VISUAL DATA MINING

Real Applications & New Approaches

Ph.D THESIS

Electronic Engineering Department
University of Valencia

January, 2014

By:
José M. Martínez Martínez
Advisors:
Emilio Soria Olivas
José D. Martín Guerrero

Ph.D THESIS

Visual Data Mining: Real Applications and New Approaches

by

José María Martínez Martínez

Supervisors
Emilio Soria Olivas
José David Martín Guerrero

Electronic Engineering Department
University of Valencia

Valencia - January, 2014

Visual Data Mining: Real Applications and New Approaches.

José María Martínez Martínez, January 2014

Dpt. Enginyeria Electrònica. Escola Tècnica Superior d'Enginyeria.

D. EMILIO SORIA OLIVAS, Doctor en Ingeniería Electrónica, Profesor Titular del Departamento de Ingeniería Electrónica de la Escola Técnica Superior d'Enginyeria de la Universitat de Valéncia, y

D. JOSÉ DAVID MARTÍN GUERRERO, Doctor por la Universitat de València, Profesor Titular del Departamento de Ingeniería Electrónica de la Escola Técnica Superior d'Enginyeria de la Universitat de Valéncia

HACEN CONSTAR QUE:

El Ingeniero en Electrónica D. José María Martínez Martínez ha realizado bajo nuestra dirección el trabajo titulado "Visual Data Mining: Real Applications and New Approaches.", que se presenta en esta memoria para optar al grado de Doctor.

Y para que así conste a los efectos oportunos, firmamos el presente certificado, en Valencia, a ____ de Enero de 2014.

Emilio Soria Olivas

José D. Martín Guerrero

J. Rafael Magdalena Benedito
Dtor del Departamento

Tesis Doctoral: VISUAL DATA MINING: REAL APPLICATION AND NEW APPROACHES

Autor: JOSÉ MARÍA MARTÍNEZ MARTÍNEZ

Directores: Dr. EMILIO SORIA OLIVAS
Dr. JOSÉ DAVID MARTÍN GUERRERO

El tribunal nombrado para juzgar la Tesis Doctoral arriba citada, compuesto por los doctores:

Presidente: ____________________

Vocal: ____________________

Secretario: ____________________

Acuerda otorgarle la calificación de ____________________

Y para que así conste a los efectos oportunos, firmamos el presente certificado.

Valencia, a

This doctoral thesis has been partially supported by *Fresenius Medical Care (FME)* through the research project entitled *“Intelligent analysis of Fresenius Medical Care data”*.

Agradecimientos

Desde estas líneas pretendo expresar mi más sincero agradecimiento a todas aquellas personas que durante este largo camino de esfuerzo, trabajo y dedicación han estado a mi lado y que, de una u otra forma, han contribuido a que esta tesis haya llegado a buen fin. Sin todas ellas, hubiese sido del todo imposible afrontar con éxito la elaboración de este proyecto, en la que tanta ilusión he puesto.

En primer lugar a los directores de esta tesis, Emilio Soria y José D. Martín, por ser los principales responsables de que este trabajo llegara a buen puerto, estando siempre a mi lado, en los buenos y no tan buenos momentos, animándome siempre a continuar. Agradecer su apoyo incondicional, sus fantásticas ideas y aportaciones a esta tesis que, especialmente sin ellos, no hubiera sido posible y no gozaría de la misma calidad. Muchas gracias por vuestra ayuda, por considerarme uno más en el IDAL, por preocuparos constantemente de mi futuro y porvenir y sobretodo por considerarme, al igual que yo os considero, un buen amigo. Gracias por estos años de amistad y trabajo que espero, y deseo, que se prolonguen durante muchos más.

También quiero agradecer al resto del IDAL (Juan, Pablo, Marcelino, Rafa, Antonio y Joan) su constante apoyo y preocupación. Gracias por haber hecho del nuestro un ambiente de trabajo excepcional en el que se ha hecho muy fácil trabajar durante todos estos años. Creo que esto se hace realidad gracias a la gran amistad que une al grupo de investigación tanto dentro como fuera del trabajo. A ti en particular, Pablo, gracias por tu apoyo, amistad, ayuda y por compartir las mismas preocupaciones y devenires en nuestro camino con destino al grado de Doctor.

En especial a mi familia. A mis padres, Pedro y Sole, por haberme hecho disfrutar de la vida, por la educación que me han dado y por inculcarme unos valores que han hecho de mí la persona que soy hoy en día. Gracias por ser un ejemplo a seguir en la vida y por vuestro apoyo y amor incondicional. A mi hermano Pedro, por creer en mi y por animarme siempre a seguir hacia adelante. Gracias por ser también un ejemplo a seguir, y por hacer desde muy pequeños que los hermanos tuviésemos una muy estrecha relación. Gracias a ti, hoy puedo decir con la boca llena que mis mejores amigos son mis dos hermanos. A mi hermano Fran, por apoyarme, entenderme como nadie y por ser parte de mi. Sin ti no sería la persona que soy hoy en día. Gracias

por tantos y tan buenos momentos que hemos pasado juntos y por estar siempre a mi lado, aunque en épocas de la vida hayas estado físicamente alejado. A mi abuela, por todo lo que se preocupa por sus nietos. A mis cuñadas, Sandra y Helena, por estar siempre ahí, cuidar de mis hermanos y demostrarme que son también familia carnal; y finalmente a mis sobrinos, Eric y Andrea, porque ellos hacen de la vida un lugar más bonito.

A mis amigos, por esos momentos de evasión en los que uno disfruta y se olvida de todo lo malo.

A las personas que, aunque no aparecen aquí con nombres y apellidos, han estado presentes de alguna forma durante el desarrollo de este trabajo y han hecho posible que hoy vea la luz. A todos mi eterno agradecimiento.

Y finalmente en especial mención a Aida, por hacerme la vida mucho más feliz y más fácil, por sufrirme y soportarme día a día, en la convivencia, durante todo el tiempo que le he dedicado a la tesis. Gracias por tu comprensión y por ser el pilar de mi vida. Porque con una mirada, una caricia o un pequeño gesto haces que me olvide de todo. Por haberme apoyado en todo momento y haber estado ahí incluso en momentos difíciles para ti. Gracias por tu altruismo para conmigo, por tu apoyo incondicional y sobre todo por ser como eres. Gracias por cambiarme la vida y hacer que no me la pueda imaginar sin estar junto a ti. No hay palabras en el mundo para agradecerte todo lo que eres y significas para mi.

Valencia, Octubre de 2013.

José M. Martínez.

A mi Familia
por estar siempre a mi lado

"An attempt at visualizing the Fourth Dimension: Take a point, stretch it into a line, curl it into a circle, twist it into a sphere, and punch through the sphere."

-Albert Einstein (1879-1955).

Contents

List of Figures

List of Tables

Summary and objectives

Data visualization has in recent years become a very active and vital area of research. It is an effective way to analyze large amounts of data to identify correlations, trends, outliers, patterns, among many other information. Raw data are often meaningless, but representing visually such data offers audiences important context for understanding the information contained in them. Due to the importance of this area of research, and its novelty, this thesis aims to discover new findings, draw conclusions and bequeath significant contributions to the scientific community in this field. To achieve this purpose, this work attempts to address two main objectives.

The first objective of this thesis is to try to develop new visualization methods for interpreting the results of several data mining algorithms. For example, cluster analysis is a big challenge in data visualization; for this reason, they both often go hand in hand. Nonetheless, there is a lack of visualization techniques associated with clustering and hierarchical clustering that provide information about the values of the centroids' attributes and the relationships among them. Thus, this thesis researches new approaches that make possible to include this information visually, as well as to find new methods for visualizing the results of several data mining algorithms, apart from those above mentioned, in order to help simplify their interpretation and to obtain a better understanding.

Another objective of the present thesis is focused on addressing different real problems of diverse nature, some of them framed in funded research projects. The solution of these problems are approached through data visualization and visual data mining in order to gain insight about the problem making possible the knowledge extraction, the discovery of hidden information, and finding patterns and relationships in data. Particularly, the present thesis focuses on the use of the well-known Self-Organizing Maps (SOMs) to solve real problems in several different fields of research, providing solutions to complex problems that would otherwise have been very difficult to solve.

Resumen y objetivos

En los últimos años, la visualización de datos se ha convertido en un área muy activa y vital de la investigación. Es una manera eficaz de analizar grandes cantidades de datos para identificar correlaciones, tendencias, valores extremos, patrones, entre otra mucha información. Los datos sin procesar a menudo carecen de sentido, pero representar dichos datos visualmente ofrece al público un contexto importante para entender la información contenida en ellos. Debido a la importancia de esta área de investigación, y a su novedad, esta tesis se centra en esta temática y pretende descubrir nuevos hallazgos, extraer conclusiones y legar contribuciones relevantes a la comunidad científica en dicho campo. Para alcanzar dicho propósito, este trabajo trata de abordar dos objetivos principales.

El primer objetivo de la presente tesis es tratar de desarrollar nuevos métodos de visualización para interpretar los resultados de varios algoritmos de minería de datos. Por ejemplo, el análisis de *clusters* o técnicas de agrupamiento es un gran desafío en la visualización de datos; por esta razón, ambos van a menudo de la mano. Sin embargo, hay una falta de técnicas de visualización asociadas al *clustering* y *clustering jerárquico* que proporcionen información sobre los valores de los atributos de los centroides y de las relaciones entre ellos. Por lo tanto, esta tesis investiga nuevas aproximaciones que hagan posible incluir esta información visualmente, además de encontrar nuevos métodos para visualizar los resultados de varios algoritmos de minería de datos, aparte de los anteriormente mencionados, con el fin de ayudar a simplificar su interpretación y para obtener una mejor comprensión.

Otro de los objetivos de esta tesis se centra en abordar diferentes problemas reales de diversa índole, algunos de ellos enmarcados en proyectos de investigación financiados. La solución de estos problemas se aborda a través de la visualización de datos y minería de datos visual con el fin de obtener una perspectiva sobre el problema, lo que hace posible la extracción de conocimiento, el descubrimiento de información oculta y encontrar patrones y relaciones entre los datos. En particular, la presente tesis se centra en el uso de los conocidos *Self-Organizing Maps (SOMs)* para resolver problemas reales en diversos campos de investigación, proporcionando soluciones a problemas complejos que de otra manera habría sido muy difícil de resolver.

Chapter 1

Introduction

Abstract

This chapter introduces general aspects about data visualization and presents the research motivation of this thesis. In addition, this chapter presents a brief review of classical techniques to visualize multidimensional data sets and defines in general terms the concept of Visual Data Mining, which is the topic covered in this thesis.

1.1 Research Motivation

In recent years, there has been tremendous growth in our capabilities to generate and collect data, mainly due to the processing power of machines and its low-cost storage (Han, 2005). Hence, organizations today have large amounts of data stored and organized, which cannot be analyzed efficiently in its entirety. However, within these large amounts of data there is a lot of hidden information of strategic importance. Therefore, the knowledge extraction of these data is of great importance, and the organizations are very interested in analyzing the data optimally. The discovery of this hidden information is possible through Data Mining (Fayyad et al., 1996), which applies Artificial Intelligence to find patterns and relationships in data making possible building models among other sophisticated techniques, that is, abstract representations of reality. Thus, the real value of data lies in the information that can be extracted from them. This information helps us to make decisions or to improve our understanding of the phenomena around us.

A useful approach may be the transformation of data into visual representations that enable a human supervisor understand more easily the process and recognize important events. This strategy of Data Mining based on visual exploration of data is called Visual Data Mining (Oliveira and Levkowitz, 2003; Berthold and J.Hand, 2002). Techniques of Visual Data Mining are very powerful, intuitive and they do not need a lot of a priori knowledge on Data Mining techniques. Moreover, Visual Data Mining can be a previous stage to the Data Mining itself. Thus, they provide a snapshot of the data set that allows analysts to extract knowledge. Its main objective is, therefore, to integrate the person in the process of data exploration and exploit their skills of visual perception and reasoning with visible objects. Data visualizations help to find trends and correlations that can lead to relevant discoveries. Representing large amounts of information in a visual form often allows the visualization of patterns that would otherwise be impossible to find in datasets. Opposed to the traditional hypothesis-and-test method of inquiry, which relies on asking the right questions, data visualizations bring themes and ideas out to the surface, where they can be easily discerned. Summarizing, visualizations allow to better understand and process enormous amounts of information quickly because it is all represented in a single image or a reduced collection of images (Kirk, 2012).

Most scientific, engineering, and business data is multi-dimensional; i.e. datasets contain typically more than three variables. A large number of representations in two and three dimensions can be carried out (classical representation as bar charts, scatter plots, boxplots, etc), but when the dimensionality is greater than three, it is very complicated to represent the obtained data without establishing any type of restriction (as keep fixed certain set of variables and representing the rest). Such a restriction leads to a partial representation of the information below certain conditions. Therefore, visualizing multi-dimensional data without restrictions has tremendous effects on science, engineering, and business decision-making. For that type of representation, which will make possible to find patterns in data sets with high dimensionality, the multi-dimensional visualizations are specially indicated. Due to these facts, the central research question in this thesis is to create new visualization methods that make possible to understand large amounts of multidimensional data, extracting information about them. In this way, data visualization will allow to detect which variables are relevant and the relationships among them. In addition to presenting new visualization techniques, a part of the research in this thesis focuses on the use of visualization techniques to solve real problems in different fields of research, providing solutions to complex problems that would otherwise have been very difficult to solve.

1.2 Knowledge Discovery in Databases (KDD) and Data Mining

This section sets out the conceptual framework of the thesis. It explains the concepts of Knowledge Discovery in Databases (KDD) and Data Mining, outlining the steps in a typical KDD process.

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Jensen and Shen, 2008). In other words, KDD prepares, sounds out and explores the data in order to extract the hidden information in them. KDD goals are (Zhang et al., 2004):

- Automatically process large amounts of raw data.
- Identify the most significant and relevant patterns.
- Presenting them as appropriate knowledge to meet the goals of the user.

An important issue regarding Data Mining and KDD is that they are frequently treated as synonyms, but Data Mining is actually part of the knowledge discovery process. That is, Data Mining encompasses a whole set of techniques designed to extract knowledge implicit in the databases, but it does not embrace everything related to the data preparation. This fact can be checked as follows, where the steps of a typical KDD process are listed (Kohonen, 2010)

- **Selection of data set:** This involve both the dependent variables and target variables, as well as the sampling of the available records in some cases. The selection includes both a filtering or horizontal merger (observations) and vertical (attributes).

- **Analysis of the properties of the data:** In particular, histograms, scatter plots, presence of outliers, missing data or null values as well as basic statistics should be at least taken into account.

- **Transformation of input data set:** There are several ways depending on the previous analysis, with the aim of preparing the input data set to implement the Data Mining technique that best fits the data and the problem. This transformation of the data includes cleaning and pre-processing of data. This is achieved by designing an appropriate strategy for managing noise, incomplete values, time sequences, extreme cases (if necessary), etc. Moreover, feature selection or dimensionality reduction techniques can be applied to the data set depending on the addressed problem.

