulczynski 2* clustering structure did in fact show that these two diseases were classified in the same cluster.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the clustering trees show that these chronic diseases form a cluster, which suggests that CSD and IBD are strongly related.

The fifth body system rule involves the chronic diseases BMD and SCZ. It seems that these chronic diseases are strongly related, as they tend to be classified in the same cluster. It also seems that these two chronic diseases tend to form a cluster with EPL. This supports part of the sixth body system rule.

The sixth body system rule involves MSS, BMD and EPL. It appears that BMD and EPL tend to form clusters, along with SCZ. However, MSS tends to form a cluster on its own, as can be seen from most of the clustering trees. It therefore seems that BMD and EPL are strongly related, but MSS does not seem to be related to BMD and EPL.

The seventh body system rule involves the chronic diseases SLE and RHA. Almost all the clustering trees show that these two diseases form a cluster, which indicate that these two chronic diseases are strongly related.

The last system body rule involves DM1 and DM2. It appears that DM1 tends to form a cluster with CRF, and that DM2 tends to form a cluster with the cardiovascular diseases CMY, IHD, HYP and DYS.

Most of the results produced in this section are in accordance with the results produced by the MDS methods in Chapter 4. This inspires even more confidence in the accuracy of the results.

It was mentioned in Chapter 2 that Figure 2.3 shows that the diseases MSS, PAR, HIV and HAE have very little co-occurrences with other chronic diseases. These observations are validated by the clustering structures. Each of the chronic diseases HAE, MSS and PAR form their own separate cluster.

The chronic diseases ADS and DBI also form a cluster in most of the clustering trees. These diseases however are not mentioned in any of the body system rules.

The banner plot displays the same information as a clustering tree, but in a different way. Banner plots based on four similarity coefficients are displayed in Figure 5.4. The agglomerative coefficient (AC) is the average of all $\{1-d(i)\}$ values, and measures the clustering structure of the data set. The

AC value can also be defined as the average width (or the percentage filled) of the banner plot. The banner plots based on the *Yule* and *Baroni-Urbani-Buser* similarity coefficients have the clearest clustering structure.

The banner plot based on the *Jaccard* coefficient shows that HYP and HYL form a cluster, and also form clusters with AST, COP, DM2, TDH, CMY, IHD and DYS. Also, notice that BMD, SCZ and EPL form a separate cluster. The chronic diseases CRF and DM1, ADS and DBI, CSD and IBD, RHA and SLE all form separate clusters. The clustering tree based on the *Jaccard* coefficient show exactly the same clustering structure. The same can be said for the other three banner plots.

The banner plot does not provide any extra information, it only displays the information in a different format. The clustering tree may become over-crowded when many objects are clustered, but the banner plot can usually plot many objects without any over-crowding.

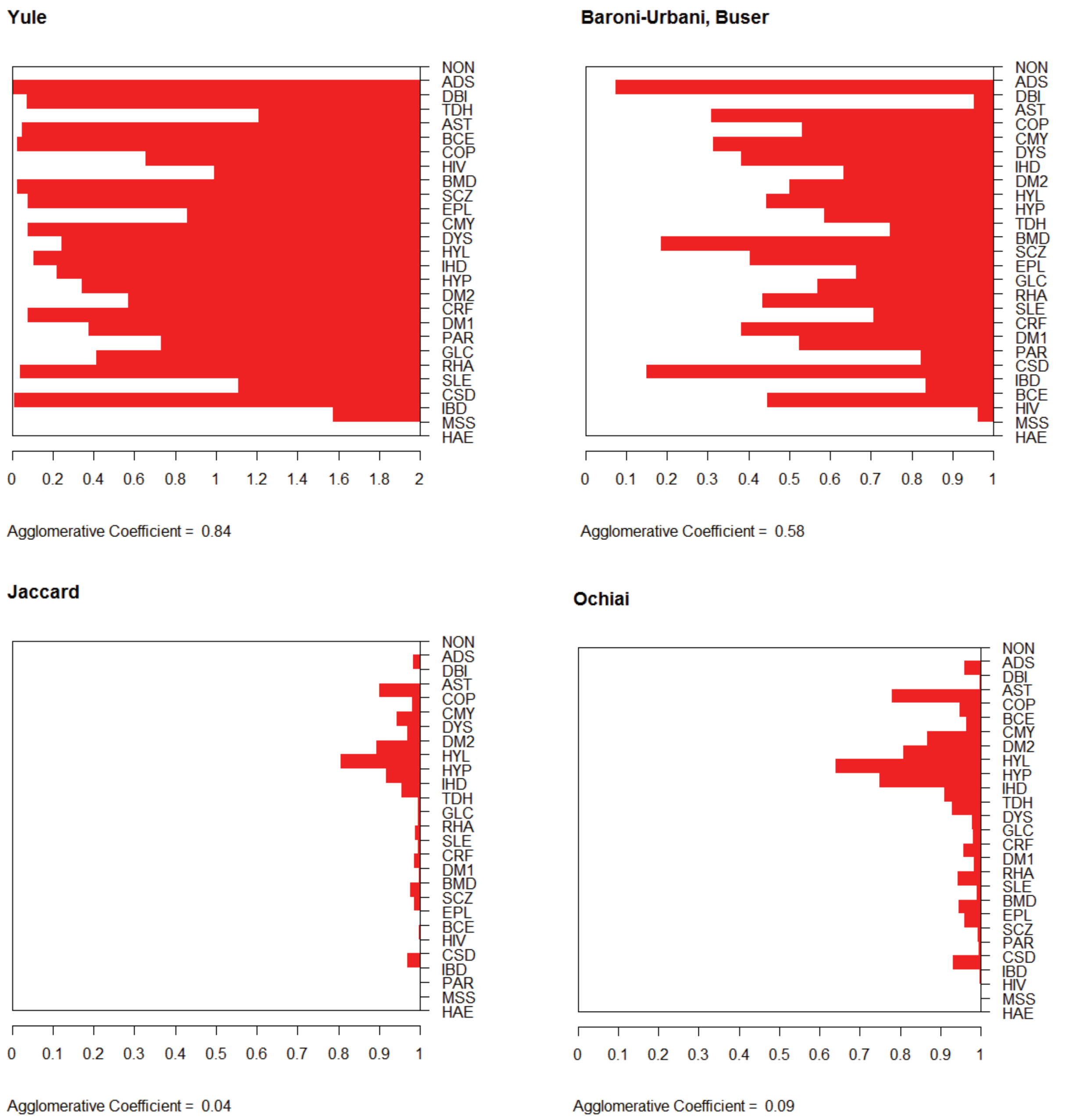


Figure 5.4: *Banner plots of the chronic diseases that occur in the Medical Scheme 55-59 data set. They contain the same information as the clustering trees. The algorithm AGNES was used to construct the banner plots. Higher agglomerative coefficient (AC) values show clearer clustering structures. The banner plots should be read from left to right. The bars represent the mergers, which commence at the between cluster dissimilarity.*

5.4.3 Application of DIANA

This divisive hierarchical clustering method can be implemented by the *R* function *diana* found in the package *cluster*. Many different similarity coefficients were used to construct dissimilarity matrices, and the similarities were transformed to dissimilarities by the same function $d_{ij} = 1 - s_{ij}$. The average linkage method was used to define the between-cluster dissimilarities.

The DIANA method is very similar to AGNES in several aspects including the output. DIANA however is a divisive hierarchical method and starts off with all objects belonging to a single large cluster, and then splits up clusters until each object is separate. It might seem that it works in the opposite direction, but the resulting output provided by DIANA may differ from the output of AGNES because the divisive algorithm is not the exact counterpart of the agglomerative algorithm.

Clustering trees are constructed in Figure 5.5 and Figure 5.6, using different similarity coefficients. Not all similarity coefficients listed in Table 3.2 will be displayed. The use of the *Simple matching* and the *Rao-Russell* similarity coefficients are unsuitable, because the Medical Scheme 55-59 data set does not consist of symmetric binary variables. The *Jaccard*, *Dice-Sorensen* and *Sokal-Sneath-Anderberg* similarity coefficients are all monotonic to each other and will show very similar clustering structures. The *Sokal-Sneath-Anderberg* similarity coefficient will therefore not be considered. Various banner plots are also constructed in Figure 5.7. These graphical representations can again be used to describe the clusters that form, just like the graphical representations of AGNES. The clustering structures will also be compared to the body system rules.

Figure 5.5 shows different clustering trees that have almost identical clustering structures. The *Yule*, *Phi* and *Baroni-Urbani-Buser* similarity coefficients have very clear clustering structures, as seen in Figure 5.6. The other similarity coefficients do not seem to have clear clustering structures. The *Jaccard* coefficient is the most widely used similarity coefficient when the data consist of asymmetric binary variables, but it does not produce a clear clustering structure of the chronic diseases. Most of the clustering trees show similar clustering structures, but the clustering tree based on the *Yule* similarity coefficient is different from the other clustering trees.

The first body system rule involves the respiratory diseases COP, AST and BCE. The clustering trees based on the *Kulczynski 2*, *Ochiai* and *Phi* similarity coefficients form clusters with these three chronic diseases. The other clustering trees show that COP and AST are located in the same cluster, but BCE tends to form a cluster with HIV. The clustering tree based on the *Yule* similarity coefficient shows that BCE, COP and HIV form a cluster and AST is located in a different cluster.

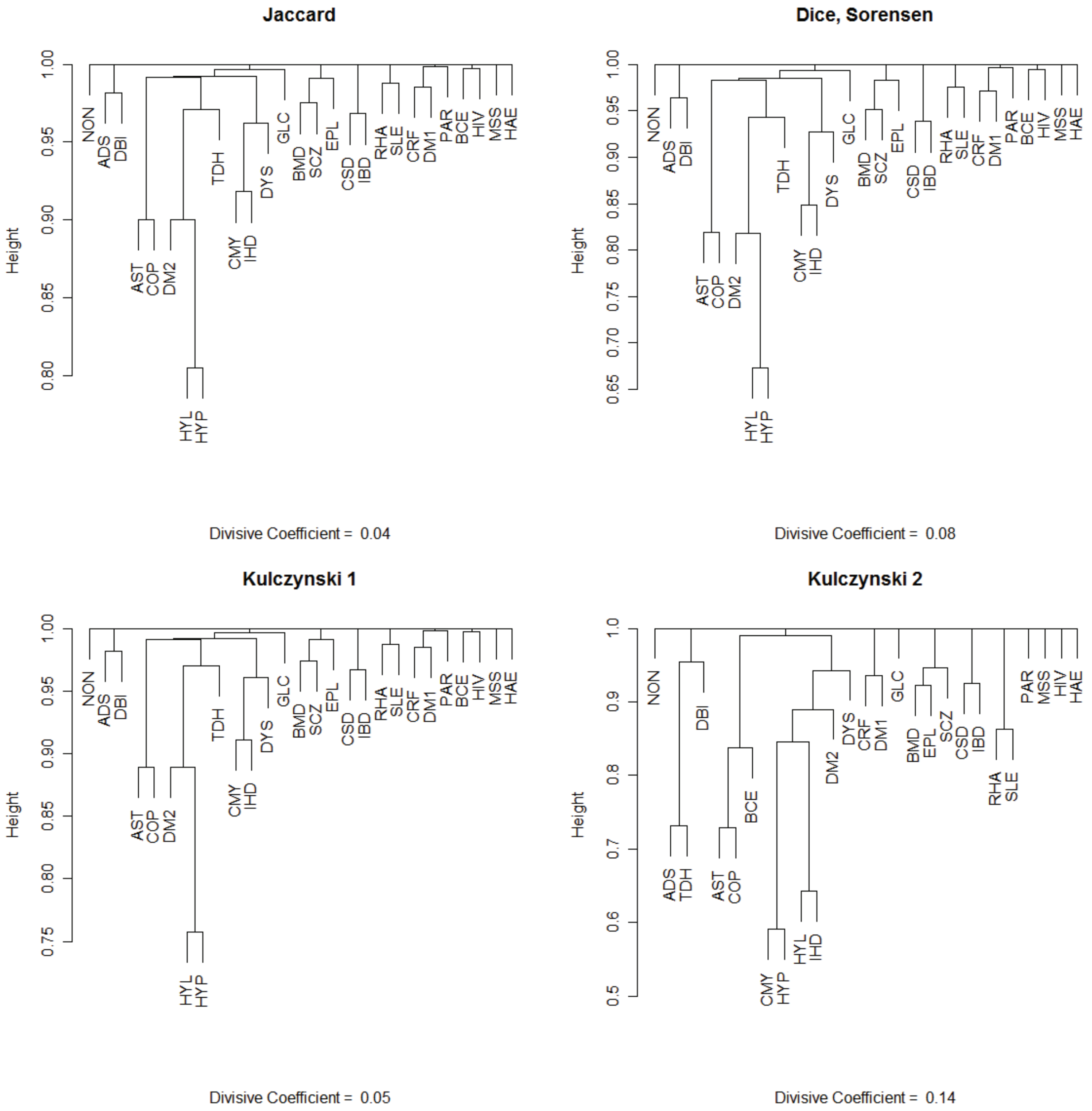


Figure 5.5: *Clustering trees of the chronic diseases of the Medical Scheme 55-59 data set using DIANA. Notice how the first three clustering trees have almost identical clustering structures.*

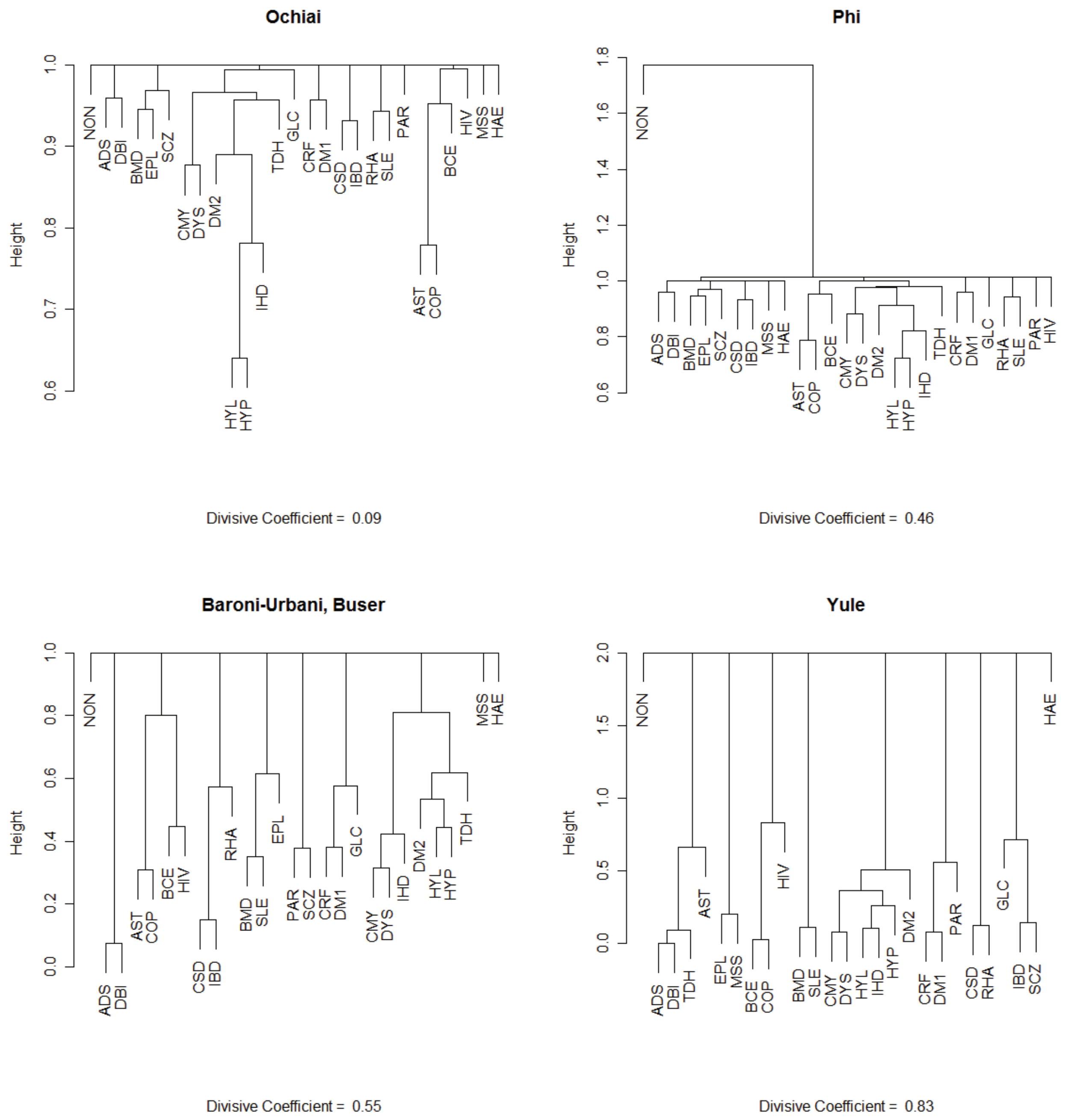


Figure 5.6: *Clustering trees of the chronic diseases of the Medical Scheme 55-59 data set using DIANA. The last three clustering trees have reasonable clear clustering structures, because of higher values of the divisive coefficient (DC).*

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. Most of the clustering trees show that these diseases are located in the same cluster. These cardiovascular diseases also seem to cluster with chronic diseases HYL, DM2 and TDH. The chronic diseases HYL and HYP appear to co-occur regularly. It seems that there is evidence based on the data to indicate that the chronic diseases CMY, IHD, DYS and HYP are strongly related.

The third body system rule involves HYP and CRF. The chronic disease CRF tends to cluster with DM1, while HYP tends to form a cluster with HYL and other cardiovascular related diseases. The chronic diseases CRF and HYP do not even seem to be broadly clustered together.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the clustering trees, except the clustering tree based on the *Yule* similarity coefficient show that these chronic diseases are strongly related. The clustering tree based on the *Yule* similarity coefficient shows that CSD and RHA form a cluster while IBD and SCZ form another cluster.

The fifth body system rule involves the chronic diseases BMD and SCZ. It seems that these two chronic diseases are strongly related. It also seems that these two chronic diseases tend to form a cluster with EPL.

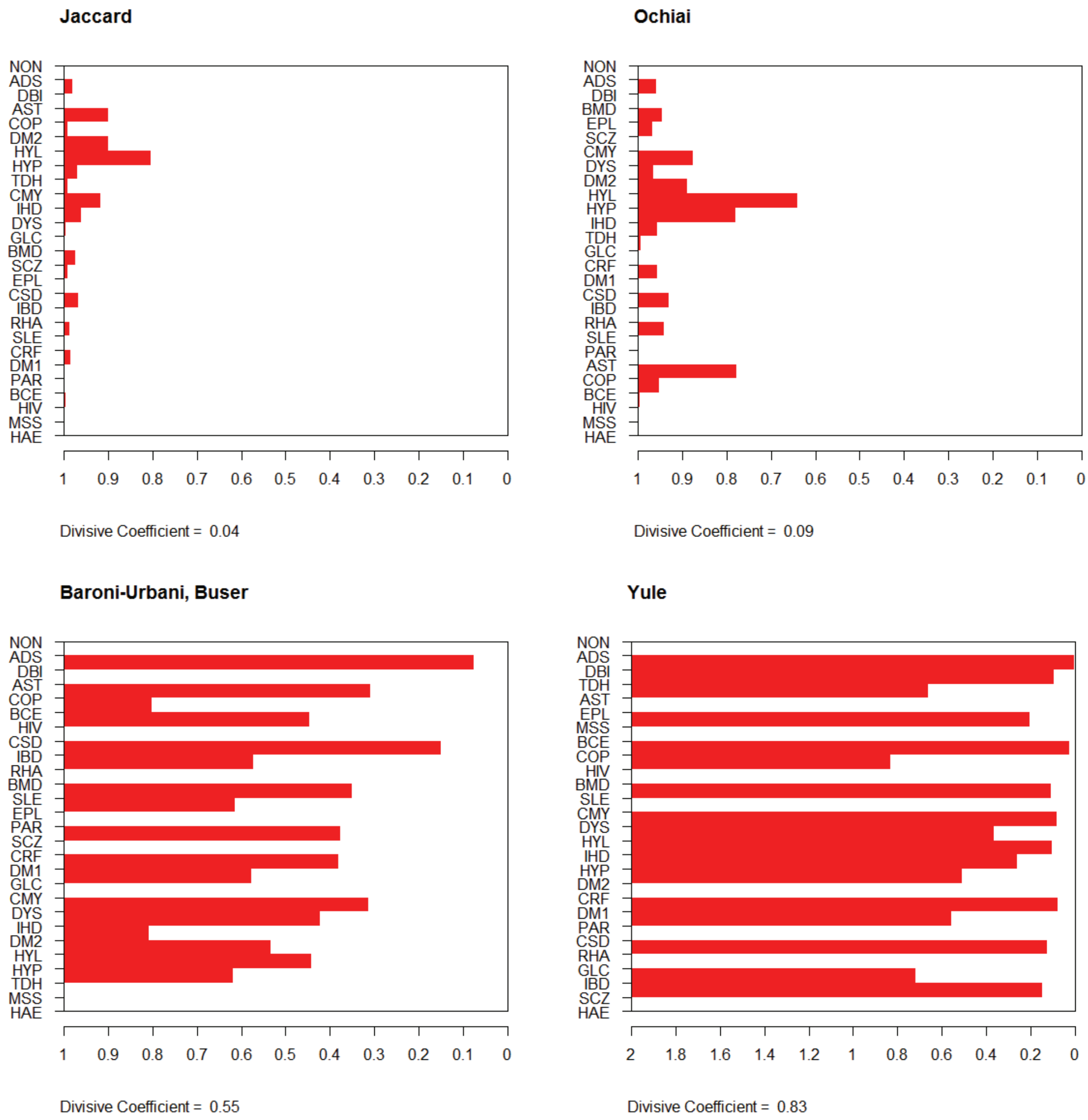
The sixth body system rule involves MSS, BMD and EPL. There is strong evidence that indicates that BMD and EPL tend to form clusters, along with SCZ. However, MSS tends to form a cluster on its own, as can be seen from most of the clustering trees.