- **Select and apply the *Data Mining* technique:** This includes the selection of the discovery task to perform, for example, classification, clustering, prediction, etc. It also includes the selection of the algorithms to use, the transformation of data to the format required by the specific Data Mining model as well as to carry out the process of Data Mining, looking for patterns that can be expressed in terms of a model or simply to express data dependencies. The obtained model depends on its purpose (classification, regression, etc.) and their representation approach (decision trees, rules, etc.). Finally, it must be specified a criterion for selecting a model within a possible set of models, as well as to specify the search strategy to use (usually is predetermined in the Data Mining algorithm).

- **Model validation:** Evaluate the results contrasting them with a previously reserved data set to validate the generality of the model. It involves the evaluation, interpretation, processing and representation of the extracted patterns. This may involve repeating the process, perhaps with other data, other algorithms, other targets and other strategies. This is a crucial step, where having domain knowledge is needed. The interpretation may benefit from visualization processes, and it is useful to remove redundant or irrelevant patterns.

- **Diffusion and use of new knowledge:** After the interpretation of the results, the knowledge discovered can be used to take actions on the model (as retrain the model with other parameters, extract/include patterns of the data set, or even train another model usually to improve it) or it can be simply stored and reported to interested persons. In this sense, KDD involves an interactive and iterative process.

If the final model did not pass this evaluation, the process could be repeated from the beginning or, if the expert considers it appropriate, from any of the above steps. This feedback can be repeated as often as deemed necessary until obtaining a valid model. Once validated the model, if it is acceptable (providing appropriate outputs and/or with acceptable margins of error) it is ready for exploitation.

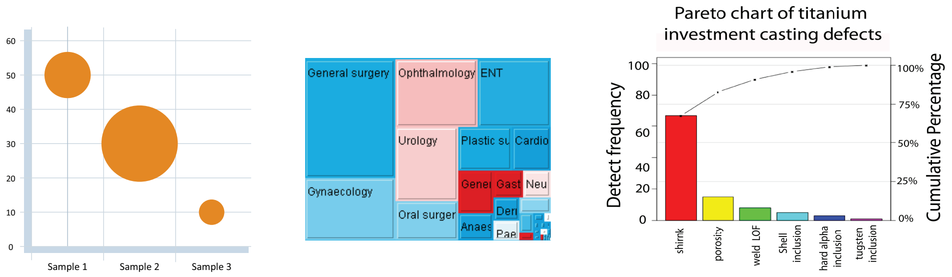
Regarding the Data Mining concept, it can be said that their foundations are found in Artificial Intelligence and Statistical Analysis. There are a number of different techniques that can be used in this framework; they include methods in a wide spectrum from statistics to neural networks. In general, methods for the extraction of knowledge are known as Machine Learning (Alpaydin, 2010). As its name suggests, the aim of these techniques is to optimize an objective or adapt to it, learning from data. As a general feature, one can say that these techniques are generic and versatile, applicable to various types of systems. Some of this Machine Learning methods make use of unsupervised learning techniques, which do not require specifying desired outputs or get reinforcements in the environment since its goal is to obtain an accurate representation of the input that conforms to the goals (Tzanako, 2002). This fits with data visualization purposes, which is the theme of this thesis.

Summarizing, KDD is the nontrivial extraction of information that lies implicitly in the data using Data Mining techniques. KDD aims to, automatically, process large amounts of data to find useful knowledge in them, in this way it will allow the user to use this valuable information for his own convenience. At the same time, there is a deep interest in presenting the results visually or at least so that their interpretation is very clear. Due to this fact, and to the difficulty of finding valuable information, Data Mining techniques have been developed in the visualization field in last decade. Visualization of data is one of the techniques that is used in the KDD process as an approach to explore the data and also to present the results. The result of the exploration should be interesting, and its quality should not be affected by higher volumes of data or noisy data. In this sense, information discovery algorithms must be highly robust. The following section discusses various aspects of data visualization as well as what it is and what it is for. It also provides a brief review of high dimensional visualization. Finally, it explains in detail the concept of Visual Data Mining, which is the focus of the research conducted in this thesis.

1.3 Data Visualization

1.3.1 Introduction

Data visualization is a new term that expresses the idea that it involves more than just representing data in a graphical form (instead of using a table). The information behind the data should also be revealed in a good display. The graphic should aid readers or viewers in seeing the structure in the data. Data visualization is the process of representing data graphically and interacting with these representations in order to gain insight into the data (Simoff et al., 2008). Data is mapped to some numerical form and translated into some graphical representation. The most recognizable and utilized form of data visualization is the basic chart: bar charts, scatter graphs, line charts, pie charts and maps are examples of simple data visualizations that have been used for decades (Harris, 1999). The first function of a good chart is to allow decision makers to examine the data and reduce the time required to extract key information. More advanced examples of data visualization include, bubble charts[1], tree maps (Shneiderman, 1991), pareto charts (Wilkinson, 2006), and many others (see Figure 1.1). These more sophisticated visualizations are designed to display data in ways tailored to a specific function or problem.



(a) *Bubble chart. Source: Wikipedia.*

(b) *Tree map. Source: Wikipedia.*

(c) *Pareto chart. Source: Wikipedia.*

Figure 1.1: *More advanced data visualization than the basic charts.*

The graphs are very useful tools because they can represent relationships between sets of objects or data. Therefore, it is not surprising that graph based Data Mining has become quite popular in the last few years. They are used for modeling com-

[1] http://office.microsoft.com/en-us/excel-help/creating-a-bubble-chart-HA001117076.aspx. (*Last checked September 2013*)

plex systems and for visualizing relationships. In statistics and data analysis, several graphics are used for different purposes. For example, dendrograms are used in hierarchical cluster analysis; trees are used in classification and regression problems; and path diagrams are used and in Bayesian networks or in order to express dependencies between different variables (Chen et al., 2008a). A good graphic is of great importance in a given problem since it may be the key to understanding the problem. There is an extensive literature that considers the problem of how to draw a graph (Battista et al., 1994, 1999).

The objective of visual exploration techniques, as introduced in section 1.1, is to integrate people into the process of data exploration. These are techniques aimed at the representation data in a visual way that allow to exploit the flexibility, creativity and knowledge about the field problem processes by humans (Keim, 2001). Different ways of visualizing a data set provide different types of information, which can help when trying to understand a model or a problem due to the fact that it is seen from different points of view.

According to (Friendly, 2008) the main goal of data visualization is to communicate information clearly and effectively through graphical means. It does not mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights about complex dataset by communicating its key-aspects in a more intuitive way. However, designers often tend to discard the balance between design and function, creating gorgeous data visualizations which fail to serve its main purpose, communicate information (Friedman, 2008).

One of the most important benefits of visualization is that it enables the access to huge amounts of data in ways that would not be otherwise possible. The knowledge encompassed in these various data sets would be nearly inaccessible to the casual, or even moderately interested viewer, if it was not visualized. But a good visualization gives access to that knowledge, and does so quickly, efficiently, and effectively. The data visualization tool to use depends on the nature of the data set and its underlying structure. According to (Bansal and Sood, 2011), data visualization tools can be classified into two main categories: a) multidimensional visualizations; b) specialized hierarchical and landscape visualizations.

The most commonly used data visualization tools are those that graph multidimensional data sets. Multidimensional data visualization tools enable users to visually compare among data dimensions. Section 1.3.2 presents a review of the most common methods. Most multidimensional visualizations are used to compare and contrast the values among the different data dimensions in the prepared data set. They are also used to investigate the relationships between two or more continuous or discrete variables in data sets. On the other hand, it can be found specialized hierarchical and landscape visualizations. Hierarchical, landscape, and other specialized data visualization tools differ from normal multidimensional tools in that they exploit or enhance the underlining structure of the data set itself (Soukup and Davidson, 2002). We are most likely familiar with an organizational chart or a family tree. Some data sets possess an inherent hierarchical structure. For example, tree visualizations can be useful for exploring the relationships between the hierarchy levels. This thesis proposes several methods that make possible to visualize the two key aspects of multidimensional and hierarchical visualizations previously mentioned, namely, the relationships between two or more variables, or dimensions, and the underlining structure of the data set making possible to obtain information about the structure of the data. This fact makes possible to extract the relationships among the variables even in different hierarchical levels.

As pointed out in (Bansal and Sood, 2011), most good data visualization allows the user some key attributes:

- Ability to compare data.
- Ability to control scale (look from a high level or drill down to detail).
- Ability to map the visualization back to the detail data that created it.
- Ability to filter data to look only at subsets or sub regions of interest at a given time.

On the other hand, data visualization can roughly be categorized into two applications (Chen et al., 2008a):

- **Exploration:** In the exploration phase, the data analyst will use many graphics that are mostly unsuitable for presentation purposes, yet may reveal very interesting and important features. The amount of interaction needed during exploration is very high. Plots must be created fast, and modifications like sorting or rescaling should happen instantaneously so as not to interrupt the line of thought of the analyst.

- **Presentation:** Once the key findings in a data set have been explored, these findings must be presented to a broader audience interested in the data set. These graphics often cannot be interactive but must be suitable for printed reproduction. Furthermore, some of the graphics for high-dimensional data are all but trivial to read without prior training, and thus probably not well suited for presentation purposes, especially if the audience is not well trained in statistics.

As a result, data visualization is used in a number of places within Data Mining (Bansal and Sood, 2011)

- As a first-pass look at the "data mountain" that provides the user some idea of where to begin mining.
- As a way to display the Data Mining results and predictive model in a way that is understandable to the end user who is not an expert in Data Mining.
- As a way of providing confirmation that the Data Mining was performed the correct way (e.g. to confirm intuitions and common sense at a very high level).
- As a way to perform Data Mining directly through exploratory analysis, allowing the end user to look for and find patterns so efficiently that it can be done in real time by the end users without using automated Data Mining techniques.

1.3.2 High dimensional visualization review

Visualization is strongly connected with data and graphs, as mentioned in the previous section. The main task of graphs in visualization is, thus, the presentation of large amounts of data to the analyst, so that the information obtained by the graphs sharpen their reasoning and facilitate the recognition of structures, patterns, novelties, anomalies, trends or correlations. They also aim to facilitate the comparison among models and the discovery of errors or unexpected details. Reasoning with visual concepts has proven to be very appropriate in this regard for their utility when directly perceiving patterns that, otherwise, could only be discovered through arduous processes.

Due to the increasing size of data sets, both in number of objects and in number of attributes, new techniques focused on high-dimensional data visualization are used.

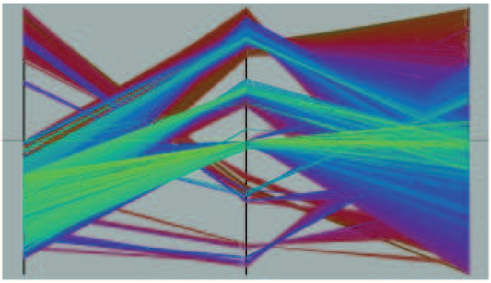
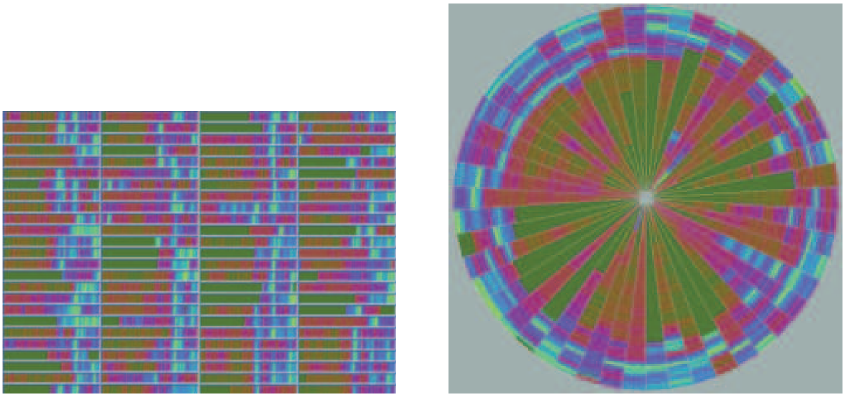


Figure 1.2: *Parallel Coordinate Visualization (from (Keim, 2002)).*

For this reason, high-dimensional data visualization is an active area of research and application.

One of the biggest challenges in high-dimensional data visualization is to find general representations of data that can present the multivariate structure of more than three variables. Different types of graphs as *mosaic plots*, *parallel coordinate plots*, *trellis displays*, and *the grand tour* have been developed over the course of the last three decades (Chen et al., 2008a). Moreover, several visualization techniques were proposed in (Keim, 2001) for high-dimensional data sets, which use various methods to visualize more than three variables simultaneously. These techniques are based on:

- **Geometric transformations, and Andrews curves or parallel coordinates technique** (Keim, 2002), where a point of N-dimensional space is represented by a poly-line. Parallel Coordinate is a technique for visualizing multidimensional data sets (Inselberg and Dimsdale, 1990). Parallel coordinate plots, as described by Inselberg (Chen et al., 2008b), escape the dimensionality of two or three dimensions and can accommodate many variables at a time by plotting the coordinate axes in parallel (Figure 1.2). They were introduced by Inselberg (1985) and discussed in the context of data analysis by Wegman (1990). Informally speaking, this technique lie in assigning to each dimension an axis, and arrange this axes in a parallel way in the plane. Each n-dimensional data $(a_1, a_2, a_3, ..., a_n)$ is a poly-line that crosses the n parallel axes at points $(p_1, p_2, p_3, ..., p_n)$.
- **Dense pixel displays** map each dimension to a colored pixel and group the pixels from each dimension into adjacent areas (Figure 1.3). In general, dense



(a) *Recursive Pattern Technique.* **(b)** *Circle Segments Technique.*

Figure 1.3: *Dense Pixel Displays (from (Keim, 2002)).*

pixel displays use one pixel per data value allowing large amounts of data to be displayed (Keim, 2002).

- **Multiple simultaneous graphics**, which allow effective visual inspection, since it is possible to interpret in a similar way and compare them. An example can be the **matrices of scatter plots**[2] or the well-known **trellis displays** (Richard A. Becker, 1996), which use a lattice like arrangement to place plots onto so-called panels (Figure 1.4). Each plot in a trellis display is conditioned upon at least one other variable. To make plots comparable, the same scales are used in all the panel plots. The simplest example of a trellis display is probably a boxplot x by y. The panel plot is probably the core of a trellis display, in which up to two variables can be plotted in the panel plot (axis variables). In principle, the panel plot can be any arbitrary statistical graphic, but usually nothing more complex than a scatter plot is chosen. A limitation with trellis displays is the fact that all variables besides the axis variables must be categorical.