The seventh body system rule involves the chronic diseases SLE and RHA. Almost all the clustering trees show that these two diseases form a cluster.

The chronic diseases ADS and DBI also form a cluster in most of the clustering trees. These diseases are however not mentioned in any of the body system rules.

Banner plots based on four different similarity coefficients are plotted in Figure 5.7. These banner plots of DIANA should be read from right to left. The divisive algorithm starts with one single cluster, and then forms smaller and smaller clusters. It works in the opposite direction to AGNES, and is not an exact counterpart.

The divisive coefficient (DC) value measures the clustering structure found in the data. The banner plots based on the *Yule* and *Baroni-Urbani-Buser* similarity coefficients have the clearest clustering structure. The banner plot based on the *Jaccard* coefficient shows that HYP and HYL form a cluster, and also form clusters with AST, COP, DM2, TDH, CMY, IHD and DYS. Notice that BMD, SCZ and EPL form a separate cluster. The chronic diseases CRF and DM1, ADS and DBI, CSD and IBD, RHA and SLE all form separate clusters. The clustering tree based on the *Jaccard* coefficient shows exactly the same clustering structure. The *Yule* banner plot is different from the other banner plots.



Jaccard

Ochiai

Baroni-Urbani, Buser

Yule

Figure 5.7: *Banner plots of the chronic diseases using DIANA. They contain the same information as the clustering trees. Larger divisive coefficient (DC) values show more clear clustering structures. The banner plots should be read from right to left.*

5.4.4 Application of PAM

The partitioning around medoids (PAM) algorithm is one of the partitioning methods. These partitioning methods are different to the hierarchical methods, AGNES and DIANA. The partitioning and hierarchical methods should not be directly compared, as they have different objectives. This partitioning clustering method can be implemented by the *R* function *pam*, found in the package: *cluster*. PAM can take a dissimilarity matrix as input.

Many different similarity coefficients were used to construct dissimilarity matrices, and the similarities were transformed to dissimilarities by the same transformation function

$$d_{ij} = (1-s_{ij})^{0.5}.$$

This transformation leads to more dissimilarity matrices with metric and Euclidean properties, as listed in Table 3.3. This is useful because the clusterings will be displayed in a clusplot that require dissimilarity matrices with metric properties, because classical scaling techniques are used to find the configuration of points that represent the chronic diseases. This also means that the clusplots and the displays produced by classical scaling in Chapter 4 will be very similar, the only difference being that the clusplot also draws ellipses to indicate certain clusters.

The partitioning methods, like PAM, need the user to specify the desired number of clusters, defined as *k*. This is a drawback of these partitioning methods because the user might not know how many clusters are present in the data. However, Kaufman & Rousseeuw (1990) suggest that the user can choose the value of *k* (number of clusters) by considering the overall average silhouette width. The overall average silhouette width gives an indication of the clustering structure found in the data. The user should run PAM with different values of *k* and the overall average silhouette width value should be noted for each run. The most appropriate *k* value is the one that provides the highest overall average silhouette width.

Figures 5.8 and 5.9 display the overall average silhouette width for different values of *k*, for the different similarity coefficients. The largest overall average silhouette width is also calculated in Table 5.1, with the corresponding value of *k* that should be used. It can be seen that most similarity coefficients have different optimal values for *k*. The greatest number of clusters is 16 and the smallest number is 8. The largest overall average silhouette width gives an indication of the quality of the clustering structure. A value of above 0.7 indicates very clear clustering structure; a value between 0.5 and 0.7 shows reasonable clustering structure, a value between 0.25 and 0.5 shows weak clustering structure and a value below 0.25 indicates that no substantial structure has been found.

Table 5.1: *Largest overall average silhouette width calculated for the different similarity coefficients. The value of the number of clusters that lead to the largest value is also given. This value corresponds to the value of k that must be used when PAM is implemented.*

Similarity coefficients	**Largest overall average silhouette width**	**Corresponding number of clusters (*k*)**
Jaccard	0.0110	16
Dice, Sorensen	0.0204	16
Sokal, Sneath, Anderberg	0.0057	16
Kulczynski 2	0.0295	14
Ochiai	0.0173	11
Phi	0.0212	13
Baroni-Urbani, Buser	0.1473	8
Yule	0.3148	15

The clustering structure based on the *Yule* similarity coefficient is the only one that produced a largest overall average silhouette width value of above 0.25. The other similarity coefficients have no substantial clustering structure. This does not mean that the *Yule* similarity coefficient is the best coefficient to use, as it still needs to be appropriate in the Medical Scheme 55-59 data context. It was previously mentioned that most authors recommend the *Jaccard* similarity coefficient when the data consist of asymmetric binary variables, as is the case with the Medical Scheme 55-59 data set. The *Jaccard* similarity coefficient shows a very low clustering structure, but it is an appropriate coefficient to use. It is interesting to note that the *Yule* similarity coefficient produced the widest range of dissimilarity values, as seen in Section 4.3.2. The *Jaccard* similarity coefficient, on the other hand, produced a much more concentrated range of dissimilarity values. It therefore seems that the range of dissimilarity values have a great influence on the quality of the clustering structure.

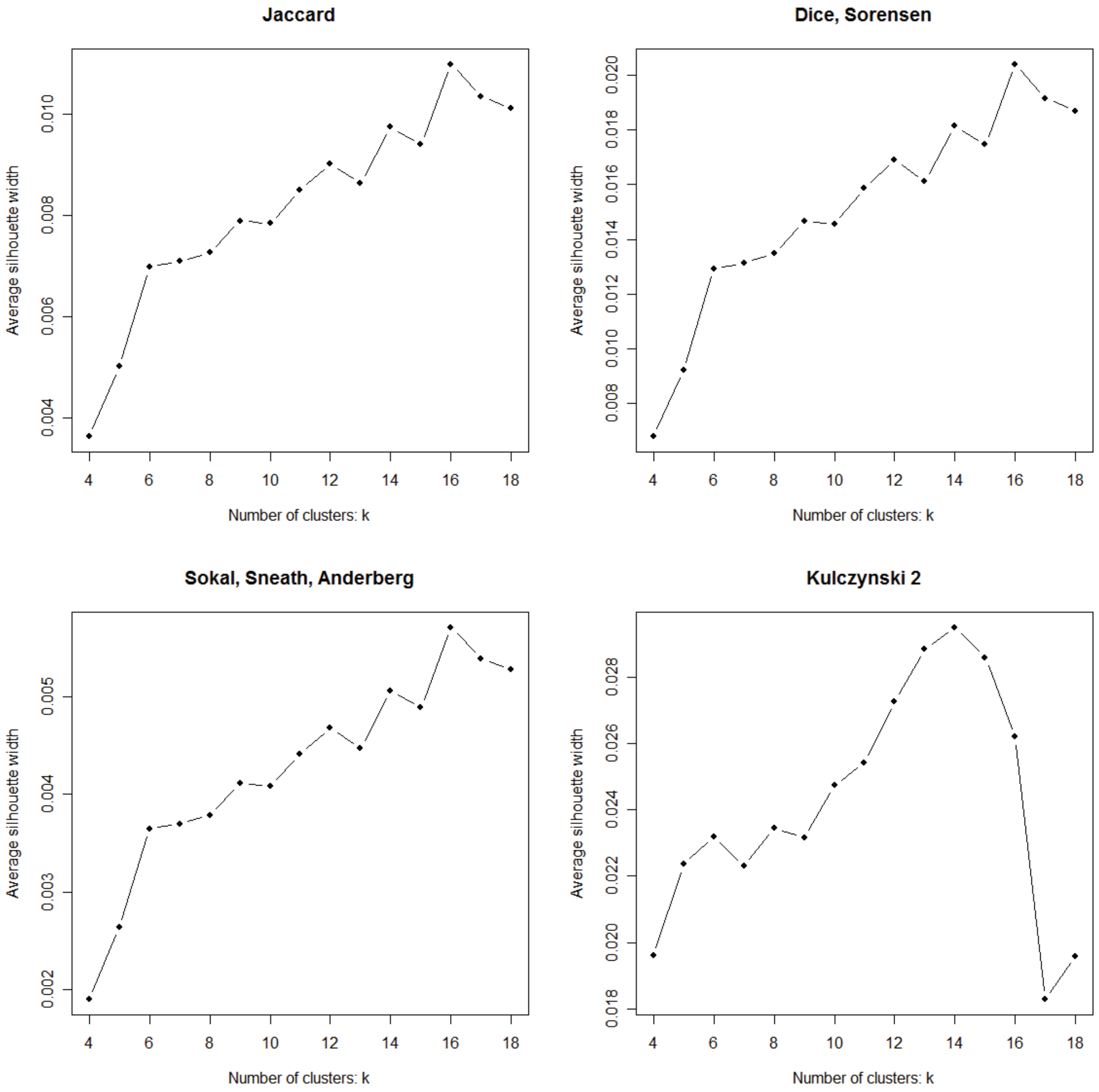


Figure 5.8: *The plot of the overall average silhouette width calculated by means of PAM, for different choices of k (number of clusters). The value of k corresponding to the highest overall average silhouette width should be the preferred choice for number of clusters. This was done for different similarity coefficients.*

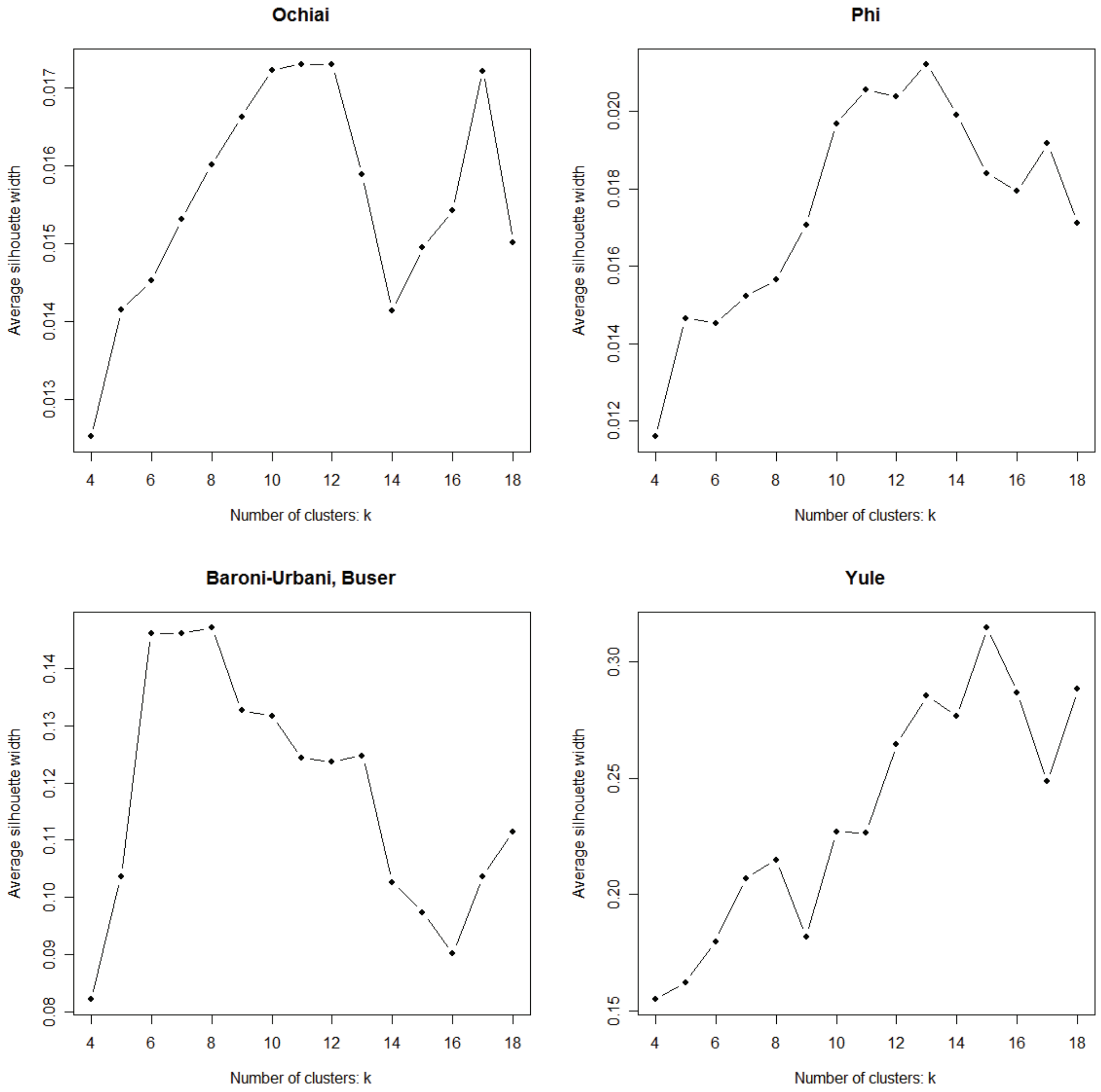


Figure 5.9: *The plot of the overall average silhouette width calculated by means of PAM, for different choices of k. The value of k corresponding to the highest overall average silhouette width should be the preferred choice for number of clusters.*

Graphical representations of the resulting clustering structures will again be constructed for the different similarity coefficients. Clusplots and silhouette plots will be constructed to display the clusters. These clustering structures can be compared to the body system rules to see if the chronic diseases are related.

Silhouette plots based on different similarity coefficients are presented in Figures 5.10 and 5.11. For each object *i*, the silhouette value *s(i)* is computed. The silhouette value *s(i)* always lies between −1 and 1. The value of *s(i)* close to −1 can be interpreted that object *i* is badly classified. The value of *s(i)* close to 1 means that the object *i* is well classified and a value close to 0 means that object *i* lies between two clusters. The silhouette of a cluster is a plot of all the *s(i)* values, ranked in decreasing order, of all the objects *i* in that cluster. The entire silhouette plot shows the silhouettes of all the clusters next to each other. The quality of the clusters can then be compared. The overall average silhouette width of the silhouette plot is the average of the *s(i)* values of all the objects in the data set.

Notice that the *Sokal-Sneath-Anderberg, Dice-Sorensen* and *Jaccard* similarity coefficients show almost identical silhouette plots. The *Yule* similarity coefficient has the clearest clustering structure, but even this clustering structure is weakly defined.

Clusplots of clustering structures produced by PAM based on different similarity coefficients are displayed in Figures 5.12 and 5.13. The following instructions in *R* were used to construct the *Jaccard* clusplot in Figure 5.12:

```
All.Dissim<-Dissim.CDL(x=final.dat[c(1:1636),c(1:28)],transformation=2)
clusplot(x=All.Dissim$Jaccard,pam(x=All.Dissim$Jaccard),diss=TRUE,k=16)$c
lustering,lines=0,color=TRUE, labels=3,main="Jaccard",asp=1)
```

The first line is used to construct all dissimilarity matrices based on the different similarity coefficients. The transformation $d_{ij} = (1-s_{ij})^{0.5}$ was used to construct the dissimilarities. The *R* function *Dissim.CDL* is provided in Appendix A. The second and third lines of the *R* instructions are the actual implementation of the *R* function *clusplot*, which is found in the package: *cluster*. It is very important to note that the *asp=1* function argument is used to ensure that the correct aspect ratio is used. This means that a unit change in the horizontal direction is equal to a unit change in the vertical direction. Otherwise distances cannot be appreciated, which is vital for interpreting the graphical displays. The other clusplots displayed in Figures 5.12 and 5.13 are constructed in a similar manner.

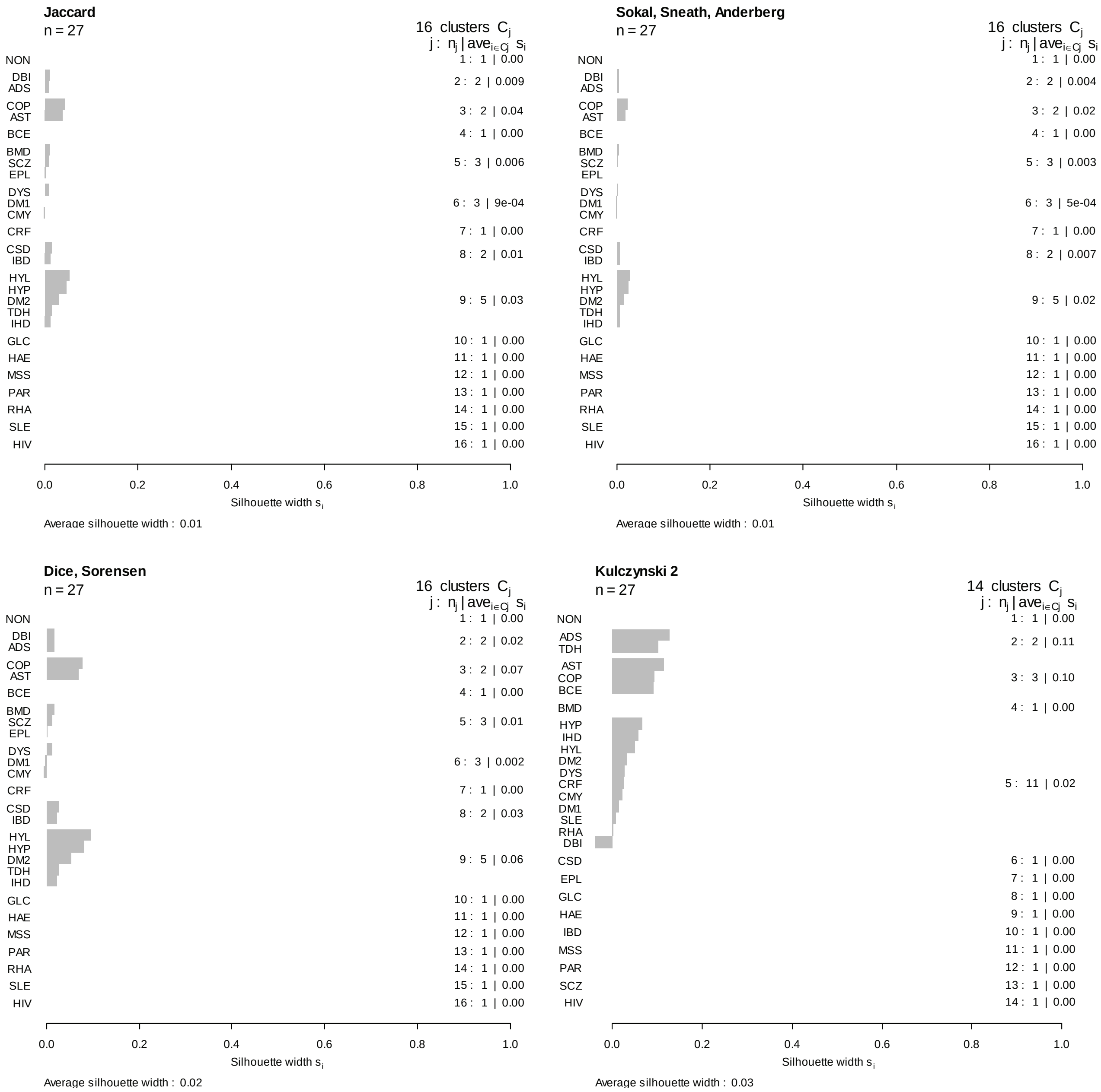


Figure 5.10: *Silhouette plots of PAM using different similarity coefficients. Large silhouette width values (values close to 1) indicate that a certain chronic disease is well classified within the corresponding cluster. Silhouette width values close to −1 indicate that the chronic disease is badly classified within the corresponding cluster. The average silhouette width gives a measure of the quality of the clustering structure. Medoids are shown as the first chronic disease in each cluster.*

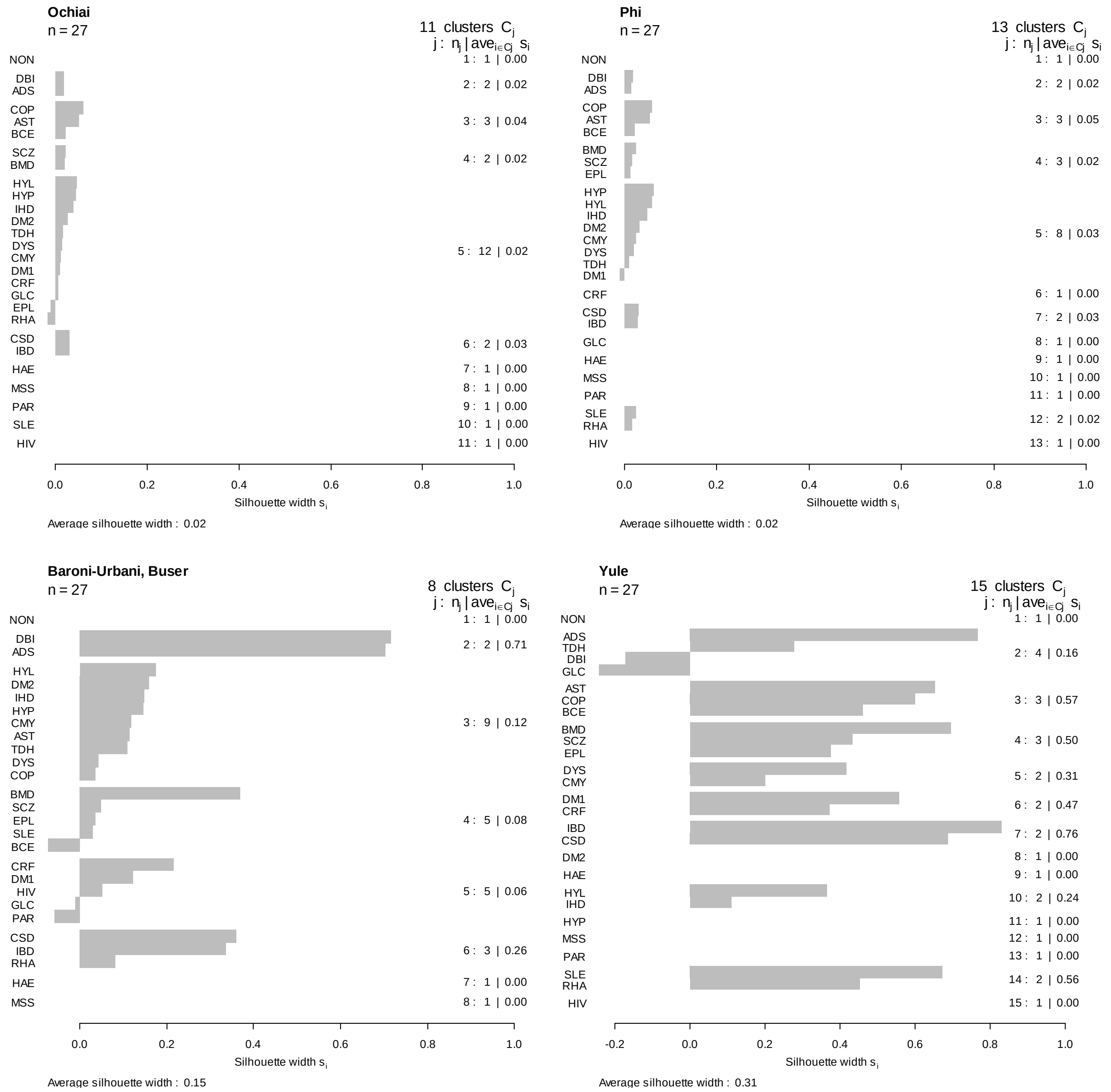


Figure 5.11: *Silhouette plots of PAM using different similarity coefficients. Large silhouette width values (values close to 1) indicate that a certain chronic disease is well classified within the corresponding cluster. Silhouette width values close to −1 indicate that the chronic disease is badly classified within the corresponding cluster. The average silhouette width gives a measure of the quality of the clustering structure. The Yule silhouette plot shows weak clustering The other silhouette plots show no substantial clustering structure. Medoids are shown as the first chronic disease in each cluster.*

The chronic diseases are represented by points in a two-dimensional display with the clusters in Figures 5.12 and 5.13. This two-dimensional space is found using classical scaling, which was discussed in Chapter 4. The points representing the various chronic diseases will have the same configurations as those displayed in Section 4.3.1.1., up to a reflection of the points in the origin. The clusplots also show what percentage of point variability is explained by the two-dimensional display. Notice that most of the clusplots explain a very small proportion of the point variability. This is a major problem as it makes interpretation of these clusplots more difficult, because there are many over-crowded areas in these clusplots. A similar situation was noticed in Section 4.3.1.1. Enlargements of these over-crowded areas were provided in Section 4.3.1.1 and can be used to identify chronic diseases in these over-crowded areas.

It is not appropriate to use dissimilarity coefficients without metric properties when classical scaling is used, because classical scaling treats the dissimilarities directly as distances. The clusplots based on dissimilarity coefficients with metric properties will have a direct distance interpretation. This means that chronic diseases that are close to each other, are more similar. The clusplots based on dissimilarity coefficients without metric properties should not be interpreted in this way. The *Yule* and *Kulczynski 2* clusplots are therefore not appropriate for use, because these dissimilarity coefficients do not have metric properties. This is unfortunate because the *Yule* clusplot explains the highest proportion of the point variability. Notice that the *Sokal-Sneath-Anderberg, Dice-Sorensen* and *Jaccard* similarity coefficients show almost identical clusplots.

The clusplots could be used to describe the relation between the chronic diseases, but interpretation should be done with reservation. Firstly, most clusplots explain a very low proportion of variability. Secondly, clusplots based on dissimilarity coefficients without metric properties are not appropriate to use. Fortunately, the silhouette plots displayed in Figures 5.10 and 5.11 can also be used to describe the clustering structure of the chronic diseases.

The first body system rule involves the respiratory diseases COP, AST and BCE. Most of the similarity coefficients used with PAM show that these chronic diseases are clustered together. Some of the silhouette plots do not show that this cluster is very well classified, but the *Yule* silhouette plot shows strong classification. A few silhouette plots show that COP and AST are located in the same cluster, but BCE tends to form another separate cluster.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. Approximately half of the silhouette plots show that these diseases are located in the same cluster. These cardiovascular diseases also seem to cluster with chronic diseases HYL, DM2 and TDH. The other half of the silhouette plots tend to cluster DYS and CMY together, and HYP and IHD together.

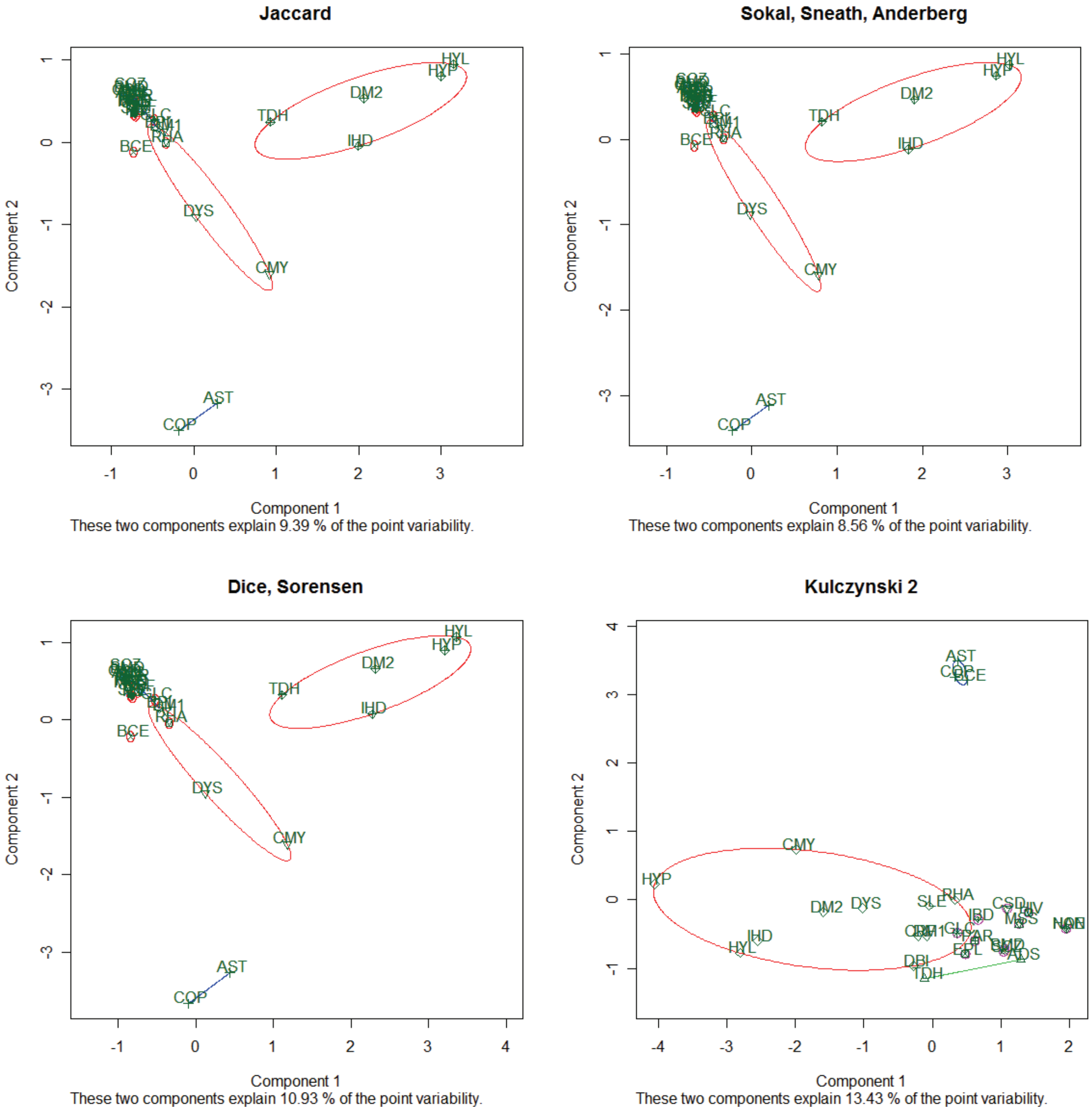


Figure 5.12: *Clusplots of PAM using different similarity coefficients. The Kulczynski 2 dissimilarity coefficient does not have metric properties. It is therefore not really appropriate to use the Kulczynksi 2 clusplot. The two-dimensional space explains very low point variability. Interpretation should be made with reservation.*

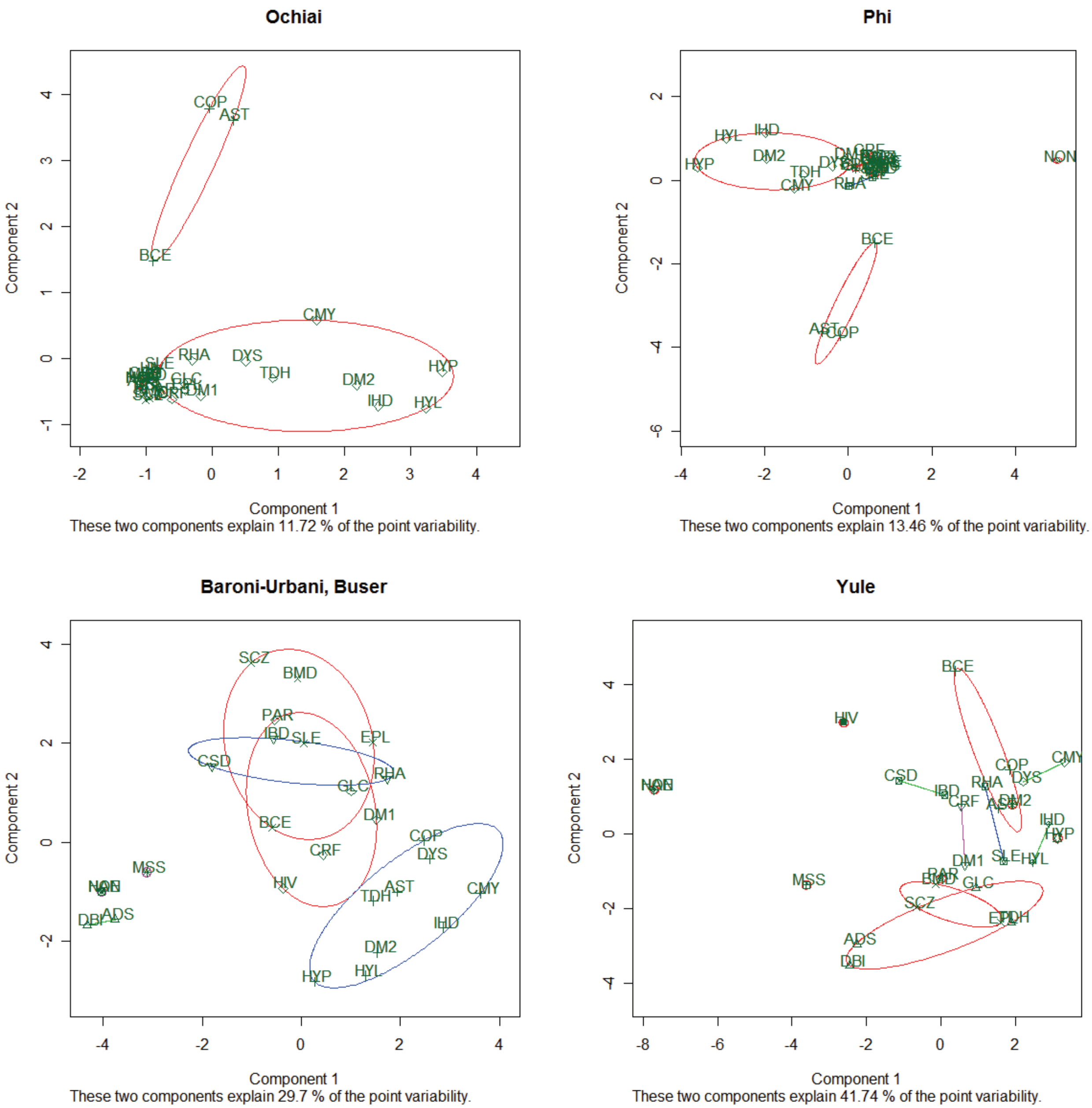


Figure 5.13: *Clusplots of PAM using different similarity coefficients. The Yule dissimilarity coefficient does not have metric or Euclidean properties. Interpretation should be made with reservation.*

The third body system rule involves HYP and CRF. It seems that these chronic diseases are not strongly related. The chronic disease CRF tends to cluster with DM1, while HYP tends to form a cluster with HYL and other cardiovascular related diseases.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the similarity coefficients used, except the *Kulczynski 2* similarity coefficient, indicate that CSD and IBD are strongly related. The *Yule* silhouette plot shows that these two chronic diseases are very well classified, because the silhouette values are large.

The fifth body system rule involves the chronic diseases BMD and SCZ. It seems that these chronic diseases are strongly related. It also seems that these two chronic diseases tend to form a cluster with EPL. The *Yule* silhouette plot also shows that these three chronic diseases are very well classified.

The sixth body system rule involves MSS, BMD and EPL. There is evidence that indicate that BMD and EPL tend to form a cluster, along with SCZ. However, MSS tends to form a cluster on its own, as can be seen from most of the silhouette plots.

The seventh body system rule involves the chronic diseases SLE and RHA. The *Yule* silhouette plot also shows that these chronic diseases are very well classified. Many other similarity silhouette plots also cluster these chronic diseases together.

The chronic diseases ADS and DBI seem to be clustered together. However, these chronic diseases are not mentioned in any of the body system rules.

The results produced in this section are very comparable to the results found when AGNES and DIANA were applied to the Medical Scheme 55-59 data set. Most of the results from Chapter 4 are also comparable to the results produced by PAM.

5.4.5 Application of FANNY

This partitioning algorithm provides fuzzy clustering. The PAM method is a crisp clustering method, because each observation can only belong to a single cluster and cannot belong to multiple clusters simultaneously. Fuzzy clustering gives each observation fractional membership in multiple clusters. The main advantage of fuzzy clustering over hard clustering is that it provides more detailed information of the clustering structure. However, more information is also much more difficult to interpret. Other disadvantages of FANNY are that it does not assign representative objects for each cluster and takes considerable computation time. This fuzzy clustering method can be implemented by the *R* function *fanny*, found in the package: *cluster*. FANNY can take a dissimilarity matrix as input. Many different similarity coefficients were used to construct dissimilarity matrices, and the similarities were transformed to dissimilarities by the same transformation function:

$$d_{ij} = (1-s_{ij})^{0.5}.$$

This transformation leads to more dissimilarity matrices with metric properties. The nearest crisp clustering will be considered for graphical output in the form of silhouette plots and clusplots.

The user must also specify the number of clusters *k*, similar to PAM. The user should run FANNY with different values of *k* and the overall average silhouette width value should be noted for each run. The most appropriate *k* value is the one that provides the highest overall average silhouette width. The largest overall average silhouette width for each similarity coefficient is also calculated in Table 5.2, with the corresponding value of *k* that should be used. It can be seen that most similarity coefficients have different optimal values for *k*. The largest overall average silhouette width gives an indication of the quality of the clustering structure.

Table 5.2: *Largest overall average silhouette width calculated for the different similarity coefficients using FANNY. The values of the number of clusters that lead to the largest value are also given. This value corresponds to the value of k that must be used when FANNY is implemented.*

Similarity coefficients	**Largest overall average silhouette width**	**Corresponding number of clusters (*k*)**
Jaccard	0.0053	3
Dice, Sorensen	0.0115	5
Sokal, Sneath, Anderberg	0.0011	3
Kulczynski 2	0.0130	2
Ochiai	0.0157	5
Phi	0.0101	5
Baroni-Urbani, Buser	0.0792	9
Yule	0.3200	9

The *Yule* similarity coefficient is the only coefficient that has an overall average silhouette width value above 0.25, which indicates that weak clustering structure was found. The other similarity coefficients have no substantial clustering structure. This does not mean that the *Yule* similarity coefficient is the best coefficient to use, as it still needs to be appropriate in the actual data context. It was previously mentioned that most authors recommend the *Jaccard* similarity coefficient when the data consists of asymmetric binary variables, as is the case with the Medical Scheme 55-59 data set. The *Jaccard* similarity coefficient shows a very poor clustering structure, but it is still an appropriate coefficient to use.

Notice that the optimal number of clusters that should be used in FANNY is far lower than the number of clusters used by PAM, displayed in Table 5.1. It must be mentioned though that the FANNY algorithm did not allow the user to choose cluster sizes above 12. Kaufman & Rousseeuw (1990, p.167) point out that the number of clusters chosen must be lower than *n*/2, because of numerical reasons and for a less extensive table of membership coefficients.

Table 5.3: *Dunn's partition coefficient, calculated for the different similarity coefficients using FANNY. The normalised version is also displayed and this should be used for comparing results. Values close to zero indicate fuzzy clusterings and values close to 1 indicate a crisp clustering structure.*

Similarity coefficients	***Dunn's partition coefficient***	***Dunn's normalised partition coefficient***
Jaccard	0.3333	0.0000
Dice, Sorensen	0.2000	0.0000
Sokal, Sneath, Anderberg	0.3333	0.0000
Kulczynski 2	0.5000	0.0000
Ochiai	0.2000	0.0000
Phi	0.2000	0.0000
Baroni-Urbani, Buser	0.1721	0.0686
Yule	0.3668	0.2878

Dunn's partition coefficients for different similarity coefficients are displayed in Table 5.3. This coefficient gives an idea of how fuzzy the resulting clustering is and has values in the range [1/*k*, 1]. A Value close to 1/*k* indicates very fuzzy clustering while a value close to 1 indicates crisp clustering. The *normalised* version is also given and always has a value in the range [0, 1], whatever the value of *k*.

Most of the similarity coefficients show very fuzzy clustering structures. This means no substantial clustering structure was found. The *Yule* similarity coefficient shows some clustering structure, but it is not very strong.

Graphical representations of the resulting clustering structures will again be constructed for the different similarity coefficients. Clusplots and silhouette plots will be constructed to display the clusters. These clustering structures can then be compared to the body system rules.

Silhouette plots, based on different similarity coefficients, are presented in Figures 5.14 and 5.15. The entire silhouette plot shows the silhouettes of all the clusters next to each other. The quality of the clusters can then be compared. The overall average silhouette width of the silhouette plot is the average of the *s(i)* values of all the objects in the data set. The *Yule* similarity coefficient has the clearest clustering structure, but the clustering structure is still weakly defined.

Clusplots of clustering structures produced by FANNY based on different similarity coefficients are displayed in Figures 5.16 and 5.17. The configurations of points are identical (up to a reflection of the points in the origin) to the configurations produced by PAM and the classical scaling MDS method. However, the ellipses produced by FANNY are different to the ellipses produced by PAM. This means that these two methods produced different clustering structures. Again, most of these clusplots are not really suitable for describing how the different chronic diseases are related. Firstly, most clusplots explain a very low proportion of variability. Secondly, the clusplots based on dissimilarity coefficients without metric properties are not really appropriate to use because classical scaling treats dissimilarities directly as distances.

The first body system rule involves the respiratory diseases COP, AST and BCE. Most of the similarity coefficients used with FANNY show that these chronic diseases are clustered together. Some of the silhouette plots do not show that this cluster is very well classified, but the *Yule* silhouette plot shows that these three chronic diseases are very well classified.