- **Iconic displays or glyphs** are the use of symbols, or icons, to map an attribute of a multidimensional data set to an attribute of the icon. These icons might include faces (Chernoff, 1973), sticks (Pickett and Grinstein, 1988), color icons (Levkowitz, 1991; Keim and Kriegel, 1994) and geometric shapes (Keim, 2002). Figure 1.5 shows examples of existing glyphs.

[2] collections of points whose coordinates are the values of the variables.

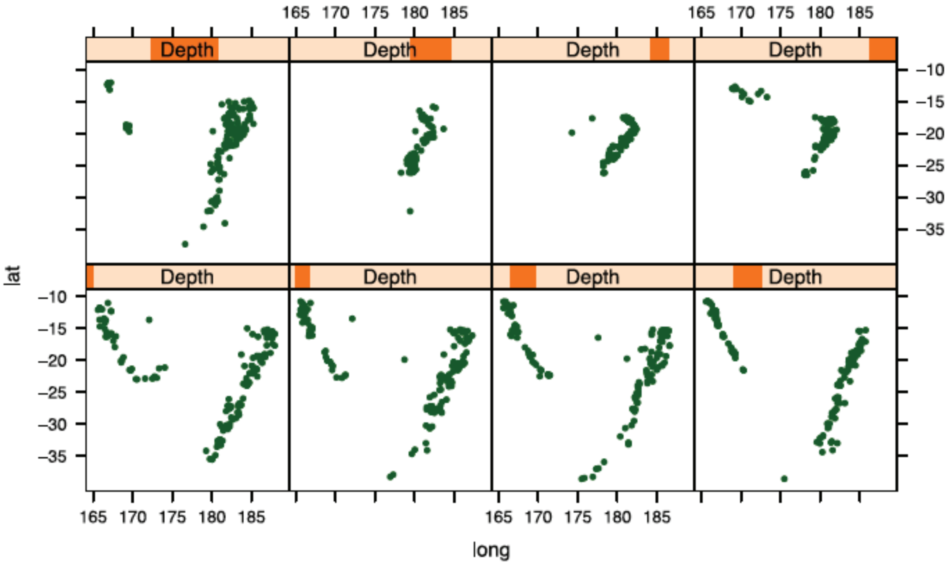


Figure 1.4: *Scatter plots of locations of earthquakes in a trellis framework. The observations in each panel are organized according to the depth at which the earthquake happened (adopted from (Chen et al., 2008a)).*

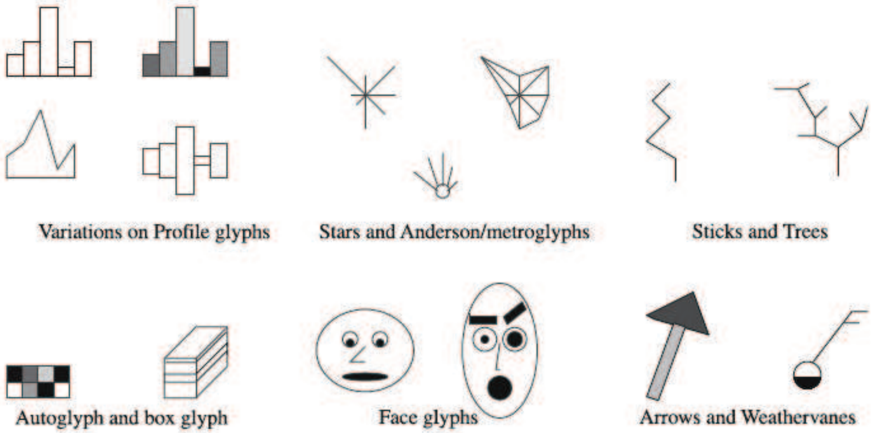


Figure 1.5: *Examples of existing glyphs (from (Ward, 2002)).*

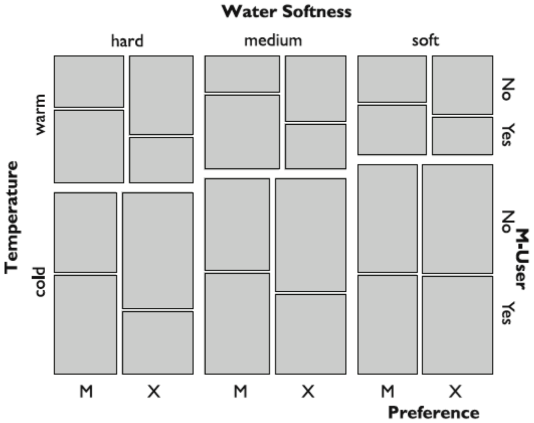


Figure 1.6: *Mosaic plot example of 4-D problem of the detergent data set ((Cox and Snell, 1981)). This data set consists of four variables: Water softness, Temperature, M-user (person used brand M before study), Preference (brand person prefers after test). The major interest of the study is to find out whether or not preference for a detergent is influenced by the brand someone uses.*

- **Mosaic plots** (Hartigan and Kleiner, 1981; Friendly, 1994; Hofmann, 2000) are graphical displays that allow to examine the relationship between two or more categorical variables (Figure 1.6). The mosaic plots begin as a square with side length equal to one. The square is first divided into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable. Then, each bar is split vertically into bars that are proportional to the conditional probabilities of the second categorical variable. If more variables must be used, further separations can be carried out. For the complete understanding of this visualization tool, a lot of training for the data analyst is required.

- **Grand Tour** (GT) of (Buja and Asimov, 1986) is a dynamic data visualization tool that allows the analyst to view hyperdimensional data from all possible angles. The idea is to project the n-dimensional data on a rotated line or a plane (Figure 1.7), and show one- or two-dimensional images of the projections for each time step or angle. (Moustafa and Hadi, 2009).

Although its appealing characteristics are beyond any doubts, the previous multidimensional visual techniques also have problems, e.g., they neither reduce the size

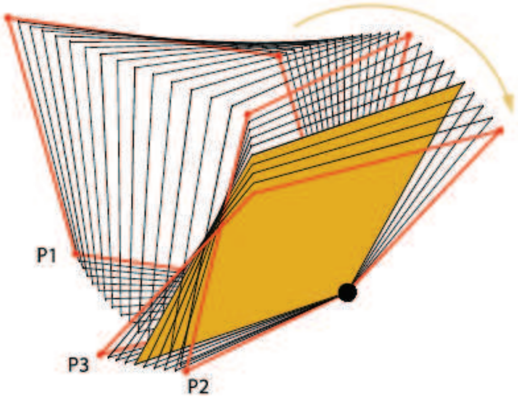


Figure 1.7: *Example path of a grand tour (adopted from (Chen et al., 2008a)).*

nor the amount of data (Kaski, 1997). Therefore, they are not always sufficient to help the analyst to reason with high-dimensional and large size data sets, of which direct representation would remain practically incomprehensible. In any case, this does not imply that these methods are not useful as auxiliary tools.

1.3.3 Visual Data Mining

Due to the inability of the visual techniques pointed out in section 1.3.2, to solve the problem of visualizing in a more intelligent way, it is necessary to use techniques that project the data into a lower dimension in an effective way. However, most of them have not been specifically designed for viewing. These intelligent, or more advanced, techniques correspond to Visual Data Mining, which are, in short, Data Mining techniques focused on visualization (Simoff et al., 2008). Thus, it is necessary to simplify the data set to compress them; for this, machine learning techniques, as vector quantization algorithms (VQ) and clustering techniques among others are used. These algorithms provide an efficient and compact representation of data, providing a set of prototype vectors that minimizes some measure of distortion.

In (Soukup and Davidson, 2002) two broad classes of visualizations are analyzed: data visualization techniques for visualizing data sets and Visual Data Mining tools for visualizing and analyzing Data Mining algorithms. As pointed out in (Soukup and Davidson, 2002), the distinction is as follows:

- **Data visualization tools** can create two- and three-dimensional pictures of data that can be easily interpreted to gain knowledge and insights into those data sets. By visually inspecting and interacting with the two- or three-dimensional visualization, it is possible to identify the interesting (nontrivial, implicit, perhaps previously unknown and potentially useful) information or patterns in the data set.

- **Visual Data Mining tools** assist users in creating visualizations of Data Mining models that detect patterns in data sets that help with decision making and predicting. With Visual Data Mining tools, it is possible to inspect and interact with the two- or three-dimensional visualization, obtained from the predictive or descriptive Data Mining model, to understand (and validate) the information discovered by the Data Mining algorithm.

In both cases, visualization is key in discovering new patterns and trends. However, the term Visual Data Mining, indicates the use of visualization techniques for inspecting, understanding, and interacting with Data Mining algorithms for better comprehension and faster time-to-insight. Unfortunately, not all models produced by Data Mining algorithms can be visualized (or wouldn't make sense to get a visualization). For instance, neural network models for classification, estimation and prediction do not lend themselves to useful visualization.

Summarizing, visualization and Visual Data Mining tools aid in the process of pattern recognition by synthesizing large quantities of complicated patterns into two- and three-dimensional pictures of data sets and Data Mining models. Visualization helps data analysts to discover, quickly and intuitively, interesting patterns and underlying information that lie in the data sets.

Chapter 2

Self-Organizing Maps: Theoretical Framework

Abstract

This chapter discusses the main theoretical aspects of the Self-Organizing Maps (SOMs), an Artificial Neural Network (ANN) used mainly for visualization purposes, due to the fact that it will be used in the next chapter to solve real problems in different fields of research.

2.1 Introduction

Self-organizing maps (SOM) (Kohonen, 1989) are one of the most popular visualization tool nowadays. The SOM is an Artificial Neural Network (ANN) proposed by Teuvo Kohonen in (Kohonen, 1982) and, since then, it has been analyzed and employed extensively. A recent overview can be found in (Kohonen, 2001; Haykin, 2009). An ANN is a computational model inspired by the basics of human brain functioning. In general, an ANN employs connections between its processing units (called *neurons*) for storing the knowledge required to perform some specified task. The fundamental characteristic of an ANN is the ability to learn from the environment and to improve its performance in accordance with a prescribed model that constitutes the learning paradigm (Haykin, 2009).

In contrast to SOMs, classical techniques can only deal with accurate visualizations of whole data sets when the number of features required is equal or lower than three; for a higher number of features to be represented, only projections onto three dimensions can be carried out, establishing restrictions (as keep fixed certain set of variables and representing the rest). Such a restriction leads to a partial representation of the information. Moreover, most of the real data sets are formed by more than three features, making graphical representations difficult. For that type of representation, which will make possible to find patterns in data sets with high dimensionality, the self-organizing maps (SOM) are specially indicated.

In particular, SOMs operate to produce a low-dimensional (typically 2D) representation of high-dimensional data by identifying data that are similar in the input space, and grouping them on a grid (Kohonen, 2001). The most appealing characteristic of SOMs is that the underlying mathematics ensure that the map is a faithful representation of the original data, e.g. two data points are represented close to each other in the resulting map when they have similar features.

The SOM and its variants have been employed very often in a wide variety of domains, such as financial (Deboeck and Kohonen, 1998), medical (Alakhdar et al., 2012), engineering applications (Kohonen et al., 1996; Díaz et al., 2012) and even in the field of animal sciences (Soria et al., 2006; Fernández et al., 2006; Magdalena et al., 2009). This neural model has as mission to find and visualize patterns in N-dimensional data sets where the key working principle is to keep a neighborhood relation between the original space of the N-dimensional data (input space) and the regular low-dimensional grid (output space).

Regarding its structure, a SOM consists of elements of process, called neurons, organized on a regular low-dimension grid (normally in two dimensions), as mentioned above. The number of neurons may vary from a few dozen up to several thousands. Each neuron is presented by a d-dimensional weight vector $m = [m_1, \ldots, m_d]$, where d is equal to the dimension of the input vectors. In its design, the first choice is related to the selection of the map type, and also with the number of neurons selected (as this will define the size of the low-dimensional grid). The first step of this algorithm is the weight initialization, which enables a great number of possibilities. Once the initial values of synaptic weights have been selected, the next step is to get them closer to the optimum values by means of an iterative procedure. This steps are put forward in the next section.

2.2 SOM algorithm

2.2.1 Size and shape

In the SOM design, the main choices are related to the selection of the map type (hexagonal or rectangular grid, which indicates the topology or neighborhood relation) and the number of neurons (this will define the size of the low-dimensional grid). These choices depend on the number of patterns considered, number of variables defining these patterns and, finally, the existing data dispersion (Vesanto et al., 1999).

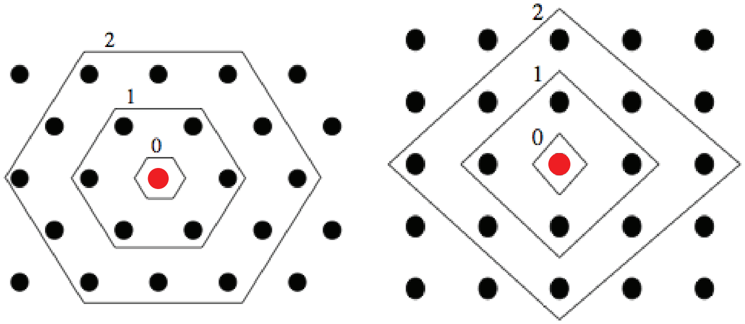
The number of neurons should usually be selected as big as possible, with the neighborhood size controlling the smoothness and generalization of the mapping. As pointed out in (Vesanto et al., 2000), the size of the map (number of neurons) can be selected as the closest integer to $5\sqrt{n}$, where n is the number of training samples. The mapping does not considerably suffer even when the number of neurons exceeds the number of input vectors, if only the neighborhood size is selected appropriately. However, as the size of the map increases e.g. to tens of thousands of neurons the training phase becomes computationally impractically heavy for most applications.

A SOM is formed of neurons located on a regular, usually 1- or 2-dimensional grid. Also higher dimensional grids are possible, but they are not generally used since their visualization is much more problematic. The neurons are connected to adjacent neurons by a neighborhood relation dictating the structure of the map. In the 2-dimensional case the neurons of the map can be arranged either on a rectangular or a hexagonal lattice, see Figure 2.1.

If the sides of the map are connected to each other, the global shape of the map becomes a cylinder or a toroid, see Figure 2.2.

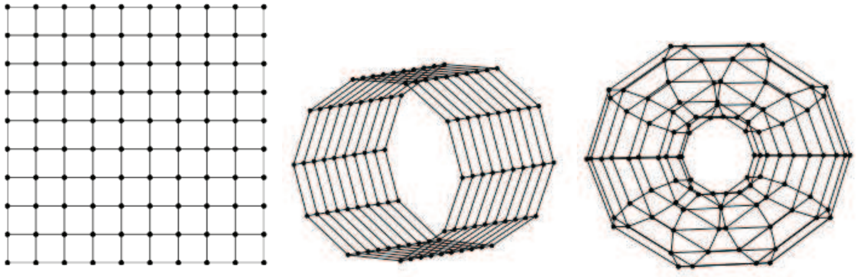
If possible, the shape of the map grid should correspond to the shape of the data manifold. Therefore, the use of toroid and cylinder shapes is only recommended if the data is known to be circular.

The use of hexagonal lattice is usually recommended, because then all 6 neighbors



(a) *Hexagonal grid* **(b)** *Rectangular grid*

Figure 2.1: *Neighborhoods (size 1, 2, 3) of the unit marked with red dot: (a) hexagonal lattice, (b) rectangular lattice.*



(a) *Sheet* **(b)** *Cylinder* **(c)** *Toroid*

Figure 2.2: *Different map shapes: sheet on the left, cylinder in the center and toroid on the right.*

of a neuron are at the same distance (as opposed to the 8 neighbors in a rectangular lattice). This way the maps become smoother and more pleasing to the eye. However, this is mostly a matter of taste.

2.2.2 Initialization

The neurons are defined by the weights; the weights represent the membership of the neuron to each of the components of the space defined by the input variables, which is known as representation space. Although the SOM is quite robust to different initialization, if it is made correctly, a faster convergence can be achieved. Basically, there are three possible types of weights' initialization (Kohonen, 2001)

- **Random initialization:** The weights are chosen randomly so as to cover the entire space of representation.
- **Sample initialization:** The initial weights are assigned arbitrarily, that is, the weight vectors are initialized with random samples drawn from the input data set.
- **Monotonous Initialization:** The initial weights are assigned according to a monotonically increasing function, usually linear, covering the entire space of representation. As the feature map usually is two-dimensional, generally a principal component analysis (PCA) is carried out to determine the first two principal directions of the input set. Then the weights are increased following the principal directions obtained with the PCA along each of the dimensions of the feature map.

2.2.3 Training

Once the initial values of synaptic weights have been selected, the next step is to get them closer to the optimum values by means of an iterative procedure.

In each training step, one sample vector x from the input data set is chosen randomly and the distances between it and all the weight vectors of the SOM are calculated using some distance measure. The neuron whose weight vector is closest to the input vector x is called the *Best-Matching Unit (BMU)*, denoted here by c:

$$\|x - m_c\| = min_i \left\{ \|x - m_i\| \right\}, \tag{2.1}$$

where $\| \quad \|$ is the distance measure, typically Euclidean distance, which is denoted as follows:

$$\|x - m\|^2 = \sum_{k \in K} (x_k - m_k)^2, \tag{2.2}$$

where K is the set of known variables of sample vector x, and x_k and m_k are the k^{th} components of the sample and weight vectors, respectively.

After finding the BMU, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space. The topological neighbors of the BMU are treated similarly. This adaptation procedure stretches the BMU and its topological neighbors towards the sample vector as shown in Figure 2.3.

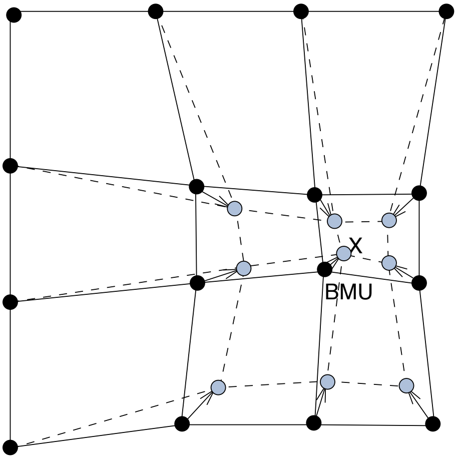


Figure 2.3: *Updating the best matching unit (BMU) and its neighbors towards the input sample marked with x. The solid and dashed lines correspond to the situation before and after updating, respectively.*

The SOM update rule for the weight vector of unit i is:

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)], \tag{2.3}$$

where t denotes time. The $x(t)$ is an input vector randomly drawn from the input data set at time t, $h_{ci}(t)$ the neighborhood kernel around the winner unit c and $\alpha(t)$ the learning rate at time t.

By means of this competitive learning dynamics, the neurons of the SOM end up covering the input space in such a way that the neighborhood relationships are mostly preserved (that is, close regions in the output map correspond to close regions in the input space). As typically the number of neurons in a SOM is smaller than the size of the input dataset, at the end of the training process a given neuron will be the BMU for a group of input vectors – that is, that neuron weight vector is the prototype that sums up the common features of the data falling into that region of the input space.

Once the map training is finished, the visualization of the two-dimensional map provides qualitative information about how the input variables are related to each other for the data set used to train the map.

2.2.4 Learning Rate

The pace of learning, that is, the speed with which the weights change, is determined by a parameter $\alpha(t)$, which is known as learning rate or adaptation rate. Depending on the application may be interesting that this speed is variable and dependent on the number of iterations. In particular, it is often desirable that the learning rate has a higher value at the beginning of the training than at the end. Thus, the learning is fast initially, and as the iterations go by, and the map is learning the underlying information in patterns, learning rate is slowing down to avoid the network to become unstable. Some of the most common expressions to determine the change in the constant learning are (Vesanto et al., 2000):

- **Exponential.** The expression of the parameter that controls the speed of learning is given by:

  $$\alpha(t) = \alpha_0 \left(\frac{\alpha_{end}}{\alpha_0} \right)^{\frac{t}{T}} \tag{2.4}$$

  where t is the current iteration and T the total number of iterations, while α_0 and α_{end} refer to the initial and final learning rate, respectively.

- **Inversely proportional.** In this case, we have the expression:

  $$\alpha(t) = \frac{A}{t + B} \tag{2.5}$$

being A and B constants that determine the initial and final learning.

- **Linear.** For the linear case, we have the next expression:

$$\alpha(t) = \alpha_0 \left(1 - \frac{t}{T}\right) \tag{2.6}$$

where t is the current iteration and T the total number of iterations.

In any case, the learning rate often takes values between 0 and 1, as it can be seen in Figure 2.4, where the learning rate for the different methods is represented.

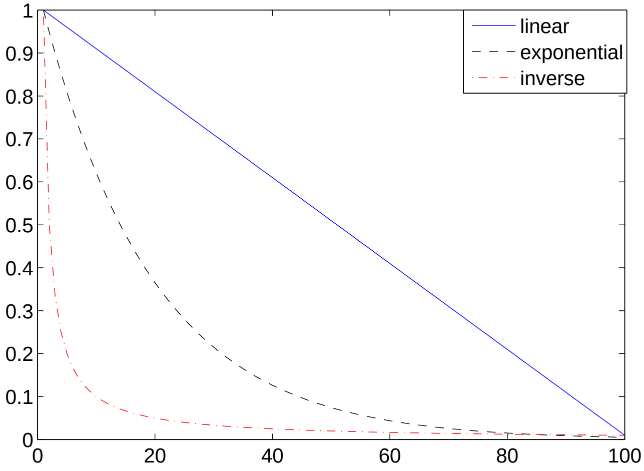


Figure 2.4: *Different learning rate function. In blue solid line the linear function, in black dashed line the exponential function and in red dot-dashed line the inversely proportional function.*

2.2.5 Neighborhood function

Neighborhood kernel is a non-increasing function of time and of the distance of unit i from the winner unit c. It defines the region of influence that the input sample has on the SOM. The neighborhood function determines how strongly the neurons are connected to each other; and it depends on a radius. Since large neighborhood radius makes the SOM more rigid, it is usually used in the beginning of training, and then it is gradually decreased to a suitable final radius, as occurred with learning rate. A suitable final radius is, for example, one.

The training is usually performed in two phases. In the first phase, relatively large initial learning rate and neighborhood radius are used. In the second phase both learning rate and the neighborhood radius are small right from the beginning, as mentioned previously. If the linear initialization procedure is used the first training phase can be skipped. This procedure corresponds to start tuning the SOM approximately to the same space as the input data and then fine-tuning the map.

The total training time, or the number of samples presented to the SOM, is an important consideration. The number of training steps should be at least 10 times the number of map units (Vesanto et al., 1999).

The more common neighborhood functions are (see Figure 2.5) (Vesanto et al., 2000):

- **Bubble function.** This is the simplest neighborhood function. It is constant over the whole neighborhood of the winner unit and zero elsewhere:

$$h_{ci}(t) = \begin{cases} 1 \text{ if } \|r_c - r_i\| \leq \sigma_t \\ 0 \text{ if } \|r_c - r_i\| > \sigma_t \end{cases} \tag{2.7}$$

  where σ_t is the neighborhood radius at time t and $\|r_c - r_i\|$ is the distance between map units c and i (BMU) on the map grid.

- **Gaussian function.** In this case, the function neighborhood has the shape of a Gaussian centered on the winning neuron. The neighborhood radius controls the width of the Gaussian, that is, determines its standard deviation:

$$h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2} \tag{2.8}$$

  where σ_t is the neighborhood radius at time t and $d_{ci} = \|r_c - r_i\|$ is the distance between map units c and i (BMU) on the map grid.

- **Cut Gaussian function.** In this case, the neighborhood function is a mixture of the two previous:

$$h_{ci}(t) = \begin{cases} e^{-d_{ci}^2/2\sigma_t^2} \text{ if } \|r_c - r_i\| \leq \sigma_t \\ 0 \text{ if } \|r_c - r_i\| > \sigma_t \end{cases} \tag{2.9}$$

- **Epanechnikov function.** This neighborhood function is given by the expression:

$$h_{ci}(t) = max\{0, 1 - (\sigma_t - d_{ci})^2\} \tag{2.10}$$

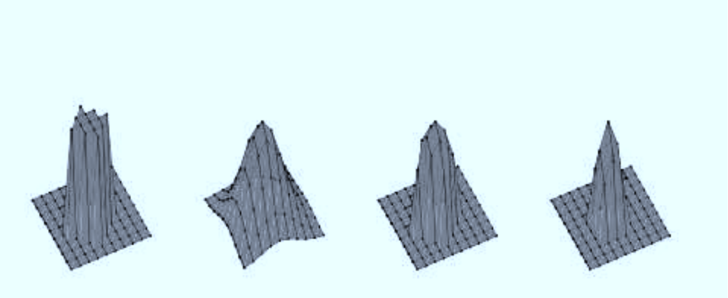


Figure 2.5: *Different neighborhood function. From the left* "Bubble", "Gaussian", "Cut Gaussian", "Epanechnikov".

In summary, the main characteristics of SOM are:

1. The mapping carried out by the SOM is non-linear, much more powerful than classical linear methods.

2. SOM has the ability to preserve topological relationships; i.e. input patterns that are similar in the original high-dimensional data space are mapped close in the *component planes* provided by the SOM visualization.

2.3 SOM visualization

2.3.1 Component planes

After the training process of a SOM, it is very easy to project the map onto the different features; these projections are called *component planes* (Kohonen, 2001), also known as component maps. The component planes can be plotted, so that information regarding each single variable can be visualized allowing the most complete visualization of reality, making the detection of relationships among the different analyzed variables possible. A component plane of a SOM is a map where, for each neuron, only one component of its weight vector (corresponding to a given input variable) is shown (see for instance Figure 3.1 in next chapter); thus in the experiments it can be displayed a total of N component planes (where N is the data dimension or the number of variables). Therefore, input patterns which are mapped onto a certain area of a component plane maintain the graphical positions in all the other component planes. Since all the component planes actually belong to the same map, a certain area of the map can be analyzed for the different features at the same time.

In the component plane i, each neuron in the SOM grid is colored based on the value of the i-th component of its weight vector; higher values are usually depicted in red and lower ones in blue. Thus, component planes tend to show some parts of the map with a similar color; it means that similar input vectors are clustered in those particular parts of the map, that is, similarly colored neurons within a plane represent a set of input vectors that are similar according to that specific variable. The same region within every plane (say, the upper left corner) identifies the same set of records, but different planes focus on a different variable. This kind of visualization allows comparing values over different features: for instance, one could easily identify regions of the input space where a given variable i_1 takes large values (a red region in plane i_1) whereas variable i_2 has small values instead (a blue region in plane i_2 – negative correlation). Therefore, the comparison between different planes allows for an intuitive grasp of the relationships existing between the variables under study. While other classical representations need to set up some thresholds to emphasize different profiles, SOM can work with continuous variables and profiles are shown with color gradient, as mentioned above. Each component plane is shown next to a colored bar which gives information about the relationship between the color and the corresponding numerical value. In order to better understand the SOM technique it can be made a simple comparison with the "social geography maps", representing the characteristics of residents in a given area (Ball and Petsimeris, 2010). These "social geography maps" are normally introduced by density population maps (in this case similar to the map of winners neurons, see next section) and then a list of "social geography maps" represents other characteristics of the residents (input vectors). These characteristics are features like income, life expectancy, education, etc. (variables that define the problem). Finally, to complete the comparison between SOM and "social geography maps", the distribution on the map of residents remains constant and the representation of the characteristics is realized through a color gradient (like the component planes of the SOM technique).

2.3.2 Winners map

A "Winners map", also known as "Hits map" or "BMU map", represents the response of the given data on Self-Organizing Maps (Vesanto et al., 2000). Traditionally the response is shown on the map by showing the BMU (Kohonen, 2001), so the map presented has the same size that the component planes of the SOM. "Winners map" are markers showing how many times each map unit, or neuron, was the BMU for each input register so that the distribution of the best matching units for a given data set is represented, and therefore which regions of the SOM contain more data points. Figure 2.6 shows an example of a conventional "Winners map", where each neuron is represented by an hexagon on the map grid. The black-filled area inside each hexagon is proportional to the number of input patterns which are most similar to this neuron. This gives a quantitative idea of the input vectors belonging to each neuron so that the largest neurons host the most of the records while the smallest ones denote those regions of the SOM that are record-less, but not always least important. In some cases, areas of the map that represent a low number of input vectors should not be neglected if the goal is related with the identification and knowledge extraction of minority groups as, for example, in Chapter 3 (Section 3.2), where the goal is to identify profiles of patients particularly dissatisfied with some of the issues related to the treatment. In this case, it should be taken into account that the number of patients analyzed is low, and it does not represent a standard patient profile. Thus, the BMU map supports the interpretation of the component planes as it highlights those regions that are mostly worthy of attention when looking for correlations across planes. Moreover, multiple "hits" can be drawn in different colors.

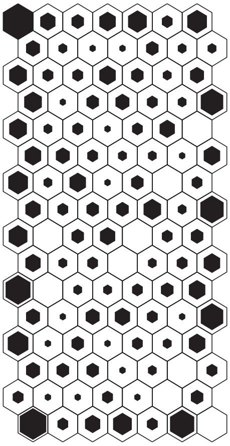


Figure 2.6: *Example of a conventional "Hits map", where each neuron is represented by a hexagon on the map grid. The colored area inside each hexagon is proportional to the number of input patterns which are best represented by this neuron.*

This makes possible to compare the different patterns associated to different classes (in a supervised problem) by the distribution of their "hits" on the map. As shown in Chapter 3 (Section 3.3), these maps provide more information than those obtained by simple labeling.