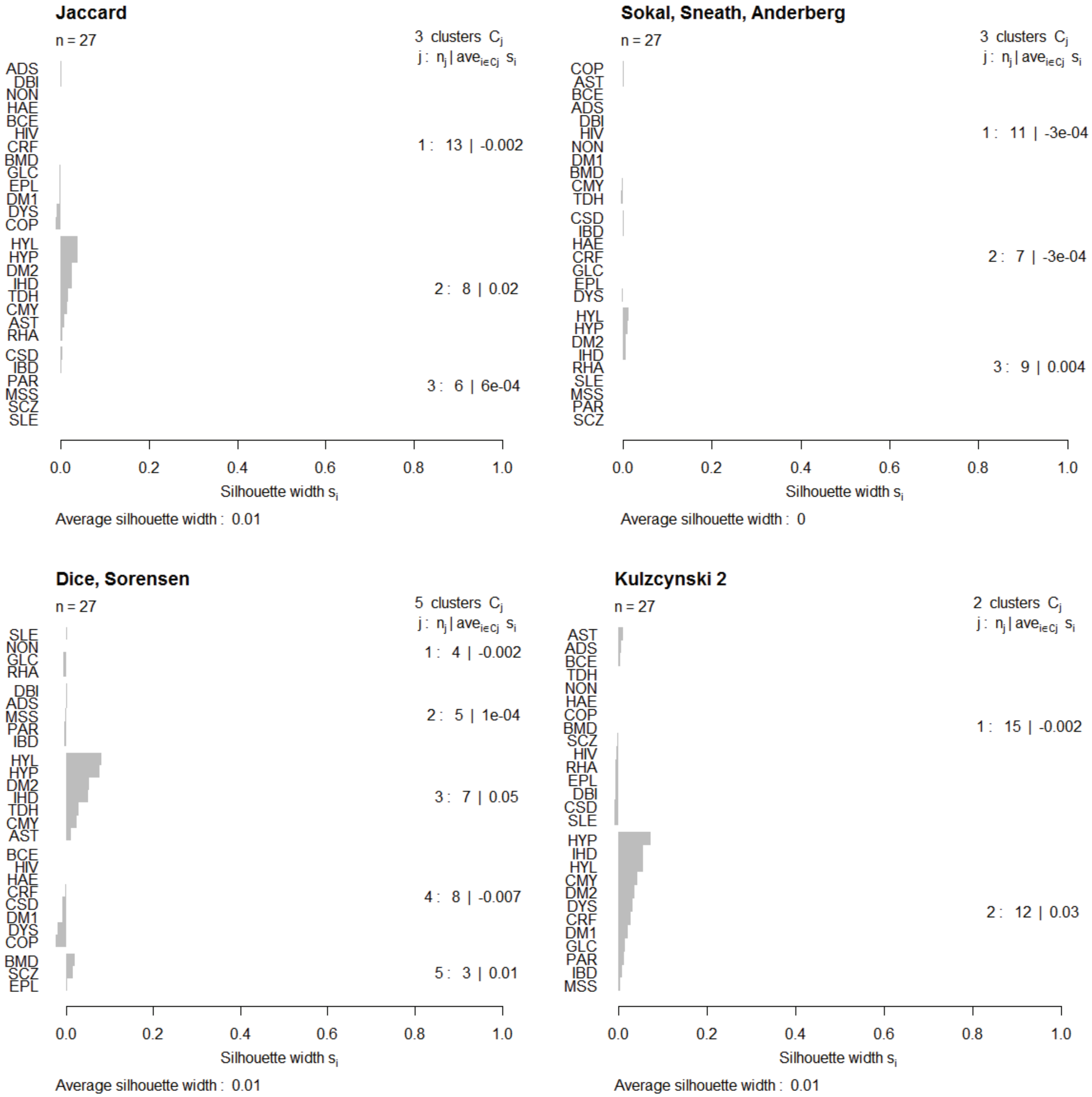


Figure 5.14: *Silhouette plots of the crisp clusterings produced by FANNY using different similarity coefficients. Large silhouette width values (values close to 1) indicate that a certain chronic disease is well classified within the corresponding cluster. Silhouette width values close to −1 indicate that the chronic disease is badly classified within the corresponding cluster. The average silhouette width gives a measure of the quality of the clustering structure.*

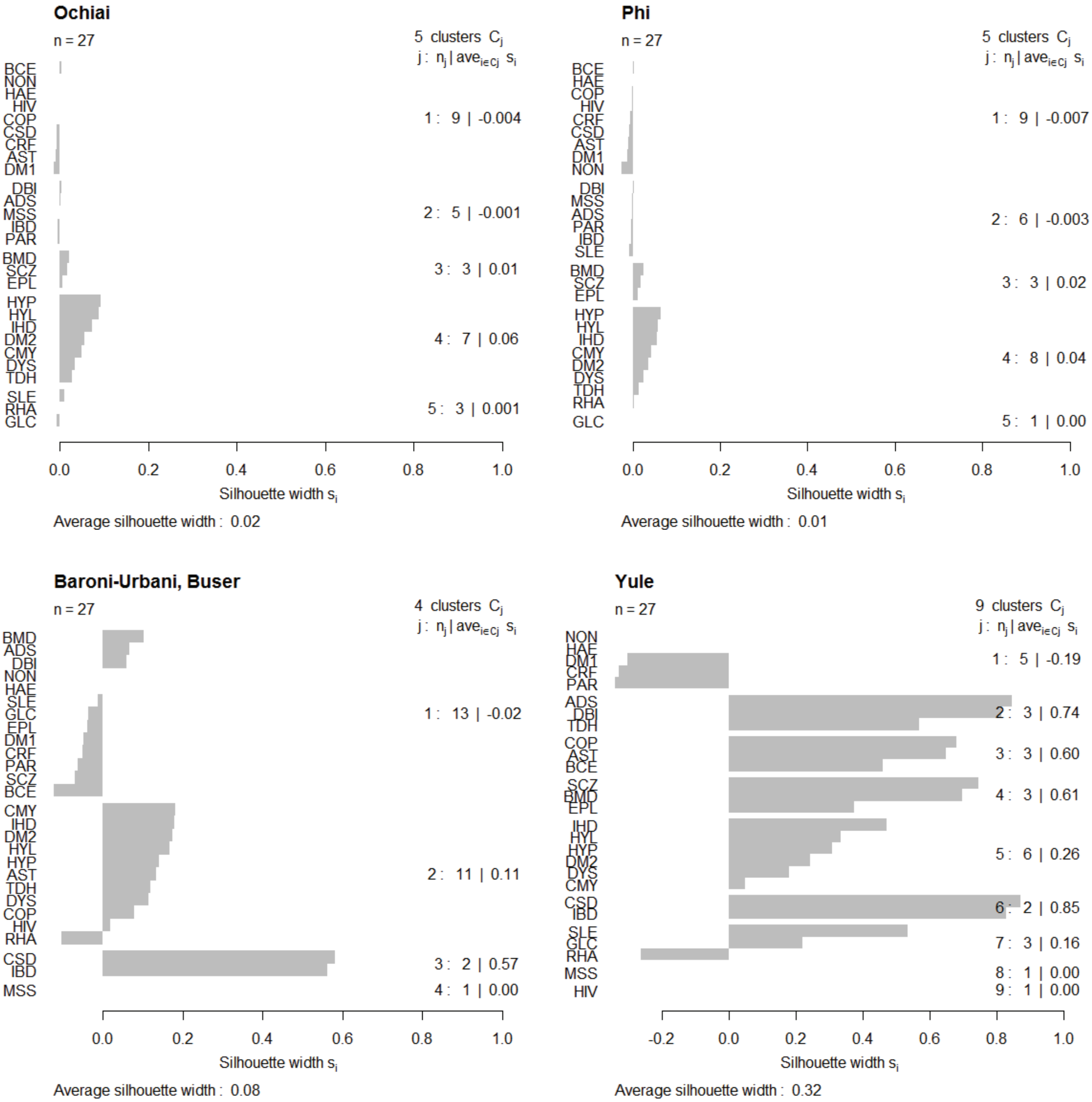


Figure 5.15: *Silhouette plots of the crisp clusterings produced by FANNY using different similarity coefficients. The average silhouette width gives a measure of the quality of the clustering structure. The Yule silhouette plot produces the clearest clustering structure.*

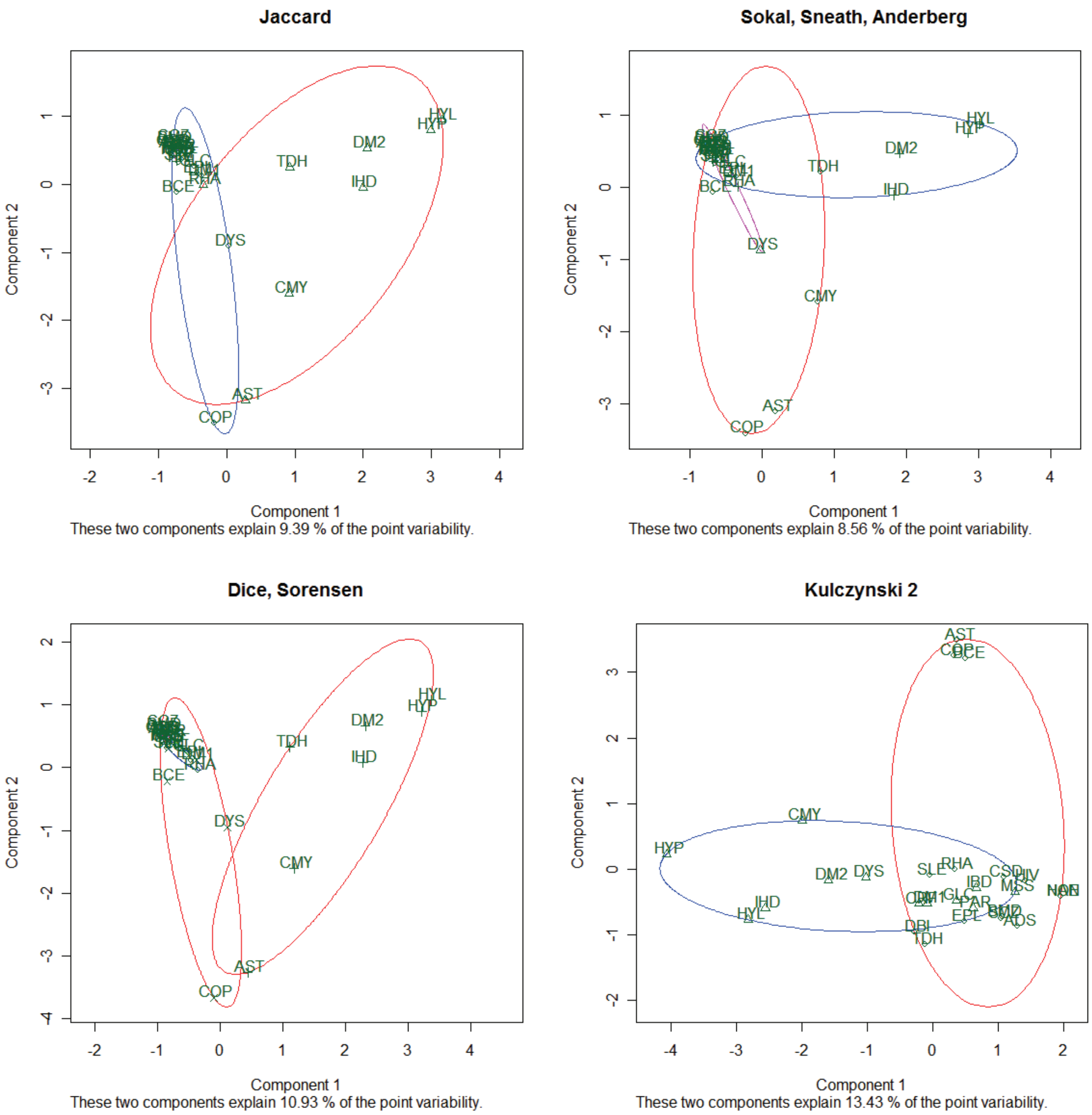


Figure 5.16: *Clusplots of the crisp clustering structure produced by FANNY using different similarity coefficients. The Kulczynski 2 clusplot is not really appropriate to use because the Kulczynski 2 dissimilarity coefficient does not have metric properties. The two-dimensional space explains very low point variability. Interpretation should be made with reservation.*

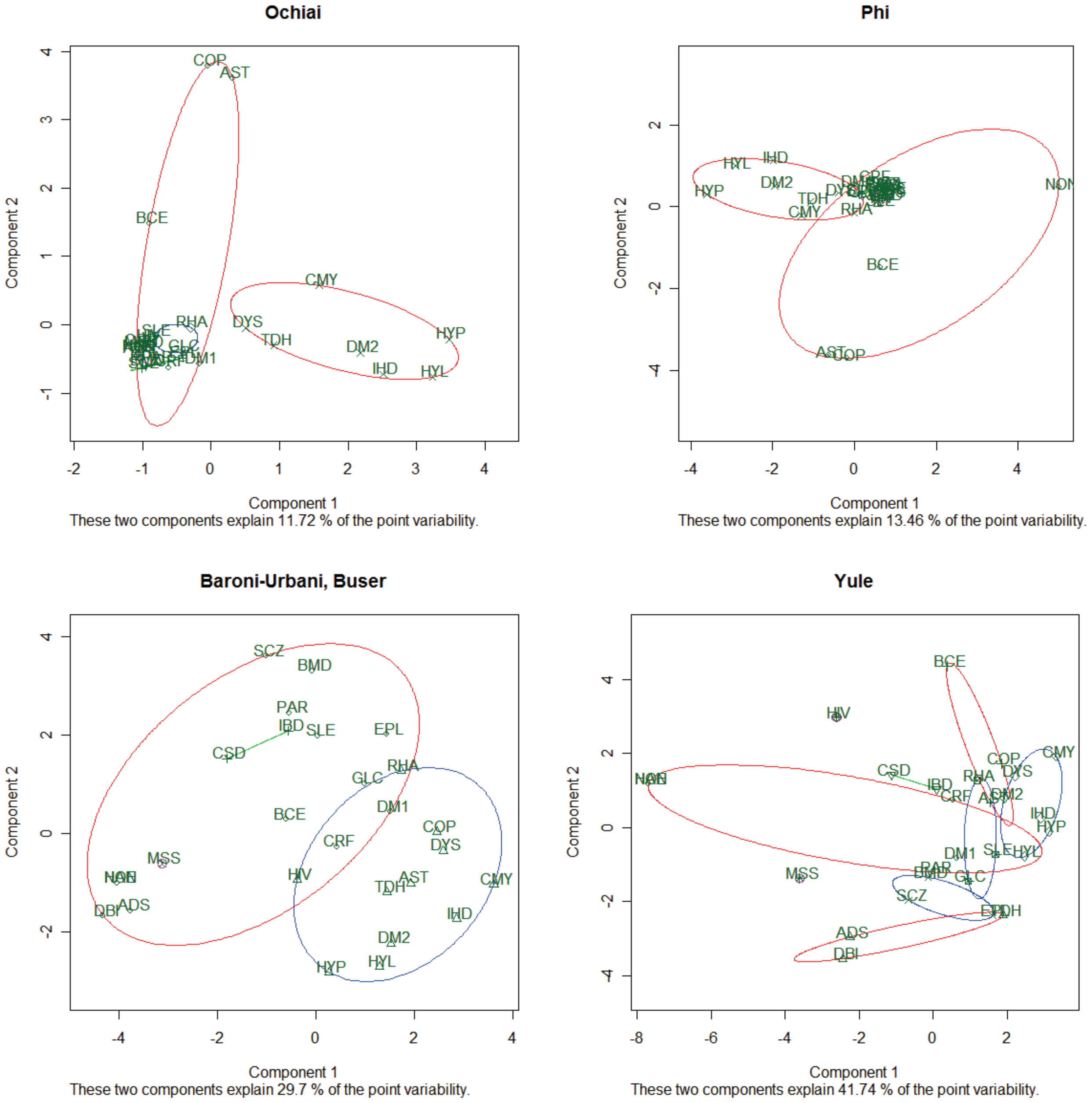


Figure 5.17: *Clusplots of FANNY using different similarity coefficients. The Yule dissimilarity coefficient does not have metric properties. Interpretation should be made with reservation.*

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. Many of the silhouette plots show that these diseases are located in the same cluster. These cardiovascular diseases also seem to cluster with chronic diseases HYL, DM2 and TDH. The *Yule* silhouette plot, which has the clearest clustering structure, shows that CMY, IHD, DYS and HYP are clustered together with HYL and DM2. These chronic diseases seem to be well classified in this cluster.

The third body system rule involves HYP and CRF. These chronic diseases do not seem to be strongly related. The chronic disease CRF tends to cluster with DM1, while HYP tends to form a cluster with HYL and other cardiovascular related diseases.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. Most of the similarity coefficients used show that CSD and IBD are strongly related. The *Yule* silhouette plot shows that these two chronic diseases are very well classified, because the silhouette values are large.

The fifth body system rule involves the chronic diseases BMD and SCZ. Again, it seems that these chronic diseases are strongly related. It also seems that these two chronic diseases tend to form a cluster with EPL. The *Yule* silhouette plot also shows that these three chronic diseases are very well classified, because the silhouette values are large.

The sixth body system rule involves MSS, BMD and EPL. The graphical representations indicate that BMD and EPL tend to form clusters, along with SCZ. However, MSS tends not to form a cluster with BMD and EPL.

The seventh body system rule involves the chronic diseases SLE and RHA. These chronic diseases seem to be strongly related, as many silhouette plots show that these two diseases are classified into the same cluster.

The last system body rule involves DM1 and DM2. The silhouette plots show that DM1 and DM2 do not form small clusters together, but are sometimes found together in a larger cluster. DM1 tends to form a cluster with CRF, and DM2 tends to form a cluster with HYP, HYL and IHD.

Most of these results are very comparable to the results produced by PAM, AGNES and DIANA.

5.5 Application of the Clustering Techniques on the Male MS (55-59) and Female MS (55-59) Data Sets

The various clustering techniques were applied to the Medical Scheme 55-59 data set in Section 5.4. The Medical Scheme 55-59 data set however does contain male and female lives that might display different clustering structures. It is important to investigate if the separate Male MS (55-59) and Female MS (55-59) data sets show similar clustering structures. The clustering structures will also be compared to the body system rules. The algorithms PAM, FANNY, AGNES and DIANA will again be used. The methods CLARA and MONA cannot be used, as they do not take dissimilarity matrices as input.

It is not practical to use all the similarity coefficients for the Male MS (55-59) and Female MS (55-59) data sets again. Many authors recommend using the *Jaccard* coefficient with asymmetric binary variables (Sibson *et al.*, 1981; Kaufman & Rousseeuw, 1990). The *Yule* similarity coefficient produced the clearest clustering structure in the previous sections. It is for these reasons that these two similarity coefficients will be applied to the Male MS (55-59) and Female MS (55-59) data sets.

It was mentioned in Chapter 2 that none of the male lives in the 55 to 59 year age band were treated for the chronic disease HAE. The dissimilarity matrices of the Male MS (55-59) data set will therefore not contain HAE and HAE will therefore not be present in any of the clustering displays of the male lives. The Female MS (55-59) data set showed that 4 lives were treated for HAE and HAE will therefore be displayed in the clustering displays of the female lives.

5.5.1 Application of AGNES

The similarities were transformed to dissimilarities by the function $d_{ij} = 1 - s_{ij}$. The average linkage method was used to define the between-cluster dissimilarities because Kaufman & Rousseeuw (1990) believed it was the most appropriate method to use. Clustering trees are constructed in Figure 5.18 using the two similarity coefficients. These graphical representations can be used to describe the clusters that seem to form based on the Male MS (55-59) and Female MS (55-59) data sets.

There are some differences in the clustering trees of the Male MS (55-59) and Female MS (55-59) data sets, but the differences do not appear to be significant.

The first body system rule involves the respiratory diseases COP, AST and BCE. The clustering of these diseases is almost identical for the Male MS (55-59) and Female MS (55-59) data sets. The *Yule* clustering trees show that these diseases form a cluster. The clustering trees based on the *Jaccard* coefficient however show that COP and AST form a cluster, and BCE forms a cluster with HIV.

The second body system rule involves the cardiovascular related diseases CMY, IHD, DYS and HYP. It seems that these chronic diseases are strongly related. The clusterings of these chronic diseases formed by the Male MS (55-59) and Female MS (55-59) data sets are very similar. It also seems that the cardiovascular and respiratory diseases are located in the same larger cluster and chronic diseases HYL, DM2 and TDH also form part of this large cluster.

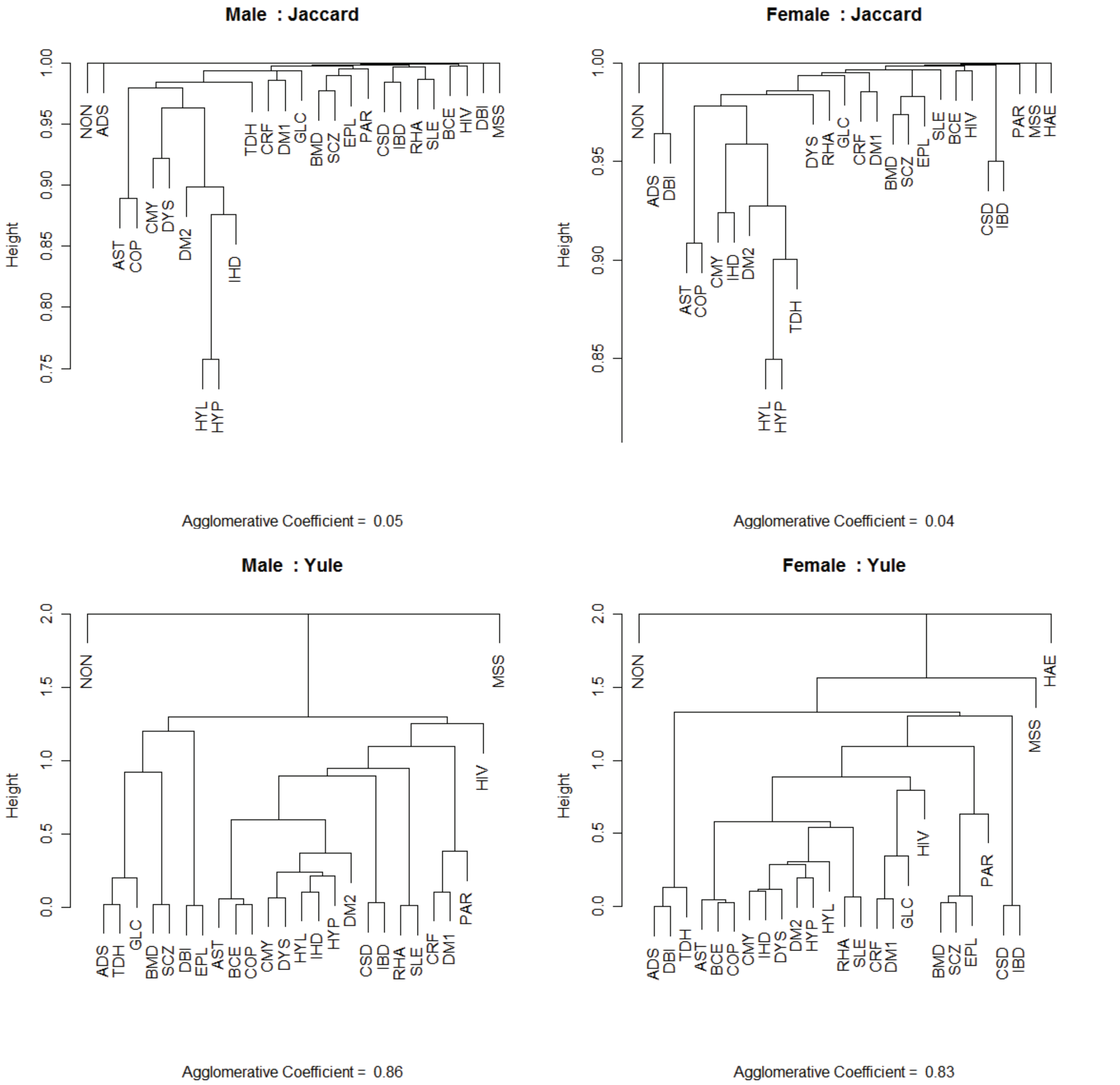


Figure 5.18: *Clustering trees produced by AGNES for the Male MS (55-59) and Female MS (55-59) data sets. The Yule clustering trees have the clearest clustering structure.*

The third body system rule involves HYP and CRF. Most of the clusterings show that these two chronic diseases are not strongly related. The chronic disease CRF tends to cluster with DM1, and not HYP. This is evident in both the Male MS (55-59) and Female MS (55-59) data sets.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the clustering trees show that CSD and IBD are located in the same cluster and are strongly related.