Chapter 3

Visual Data Mining with Self-Organizing Maps (SOMs)

Abstract

This chapter illustrates the usefulness of the visualization techniques on real problems analyzed by the author of this thesis during the last years. In particular, it shows the use of the well-known Self-Organizing Maps (SOMs) to solve real problems in different fields of research, providing easily interpretable solutions to complex problems thanks to the intuitiveness of the SOMs that would otherwise have been very difficult to solve. The first problem to address is about the study of Balanced Scorecard (BSC), which is a validated tool to monitor enterprise performances against specific objectives. Herein the use of SOMs is proposed as an innovative approach to extract information/knowledge from the BSC data and to present it in an easy-readable informative form. The second problem uses SOMs to evaluate Patients Satisfaction Surveys (PSS), whose evaluation has become an important indicator for assessing health care quality. The aim of this work is to test and validate a methodology for identification of areas of potential improvement for specific patient groups. The third real problem studied in this chapter is framed in the field of cardiology. This study proposes a new methodology in order to obtain visual information among four important groups of patients:VF (Ventricular Fibrillation), VT (Ventricular Tachycardia), HP (Healthy Patients) and

AHR (other Anomalous Heart Rates and Noise) since methods used up to now do not provide insight into the problem (such methods only attempt to classify the different groups of patients). The fourth problem addressed shows the application of SOMs in a physiotherapy problem by means of the valuation analysis of the knee in athletes in the pre- and post-surgery of the anterior cruciate ligament, studying variables of force and measurements at different distances of the knee. Finally, the last study of this chapter proposes the use of SOMs for evaluating data about comfort in footwear provided by Instituto Tecnológico del Calzado y Conexas (INESCOP)

3.1 Use of SOMs for Balanced Scorecard analysis to monitor the performance of dialysis clinic chains

3.1.1 Introduction

Cases of End Stage Renal Disease (ESRD) requiring dialysis treatment are becoming more frequent worldwide (Udani et al., 2011; Schieppati and Remuzzi, 2005). In the particular case of Europe, prevalence and incidence of ESRD are increasing making the kidney disease emerge as a crucial health, social, and economic concern (Lameire et al., 2005). Therefore, governments and healthcare companies must aim to apply adequate strategies that allow the highest quality at the best cost, and that ensure the patient well-being as a mandatory duty. Currently, *Fresenius Medical Care (FME)* European activities involve more than 400 dialysis centers located in 19 countries (Stopper et al., 2007). This wide distribution implies a high variability concerning both the different governments' healthcare systems and the heterogeneity of the dialytic population (de Francisco et al., 2010). To maximize the results, *FME* has chosen a continuous quality improvement strategy which combines clinical enhancements with management benchmarks (Stopper et al., 2007). To address this purpose, a robust warehouse of clinical, operational and financial data was designed in the past years and continuously implemented (Marcelli et al., 2001; Steil et al., 2004). This framework represents the fundamental decision support for the use of a continuous performance monitoring system.

Over the years, several methodologies have been presented for business performance measurement, including techniques based both on reliable mathematical algorithms and on more empirical approaches (Charnes et al., 1978; Seiford, 1996; Kaplan and Norton, 1996; Bourne et al., 2003; Marr and Schiuma, 2003).

Given the complexity of reconciling clinical and financial requirements, *FME* has been adopting since 2007 the *Balanced Scorecard* (BSC) approach in order to successfully align the business strategy of the organization at any level (as extensively described in (Stopper et al., 2007)).

The BSC was first introduced in 1992 as a powerful tool to monitor and align the performance of all the branches of a single enterprise (Kaplan and Norton, 1992). This methodology differentiates from the traditional accounting measures in that it combines financial measures (which give information about actions taken in the past) and operational measures which will drive the future performance. The operational measures include aspects like customer satisfaction, internal processes, and innovations. This gives a balanced view of the general efficiency of a company – not only of its productivity – providing cues for the future plans of business. In general, a BSC identifies main areas of business in which the executives should address the improvement effort of the company and that are usually named *perspectives*. The perspectives have the aim to guide the BSC implementers in the selection of those *Key Performance Indicators* (KPIs) that are crucial for tracking the whole company growth and that, therefore, embrace both financial and non-financial topics. Basically, the use of a BSC implies the definition of strategic KPIs describing the company requirements of efficiency and the evaluation of each of them with reference to a real or hypothetical best standard model. The final score for every KPI draws a picture of the whole company achievements. Over the years, the BSC has been modified to be adopted as a conceptual framework in totally different organizations, including healthcare organizations (Zelman et al., 2003; Inamdar et al., 2002).

In the particular case of *FME*, the use of this technique implies the selection of long-term objectives, clear perspectives, and specific KPIs in conjunction with the respect of both internal requirements and external standards (i.e. European Best Practice Guidelines). According to this framework, each clinic is (on a monthly basis) associated with 29 values, each representing the score of that clinic with respect to one single KPI (see Section 3.1.2 and (Stopper et al., 2007)).

The use of the BSC within *FME* is now fully established as the main instrument for performance monitoring: monthly reports are generated that detail the scores of single KPIs, focusing on groups of clinics based on their geographical area, on single clinics, and even on single patients. The management, based on such reports, can ask for more information on the clinics that show defective performance under some KPI, and take corrective actions if needed; on the other hand, clinics whose KPI scores are in the excellence area may receive incentives. Currently, BSC reports make possible to look at the monthly trend of single KPIs (or perspectives) separately, or alternatively at the whole set of KPIs within the same month. However, there is no simple way for the management to compare data for a group of selected KPIs over a large time window. In particular, one cannot, based on the information provided by standard reports, identify groups of clinics based on their performance on combinations of KPIs – that is, clinics with a well-defined behavior, characterized by correlated scores for a set of KPIs. In other words, the relations existing between KPIs cannot be easily extracted with the currently employed analytical tools. Moreover, even if such information was to be processed with ad-hoc analyses, it would still remain difficult to effectively convey the results to the management, as the nude numbers do not offer an intuitive depiction of relations across KPIs, especially when dealing with high-dimensional data: in this case, "a picture is worth a thousand *numbers*". Thus, there is a need for analytical techniques that can easily extract interesting correlation patterns on groups of KPIs, and at the same time offer an effective visualization of such complex information. To this end, it is proposed the use of *Self-Organizing Maps* (SOMs).

SOMs have already been applied in the healthcare field, for population studies (Basara and Yuan, 2008), clinical diagnosis (Nelson et al., 2004; Makinen et al., 2008), and for organization (Lloyd-Williams and Williams, 1996) or economic considerations (Montefiori and Resta, 2008; Resta, 2011). Herein, in this work it is proposed the application of SOMs to the *FME* BSC for cost-benefit analyses and for company efficiency evaluations. That is, to demonstrate that SOMs, providing a compact and unbiased representation of complex datasets, are a valuable method to complete the analysis performed on BSC with traditional statistics. Indeed, SOMs gave more insights on the role of the different KPIs in driving the clinic performance and highlighted unpredicted relations and dynamics existing among the KPIs.

Beyond its consolidated reliability, the choice of SOM was driven by two main advantages of this technique. First, SOMs enable to summarize large collections of complex data in a compact and easily interpretable graphical representation, as previously mentioned. Moreover, the modeling approach of SOMs is unsupervised, meaning that no *a priori* hypotheses need to be injected by the user. Results are, therefore, data-driven and unbiased. This allows to infer unanticipated relationships between variables to freely emerge. It must be noted that the SOM is by no means the only technique proposed in the literature for performing and data visualization (Lee and Verleysen, 2010): other popular methods include Principal Component Analysis (Pearson, 1901), Sammon's nonlinear mapping (Sammon Jr, 1969), Isomap (Tenenbaum et al., 2000a), and Locally Linear Embedding (Roweis and Saul, 2000), just to name a few of them. However, the simplicity and intuitiveness of the SOM makes it preferable for the aims of the work, as the ability to produce an easily-readable map that immediately conveys crucial information content in the data is a key factor for supporting the use of SOMs in the everyday management practice, which was in fact one of the main goals of this work. Besides the "static" analysis of the SOM (the analysis of the relationships among variables), dynamic analysis of clinics from different countries (Turkey, Italy and Portugal) was carried out. The temporal evolution of single clinics over the considered variables can be analyzed by computing their trajectories on the SOM. Thus, it can be observed if a clinic presents an improving trend or not (depending on the value of its KPIs). Furthermore, it has attempted to predict the future state of the clinics (whether they will improve, worsen or remain stable) using Markov chains (Iosifescu, 2007).

3.1.2 Balanced Scorecard and Key Performance Indicators Definition

The *FME* Balanced Scorecard framework is based on four main perspectives (i.e. relevant topics of business or area of improvement): i) patients, ii) employees, iii) shareholders, and iv) the community. For each perspective, specific quality goals (KPIs) have been defined by FME. For each KPI a target of excellence has been selected, according to healthcare, financial, and managerial guidelines, so that each KPI value can be scored with reference to its excellence target. More precisely, the numeric raw data concerning every single KPI are collected every month (extracted from the *FME* clinical/financial data warehouse) and elaborated, so that a score can be associated to the final value of each indicator. The closer the KPI value is to the target of excellence, the higher its score will be for that month. On the whole, the four perspectives embrace 29 KPIs that selectively describe i) patients' outcomes (i.e. satisfaction, compliance and prolonged life expectancy); ii) personnel qualification and its continuous professional growth; iii) financial control and company development; iv) enterprise's social responsibilities (e.g. energy savings and preservation of the environment). Table 3.1 lists the 29 KPIs grouped based on the perspective they belong to. An exhaustive description of the implementation process of the *FME* BSC is reported in (Stopper et al., 2007).

BSC data are collected for 19 European countries (450 clinics). The present work is focused on data from the Turkish, Italian and Portuguese clinics (46, 30 and 33 clinics respectively) monitored from January, 2008 to April, 2010), which have been chosen as a case study to demonstrate the potentialities of the SOM approach on this kind of data because they represent different standard clinics' profiles.

Table 3.1: *KPIs, and corresponding perspectives, as defined in the FME BSC. Notice that for KPIs marked with a *, high scores will correspond to good scores which in turn are obtained for low raw values on those parameters: for instance, a high score in the HepB infection risk KPI (9th KPI) does not mean that the risk is high, but rather that the performance regarding such aspect is a good one - that is, the infection risk is low.*

Patient perspective	Employee perspective	Shareholder perspective	Community perspective
1. High Flux Dialysis	11. Turnover of Personnel*	16. Treatment growth	22. Accidents to employees (per 1.000)*
2. HDF Online Dialysis	12. Absenteeism*	17. Patient Growth	23. Patient education and support program
3. eKt/V	13. Overtime	18. New Patient inflow	24. ISO 9001 and equivalent
4. Hgb	14. Employee Satisfaction Survey	19. Scheduling Efficiency	25. ISO 14001 Certification
5. Vascular access (native fistula)	15. Training hours	20. Personnel costs*	26. Compliance program
6. Treatment Adequacy		21. Other costs*	27. Contaminated waste: kg per treatment*
7. Reporting Compliance			28. Electricity consumption: kWh per treatment
8. Patient Satisfaction Survey			29. Water consumption: liter per treatment
9. Patients at risk for HepB infection*			
10. Seroconversion HepB-C*			

3.1.3 Clinic data

The data is composed by a set of 29 different Key Performance Indicators (KPIs) measured monthly in each clinic for a period of 28 months.

The SOM algorithm takes, as input data, N-dimensional real-valued vectors $\boldsymbol{x}_i$:

$$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\},$$

where

$$\boldsymbol{x}_i = (v_1, \ldots, v_N) \in \mathbb{R}^N \text{ for } i = 1, \ldots, m$$

In this case, each vector $\boldsymbol{x}_i$ represents a KPI record, that is, it collects KPI scores for one clinic in one month. As mentioned previously, this study is focused on data from the Turkish, Italian and Portuguese clinics, so three different data sets were built. Turkey data set consisted of 46 clinics, each monitored over 28 consecutive months, so a total of $m = 1,288$ KPI records was available. Italy data set consisted of 30 clinics (840 input vectors) and Portugal data set consisted of 33 clinics (924 input vectors). Each input vector contains one score, v_l, for each KPI, with N denoting here the number of considered KPIs; thus, $N \leq 29$ ($N = 29$ when all the KPIs are taken into account). Notice that KPI scores can take values between 0 (worst performance) and 100 (best performance). Before being processed by the SOM algorithm, data were preprocessed so that those KPIs for which no data were available, or those being almost constant (i. e., low standard deviation) were discarded; in fact, a constant KPI means that all clinics in every month perform the same on that parameter, and therefore that KPI cannot be used to distinguish among clinics having different behaviors. The other KPI data were then normalized to mean $= 0$ and standard deviation $= 1$.

3.1.4 Methodology

Use of SOMs for KPIs analysis

In this work, the SOM analysis was carried out in MATLAB using the *SOM toolbox* (Vesanto et al., 2000). In order to ensure that a good quality map is produced, for each analysis 4008 different maps were produced. These differed in the way the weight vectors were initialized (either randomly or linearly), in the type of algorithm used for training (either sequential or batch), and in the neighborhood function employed in weight updates (Gaussian, Cut Gaussian, Bubble, Epanechnikov). Moreover, the random initialization was fulfilled 100 times for each combination of parameters. Regarding the size of the map, the total training time, the learning rate and the neighborhood radius, the default number of these parameters that SOM toolbox documentation considers the most appropriate was selected[1]. For the size of the map (number of neurons), the integer closest to $5\sqrt{n}$ was selected, where n is the number of training samples (Vesanto et al., 2000). For the total training time (or, the number of samples presented to the SOM) the default value is 10 times the number of map units. Learning rate begins from 0.5 in the first phase, and from 0.05 in the second phase. Neighborhood radius starts from $max(mapsize)/4$ and goes down to one fourth of that (unless this would be less than 1). On second phase, neighborhood radius starts from where it stopped in first phase, and goes down to 1. The length of second phase is 4 times that of the first phase. Among these maps, the best one was then selected. The quality of a SOM is typically evaluated based on two error measures: the quantization error, qe, (telling how well the prototypes represent the input data) and the topographic error, te (measuring the degree of preservation of neighborhood relations). Finally, it was decided to choose as final map the one with the minimum topographic error, because this measure gives the percentage of data vectors for which the BMU and the second-BMU are not neighboring map units (Kiviluoto, 1996), that is, it quantifies the topology preservation or the neighborhood preservation. Therefore, importance is given to the fact that similar input patterns are placed close on the SOM.

Once this final SOM has been selected, a number of operations for both visualizing the SOM, and extracting additional information from it, can be performed. First of all, the *component planes* of the SOM can be plotted, so that information regarding each single KPI can be visualized.

Analysis of Clinics temporal evolution

A dynamic analysis of clinics from the three studied countries (Turkey, Italy and Portugal) was carried out. The temporal evolution of single clinics over the considered variables can be analyzed by computing their trajectories on the SOM. The neurons that represent the different clinics (each neuron represents similar temporal instants,

[1] http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml

or similar situations, of the clinics) were identified, using this information to include the dynamic behavior within the SOMs (analyzing the evolution of a given clinic within the SOM). The achieved trajectories using that strategy did not provide a straightforward representation, in addition to be complex due to the high number of transitions they presented. Hence, a clustering of the vectors that define the neurons of the SOM was carried out in order to identify the most relevant areas of the map, and then, analyzing whether a clinic changes between different areas as the time increases or stays in the same area. Usually, 4-5 clusters defined the map sufficiently well (the number of clusters was set to $k = 5$ for Turkey and Portugal, and $k = 4$ for Italy for better readability of the results) as shown in section 3.1.5. This clustering step defines macro-regions over the SOM, which can then be characterized, for instance, as "positive" or "negative" clusters (that is, clusters collecting KPI records having overall high scores vs those characterized by lower scores) based on the weight vector values of neurons falling inside them; in this way, one can get a high-level categorization of larger sets of KPI records. Finally, as the data is temporal in nature (records are taken on a monthly basis), it is investigated how the KPI scores for individual clinics evolve in time. This was done by identifying, for a given clinic, the sequence of its BMUs, and then visualizing that evolution as trajectories superimposed on the SOM, so that temporal trends could be intuitively inferred. The performance evolution for groups of clinics was modeled under a probabilistic framework by resorting to Markov chain properties. These allow a study of the probability of transitioning between performance clusters as time progresses for the identification of the performance level that is expected to become dominant over time.

A Markov chain is defined as a sequence of random variables $\{X_t\}_{t\geq 0}$ taking values on a set of states C, with transition matrix $\mathbf{P}$ (Eq. 3.1) storing the probability of going from state i to state j in one step, for every pair i, j of states in C (Iosifescu, 2007).

$$\mathbf{P} = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} \\ P_{2,1} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ P_{j,1} & \dots & \dots & P_{i,j} \end{bmatrix} \tag{3.1}$$

In our context, states are clusters, and $P_{i,j}$ is computed by counting the number of transitions from cluster i to cluster j collectively occurring in the clinic trajectories and then turning such counts into probabilities so that $\sum_j P_{ij} = 1 \forall i$. The main diagonal corresponds to a stable situation since there are no transitions among the different groups or clusters. In a Markov chain, the Markov property is assumed, i.e., the probability of being in any state of the chain at a given time step t only depends on the state at time $t-1$. For this reason, the probability of going from any state i to any state j in exactly t steps can be easily computed as $(\mathbf{P}^t)_{ij}$; by plotting such transition probabilities as they evolve in time, one can see where the probabilities tend to concentrate as time passes.

3.1.5 Results

The purpose of the present work is to demonstrate the power and the reliability of the SOM technique for the analysis of data concerning dialysis clinic performances by discussing some representative results obtained for *FME* BSC data. To this end, it has been performed the SOM analysis separately for the different countries (Turkey, Italy and Portugal).

It is important to stress that the easy-readable representation of multidimensional datasets provided by the SOMs preserves their main information content. Hence SOMs offer a less complex but highly informative visualization of multivariable data that supports the BSC users and managers in the interpretation of the performance measurements. The main goal of this work is to provide evidence that, given its flexibility and reliability, the SOM methodology can be successfully applied for the interpretation of data coming from the BSCs of any healthcare enterprise, with the potential of becoming a standard analytic tool for efficiency measurement.

As expected, SOMs enabled to highlight both unexpected and more predictable correlations between KPIs. Interestingly, some other correlations that were expected to be found were not confirmed by the SOM analysis, and this underlines that no preliminary assumption biased the outcomes.

Analysis of Italy KPIs

It must be taken into account that for Italy, one KPI was not measured (*Reporting Compliance* [KPI 7]. Moreover, KPIs almost constant (low standard deviation) were discarded, as mentioned in Section 3.1.3. Based on this criterion, the following KPIs were discarded: *High Flux Dialysis* (KPI 1), *Overtime* (KPI 13), *Patient education and support program* (KPI 23), *Electricity consumption* (KPI 28) and *Water consumption* (KPI 29).

The component planes for the relative SOM trained with the remainder KPIs are shown in Figure 3.1. In interpreting these and the following maps, one must keep in mind that a SOM groups similar KPI records (that is, vectors of KPI scores for one clinic in one month) by assigning them to the same neuron in the grid; each neuron is represented by a weight vector that is the prototypical KPI record residing in that neuron. Each component plane of a SOM acts as a separate filter on the SOM itself, showing only information pertaining to one KPI. In other words, the *i*-th component plane shows only the *i*-th component of its weight vector. Although multiple component planes are present, these all correspond to the same SOM, and thus the same hexagon (neuron) in all planes represents the same set of KPI records. Therefore, if a region is found that has a reddish coloring in component plane 1, and bluish coloring in component plane 2, then that region contains KPI records characterized by high scores on KPI 1 and low scores on KPI 2.

In the component planes it can be observed:

- Constant KPIs: the KPIs 2 (*HDF Online Dialysis*) and 12 (*Absenteeism*) are relatively constant with high values. The KPI 26 (*Compliance program*) presents a small area of variation (in the central-right part of the map). This variation is quite low in comparison with the variation observed in the rest of KPIs. In the study of the temporal evolution, the influence of these KPIs would be negligible because its value remain constant.

- Comparing among all KPIs, the KPI 27 (*Contaminated waste*) has the lowest values in the global map; it means that Italian clinics have poor management regarding topic contamination. As it will be seen in the following subsections, the other two studied countries show better values in this KPI; especially Turkish clinics.

- The map is split into two areas according to the KPIs 24 (*ISO 9001 and equivalent*) and 25 (*ISO 14001 Certification*) because they are binary variables, that is, either they are ISO compliant or not; the area corresponding to the upper-left corner is not ISO compliant. It was considered removing them from the

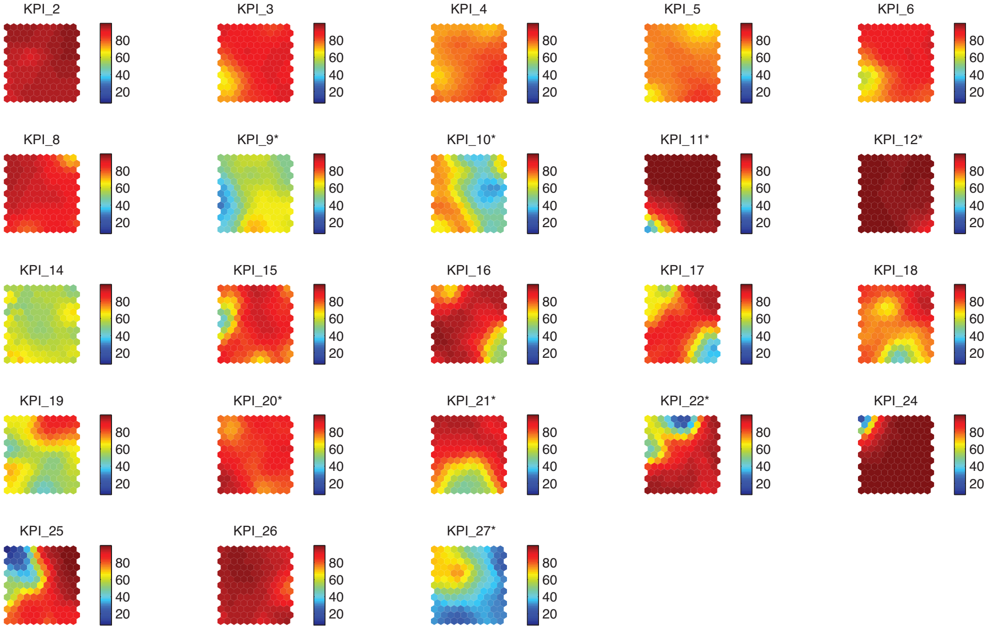


Figure 3.1: *Component planes for the SOM analysis obtained for Italy after training the model without the lost KPI (7) and without the constant KPIs (1, 13, 23, 28 and 29). Notice that color scales are normalized to the same interval, where 100 corresponds to the largest value over all weight vectors, and 0 to the smallest one; therefore, the maximum value for one KPI might not correspond to dark red for all planes. Also notice that for KPIs marked with a * (also in Table 3.1), red regions correspond to good scores which in turn are obtained for low raw values on those parameters: for instance, a red region in the HepB infection risk component (KPI 9) does not mean that the risk is high, but rather that the performance regarding such aspect is a good one - that is, the infection risk is low.*

training and building a separate map, but finally it was thought that including them in the training all the KPIs could be related (if these KPIs were removed the map would have to be analyzed according to 4 options: $KPI\_24 = 0/1$ or $KPI\_25 = 0/1$). On the other hand, the behavior of each one of these KPIs is not completely the same. In Turkish and Portuguese clinics the values of KPIs 24 and 25 are almost equal, that is, they follow the same trend (see the following subsections).

- In general terms, the KPI 9 (*Patients at risk for HepB infection*), KPI 10 (*Seroconversion HepB-C*), KPI 14 (*Employee Satisfaction Survey*) and KPI 19 (*Scheduling Efficiency*) present low values compared to other KPIs (without taking into account the KPI 27, which as above mentioned is the one with the lowest values).

- Correlation between pairs of KPIs are verified. These correlations are given between KPIs 3 and 6 (*eKt/V* and *Treatment Adequacy*), KPIs 16 and 17 (*Treatment Growth* and *Patient Growth*), KPIs 4 and 5 (*Hemoglobin* and *Vascular acces*). As expected, a positive correlation emerged between the scores for *eKt/V* (KPI3) and those for *Treatment Adequacy* (KPI 6), as shown by a strikingly similar distribution of values over the two maps. This direct correlation was predictable since *eKt/V*, also defined as *dialysis dose* (Gotch and Sargent, 1985; Daugirdas, 1993), is one of the main parameters used for treatment efficacy evaluation; then, *eKt/V* and *Treatment Adequacy* KPIs measure different aspects of the same target. This result confirms the reliability of SOMs in extracting relations that actually exist within the dataset. The effectiveness of the SOMs in highlighting real correlations is also stressed in the same panel when the *Treatment Growth* and *Patient Growth* maps are compared (KPIs 16 and 17). A predictable strong correlation between the two KPIs is, in fact, showed since an increase of treatments usually can be associated with an increase in patients. Another interesting outcome applies to the *Hemoglobin* and *Vascular Access* (KPIs 4 and 5). Indeed, a less sharp but still quite predictable correlation can be seen when comparing these two maps. It has already been suggested that dialysis performed through temporary subcutaneous catheters or grafts raises blood loss, inflammation and infection events, all conditions which tend to deteriorate the patients' anemic status and quality of life (Stevenson et al., 2002; Wasse et al., 2007). Moreover, vascular accesses other than the permanent arteriovenous fistula increase the risk of resistance to the erythropoiesis stimulating agents' therapy (such as recombinant human erythropoietin) as well (Greenwood et al., 2003; Goicoechea et al., 2001). Ultimately, this results in poor hemoglobin plasma levels.

- After this correlation assessment, it can be further examined the maps in Figure 3.1 by analyzing the overall information therein contained. Focusing on the lower left corner of the maps, it can be noticed that for the lowest values of *eKt/V* and *Treatment Adequacy* (KPIs 3 and 6) highest *Patient Satisfaction* values can be found (KPI 8). This surprising outcome could be explained by the fact that the goodness of the treatment implies conditions (such as longer or more frequent sessions) that tend to decrease the patients' compliance. This strongly suggests that the patients do not always perceive an effective treatment as the key factor for a satisfactory quality of life.

- Comparing KPIs across the two different perspectives (patient and shareholders), it is of interest underlining that both the *Patient Growth* and the *Treatment Growth* (KPIs 16 and 17) tend to be in general independent of medical performances: this suggests that a large fraction of the considered clinics is able to deal with an increased work load while keeping high medical standards. Moreover, these two KPIs are correlated with KPI 20 (*Personnel costs*). It can be seen that the component plane corresponding to *Personnel costs* follows the same trend, but in a smoother way; that is, when this KPI takes low values, the same occurs for the other two KPIs. This makes sense because an increasing number of patients and treatments often lead to increased costs, and it may also entail increased personnel costs in many cases. On the other hand, a high rate of new patient income (KPI 18; see left side of the component plane) corresponds in some cases to a higher risk of hepatitis B infection (KPI 9; note that high risk is indicated by low KPI scores, blue colors), and to suboptimal scores in the Hemoglobin KPI (KPI 4). It is is clearly noted that low values in KPI 18 (*New patient inflow*, see the bottom of component plane) entails high values in KPI 9 (*Patients at risk for HepB infection*), demonstrating the previous statement. These partial inverse correlations find a possible explanation in that when new patients are accepted in a *FME* clinic to start the therapy they might display conditions that require time to be corrected.

Analysis of the Italian clinics evolution

In order to check the temporal evolution of the clinics (in total 30 clinics for Italy), the trajectories of the prototypes corresponding to each clinic inside the SOM map have been plotted. Each trajectory represents the the current value of the KPIs of a given clinic at several temporal instants (the current value of each KPI can be read by projecting the points of the trajectory over each KPI map of Figure 3.1). Figure 3.2 shows some of the obtained trajectories. Each trajectory starts at the smallest green point, and finishes at the biggest one. It must be emphasized that these trajectories can be superimposed onto any component of the obtained map (Figure 3.1).

Due to the complexity of some trajectories and to their numerous transitions, it was decided to apply the *k-means* algorithm (Forgy, 1965) in order to separate different areas (*clusters*) inside the map. These areas or clusters are represented in Figure 3.3a. Moreover, to facilitate an intuitive understanding of the meaning of a trajectory, it can be useful to project it on the SOM after this has been submitted to a clustering step. Figure 3.3b reports the trajectory of one clinic as an example; in this case, four clusters were found on the SOM. The considered clinic had an almost stable behavior over time, as its KPI records remained inside one cluster; in other words, its performance did not change much during the 28-month window that was considered.