The fifth body system rule involves the chronic diseases BMD and SCZ. These two chronic diseases seem to be strongly related, for both male and female lives. The sixth body system rule involves MSS, BMD and EPL. There is evidence that indicates that BMD and EPL form clusters, along with SCZ. However, MSS tends to form a cluster on its own.

The seventh body system rule involves the chronic diseases SLE and RHA. It seems that these chronic diseases are strongly related. However, the *Jaccard* clustering tree for male lives does not cluster these diseases together. It clusters RHA together with cardiovascular and respiratory diseases, while SLE is clustered with BMD, SCZ and EPL. Overall, it seems that this rule is supported by the Male MS (55-59) and Female MS (55-59) data sets.

5.5.2 Application of DIANA

The similarities were transformed to dissimilarities by the same function $d_{ij} = 1 - s_{ij}$. The average linkage method was again used to define the between-cluster dissimilarities. The DIANA method is very similar to AGNES in several aspects including the output. DIANA may differ from AGNES because the divisive algorithm is not the exact counterpart of the agglomerative algorithm.

Clustering trees are constructed in Figure 5.19 using the two similarity coefficients. These graphical representations can be used to describe the clusters that seem to form based on the Male MS (55-59) and Female MS (55-59) data sets. There are some differences in the clustering trees of the Male MS (55-59) and Female MS (55-59) data sets, but the differences are not great. These differences however are more apparent than the differences that were observed when AGNES was used.

The first body system rule involves the respiratory diseases COP, AST and BCE. The clustering of these diseases is almost identical for the Male MS (55-59) and Female MS (55-59) data sets. The clustering trees based on the *Jaccard* coefficient however show that COP and AST form a cluster, and BCE forms a single cluster. The *Yule* clustering trees show a difference between male and female lives. COP and AST form a cluster in the Female MS (55-59) data, but COP and BCE form a cluster in the Male MS (55-59) data.

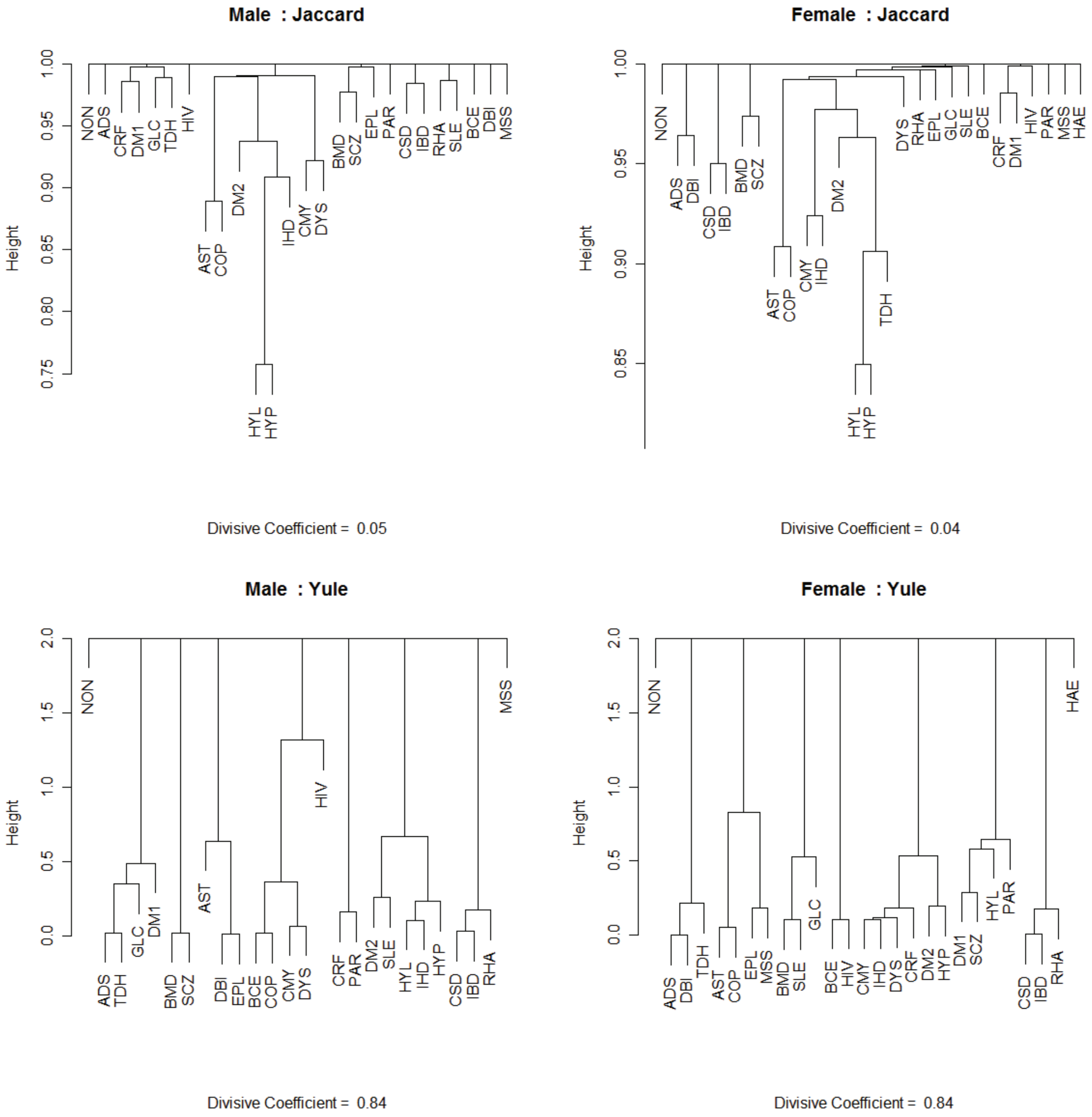


Figure 5.19: *Clustering trees produced by DIANA for the Male MS (55-59) and Female MS (55-59) data sets. The Yule clustering trees have the clearest clustering structure.*

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. It seems that these chronic diseases are strongly related. The *Yule* clustering tree for male lives is the only one that does not show that these chronic diseases fall into the same cluster. The clustering of these diseases formed by the Male MS (55-59) and Female MS (55-59) data sets is very similar.

The third body system rule involves HYP and CRF. The *Jaccard* clustering trees show that chronic disease CRF clusters with DM1, and not HYP. The *Yule* clustering tree for female lives shows that CRF and HYP are broadly clustered together.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. All of the clustering trees show that these chronic diseases are strongly related.

The fifth body system rule involves the chronic diseases BMD and SCZ. The clustering trees for the Male MS (55-59) and Female MS (55-59) data sets seem to indicate that BMD and SCZ are indeed strongly related. The *Yule* clustering tree for female lives however shows that BMD and SLE, and not SCZ, are clustered together.

The sixth body system rule involves MSS, BMD and EPL. Only the *Jaccard* clustering tree for male lives shows that BMD and EPL are in the same cluster. The *Yule* clustering tree for female lives shows that MSS and EPL are in the same cluster. The Male MS (55-59) and Female MS (55-59) data sets do not seem to show that these chronic diseases are strongly related.

The seventh body system rule involves the chronic diseases SLE and RHA. The *Jaccard* clustering tree for male lives is the only clustering tree that clusters these diseases together.

5.5.3 Application of PAM

The similarities were transformed to dissimilarities by the transformation function

$$d_{ij} = (1-s_{ij})^{0.5}.$$

The number of clusters k, must first be chosen. PAM should be run with different values of k and the overall average silhouette width value should be noted for each run. The most appropriate k value is the one that provides the highest overall average silhouette width. The overall average silhouette width gives an indication of the clustering structure found in the data.

The largest overall average silhouette width is calculated in Table 5.4, with the corresponding value of k that should be used. This is done for the Male MS (55-59) and Female MS (55-59) data sets respectively. The optimal number of clusters to be used with PAM is slightly different for the Male MS (55-59) and Female MS (55-59) data sets. The largest overall average silhouette width values for the Male MS (55-59) and Female MS (55-59) data sets are very similar. The Yule similarity coefficients have the clearest clustering structure.

Table 5.4: *Largest overall average silhouette width calculated for the two similarity coefficients. This is done for the Male MS (55-59) and Female MS (55-59) data sets. The value of the number of clusters that lead to the largest value is also given. This value corresponds to the value of k that must be used when PAM is implemented.*

Similarity coefficients	**Largest overall average silhouette width**	**Corresponding number of clusters (*k*)**
Male: Jaccard	0.0116	15
Female: Jaccard	0.0113	16
Male: Yule	0.3049	14
Female: Yule	0.3178	13

Silhouette plots are displayed in Figure 5.20. These graphical representations can be used to compare male and female clustering structures. They will also be used to check if the data supports the body system rules.

The first body system rule involves the respiratory diseases COP, AST and BCE. The silhouette plot of these diseases is almost identical for the Male MS (55-59) and Female MS (55-59) data sets. The silhouette plots based on the *Jaccard* coefficient however show that COP and AST form a cluster, and BCE forms a single cluster. The *Yule* silhouette plots show a difference between male and female lives. COP and AST form a cluster in the Male MS (55-59) data, but COP, AST and BCE form a cluster in the Female MS (55-59) data. Overall, it seems that these chronic diseases are strongly related.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. The chronic diseases do not all occur in the same cluster, for any of the silhouette plots. Pairs of these chronic diseases tend to occur in the same clusters. The male lives show that CMY and DYS fall in the same cluster and IHD and HYP tend to fall in another separate cluster. Female lives show that IHD and DYS fall in the same cluster.

The third body system rule involves HYP and CRF. The *Yule* silhouette plots show that the chronic disease CRF clusters with DM1, and not HYP. This classification is also very strong.

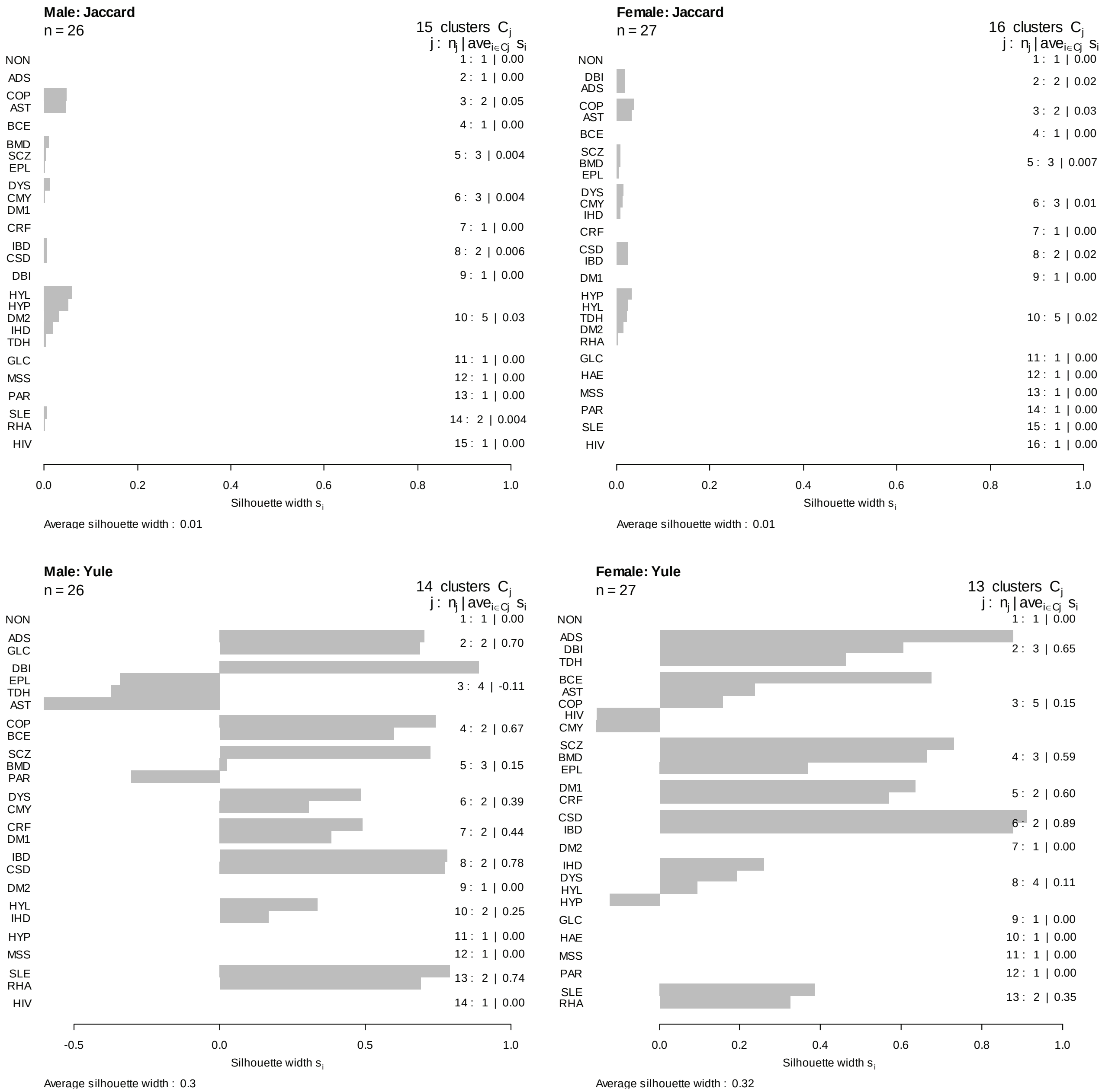


Figure 5.20: *Silhouette plots produced by PAM for the Male MS (55-59) and Female MS (55-59) data sets. Large silhouette width values (values close to 1) indicate that a certain chronic disease is well classified within the corresponding cluster. Silhouette width values close to −1 indicate that the chronic disease is badly classified within the corresponding cluster. The average silhouette width gives a measure of the quality of the clustering structure. Medoids are shown as the first chronic disease in each cluster.*

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. The *Yule* silhouette plots show that these two diseases are very well classified in the same cluster.

The fifth body system rule involves the chronic diseases BMD and SCZ. It seems that these chronic diseases are strongly related for both the male and female lives. The *Yule* silhouette plot for female lives shows that BMD and SCZ are well classified in the same cluster with EPL.

The sixth body system rule involves MSS, BMD and EPL. There is strong evidence to support the notion that BMD and EPL are related, as they tend to occur in the same cluster. However, MSS tends to form a single cluster.

The seventh body system rule involves the chronic diseases SLE and RHA. The *Yule* silhouette plots show that these two diseases are well classified in the same cluster. It therefore seems that these chronic diseases are strongly related.

It is interesting to note that CRF and DM1 usually form a cluster, and the *Yule* silhouette plots show that these two diseases are well classified in this cluster.

5.5.4 Application of FANNY

This partitioning algorithm provides fuzzy clustering. The PAM method is a crisp clustering method because each observation can only belong to a single cluster and cannot belong to multiple clusters simultaneously. Fuzzy clustering gives each observation fractional membership in multiple clusters. Similarities were transformed into dissimilarities by the same transformation function:

$$d_{ij} = (1 - s_{ij})^{0.5}.$$

The nearest crisp clustering will be considered for graphical output in the form of silhouette plots.

The number of clusters *k*, must first be chosen. FANNY should be run with different values of *k* and the overall average silhouette width value should be noted for each run. The most appropriate *k* value is the one that provides the highest overall average silhouette width. The largest overall average silhouette width is calculated in Table 5.5, with the corresponding value of *k* that should be used. This is done for the Male MS (55-59) and Female MS (55-59) data sets.

The optimal number of clusters to be used with FANNY is slightly different for the Male MS (55-59) and Female MS (55-59) data sets. The largest overall average silhouette width values for the Male MS (55-59) and Female MS (55-59) data sets are very similar. The Yule similarity coefficient gives the clearest clustering structure. Silhouette plots are displayed in Figure 5.21.

Table 5.5: *Largest overall average silhouette width calculated for the two similarity coefficients. This is done for the Male MS (55-59) and Female MS (55-59) data sets. The values of the number of clusters that lead to the largest value are also given. This value corresponds to the value of k that must be used when FANNY is implemented.*

Similarity coefficients	**Largest overall average silhouette width**	**Corresponding number of clusters (*k*)**
Male: Jaccard	0.0063	4
Female: Jaccard	0.0037	3
Male: Yule	0.3411	9
Female: Yule	0.3330	9

The first body system rule involves the respiratory diseases COP, AST and BCE. It seems that these chronic diseases are strongly related. All of the silhouette plots of the Male MS (55-59) and Female MS (55-59) data sets, displayed in Figure 5.21, show that these chronic diseases are classified into the same cluster. The *Yule* silhouette plot for female lives shows that COP, AST and BCE are very well classified in the same cluster.

The second body system rule involves cardiovascular related diseases CMY, IHD, DYS and HYP. It seems that these chronic diseases are related. The *Yule* silhouette plot for female lives shows that CMY, IHD, DYS, HYP, HYL and DM1 are reasonably well classified in the same cluster. The other silhouette plots do not show that CMY, IHD, DYS and HYP fall into a single cluster, but in two separate clusters.

The third body system rule involves HYP and CRF. It seems that these two chronic diseases are weakly related. The silhouette plots based on the Male MS (55-59) and Female MS (55-59) data sets show that these chronic diseases do not occur in the same cluster.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. Most of the silhouette plots show that these chronic diseases are strongly related, only the *Jaccard* silhouette plot for male lives does not classify CSD and IBD into the same cluster. The *Yule* silhouette plots shows that these two diseases are very well classified in the same cluster with RHA.

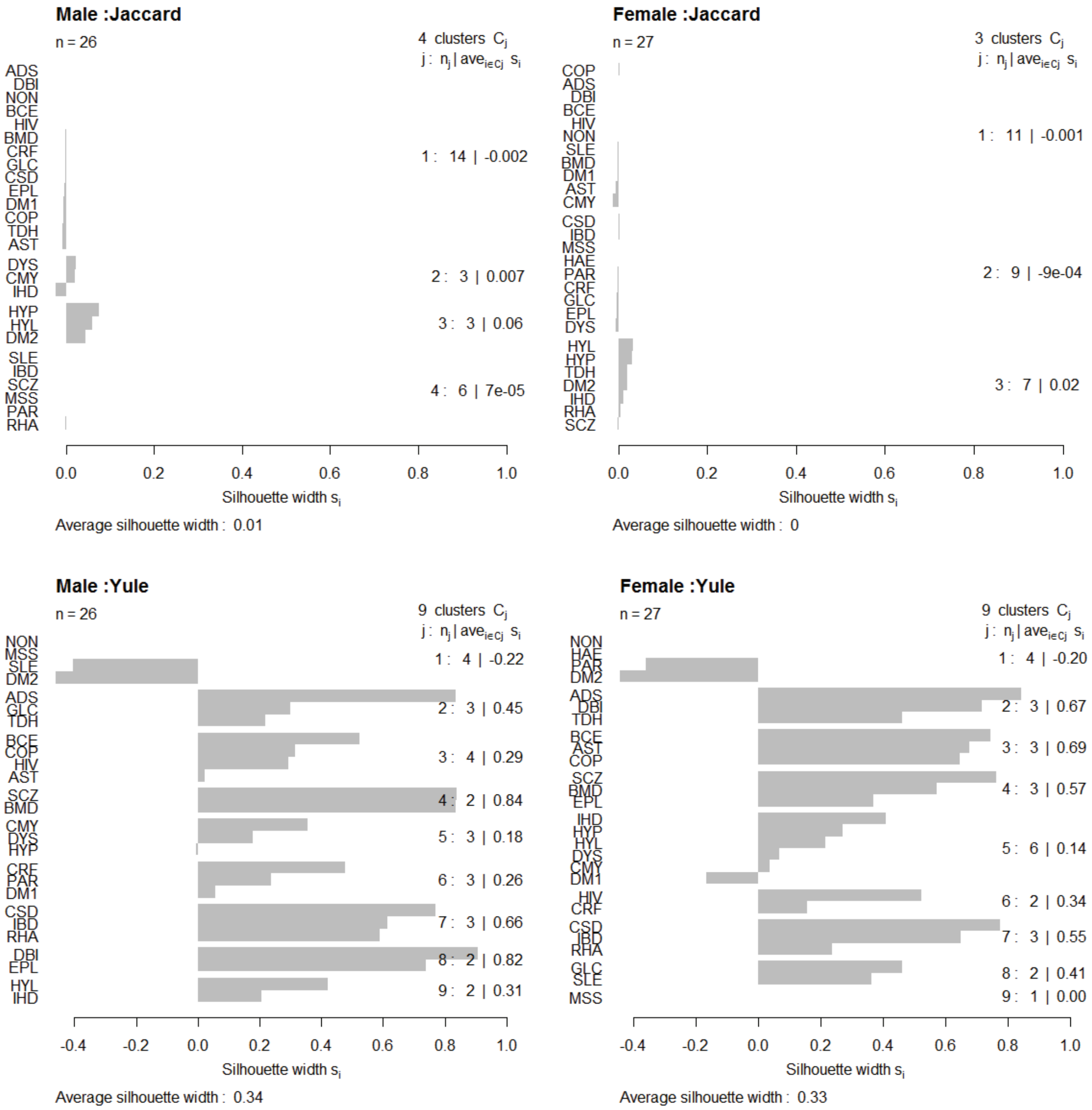


Figure 5.21: *Silhouette plots produced by FANNY for the Male MS (55-59) and Female MS (55-59) data sets. Large silhouette width values (values close to 1) indicate that a certain chronic disease is well classified within the corresponding cluster. Silhouette width values close to −1 indicate that the chronic disease is badly classified within the corresponding cluster. The average silhouette width gives a measure of the quality of the clustering structure. Medoids are shown as the first chronic disease in each cluster.*

The fifth body system rule involves the chronic diseases BMD and SCZ. The *Yule* silhouette plot for female lives shows that BMD and SCZ are well classified in the same cluster with EPL. The *Yule* silhouette plot for male lives shows that BMD and SCZ are also well classified in the same cluster. However, the *Jaccard* silhouette plots do not classify BMD and SCZ into the same cluster. It must be noted though that the silhouette plots based on the *Jaccard* coefficient have a very weak clustering structure.

The sixth body system rule involves MSS, BMD and EPL. Is does not seem that these chronic diseases are strongly related. None of the silhouette plots in Figure 5.21 shows that these three chronic diseases are located in the same cluster. However, the *Yule* silhouette plot for female lives shows that BMD and EPL are well classified in the same cluster.

The seventh body system rule involves the chronic diseases SLE and RHA. Only the *Jaccard* silhouette plot for male lives shows that these two diseases are classified in the same cluster. It therefore seems that these two chronic diseases are only weakly related, based on Figure 5.21.

5.6 Summary

The clustering algorithms of Kaufman & Rousseeuw (1990) were implemented to display the clustering structures of the various chronic diseases, and to investigate which of the chronic diseases mentioned in the same body system rule are related. This was done for the Medical Scheme 55-59 data set and for the Male MS (55-59) and Female MS (55-59) data sets. Note that only lives between the ages 55 to 59 were considered in this cluster analysis.

The clustering of the chronic diseases was not a straightforward task. The Medical Scheme 55-59 data set could not be directly used as an input structure for the different clustering algorithms. The reason is that each row of the Medical Scheme 55-59 data set does not represent a single life, but represents multiple lives that have the same unique combination of chronic diseases. It will therefore be wrong to use the Medical Scheme 55-59 data set directly as input structure. One possibility to overcome this problem is to expand the multiple lines into individual lines and then proceed accordingly, but this also leads to problems. The expanded matrix will then contain 2 199 367 rows, which is impractical to use. Furthermore, it must be remembered that the Medical Scheme 55-59 data set consists of binary data, which means that only clustering algorithms suited for binary data and clustering algorithms that accept a dissimilarity matrix **D** as input structure, can be used on the Medical Scheme 55-59 data set. This meant that the model-based clustering algorithms, the *K-means* partitioning method and CLARA could not be used to cluster the chronic diseases. The clustering method MONA does not take a dissimilarity matrix as input structure, but it can be used on binary data. The only practical requirement for the use of MONA on the Medical Scheme 55-59 data set is that the multiple lines must be expanded into individual lines and then this expanded data matrix must be used as input structure. However, the expanded matrix will then contain 2 199 367 rows, which is impractical to use. Another problem is that the association measure used in MONA assumes that the data set consists of symmetric binary variables, but the Medical Scheme 55-59 data set consists of asymmetric binary variables and it is therefore not really appropriate to use MONA. However, four of the clustering algorithms of Kaufman & Rousseeuw (1990): PAM, FANNY, AGNES and DIANA were applied to the Medical Scheme 55-59 data set. The dissimilarity matrices, calculated in Chapter 3, were used as input structures for these clustering algorithms.

The partitioning methods, PAM and FANNY, only performed reasonably well for the Medical Scheme 55-59 data set. The clusplots produced by PAM and FANNY were not suitable for interpretation, because of certain over-crowded areas of points. The reason for this is that PAM and FANNY use classical scaling to find a configuration of points, and it was seen in Chapter 4 that the classical scaling method was not really suitable for the Medical Scheme 55-59 data set. Another problem with the clusplots is that only dissimilarity coefficients with metric properties should be considered, as classical scaling is used. Furthermore, the PAM and FANNY methods require the user to specify the number of clusters. The number of clusters for the Medical Scheme 55-59 data set was unknown and hence a subjective choice had to be made. Kaufman & Rousseeuw (1990) recommend choosing a certain number of clusters that produce the clearest clustering structure, but other choices are also possible. It seems however that the hierarchical algorithms, AGNES and DIANA, performed very well on the Medical Scheme 55-59 data set. These two methods do not require the user to specify the number of clusters and dissimilarity coefficients without metric properties can also be considered.

Several different dissimilarity coefficients were used in this chapter, hoping for some robustness against a specific choice. Dissimilarity coefficients that were usually used with symmetric binary variables, like the *Rao-Russell* and the *Simple matching* coefficients, did not perform well for this age band and are not recommended for the other age bands. The *Jaccard* dissimilarity coefficient is the most used coefficient when binary variables are asymmetric, and it did perform reasonably well. The clustering structures based on the *Yule* dissimilarity coefficient were clearer than the clustering structures based on the other dissimilarity coefficients. The *Jaccard, Kulczynski 1, Dice-Sorensen* and *Sokal-Sneath-Anderberg* dissimilarity coefficients produced very similar clustering structures.

Most of the clustering structures produced in this chapter were not clear. It must be remembered though that the clarity depends on the range of dissimilarity values produced by the various dissimilarity coefficients, which depends on the actual Medical Scheme 55-59 data set. The clustering structures based on the Medical Scheme 55-59 data set were compared to the various body system rules. Most of the chronic diseases mentioned in the same body system rule seem to be related and are supported by the clustering structures found in the Medical Scheme 55-59 data set. The chronic diseases mentioned in the second (CMY, IHD, DYS and HYP), fourth (CSD and IBD) and the fifth (BMD and SCZ) body system rules were strongly related and are supported by most clustering algorithms and most similarity coefficients. The chronic diseases mentioned in the first (COP, AST and BCE), sixth (MSS, BMD and EPL) and seventh (SLE and RHA) body system rules seem to be reasonably related, but the relation is not very strong. It was found in most clustering structures that MSS tends to form its own cluster. The clustering structures showed that DM1 and DM2 do not form a small cluster, but are sometimes found together in a larger cluster. The third body system rule involves HYP and CRF. Very few clustering structures did in fact show that these two diseases were classified in the same cluster.

It was mentioned in Chapter 2 that the histograms in Figure 2.4 of MSS, PAR, HIV and HAE show that these diseases show very little co-occurrence with other diseases. These observations are validated by the clustering structures. Each of the chronic diseases HAE, MSS and PAR tend to form their own separate cluster.

The clustering structures based on the Male MS (55-59) and Female MS (55-59) data sets were compared to the various body system rules. The clustering structures based on male and female lives were very similar, and were similar to the clustering structures based on the Medical Scheme 55-59 data set.

It was interesting to note that cardiovascular and respiratory chronic diseases also tend to appear together in a larger cluster. The chronic diseases HYL, DM2 and TDH also tend to form part of this larger cluster. The chronic diseases ADS and DBI tend to form a separate cluster in most of the clustering structures, even though they are not mentioned in any of the body system rules.

Most of the conclusions presented in this chapter agree with the conclusions reached in Chapter 4, even though the analyses used in the two chapters are different. This inspires more confidence in the accuracy of the results.

Chapter 6

Conclusions

The Medical Scheme 55-59 data set was described in this study using univariate and multivariate statistical techniques. It is very important to note that the results only apply to the age band of 55 to 59 years, and should not be generalised for all ages. However, it will still be appropriate to use the same methodology for the other age bands in the future.

It was seen that approximately 72% of lives between ages 55 and 59 were not treated for any chronic disease or HIV and approximately 17% of lives were treated only for HIV or only for one chronic disease. This means that approximately 11% of the lives were treated for multiple chronic diseases.

Histograms and tables were used to describe the chronic disease rates and average costs. It was seen that the chronic diseases HAE, DBI, ADS, BCE and MSS displayed the lowest chronic disease rates. On the other hand, the chronic diseases CMY, IHD, AST, TDH, DM2, HYL and HYP occurred much more often. Chronic disease rates were also used to make a direct comparison between male and female lives, with regard to the occurrence of chronic diseases. It was found that female lives displayed higher chronic disease rates for DBI, MSS, BCE, SLE, SCZ, BMD, RHA and TDH relative to the male lives. The male lives however displayed higher chronic disease rates for CRF, PAR, HIV, IHD and HYL. The average costs related to lives treated for a single chronic disease and lives not treated for any chronic disease were also described. It was found that the chronic diseases CRF, MSS and DM1 had very high average costs for both male and female lives. The male and female lives displayed very low average costs for HYP, HYL and TDH. Lives not treated for any chronic disease or HIV also displayed very low average costs. Unfortunately, the average costs cannot be allocated correctly between chronic diseases when the chronic diseases co-occur, which meant that all information regarding average costs of lives treated for multiple chronic diseases had to be discarded, which may limit the relevance of the average cost results.

The chronic disease rates provide information of the occurrence, but not the co-occurrence, of the various chronic diseases. Histograms were used to display the breakdown of the number of lives treated for 1, 2, 3,…, 8 chronic diseases, where the breakdown was done for each chronic disease. This provided information about the co-occurrences of certain diseases. The chronic diseases CMY, IHD, DM2, HYL, DYS, BCE, SLE, ADS and DBI tend to co-occur often with other chronic diseases. The chronic diseases MSS, PAR, HIV and HAE however showed much less co-occurrence with other chronic diseases. The usefulness of this information however was limited, as it could not be used to describe the relationships among the various chronic diseases. For example, the histograms showed that the chronic disease HYL co-occurred often, but the histograms did not show with which chronic diseases HYL co-occurred. It was therefore necessary to use MDS methods and clustering techniques to describe the relationships among the chronic diseases.

Various MDS methods and various clustering techniques were applied to the Medical Scheme 55-59 data set. The different analyses however produced reasonably similar conclusions, which may inspire more confidence in the accuracy of the results. The results were compared to the various body system rules.

The first body system rule involves the respiratory diseases COP, AST and BCE. It was found that COP and AST are strongly related with each other, but only reasonably related with BCE. This was the case for most of the MDS methods and clustering techniques that were used in this study. It seems therefore that COP and AST co-occur more often.

The second body system rule involves the cardiovascular related diseases CMY, IHD, DYS and HYP. It seems that these chronic diseases are strongly related, based on the results produced by the MDS methods and clustering techniques. It was also found in many cases that HYP and HYL were very strongly related.

The third body system rule involves HYP and CRF. It seems that HYP and CRF are only weakly related. This was the case for almost all of the results produced by the MDS methods and clustering techniques.

The fourth body system rule involves the gastro-intestinal conditions CSD and IBD. It seems that these two chronic diseases are very strongly related. It was only occasionally found that these two diseases were not classified in the same cluster.

The fifth body system rule involves the chronic diseases BMD and SCZ. It was found that BMD and SCZ are strongly related with each other. This was the case for most of the MDS methods and clustering techniques used in this study.

The sixth body system rule involves MSS, BMD and EPL. It seems that BMD and EPL are reasonably related to each other, but it does not seem that MSS is related to either BMD or EPL. Many of the clustering techniques showed that MSS tends to form its own separate cluster, and most of the MDS methods also showed that MSS was not located close to other diseases.

The seventh body system rule involves the chronic diseases SLE and RHA. The MDS methods and clustering techniques showed slightly different results for these two chronic diseases. The MDS methods suggested that there might be some weak relation between SLE and RHA. The clustering techniques tended to show that SLE and RHA are reasonably related, as these diseases sometimes occurred in the same cluster.

The last body system rule, involving the Diabetes Mellitus diseases DM1 and DM2, was the only rule that was actually applied to the Medical Scheme 55-59 data set. This means that the Medical Scheme 55-59 data set showed no co-occurrence between these two diseases. It is therefore expected that these two diseases should not be located very close to each other in the MDS plots or classified in the same small cluster. It might however be possible for these two diseases to occur together in a larger cluster with other chronic diseases, because of their co-occurrence with other related diseases. The clustering structures did indeed mostly show that DM1 and DM2 do not form a small cluster, but are sometimes found together in a larger cluster. It seems though that DM1 and CRF are strongly related.

It was interesting to note that some other chronic diseases not mentioned in the same body system rule also seem to be strongly related. The cardiovascular chronic diseases HYP, DYS, IHD and CMY seem to have a strong relation with chronic diseases AST, COP, DM2, TDH and HYL. The chronic diseases BMD, EPL, SCZ and PAR also seem to be related, especially BMD and SCZ. The chronic diseases ADS and DBI also showed a strong relation, even though they are not mentioned in any of the body system rules.

The previous results regarding the body system rules are based on the Medical Scheme 55-59 data set. However, various MDS methods and various clustering techniques were also applied to the Male MS (55-59) and the Female MS (55-59) data sets, and these results were also compared to the various body system rules. The MDS plots and clustering structures based on male and female lives were similar, and the results produced were also similar to those based on the Medical Scheme 55-

59 data set. The relationships among chronic diseases mentioned in the same body system rule were also very similar to those described previously for the Medical Scheme 55-59 data set.

Various dissimilarity coefficients, based on binary data, were used in the MDS methods and clustering algorithms as input structures. It was more appropriate to consider various dissimilarity coefficients, and not just one, as different dissimilarity coefficients produced different results. However, it was found that some of the dissimilarity coefficients were more appropriate for use than others. It was seen that the *Rao-Russell* and the *Simple matching* dissimilarity coefficients are unsuitable for use on the Medical Scheme 55-59 data set. The MDS plots based on these dissimilarity coefficients showed that chronic diseases with similar chronic disease rates were located close to each other. This is undesirable, as chronic diseases should be located close to each other when they tend to co-occur often, and not because the chronic diseases have similar chronic disease rates. The Medical Scheme 55-59 data set consists of asymmetric binary variables, but the *Rao-Russell* and the *Simple matching* dissimilarity coefficients are usually used when the data consists of symmetric binary variables. Therefore, only dissimilarity coefficients that are based on asymmetric binary variables should be used with the Medical Scheme 55-59 data set. The other dissimilarity coefficients used in this study produced much more suitable results. It was found that the *Jaccard, Dice-Sorensen, Kulczynski 1* and *Sokal-Sneath-Anderberg* dissimilarity coefficients produced very similar results. It may therefore be more practical to only use the *Jaccard* dissimilarity coefficient, and not all four, in the future. Many authors recommend using the *Jaccard* dissimilarity coefficient when the data consist of asymmetric binary variables, which is the case with the Medical Scheme 55-59 data set. The *Ochiai*, *Baroni-Urbani-Buser*, *Phi* and *Yule* dissimilarity coefficients also seem to work well generally. The *Yule, Kulczynski 2* and *Baroni-Urbani-Buser* dissimilarity coefficients produced the clearest clustering structures, but these coefficients do not have metric properties. This means that the *Yule; Kulczynski 2* and *Baroni-Urbani-Buser* dissimilarity coefficients cannot be used for all MDS methods and clustering techniques, only some. The *Jaccard, Ochiai* and *Phi* dissimilarity coefficients have metric and Euclidean properties when the transformation $d_{ij} = (1-s_{ij})^{0.5}$ is used. Unfortunately, these three dissimilarity coefficients tend to produce clustering structures that are not very clear.

The classical scaling, metric least squares scaling and nonmetric MDS methods were used in this study, but it seems that the nonmetric MDS method was most useful. It seems that the nonmetric MDS method should be preferred over the classical scaling method for the Medical Scheme 55-59 data set, because the classical scaling plots explained a very small proportion of the point variability. This caused an over-crowding of points in the classical scaling plots which made interpretation very difficult. Nonmetric MDS also has the advantage of making minimal assumptions about how distances and dissimilarities are related. The metric MDS methods should only consider dissimilarity coefficients with metric or Euclidean properties, because it is not appropriate to use dissimilarities directly as distances when the dissimilarities do not have metric properties. Nonmetric MDS does not treat dissimilarities as distances, and can therefore also use dissimilarities without metric properties. The nonmetric MDS method was also less subjective than the metric least squares scaling method as no different weights had to be used to produce reasonably clear clustering structures. However, the nonmetric MDS method takes more computation time than metric least squares scaling as optimal disparities and distances had to be computed in the SMACOF algorithm. The classical scaling method is least computer intensive. The SMACOF algorithm performed very well in this study, and is also recommended for future use. It is also recommended that random initial configurations should be used for the nonmetric MDS minimising algorithms, as this will improve the likelihood that a global minimum, and not just a local minimum, will be found. The disadvantage of this approach however is the increased computation time. Fortunately, the computation time was not a major concern with the Medical Scheme 55-59 data set as only 27 objects are displayed in the plots.

The Medical Scheme 55-59 data set could not be directly used as an input structure for the different clustering algorithms. The reason is that each row of the Medical Scheme 55-59 data set does not represent a single life, but represents multiple lives that have the same unique combination of chronic diseases. It will therefore be wrong to use the Medical Scheme 55-59 data set directly as input structure. One possibility to overcome this problem is to expand the multiple lines into individual lines and then proceed accordingly, but this also leads to problems. The expanded matrix will then contain 2 199 367 rows, which is impractical for use. Furthermore, it must be remembered that the Medical Scheme 55-59 data set consists of binary data, which means that only clustering algorithms that accept binary data and clustering algorithms that accept a dissimilarity matrix **D** as input structure, can be used on the Medical Scheme 55-59 data set. This meant that the model-based clustering algorithms, the *K-means* partitioning method and CLARA could not be used to cluster the chronic diseases. The clustering method MONA does not take a dissimilarity matrix as input structure, but it can be used on binary data. The only practical requirement for the use MONA on the Medical Scheme 55-59 data set is that the multiple lines must be expanded into individual lines and then to use this expanded data matrix as input structure. However, the expanded matrix will then contain 2 199 367 rows, which is impractical for use. Another problem is that the association measure used in MONA assumes that the data set consists of symmetric binary variables, but the Medical Scheme 55-59 data set consists of asymmetric binary variables and it is therefore not really appropriate to use MONA. However, four of the clustering algorithms of Kaufman & Rousseeuw (1990): PAM, FANNY, AGNES and DIANA were applied to the Medical Scheme 55-59 data set. Dissimilarity matrices were used as input structures for these clustering algorithms.

The partitioning methods, PAM and FANNY, performed reasonably well for the Medical Scheme 55-59 data set while the hierarchical algorithms, AGNES and DIANA, performed very well. The clusplots produced by PAM and FANNY were not really suitable for interpretation, because of certain over-crowded areas of points. The reason for this is that PAM and FANNY use classical scaling to find a configuration of points. Another problem with the clusplots is that only dissimilarity coefficients with metric properties should be considered, as classical scaling is used. Further, the PAM and FANNY methods require the user to specify the number of clusters. The number of clusters for the Medical Scheme 55-59 data set was unknown and hence a subjective choice had to be made. However, the hierarchical algorithms, AGNES and DIANA, performed very well. The user does not have to specify the number of clusters to be used and dissimilarity coefficients without metric properties can also be considered.

The results produced in this study only apply to the age band of 55 to 59 years, and should not be generalised for all ages. However, it will still be appropriate to use the same methodology for the other age bands in the future. It may also be more appropriate and practical to use only some of the MDS methods and clustering techniques that performed well on the Medical Scheme 55-59 data set. It seems that the nonmetric MDS method, AGNES and DIANA should be the preferred choices. It was also found that the *Jaccard*, *Ochiai*, *Baroni-Urbani-Buser*, *Phi* and *Yule* dissimilarity coefficients were most suitable for use on the Medical Scheme 55-59 data set. These dissimilarity coefficients may therefore be the preferred choices should similar analyses of the other age bands be carried out in the future.

Appendix A

The *R* functions that were constructed for use in this study are provided in this Appendix. The statistical computer package *R* is available from `http://www.cran.r-project.org.`

Dissim.CDL

This function is used to construct a dissimilarity matrices based on different choices of similarity coefficients, represented in Table 3.2.