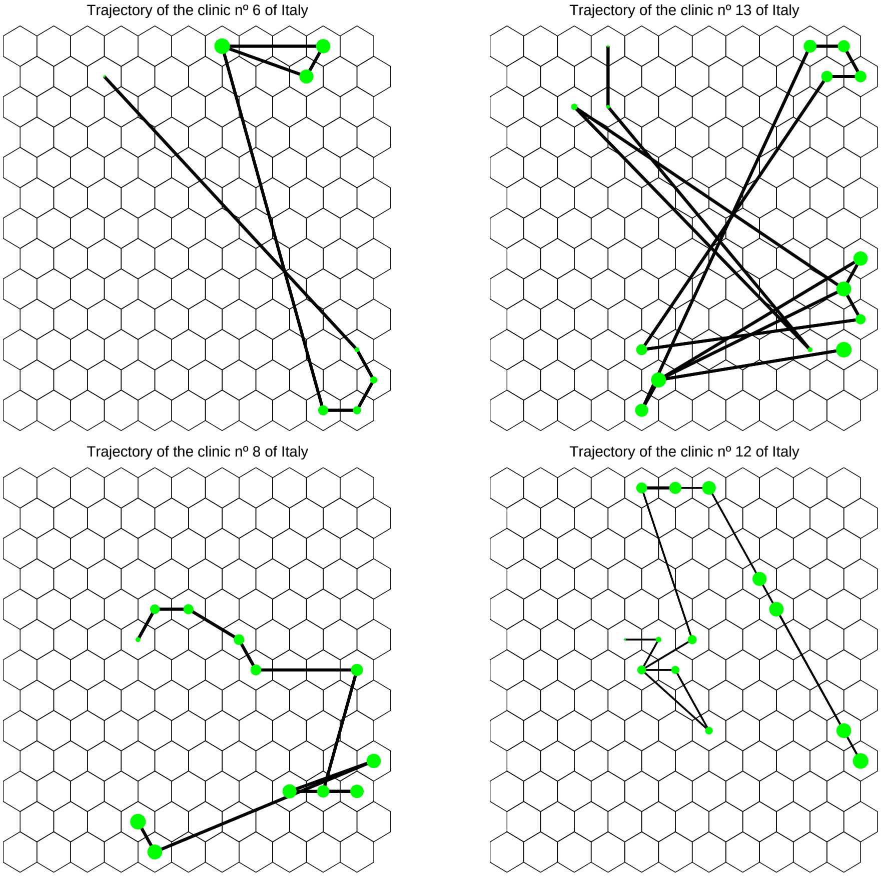
The number of clusters was chosen according to these reasons:

1. Avoiding a too high number of clusters since it entail to complicate the projected trajectories.

2. Special or relevant areas should be marked inside the map.

Actually, four clusters were selected. If the different areas obtained in Figure 3.3a are compared with the KPIs component planes (Figure 3.1), the next conclusions about the clustering are obtained:

1. **Cluster 1.** Upper-right corner; it is the cluster labeled as “good”. It contains ISO compliant clinics (KPIs 24 and 25 equal to 100), which in general terms have the highest KPI values except in the KPI 27 (*Contaminated waste*), which presents low value. The KPIs 4, 5 and 9 (*Hemoglobin*, *Vascular access* and



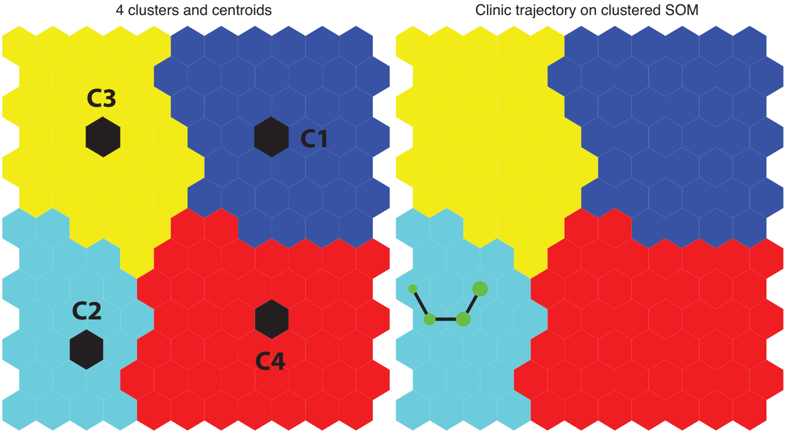
Trajectory of the clinic nº 6 of Italy

Trajectory of the clinic nº 13 of Italy

Trajectory of the clinic nº 8 of Italy

Trajectory of the clinic nº 12 of Italy

Figure 3.2: *Trajectories of several clinics of Italy throughout the time in the SOM grid. Each trajectory starts at the smallest green point and finishes at the biggest one.*



(a) *Clustered map*

(b) *Superimposed trajectory on clustered map*

Figure 3.3: *Clustering obtained after applying k-means algorithm (a) and clustered map together with one clinic trajectory (b) in the SOM map for Italy. Black hexagons represent the cluster centroids or prototypes for each cluster.*

Patient education and support program) present medium values for the clinics in this cluster.

2. **Cluster 2.** Lower-left corner; it groups clinics that are also ISO compliant but have the lowest values of KPIs 3, 5, 6, 9 (this four KPI belongs to the patients group: *eKt/V*, *Vacular access*, *Treatment adequacy* and *Patients at risk for HepB infection*), 11 and 27 (*Turnover of personnel* and *Contaminated waste*). The KPIs 10, 14, 19 and 21 (*Seroconversion HepB-C*, *Employee satisfaction survey*, *Scheduling efficiency* and *Other costs*) present medium values for the clinics in this cluster.

3. **Cluster 3.** Upper-left corner; it represents clinics that are not ISO compliant and have intermediate values in the KPIs 16, 17, 19 20, 21, and 27 (*Treatment growth*, *Patient growth*, *Scheduling efficientcy*, *Personnel costs*, *Other costs* and *Contaminated waste*).

4. **Cluster 4.** Lower-right corner; it contains clinics that are ISO compliant, but whose KPIs 16 and 17 (*Treatment growth* and *Patient growth*) present the lowest values. They also have intermediate values of KPIs 9, 10, 14, 19 and 21

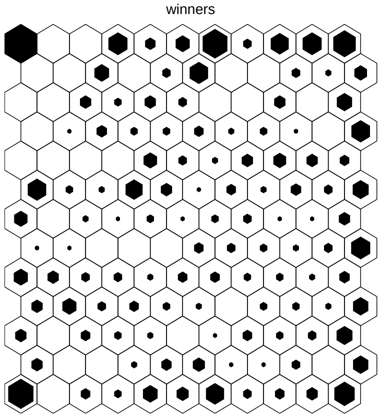


Figure 3.4: *SOM grid of Italy showing the winner neurons after training.*

(*Patients at risk for HepB infection*, *Seroconversion HepB-C*, *Employee satisfaction survey*, *Scheduling efficiency* and *Other costs*); and low values of the KPI 27 (*Contaminated waste*).

According to the distribution of the different records (840, one for each clinic at each month) it can be observed by means of the winners map (Figure 3.4) that they spread out in the map homogeneously. The winners map shows the BMU (best matching unit) for each input record, where the black area in each neuron is proportional to the number of BMUs in this neuron (as mentioned in Chapter 2).

In order to gain more insight about each one of the previous clusters, its different "prototypes" or cluster centroids (represented as black hexagons in Figure 3.3a) are represented in Figure 3.5 by means of the well-known parallel coordinates plot. This plot represents separately, for each cluster centroid, the value of each one of its components (KPI score values). The Figure 3.5 represents clearly the difference among the clusters prototypes for the several KPIs. It can be observed that some KPIs (2, 8, 11, 12, 13, 14, and 26) have no influence in distinguishing the prototypes (they present similar values for all clusters). The biggest differences between the group 1 (cluster labeled as "good", as discussed below) and the rest of groups are in KPIs 18, 19 and 22 (*New patient inflow*, *scheduling efficiency* and *Accidents to employee*), which indicates the key role of these KPIs. This KPIs are not related to medical perspectives, but they are related to shareholder and community perspectives.

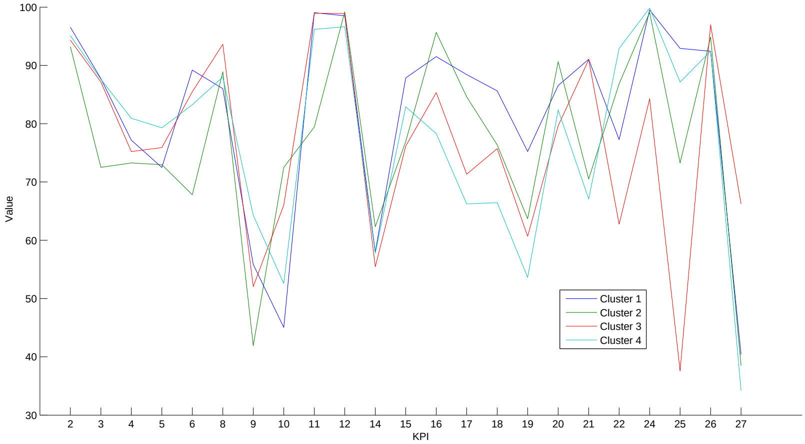


Figure 3.5: *Parallel coordinates plot of Italy centroids.*

It was considered to assign a label to each one of the clusters in order to find out what entails qualitatively that one clinic moves from one cluster to another. For this purpose, a global score for every cluster centroid was determined. This value is computed taking into account the number of measured KPIs and their relative weight (according to the relative importance assigned to each KPI by FME). In other words, only the measured KPIs are being used to normalize and obtain the score value. Due to the normalization in the computation of this parameter, the dependence with the number of the measured KPIs is removed, allowing a fair comparison among countries with a different number of measured KPIs. Table 3.2 shows the computed score for the prototypes or centroids of the different clusters.

Table 3.2: *Prototypes scores of Italy.*

Italy Centroids Score			
C1	C2	C3	C4
81.91	77.06	78.16	77.67

As it can be observed in Table 3.2, the highest score is obtained for cluster 1 (81.91%) whereas the score for the other clusters are more similar among them (they are all in the range around 78%) and also lower than the obtained for cluster 1.

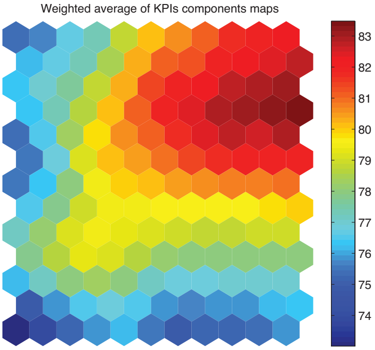


Figure 3.6: *Average of the KPIs component planes of Italy (Fig 3.1) according to the weight established by FME.*

According to this table, cluster 1 was classified as "good" and the rest as "conflictive" clusters. It can be seen that the difference between a "bad" cluster and a "good" one is not so relevant in this country, but this is due to the fact that all the clinics have a very high score, so a difference about 5% is important in this case.

In order to confirm the previous decision it was considered to average all the component planes (KPIs) of the SOM (Figure 3.1), according to the weight established by *FME*. The result is shown in Figure 3.6. Computing the average SOM (that is, a map where each neuron has a color that is proportional to the average of its weight vector, as shown in Figure 3.6) helps characterize the previous clustering as it contains mean values of KPI records corresponding with clinic performance. Additional insights into possible causes of an observed behavior can be obtained by tracing back other relevant features of the considered clinic (e.g. number of patients, clinic operating).

The component planes (Figure 3.1) do show that the potential of the SOM technique applied to the BSC analysis does not exhaust with the information provided by component planes. As previously noted, additional insights on the dataset can be gained by clustering the SOM itself, therefore identifying macro-groups of KPI records that, on average, share similar features. This allows detecting high-level trends in the data, such as groups of clinics that consistently have top performances. This is especially useful when paired with an analysis of the temporal evolution of records: it was shown that it is possible to reconstruct the trajectory of KPI records for one selected clinic over time (in the SOM), so that relative improvements can be tracked. Moreover, this work have been focused on developing a deeper analysis of cluster transitions, in a probabilistic framework, based on such record trajectories by means of Markov chains: the aim is to infer from the observed, past records which behavior can be reasonably expected from clinics in future months. Such prediction might then be used to take corrective actions to make sure that clinics always maintain a high standard of performance. This approach is detailed as follows.

The first part of the above mentioned approach is to compute the transitions matrix (in percentage) (Eq. 3.2), introduced in section 3.1.4, in order to extract patterns of behavior of these transitions. After carrying out the clustering, the transitions of the clinics throughout the time have been reduced to the four areas previously characterized in Figure 3.3a. In this matrix (Eq. 3.2) the term $P_{i,j}$ corresponds to the percentage of transitions from the cluster i to the cluster j with regard to the total number of transitions; and the main diagonal corresponds to a stable situation since there are no transitions among the different groups or clusters, as mentioned previously. For this reason, it is interesting to analyze the clusters labeled as “problematic” (cluster 2, 3 and 4: 2nd, 3rd and 4th rows and column except the term belonging to the diagonal):

$$\begin{bmatrix} 28.889 & 0.741 & 0.864 & 2.346 \\ 0.864 & 213.333 & 0.617 & 0.741 \\ 1.728 & 0.864 & 18.889 & 1.975 \\ 2.840 & 0.864 & 0.864 & 23.580 \end{bmatrix} \tag{3.2}$$

The following conclusions can be extracted from this matrix:

- The highest values of this matrix are found in the diagonal, which means that the transitions among clusters are not the most common scenario.

- The greatest number of transitions between clusters takes place from cluster 4 to cluster 1 (“good” transition) and from 1 to 4 (“bad” transition).

From the transitions matrix it can be determined the transition probability (Markov chains) among states, obtaining the following probability matrix (Eq. 3.3):

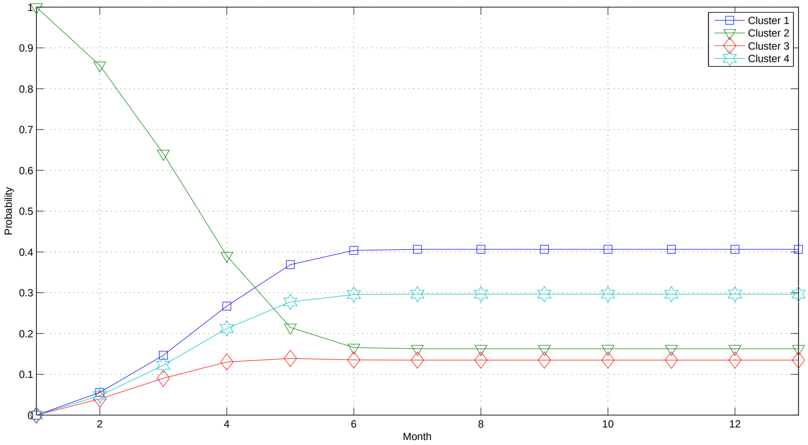


Figure 3.7: *Probabilities of changing from the second cluster to the rest.*

$$\begin{bmatrix} 0.880 & 0.023 & 0.026 & 0.071 \\ 0.056 & 0.857 & 0.040 & 0.048 \\ 0.74 & 0.037 & 0.805 & 0.084 \\ 0.101 & 0.031 & 0.031 & 0.838 \end{bmatrix} \tag{3.3}$$

Using the previous matrix (Eq. 3.3) it is possible to calculate the probability of changing from one cluster to another in the course of time, i.e. the temporal evolution of the clinics. Figures 3.7, 3.8 and 3.9 describe these probabilities starting from clusters 2, 3 and 4 respectively. Despite the starting clusters are labeled as “bad”, the probability of changing to cluster 1 (labeled as “good”) is always the highest starting from the month 5. Figure 3.10 shows the temporal evolution computed starting from the cluster 1. As it can be observed, the highest probability corresponds to remaining in the same cluster without transitioning (remaining in a “good” state). Thus, for Italian clinics it is always most probable improving than worsening since it is most likely to change from any cluster to the first one (labeled as “good”) or remain in it.