<u>Arguments of the function</u>:

x : Data matrix where the last column represent the number of lives that are diagnosed with a certain combination of chronic diseases. The other previous columns represent the chronic diseases which are coded binary variables where "1" indicates the presence of a certain chronic disease and "0" indicate its absence.

transformation : Choice of transformation. Select transformation = 1 to use transformation d(i,j)=1-s(i,j) and select transformation = 2 to use the transformation d(i,j)=(1-s(i,j))$^{0.5}$.

```
function (x=final.dat[c(1:886),c(1:28)],transformation=1)
{
#This function constructs a dissimilarity matrix of the chronic
diseases (variables)
#based on different choices of similarity coefficients, represented
in Table 3.2
#x represent the data matrix where the last column represent the
number of lives
#diagnosed with a certain combination of chronic diseases.
#The other previous columns represent the chronic diseases which
are binary variables
#The binary variables must be coded as follows
# "1" indicate the presence of a certain chronic disease and "0"
indicate absence
#transformation=1 means the transformation d(i,j)=1-s(i,j) is used
#transformation=2 means the transformation d(i,j)=(1-s(i,j))^0.5 is
used
p<-ncol(x)
tot.obj<-sum(x[,p])
tot.var<-p-1
```

```
n<-nrow(x)
check<-apply(x,2,max)
if(max(check[1:tot.var])>1)stop("The datamatrix x is in the wrong
format")
if(max(check[p])<1)stop("The datamatrix x is in the wrong format")
if(transformation>2)stop("Value for 'transformation' is wrong")
#######Chronic diseases that do not occur in x can not be displayed
as it will
##cause some similarity coefficients to be NaN. For example
denomirator=0
ch.no.occur<-c(1:tot.var)[check==0]
if(length(ch.no.occur)>0.5){
x<-x[1:n,-ch.no.occur]
p<-ncol(x)
tot.var<-p-1
}
else{}
#The a, b, c, d values as indicated in Table 3.1 must first be
determined for all
#pairs of chronic diseases.

A.mat<-matrix(0,nrow=tot.var,ncol=tot.var)
rownames(A.mat)<-colnames(x[,1:(p-1)])
colnames(A.mat)<-colnames(x[,1:(p-1)])

B.mat<-matrix(0,nrow=tot.var,ncol=tot.var)
rownames(B.mat)<-colnames(x[,1:(p-1)])
colnames(B.mat)<-colnames(x[,1:(p-1)])

C.mat<-matrix(0,nrow=tot.var,ncol=tot.var)
rownames(C.mat)<-colnames(x[,1:(p-1)])
colnames(C.mat)<-colnames(x[,1:(p-1)])

D.mat<-matrix(0,nrow=tot.var,ncol=tot.var)
rownames(D.mat)<-colnames(x[,1:(p-1)])
colnames(D.mat)<-colnames(x[,1:(p-1)])

for(i in 1:tot.var){
for(j in 1:tot.var){
a<-0
b<-0
c<-0
d<-0
aa<-0
bb<-0
cc<-0
dd<-0
for(k in 1:n){
if((x[k,i]==1)&&(x[k,j]==1)){aa<-x[k,p]}
else{aa<-0}
if((x[k,i]==1)&&(x[k,j]==0)){bb<-x[k,p]}
else{bb<-0}
if((x[k,i]==0)&&(x[k,j]==1)){cc<-x[k,p]}
else{cc<-0}
if((x[k,i]==0)&&(x[k,j]==0)){dd<-x[k,p]}
else{dd<-0}

a<-a+aa
b<-b+bb
c<-c+cc
d<-d+dd
}
if(!((a+b+c+d)==tot.obj)){stop(paste("Total A,B,C,D is
wrong",i,j))}
A.mat[i,j]<-a
B.mat[i,j]<-b
C.mat[i,j]<-c
D.mat[i,j]<-d
}}

#Now to compute dissimilarities from matrices A, B, C and D
#Dissimilarity is transformed immediately. d(i,j)=1-s(i,j)
Jaccard<-1-((A.mat)/(A.mat+B.mat+C.mat))

Dice<-1-(2*A.mat/(2*A.mat+B.mat+C.mat))

Kulczynski1<-1-(A.mat/(B.mat+C.mat))
diag(Kulczynski1)<-0

Ochiai<-1-(A.mat/sqrt((A.mat+B.mat)*(A.mat+C.mat)))

Phi<-1-((A.mat*D.mat-
B.mat*C.mat)/sqrt((A.mat+B.mat)*(A.mat+C.mat)*(B.mat+D.mat)*(C.mat+
D.mat)))

Baron<-1-
((A.mat+sqrt(A.mat*D.mat))/(sqrt(A.mat*D.mat)+A.mat+B.mat+C.mat))

Kulczynski2<-1-0.5*(A.mat/(B.mat+A.mat)+A.mat/(A.mat+C.mat))

Rao<-1-(A.mat)/(tot.obj)

Simple<-1-(A.mat+D.mat)/(tot.obj)
```

```
Yule<-1-((A.mat*D.mat-B.mat*C.mat)/(A.mat*D.mat+B.mat*C.mat))

Sokal<-1-((A.mat)/(A.mat+2*B.mat+2*C.mat))

if(transformation==2){
#Dissimilarity is transformed: d(i,j)=(1-s(i,j))^0.5

Jaccard<-sqrt(Jaccard)
Dice<-sqrt(Dice)
Kulczynski1<-sqrt(Kulczynski1)
Kulczynski2<-sqrt(Kulczynski2)
Phi<-sqrt(Phi)
Simple<-sqrt(Simple)
Ochiai<-sqrt(Ochiai)
Rao<-sqrt(Rao)
Baron<-sqrt(Baron)
Yule<-sqrt(Yule)
Sokal<-sqrt(Sokal)
}
else{}

rownames(Phi)<-colnames(x[,1:(p-1)])
colnames(Phi)<-colnames(x[,1:(p-1)])
rownames(Simple)<-colnames(x[,1:(p-1)])
colnames(Simple)<-colnames(x[,1:(p-1)])
rownames(Rao)<-colnames(x[,1:(p-1)])
colnames(Rao)<-colnames(x[,1:(p-1)])
rownames(Jaccard)<-colnames(x[,1:(p-1)])
colnames(Jaccard)<-colnames(x[,1:(p-1)])
rownames(Baron)<-colnames(x[,1:(p-1)])
colnames(Baron)<-colnames(x[,1:(p-1)])
rownames(Yule)<-colnames(x[,1:(p-1)])
colnames(Yule)<-colnames(x[,1:(p-1)])
rownames(Sokal)<-colnames(x[,1:(p-1)])
colnames(Sokal)<-colnames(x[,1:(p-1)])
rownames(Ochiai)<-colnames(x[,1:(p-1)])
colnames(Ochiai)<-colnames(x[,1:(p-1)])
rownames(Kulczynski1)<-colnames(x[,1:(p-1)])
colnames(Kulczynski1)<-colnames(x[,1:(p-1)])
rownames(Kulczynski2)<-colnames(x[,1:(p-1)])
colnames(Kulczynski2)<-colnames(x[,1:(p-1)])
rownames(Dice)<-colnames(x[,1:(p-1)])
colnames(Dice)<-colnames(x[,1:(p-1)])
return(list(A.mat=A.mat,B.mat=B.mat,C.mat=C.mat,D.mat=D.mat,
Phi=Phi,Simple=Simple,Rao=Rao,Jaccard=Jaccard,Baron=Baron,Yule=Yule
,Sokal=Sokal,Ochiai=Ochiai,Kulczynski1=Kulczynski1,Kulczynski2=Kulc
zynski2,Dice=Dice))}
```

classical.scaling

This function is used to display the classical scaling configuration in two dimensions. The display will ensure that a unit length in the horizontal direction is the same as a unit length in the vertical direction.

Arguments of the function:

D : Dissimilarity matrix produced by function *Dissim.CDL*.

rotation : Choice to rotate the display 180 degrees. Choose rotation=-1 to rotate the plot.

type : Type of plot. Only use type=”p”.

print : Indicates whether numerical output should be produced.

plot : Indicates whether the display should be plotted.

pch : Type of symbol used to display samples. Enter *?par* for more details.

xlab : Title of x-axis label.

ylab : Title of y-axis label.

main : Title of the plot.

sub : Sub print of plot. Enter *?par* for more details.

col : Colour of samples. Enter *?par* for more details.

cex: A numerical value giving the amount by which symbols should be scaled. Enter *?par* for more details.

```
function (D,label=TRUE,rotasion=1,
type="p",print=FALSE,pch=20,xlab="X1",ylab="X2",main="Classical
scaling (Jaccard)",sub="",col="black",cex=0.5,plot=TRUE) {
#Implementation of the practical algorithm of Cox and Cox (1994)
p.38
#D is a dissimilarity matrix used as input
#This function produces a two dimensional display of objects
contained in dissimilarity matrix D
if(!(is.matrix(D))) stop("D must be a matrix!")
if(!(abs(rotasion)==1))stop("rotasion can only be -1 or 1")
n<-nrow(D)
DD<-D^2
#This is matrix containing squared dissimilarities.
A<-(-0.5*DD)
mat1<-matrix(1,nrow=n,ncol=n)
```

```
H<-(diag(1,n)-((1/n)*mat1))
B<-H%*%A%*%H
#Now to find the coordinates from B
coordinaat.x<-function(B,p,rotasie=-1)
{
V<-as.matrix(rotasie*svd(B)$v)
d<-svd(B)$d
V1<-V[,c(1:p)]
L<-diag(sqrt(d[1:p]))
X<-matrix(0,nrow=length(B),ncol=p)
X<-V1%*%L
Ant<-list(X=X,Eievektore=V,lambda=d)
Ant
}
Koord<-coordinaat.x(B,p=2,rotasie=rotasion)
X<-Koord$X
d<-Koord$lambda
d<-zapsmall(d)
Eigenvectors<-Koord$Eievektore
X<-matrix(X,nrow=n,ncol=2,dimnames=list(1:n,1:2))

dimnames(X)[[1]]<-dimnames(D)[[1]]
Measure<-sum(d[1:2])/sum(d)
par(pty="s")
if(plot){
MDS.plot<-function
(X,type="p",label=TRUE,pch=20,xlab="X1",ylab="X2",
main="",sub="",col="black",cex=1,pos=1)
{
library(MASS)
on.exit(detach(package:MASS))
eqscplot(x=X[,1],y=X[,2],type="n")
usr<-par("usr")
yas.lengte<-sqrt((usr[4]-usr[3])^2)
xas.lengte<-sqrt((usr[2]-usr[1])^2)
eqscplot(x=X[,1],y=X[,2],type=type,
xlim=c(usr[1]-0.08*xas.lengte,usr[2]+0.08*xas.lengte),
ylim=c(usr[3]-0.08*yas.lengte,usr[4]+0.08*yas.lengte),
pch=pch,xlab=xlab,ylab=ylab,main=main,sub=sub)
if(label){
text(x=X[,1],y=X[,2],pos=pos,labels=dimnames(X)[[1]],col=col,
cex=cex,offset=0.2)
}}
MDS.plot(X=X,type=type,label=label,pch=pch,xlab=xlab,ylab=ylab,
main=main,sub=paste("These two dimensions
explain",round(100*Measure,2),
"% of the point variability"),col=col,cex=cex)
cc<-locator(n=2)
x.min<-min(cc$x)
x.max<-max(cc$x)
y.min<-min(cc$y)
y.max<-max(cc$y)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)

X.mag<-cbind(X,X)
for(i in 1:n){
if(X[i,1]<=x.max){X.mag[i,1]<-1}
else{X.mag[i,1]<-0}
if(X[i,1]>=x.min){X.mag[i,2]<-1}
else{X.mag[i,2]<-0}
if(X[i,2]<=y.max){X.mag[i,3]<-1}
else{X.mag[i,3]<-0}
if(X[i,2]>=y.min){X.mag[i,4]<-1}
else{X.mag[i,4]<-0}
}
X.magnified<-X[apply(X.mag,1,sum)==4,]
win.graph()
MDS.plot(X=X.magnified,type=type,label=FALSE,pch=pch,xlab=xlab,ylab
=ylab,
main=main,sub="Enlargement of specific area",col=col,cex=cex)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)
identify(X.magnified,labels=rownames(X.magnified),offset=0.3,cex=ce
x)}
Ant<-
list(Measure.of.variation=Measure,Configuration=X,Eigenvalues=d,
Cumulative.eigenvalues=cumsum(d)/sum(d),Eigenvectors=Eigenvectors)
if(print){Ant}}
```

SMACOF.metric

This function uses the SMACOF algorithm (Borg and Groenen, 2006, p.192) to produce metric MDS two-dimensional display.

Arguments of the function:

delta : Dissimilarity matrix.

X : Initial configuration of points in two dimensions. The initial configuration produced by classical scaling is used as default.

Weightsmethod : The values of weight that should be used. Select Weightsmethod = 1 to use the weights $w_{ij} = 1$ if $i \neq j$ and $w_{ij} = 0$ if $i = j$. Select Weightsmethod = 2 to use the weights $w_{ij} = \delta_{ij}^{q}$ for different values of *q*.

maxit : The maximum number of iterations before the algorithm stops.

epsil : The minimum difference in stress values between iterations reached before algorithm stops.

label : To indicate whether the samples should be labeled.

Main : Title of the plot.

q : The value of weights $w_{ij} = \delta_{ij}^{q}$.

normalise.stress : To indicate if Normalised Stress (TRUE) should be used or Raw Stress (FALSE).

Bubble.plot : To indicate whether a bubble plot (TRUE) should be drawn or a normal MDS (FALSE) display

identify.points : To indicate whether the points should be identified individually or automatically according to the value of *pos*.

Residual.plot : To indicate whether a residual plot should be drawn.

cex: A numerical value giving the amount by which text should be scaled. Enter *?par* for more details.

pch : Type of symbol used to display samples. Enter *?par* for more details.

pos : Position specified for the label of samples. Values of `1`, `2`, `3` and `4`, respectively indicate positions below, to the left of, above and to the right of the sample points.

offset : when `pos` is specified, this value gives the offset of the label from the sample point in fractions of a character width.

cex.points : A numerical value giving the amount by which text should be scaled.

```
function (delta,X=cmdscale(d=delta),Weightsmethod=1,
maxit=300,epsil=0.000001,label=TRUE,Main="Jaccard",
q=0,normalise.stress=TRUE,Bubble.plot=FALSE,identify.points=FALSE,
Residual.plot=TRUE,cex=0.6,pch=20,pos=3,offset=0.1,cex.points=1)
{
#This function uses SMACOF algorithm (Borg and Groenen, 2006,
p.192) to
#produce metric MDS two-dimensional display
```

```
#delta must be a dissimilarity matrix
#X is the initial configuration. The configuration uses classical
scaling
# configuration as default
#Bubble.plot=TRUE means Bubble.plot will be drawn
#Identify.points=TRUE  means the user click on points in plot to be
identified
#Identify.points=FALSE labels are drawn automatically according to
pos

delta<-as.matrix(delta)
n<-nrow((delta))

if(Weightsmethod==1){
W=matrix(1,nrow=nrow(delta),ncol=nrow(delta)) - diag(nrow(delta))
}
else{W<-delta^(q)}

if(min(W)<0)stop("W must only contain non-negative elements!!")
if(sum(delta-t(delta))>0.001) stop("delta must be symmetrical!!")
if(sum(diag(delta))>0.0000000001) stop("delta must have zero’s on
diagonal!!")
if(min(delta)<0) stop("delta must contain non-negative
dissimilarities!!")

#Calculation of V and inverse of V
V<-matrix(0,nrow=n,ncol=n)
V<-(-W)
for(i in 1:n){V[i,i]<-sum(W[i,-i])}
if(round(sum(W-matrix(1,nrow=n,ncol=n)),10)==0){V.inv<-(1/n)}
else{V.inv<-solve((V+matrix(1,nrow=n,ncol=n)))}
#####################################################

matrix.2.vector<-function (mat=matrix(c(1:16),nrow=4,ncol=4))
{
#Coverts lower diagonal values to vector
n<-nrow(mat)
if(!(nrow(mat)==ncol(mat)))stop("mat must be a square matrix")
length<-n*(n-1)/2
vec<-0
for(j in 1:(n-1)){
for(i in 2:n){
if(j<i){
vv<-mat[i,j]
vec<-c(vec,vv)
}
else{}
}}
vec<-vec[-1]
if(!(length==length(vec)))stop("Error. Length is wrong")
return(vec)}
#################################################
vector.2.matrix<-function (vec,size)
{
#Contructs a symmetrical matrix from vector. The diagonal elements
are zero.
#size is the number of rows of desired square matrix
mat <- matrix(0, size, size)
tel <- 1
for(i in 1:(size-1)){
for(j in (i+1):size){
mat[i,j] <- vec[tel]
tel <- tel + 1
}}
mat + t(mat)
}
B<-matrix(0,nrow=n,ncol=n)
##############################################
Calc.disparities.and.stress<-function(delta=delta,W=W,n=n,
X=X,norm.stress=normalise.stress){
#Function to calculate optimal disparities which are also
normalised
#It also calculates Stress
#Distances
d<-(as.vector(dist(X,method = "euclidean",
upper=FALSE,diag=FALSE)))
Dist.mat<-vector.2.matrix(d,size=n)
delta.vec<- matrix.2.vector(delta)
w.vec<- matrix.2.vector(W)
#Now to calculate Stress
Mat<-matrix(c(delta.vec,d,w.vec),byrow=FALSE,nrow=length(d))
colnames(Mat)<-c("Dissimilarities","Distances","Weights")

Term1<-sum((Mat[,1]^2)* Mat[,3] )
Term2<-sum((Mat[,2]^2)* Mat[,3] )
Term3<-sum(Mat[,1] * Mat[,3]* Mat[,2])
Stress<-Term1+Term2-2*Term3
if(norm.stress){Stress<-Stress/Term1}
else{Stress<-Stress}
return(list(Dist.mat=Dist.mat,Stress=Stress,Mat=Mat))}
#################################
Initial<-Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X)
#Initial stress
S0<-Initial$Stress
```

```
S.temp<-rep(0,2)
S.temp[1]<-S0
Y<-X
k<-c(0)
S.vek<-S0
###################################
####Start of Iterative process

while(k<maxit&&(S.temp[1]-S.temp[2])>epsil){
k<-k+1
S.temp[1]<-S0

D.X<-as.matrix(dist(Y,method = "euclidean",
upper=TRUE,diag=TRUE))

#Calculation of matrix B, which is used to update X
B<-matrix(0,nrow=n,ncol=n)
for(i in 1:n){
   for(j in 1:n){
if(!(round(D.X[i,j],15))==0){
B[i,j]<-(-1)*(W[i,j]*delta[i,j])/D.X[i,j]
}
else{}
}}
for(m in 1:n){B[m,m]<-(-1)*sum(B[m,-m])}
if(length(V.inv)==1){X.Updated<-V.inv*B%*%Y}
else{X.Updated<-V.inv%*%B%*%Y}
S0<-
Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X.Updated)$Stress
S.temp[2]<-S0
S.vek<-c(S.vek,S0)
Y<-X.Updated
rownames(Y)<-rownames(delta)
}
##############################
#####End of Iteration
S.verskil.vek<-S.vek[-length(S.vek)]-S.vek[-1]
Table8.4<-matrix(0,nrow=k+1,ncol=3)
colnames(Table8.4)<-c("k","Stress of k th iteration","Stress
difference")
Table8.4[,1]<-0:k
Table8.4[,2]<-S.vek
Table8.4[-1,3]<-round(S.verskil.vek,8)
MDS.plot<-function
(X,type="p",label=TRUE,pch=20,xlab="X1",ylab="X2",
main="",sub="",col="black",cex=1,pos=1)
{
```

```
library(MASS)
on.exit(detach(package:MASS))
eqscplot(x=X[,1],y=X[,2],type="n")
usr<-par("usr")
yas.lengte<-sqrt((usr[4]-usr[3])^2)
xas.lengte<-sqrt((usr[2]-usr[1])^2)

eqscplot(x=X[,1],y=X[,2],type=type,
xlim=c(usr[1]-0.08*xas.lengte,usr[2]+0.08*xas.lengte),
ylim=c(usr[3]-0.08*yas.lengte,usr[4]+0.08*yas.lengte),
pch=pch,xlab=xlab,ylab=ylab,main=main,sub=sub)

if(label){
text(x=X[,1],y=X[,2],pos=pos,labels=dimnames(X)[[1]],col=col,
cex=cex,offset=0.2)}}
############################
Dist.mat<-
Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X.Updated)$Dist.m
at
Squared.error<-(delta-Dist.mat)^2
dimnames(Squared.error)<-dimnames(delta)
SPP<-apply(Squared.error,1,mean)
min.SPP<-min(SPP)
max.SPP<-max(SPP)
SPP.cex<-1.5-(SPP-min.SPP)/(max.SPP-min.SPP)

if(Bubble.plot){cex.points<-SPP.cex}
else{cex.points<-cex.points}
if(normalise.stress){Subtext<-paste("Normalised Stress =
",round(S0,5))}
else{Subtext<-paste("Raw Stress = ",round(S0,5))}
if(Weightsmethod==1){Subtext<-Subtext}
else{Subtext<-paste(Subtext,"  Weights: q =",q)}
if(identify.points){
par(pty="s")
MDS.plot(Y,main=Main,label=FALSE,pch=pch,cex=cex,pos=pos,offset=off
set,
sub=Subtext,cex.points=cex.points)
identify(Y,labels=rownames(Y),offset=0.3,cex=cex)
}
else{
MDS.plot(Y,main=Main,label=TRUE,pch=pch,cex=cex,pos=pos,offset=offs
et,
sub=Subtext,cex.points=cex.points)
}
cat("Is an enlargement required? Click on 2 opposite corners of
area to be enlarged. First decide:y/n\n")
```

```
Enlarge<-readline()
if(Enlarge=="y"){
cc<-locator(n=2)
x.min<-min(cc$x)
x.max<-max(cc$x)
y.min<-min(cc$y)
y.max<-max(cc$y)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)
X.mag<-cbind(Y,Y)
for(i in 1:n){
if(Y[i,1]<=x.max){X.mag[i,1]<-1}
else{X.mag[i,1]<-0}

if(Y[i,1]>=x.min){X.mag[i,2]<-1}
else{X.mag[i,2]<-0}

if(Y[i,2]<=y.max){X.mag[i,3]<-1}
else{X.mag[i,3]<-0}

if(Y[i,2]>=y.min){X.mag[i,4]<-1}
else{X.mag[i,4]<-0}
}
X.magnified<-Y[apply(X.mag,1,sum)==4,]
cex.points<-cex.points[apply(X.mag,1,sum)==4]
win.graph()
MDS.plot(X=X.magnified,label=FALSE,pch=pch,
main=Main,sub="Enlargement of specific
area",col=col,cex=cex,offset=offset,
cex.points=cex.points)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)
identify(X.magnified,labels=rownames(X.magnified),offset=0.3,cex=ce
x)
}
else{}

if(Residual.plot){
win.graph()
Mat<-
Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X.Updated)$Mat
Ordered.Mat<-Mat[order(Mat[,1]),]
if(Weightsmethod==1){Subtext<-""}
else{Subtext<-paste("q = ",q)}

plot(x=c(0,Ordered.Mat[,1]),y=c(0,Ordered.Mat[,1]),main=Main,xlab="
Dissimilarities",
ylab="Distances",type="n",pch=1,xlim=c(0,max(Ordered.Mat[,1:2])),
ylim=c(0,max(Ordered.Mat[,1:2])),sub=Subtext)
abline(a=0,b=1)
points(x=c(0,Ordered.Mat[,1]),y=c(0,Ordered.Mat[,2]),pch=1)
points(x=c(0,Ordered.Mat[,1]),y=c(0,Ordered.Mat[,1]),pch=20)
legend(x=0,y=max(Ordered.Mat[,2]),
legend=c("Dissimilarities","Distances"),pch=c(20,1),cex=1)
}
else{}
Ans<-
list(Final.Configuration=Y,Summary=Table8.4,Number.of.iterations=k,
Squared.error=Squared.error)
Ans}
```

SMACOF.Nonmetric

This function uses the SMACOF algorithm (Borg and Groenen, 2006, p.205) to produce nonmetric MDS two-dimensional display.

Arguments of the function:

delta : Dissimilarity matrix.

X : Initial configuration of points in two dimensions. The initial configuration produced by classical scaling is used as default.

maxit : The maximum number of iterations before the algorithm stops.

epsil : The minimum difference in stress values between iterations reached before algorithm stops.

Shepard : To indicate whether a Shepard diagram should be drawn.

Main : Title of the plot.

normalise.stress : To indicate if Normalised Stress (TRUE) should be used or Raw Stress (FALSE).

Bubble.plot : To indicate whether a bubble plot (TRUE) should be drawn or a normal MDS (FALSE) display

identify.points : To indicate whether the points should be identified individually or automatically according to the value of *pos*.

W : Matrix of weights. The weights may not contain negative values. The default weights used is selected as $w_{ij} = 1$ if $i \neq j$ and $w_{ij} = 0$ if $i = j$.

cex: A numerical value giving the amount by which text should be scaled. Enter *?par* for more details.

pch : Type of symbol used to display samples. Enter *?par* for more details.

pos : Position specified for the label of samples. Values of 1, 2, 3 and 4, respectively indicate positions below, to the left of, above and to the right of the sample points.

offset : when pos is specified, this value gives the offset of the label from the sample point in fractions of a character width.

cex.points : A numerical value giving the amount by which text should be scaled.

```
function
(delta,X=cmdscale(d=delta),maxit=50,epsil=0.000001,Shepard=TRUE,
Main="Jaccard",normalise.stress=TRUE,Bubble.plot=FALSE,
identify.points=FALSE,display.plots=TRUE,
W=matrix(1,nrow=nrow(delta),ncol=nrow(delta)) - diag(nrow(delta)),
cex=0.6,pch=20,pos=3,offset=0.1,cex.points=1)
 {
#This function uses SMACOF algorithm (Borg and Groenen, 2006,
p.205) to
#produce nonmetric MDS two-dimensional display
#delta must be a dissimilarity matrix
#X is the initial configuration. The configuration uses classical
scaling
# configuration as default
#Bubble.plot=TRUE means Bubble.plot will be drawn
#Identify.points=TRUE  means the user click on points in plot to be
identified
#Identify.points=FALSE labels are drawn automatically according to
pos
n<-nrow((delta))
if(min(W)<0)stop("W must only contain non-negative elements!!")
if(sum(delta-t(delta))>0.001) stop("delta must be symmetrical!!")
if(sum(diag(delta))>0.0000000001) stop("delta must have zero’s on
diagonal!!")
if(min(delta)<0) stop("delta must contain non-negative
dissimilarities!!")
#Calculation of V
V<-matrix(0,nrow=n,ncol=n)
V<-(-W)
for(i in 1:n){
V[i,i]<-sum(W[i,-i])
}
if(round(sum(W-matrix(1,nrow=n,ncol=n)),10)==0){
```

```
V.inv<-(1/n)
}
else{
V.inv<-solve((V+matrix(1,nrow=n,ncol=n)))
}
#####################################################
matrix.2.vector<-function (mat=matrix(c(1:16),nrow=4,ncol=4))
{
#Coverts lower diagonal values to vector
n<-nrow(mat)
if(!(nrow(mat)==ncol(mat)))stop("mat must be a square matrix")
length<-n*(n-1)/2

vec<-0
for(j in 1:(n-1)){
for(i in 2:n){
if(j<i){
vv<-mat[i,j]
vec<-c(vec,vv)
}
else{}
}}
vec<-vec[-1]
if(!(length==length(vec)))stop("Error. Length is wrong")
return(vec)}

#################################################
vector.2.matrix<-function (vec,size)
{
#Contructs a symmetrical matrix from vector. The diagonal elements
are zero.
#size is the number of rows of desired square matrix
mat <- matrix(0, size, size)
tel <- 1
for(i in 1:(size-1)){
for(j in (i+1):size){
mat[i,j] <- vec[tel]
tel <- tel + 1
}}
mat + t(mat)
}
#############################################
#Now using function that calculate disparities (pseudo distances)
by using isotonic #regression function isoreg(). The function also
calculates Stress
Calc.disparities.and.stress<-function(delta=delta,W=W,n=n,
X=X,norm.stress=normalise.stress){
```

```
#Function to calculate optimal disparities which are also
normalised
#It also calculates Stress

#Increasing order of dissimilarity values
deltas<-matrix.2.vector(delta)
order.delta.nrs<-rank(deltas,ties.method="first")
ordered.deltas<-sort(deltas)

#Distances which is then ordered
d<-(as.vector(dist(X,method = "euclidean",
upper=FALSE,diag=FALSE)))
ordered.d<-d[order(deltas)]

#Isotonic regression to calculate ordered disparities
ir<-isoreg(ordered.d)
ordered.disparities<-ir$yf

#Order the weights as vector
wij<-matrix.2.vector(W)
ordered.wij<-wij[order(deltas)]

#Normalise disparities to size n(n-1)/2
ordered.disparities<-sqrt(n*(n-
1)/2)*ordered.disparities/(sqrt(sum(ordered.disparities*ordered.dis
parities*ordered.wij)))

#Now re-order disparities and distances to match with initial
unordered dissimilarities
disparities<-ordered.disparities[order.delta.nrs]

#Make matrices of vectors containing the  distances and disparities
Dist.mat<-vector.2.matrix(d,size=n)
D.hat<-vector.2.matrix(disparities,size=n)

#Now to calculate Stress
Mat<-
matrix(c(deltas,d,disparities,wij),byrow=FALSE,nrow=length(deltas))
colnames(Mat)<-
c("Dissimilarities","Distances","Disparities","Weights")

Term1<-sum((Mat[,3]^2)* Mat[,4] )
Term2<-sum((Mat[,2]^2)* Mat[,4] )
Term3<-sum(Mat[,4] * Mat[,3]* Mat[,2])

Stress<-Term1+Term2-2*Term3
if(norm.stress){
```

```
Stress<-Stress/Term1
}
else{Stress<-Stress}
return(list(Dist.mat=Dist.mat,D.hat=D.hat,Stress=Stress,Mat=Mat))
}
##########################################
######Now using previous function to start

Initial<-Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X)
D.hat<-Initial$D.hat
#Initial stress
S0<-Initial$Stress
S.temp<-rep(0,2)
S.temp[1]<-S0

Y<-X
k<-c(0)
S.vek<-S0

##########################################
#Start of iterative process
while(k<maxit&&(S.temp[1]-S.temp[2])>epsil){
k<-k+1
rownames(Y)<-rownames(delta)
S.temp[1]<-S0

List<-Calc.disparities.and.stress(delta=delta,W=W,n=n,X=Y)

D.hat<-List$D.hat
D.X<-as.matrix(dist(Y,method = "euclidean",
upper=TRUE,diag=TRUE))

#Calculation of matrix B, which is used to update X
B<-matrix(0,nrow=n,ncol=n)
for(i in 1:n){
   for(j in 1:n){
if(!(round(D.X[i,j],15))==0){
B[i,j]<-(-1)*(W[i,j]*D.hat[i,j])/D.X[i,j]
}
else{}
}}
for(m in 1:n){
B[m,m]<-(-1)*sum(B[m,-m])
}
if(length(V.inv)==1){X.updated<-V.inv*B%*%Y}
else{X.updated<-V.inv%*%B%*%Y}
```

```
S0<-
Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X.updated)$Stress
S.temp[2]<-S0
S.vek<-c(S.vek,S0)
Y<-X.updated
rownames(Y)<-rownames(delta)
}
#####################################
#End of iteration
S.verskil.vek<-S.vek[-length(S.vek)]-S.vek[-1]
Table8.4<-matrix(0,nrow=k+1,ncol=3)
colnames(Table8.4)<-c("k","Stress of k th iteration","Stress
difference")
Table8.4[,1]<-0:k
Table8.4[,2]<-S.vek
Table8.4[-1,3]<-round(S.verskil.vek,8)
Min.Stress<-
Calc.disparities.and.stress(delta=delta,W=W,n=n,X=X.updated)
Dist.mat<-Min.Stress$Dist.mat
D.hat<-Min.Stress$D.hat
Normalising.constant<-(n*(n-1)/2)
Squared.error<-((D.hat-Dist.mat)^2)
dimnames(Squared.error)<-dimnames(delta)
SPP<-apply(Squared.error,1,mean)*(n/(n-1))
min.SPP<-min(SPP)
max.SPP<-max(SPP)
SPP.cex<-1.5-(SPP-min.SPP)/(max.SPP-min.SPP)
if(display.plots){
if(Bubble.plot){cex.points<-SPP.cex}
else{cex.points<-cex.points}
if(normalise.stress){Subtext<-paste("Normalised Stress =
",round(S0,5))}
else{Subtext<-paste("Raw Stress = ",round(S0,5))}
MDS.plot<-function
(X,type="p",label=TRUE,pch=20,xlab="X1",ylab="X2",
main="",sub="",col="black",cex=1,pos=1)
{
library(MASS)
on.exit(detach(package:MASS))
eqscplot(x=X[,1],y=X[,2],type="n")
usr<-par("usr")
yas.lengte<-sqrt((usr[4]-usr[3])^2)
xas.lengte<-sqrt((usr[2]-usr[1])^2)

eqscplot(x=X[,1],y=X[,2],type=type,
xlim=c(usr[1]-0.08*xas.lengte,usr[2]+0.08*xas.lengte),
ylim=c(usr[3]-0.08*yas.lengte,usr[4]+0.08*yas.lengte),
pch=pch,xlab=xlab,ylab=ylab,main=main,sub=sub)

if(label){
text(x=X[,1],y=X[,2],pos=pos,labels=dimnames(X)[[1]],col=col,
cex=cex,offset=0.2)
}}
if(identify.points){
par(pty="s")
MDS.plot(Y,main=Main,label=FALSE,pch=pch,cex=cex,pos=pos,offset=off
set,
sub=Subtext,cex.points=cex.points)
identify(Y,labels=rownames(Y),offset=0.3,cex=cex)
}
else{
MDS.plot(Y,main=Main,label=TRUE,pch=pch,cex=cex,pos=pos,offset=offs
et,
sub=Subtext,cex.points=cex.points)
}
cat("Is an enlargement required? Click on 2 opposite corners of
area to be enlarged. First decide:y/n\n")
Enlarge<-readline()
if(Enlarge=="y"){
cc<-locator(n=2)
x.min<-min(cc$x)
x.max<-max(cc$x)
y.min<-min(cc$y)
y.max<-max(cc$y)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)
X.mag<-cbind(Y,Y)
for(i in 1:n){
if(Y[i,1]<=x.max){X.mag[i,1]<-1}
else{X.mag[i,1]<-0}
if(Y[i,1]>=x.min){X.mag[i,2]<-1}
else{X.mag[i,2]<-0}
if(Y[i,2]<=y.max){X.mag[i,3]<-1}
else{X.mag[i,3]<-0}
if(Y[i,2]>=y.min){X.mag[i,4]<-1}
else{X.mag[i,4]<-0}
}
X.magnified<-Y[apply(X.mag,1,sum)==4,]
cex.points<-cex.points[apply(X.mag,1,sum)==4]
win.graph()

MDS.plot(X=X.magnified,label=FALSE,pch=pch,
main=Main,sub="Enlargement of specific
area",col=col,cex=cex,offset=offset,
cex.points=cex.points)
segments(x0=x.min,y0=y.min,x1=x.max,y1=y.min,lty=2)
segments(x0=x.max,y0=y.min,x1=x.max,y1=y.max,lty=2)
segments(x0=x.max,y0=y.max,x1=x.min,y1=y.max,lty=2)
segments(x0=x.min,y0=y.max,x1=x.min,y1=y.min,lty=2)
identify(X.magnified,labels=rownames(X.magnified),offset=0.3,cex=ce
x)
}
else{}

if(Shepard){
win.graph()
Mat<-Calc.disparities.and.stress(delta=delta,W=W,n=n,X=Y)$Mat
Ordered.Mat<-Mat[order(Mat[,1]),]
plot(x=Ordered.Mat[,1],y=Ordered.Mat[,2],main=Main,xlab="Dissimilar
ities",
ylab="Distances and Disparities",type="p",pch=1)
points(x=Ordered.Mat[,1],y=Ordered.Mat[,3],pch=20)
lines(x=Ordered.Mat[,1],y=Ordered.Mat[,3])
legend(x=Ordered.Mat[1,1],max(Ordered.Mat[,2]),
legend=c("Distances","Disparities"),pch=c(1,20),cex=1)}
else{}}
else{}
Ans<-list(Final.Configuration=Y,X.initial=X,
Minimum.Stress=round(S.vek[k],6),Number.of.iterations=k,
Summary=Table8.4,Squared.error=Squared.error,SPP=round(SPP,3),
mean.SPP=round(mean(SPP),5))
return(Ans)}
```

SMACOF.Nonmetric.random.starts

This function uses random starting configurations to produce a final nonmetric MDS display that minimised the normalised stress value. The random configurations is drawn from a uniform (0, 1) distribution. The configuration produced by classical scaling is also used as an initial configuration. This function uses *SMACOF.Nonmetric*.

<u>Arguments of the function</u>:

D :	Dissimilarity matrix.

r :	The number of random initial configurations.

main :	Title of the plot.

Bubble :To indicate if a bubble plot (TRUE) or normal MDS display (FALSE) should be drawn.

iden : To indicate whether the points should be identified individually or automatically according to the value of *pos*.

```
function (D,r=200,Bubble=TRUE,iden=FALSE,main="Jaccard")
{
#This function uses SMACOF.Nonmetric function
#Various initial configurations are used, and minimum Normalised
Stress value is noted
#The initial configuration with lowest Normalised Stress will be
used
#Initial configuration produced by classical scaling is also
considered
#Random configurations are generated using uniform(0,1)
distribution
#D is dissimilarity matrix
#r is the number of random starting configurations
#First find configuration produced by classical scaling
X.ini<-cmdscale(d=D)
rownames(X.ini)<-rownames(D)
K<-SMACOF.Nonmetric(delta=D, X =X.ini, maxit = 150, epsil =0.0001,
display.plots = FALSE)
Min.Stress<-K$Minimum.Stress

######################################3
#Start of random configurations
for(i in 1:r){
X.ini<-matrix(runif(n=2*nrow(D)),ncol=2)
#Random initial configuration drawn from uniform (0,1)
distribution.
rownames(X.ini)<-rownames(D)
KK<-SMACOF.Nonmetric(delta=D, X =X.ini, maxit = 150, epsil =0.0001,
display.plots = FALSE)
Min.Stress.1<-KK$Minimum.Stress
if(Min.Stress.1<Min.Stress){
K<-KK
Min.Stress<-Min.Stress.1
}
else{}}
####################################3
#End of random configurations
#The initial configuration with lowest stress have now been found
#and will be used next
SMACOF.Nonmetric(delta=D, X =K$X.initial, maxit =400, epsil
=0.0000001,display.plots =TRUE, Bubble.plot=Bubble,
identify.points=iden,Main=main)}
```

Appendix B

Table B: *The first 30 rows of the Medical Scheme 55-59 data set.*

NON	ADS	AST	BCE	BMD	CMY	COP	CRF	CSD	DBI	DM1	DM2	DYS	EPL	GLC	HAE	HYL	HYP	IBD	IHD	MSS	PAR	RHA	SCZ	SLE	TDH	HIV	Nr.Lives	Hosital	Medical	Related
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	953	384.17	589.07	89.98
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	24 960	184.07	74.59	52.36
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	20	0.00	708.90	60.73
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	147	371.61	101.01	55.48
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	22	0.00	594.82	97.38
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	323	130.17	525.00	175.67
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	62	325.53	627.34	10.65
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4 626	177.92	238.36	69.44
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	496	350.39	543.05	48.74
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	23	0.00	130.42	0.00
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	3	0.00	163.21	0.00
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	9	0.00	119.56	28.15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	343	172.83	750.28	125.00
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	59	61.79	1 016.37	141.97
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	5	0.00	1 526.10	280.60
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	126	793.38	6 967.87	131.18
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	13	0.00	8 230.02	87.87
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	897	931.07	200.39	54.54
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	165	11.33	204.17	34.73
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	43	0.00	484.61	55.21
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	12	332.25	687.16	44.20
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	306	93.87	340.44	35.39
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	38	128.46	380.69	113.25
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	30	2 306.30	635.67	5.68
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	14	0.00	1 464.67	113.57
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	12	0.00	386.94	11.66
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	5	2 792.00	330.67	0.00
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	125 795	185.62	153.75	60.09
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	111	224.44	662.49	44.28
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	16 327	227.1	210.74	49.53

References

AMENDMENT TO THE GENERAL REGULATIONS MADE IN TERMS OF THE MEDICAL SCHEMES ACT OF 1998 (ACT 131 OF 1998), South Africa, Government Gazette, Vol. 449, No. 24007, Regulation Gazette No. 7496.

ANDERBERG, M.R. (1973), *Cluster analysis for Applications,* New York: Academic Press. 359p.

BANFIELD, J.D. & RAFTERY, A.E. (1993), Model-based Gaussian & non-Gaussian clustering, *Biometrics*, Vol. 49, 803-821.

BORG, I. & GROENEN, P.J.F. (2005), *Modern multidimensional scaling : Theory and Applications, second edition*, New York: Springer Press. 614p.

COUNCIL FOR MEDICAL SCHEMES (2005), *Report on the First Shadow Submissions by Medical Schemes to the Risk Equalisation Fund. Prepared for the Department of Health & the Council for Medical Schemes, 31 October 2005.* Available from `http://www.medicalschemes.com`.

COUNCIL FOR MEDICAL SCHEMES (2006), *Guidelines for the Identification of Beneficiaries with REF Risk Factors in Accordance with the REF Entry & Verification Criteria, Version 2. Applicable to all REF cases from 1 January 2007.* Available from `http://www.medicalschemes.com`.

COX, T.F. & COX, M.A.A. (2001), *Multidimensional scaling, second edition*, New York: Chapman & Hall/CRC Press. 308p.

DASGUPTA, A. & RAFTERY, A.E. (1998), Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering, *Journal of the American Statistical Association*, Vol. 93, 294-302.

FRALEY, C. & RAFTERY, A.E. (2002), Model-based clustering, discriminant analysis and density estimation, *Journal of the American Statistical Association*, Vol. 97, 611-631.

GARDNER, S. (2001), *Extensions of biplot methodology to discriminant analysis with applications of non-parametric principal components*, Unpublished PhD thesis, Stellenbosch University. 535p.

GORDON, A.D. (1999), *Classification, second edition*, London: Chapman & Hall/CRC Press. 256p.

GOWER, J.C. (1985), Measures of similarity, dissimilarity and distance. In S. Kotz, N.L. Johnson & C.B. Read (Eds), *Encyclopedia of statistical sciences*, Vol.5, 397-405.

GOWER, J.C. & HAND, D.J. (1996), *Biplots*, London: Chapman & Hall. 277p.

GOWER, J.C. & LEGENDRE, P. (1986), Metric & Euclidean properties of dissimilarity coefficients, *Journal of Classification*, Vol. 3, 5-48.

HEISER, W.J. (1991), A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative, *Psychometrika*, Vol. 56, 7-27.

HUBALEK, Z. (1982), Coefficients of association and similarity based on binary (presence-absence) data: an evaluation, *Biological Reviews of the Cambridge Philosophical Society*, Vol. 57, 669-689.

JOHNSON, R.A. & WICHERN, D.W. (2002), *Applied Multivariate Statistical Analysis, fifth edition*, New Jersey: Prentice-Hall. 767p.

KAUFMAN, L. & ROUSSEEUW, P.J. (1990), *Finding groups in data: An introduction to cluster analysis*, New York: Wiley. 331p.

MACNAUGHTON-SMITH, P., WILLIAMS, W.T., DALE, M.B., & MOCKETT, L.G. (1964), Dissimilarity analysis: A new technique of hierarchical sub-division, *Nature*, Vol.202, 1034-1035.

MCLEOD, H. (2007), leader of the REF Study 2005 consortium, [Personal Communication].

PARKIN, N. & MCLEOD, H. (2001), *Risk equalisation methodologies: an international perspective*, Centre for Actuarial Research, University of Cape Town.

PISON, G., STRUYF, A. & ROUSSEEUW, P.J. (1998), Displaying a clustering with CLUSPLOT, *Computational Statistics & Data Analysis*, Vol.30, 381-392.

RAFTERY, A.E. & DEAN, N. (2006), Variable selection for Model-Based Clustering, *Journal of the American Statistical Association*, Vol. 101, 168-178.

RETAP (2007), *Methodology for the Determination of the Risk Equalisation Fund Contribution Table 2007 [Base 2005, Use 2007]. Recommendations by the Risk Equalisation Technical Advisory Panel to the Council for Medical Schemes. Recommendations Report No. 9, 17 April 2007*. Available from `http://www.medicalschemes.com.`

SAMMON, J.W. (1969), A non-linear mapping for data structure analysis, *IEEE Transactions on Comput*ers, Vol.18, 401-409.

SIBSON, R., BOWYER, A. & OSMOND, C. (1981), Studies in the robustness of multidimensional scaling: Euclidean models and simulation studies, *Journal of Statistical Computation & Simulation*, Vol. 13, 273-296.

S-PLUS ® 8 GUIDE TO STATISTICS, VOLUME 2 (2007), Insightful Corporation, Seattle, Washington.

STRUYF, A., HUBERT, M., & ROUSSEEUW, P.J. (1997), Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data Analysis*, Vol. 26, 17-37.