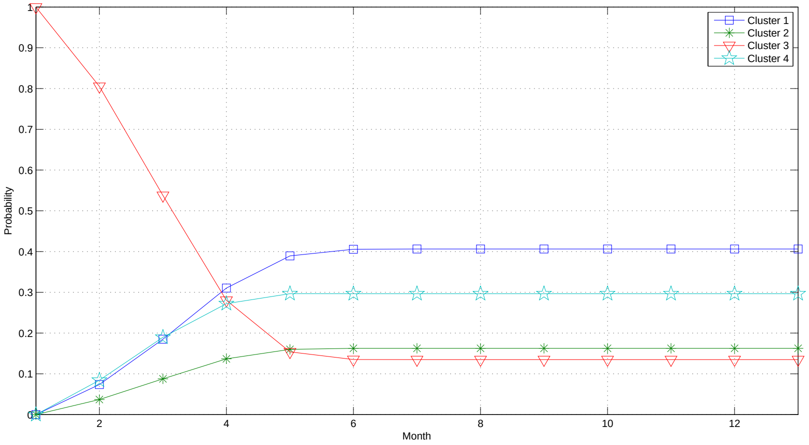


Figure 3.8: *Probabilities of changing from the third cluster to the rest.*

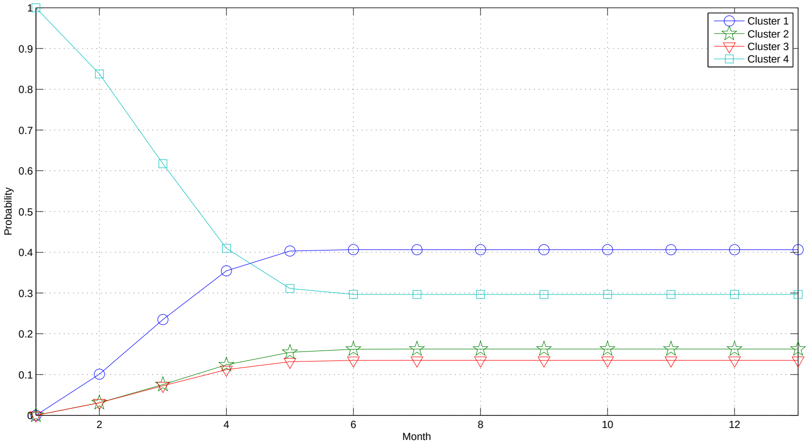


Figure 3.9: *Probabilities of changing from the fourth cluster to the rest.*

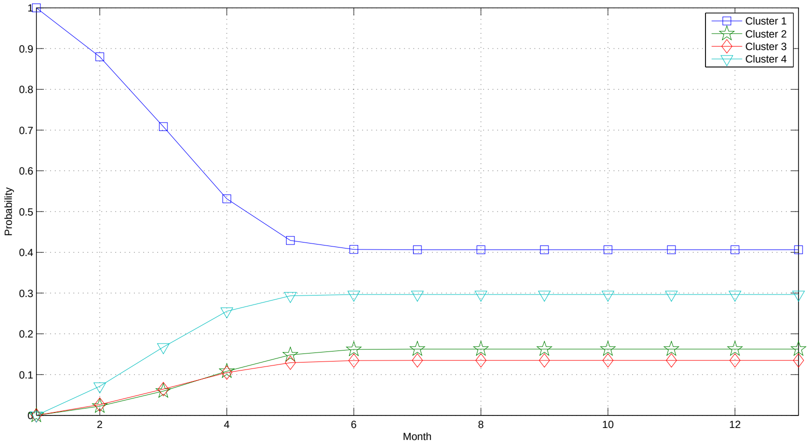


Figure 3.10: *Probabilities of changing from the first cluster to the rest.*

Analysis of Turkey KPIs

As in the case of Italy, the main objective of this study is to analyze with SOM techniques the data corresponding to Turkey in order to extract conclusions about KPIs and clinics. For this purpose the same procedure prescribed in the previous section have been applied.

The first step was to remove the KPIs whose value were almost constant. These KPIs do not contribute any information about the temporal evolution of the clinics. Moreover, information about some KPIs is not available. So finally, the KPIs *Reporting Compliance* (KPI 7), *Employee Satisfaction Survey* (KPI 14), *Absenteeism* (KPI 12), *Overtime* (KPI 13), *Patient education and support program* (KPI 23), *Compliance program* (KPI 26), *Electricity consumption* (KPI 28) and *Water consumption* (KPI 29) were not included in the model. The Figure 3.11 shows the map after training the SOM model with the rest of KPIs.

In this component planes one can observe the following facts:

- Constant KPIs: The map shows several KPIs which remain relatively constant. These KPIs present high values and small variations, because of this they were not removed in the previous step. These KPIs are the 8th (*Patient Satisfaction Survey*), and 10th (*Seroconversion HepB-C*), which take very high values; these KPIs will not have any impact in the study of the temporal evolution since their values are constant in the different areas inside the SOM.

- When comparing the different KPIs among themselves, the number 2 (*HDF Online Dialysis)* is which shows the minimum values of all the considered KPIs. Therefore, certain actions must be considered for this area of business since it has low values for all clinics (whole map).

- The map is split into two areas according to the KPIs 24 (*ISO 9001 and equivalent*) and 25 (*ISO 14001 Certification*) as in the case of Italy, but in this case, the area covered by the clinics that are not ISO compliant is larger. Moreover, the behavior of each one of these KPIs is practically the same. As mentioned previously, in Italian clinics the values of KPIs 24 and 25 did not follow exactly the same trend.

- The lower right corner can be considered as "trouble spot" since it presents low values for KPIs 5, 9, 11, 19, 24, 25 (*Vascular acces*, *Patients at risk for*

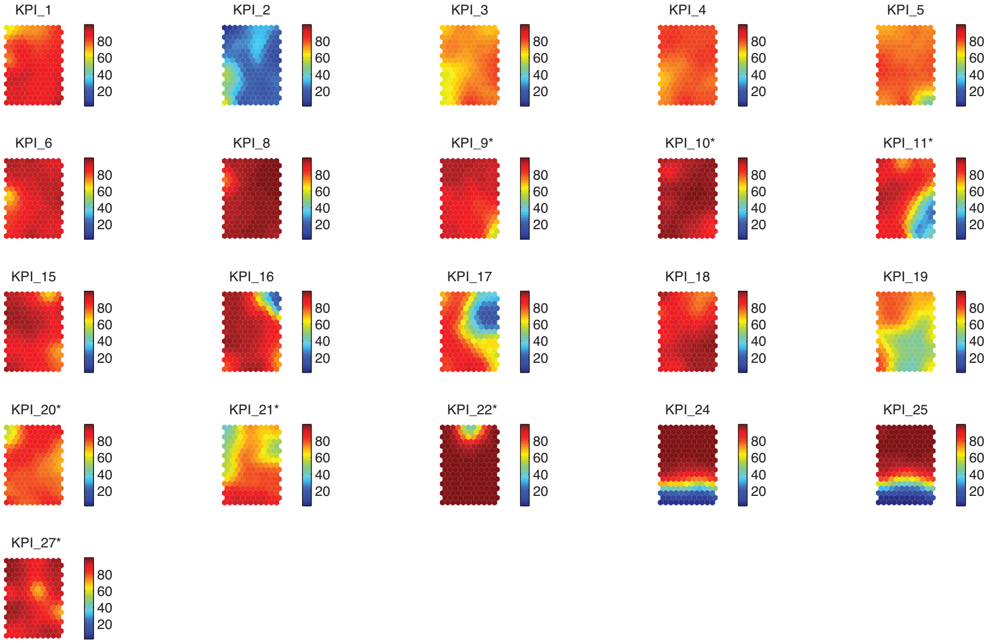


Figure 3.11: *Component planes for the SOM analysis obtained for Turkey after training the model without the lost KPIs (7 and 14) and without the constant KPIs (12, 13, 23, 26, 28 and 29). Notice that color scales are normalized to the same interval, where 100 corresponds to the largest value over all weight vectors, and 0 to the smallest one; therefore, the maximum value for one KPI might not correspond to dark red for all planes. Also notice that for KPIs marked with a * (also in Table 3.1), red regions correspond to good scores which in turn are obtained for low raw values on those parameters: for instance, a red region in the HepB infection risk component (KPI 9) does not mean that the risk is high, but rather that the performance regarding such aspect is a good one - that is, the infection risk is low.*

HepB infection, *Turnover of personnel*, *Scheduling efficiency*, *ISO 9001* and *ISO 14001*) compared with the rest of the map. As it will be seen later, this area can be catalogued as "bad".

- As special cases of KPIs behavior it is found a group of KPIs (5 [*Vascular access*], 9 [*Patients at risk for HepB infection*] and 11 [*Turnover of Personnel*]). This group presents high values in almost the whole map but at the same time it presents a common area with low values (lower right corner); this entail that if one KPI takes a low value then the other two also take low values.

- The KPIs 16 (*Treatment growth*) and 17 (*Patient Growth*) have in common an area with low values at the upper right corner of the map. It can be said that they are correlated, although not in the same degree as in Italy, where this correlation was much more evident since the components of such maps were much more similar. Comparing with the rest of the KPIs, the 16th and 17th do not seem to be correlated in any other area of the map. Thus, it can be said that both the *Patient Growth* and the *Treatment Growth* (KPIs 16 and 17) tend to be in general independent of medical performances, as in the previous section.

- As in the case of Italy, it can be seen that when the KPI 6 (*Treatment adequacy*) decreases its value, the same thing happens for the KPI 3 (*eKt/V*) (left side in the middle of both component planes). As discussed in the study of Italy, this fact is because *eKt/V*, also defined as *dialysis dose*, is one of the main parameters used for treatment efficacy evaluation. In addition, the same applies to the KPI 4 (*Hemoglobin*), it decreases its value on the left side in the middle of such component plane. This is due to the fact that good hemoglobin levels in patients are also indicative of treatment adequacy.

Analysis of the Turkish clinics evolution

Figure 3.12 represents some trajectories of different clinics from Turkey. As in the case of Italy, this trajectories can be superimposed to any component obtained in the previous map (Figure 3.11) to figure out the particular behavior of each clinic for any given KPI.

Again, due to the numerous trajectories observed, the map was divided into several

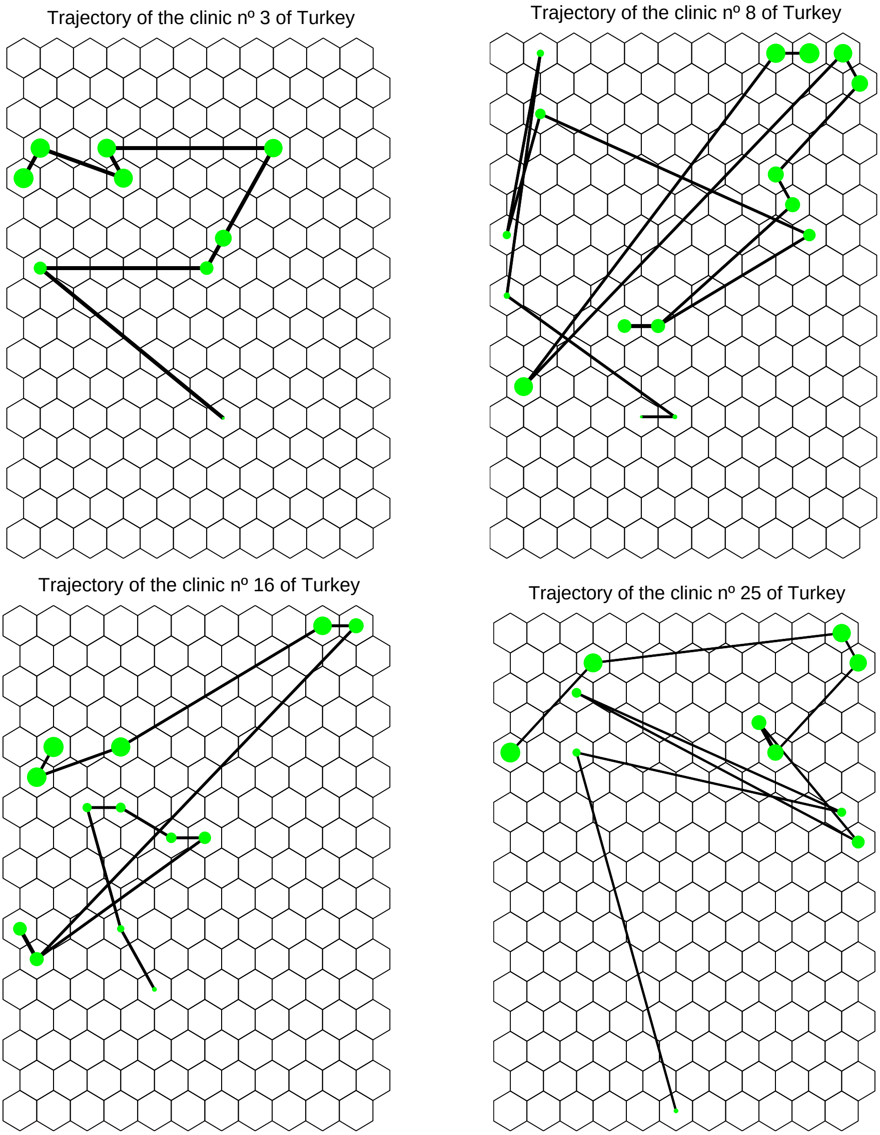


Figure 3.12: *Trajectories of several clinics of Turkey throughout the time over the SOM map. Each trajectory starts at the smallest green point and finishes at the biggest one.*

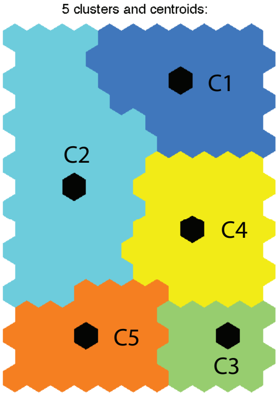


Figure 3.13: *Clustering obtained after applying k-means algorithm in the SOM map for Turkey. Black hexagons represent the cluster centroids or prototypes.*

clusters to make the analysis easier (Figure 3.13). Each cluster was classified in accordance with the global score of its prototype, computed as in the case of Italy. Also in this case it can be useful to project the trajectories on the clustered map as presented in Italy study (Figure 3.3b). The score obtained for the prototypes of the clusters are presented in Table 3.3. As it can be observed all the scores present high values.

Table 3.3: *Prototypes scores of Turkey.*

Turkey Centroids Score				
C1	C2	C3	C4	C5
80.95	83.79	69.45	78.16	77.59

According to the Table 3.3 the cluster 3 was labeled as “bad”. As it can be observed, the lowest score is the obtained for cluster 3 (69.45%). Moreover, the score for the other clusters are in the same range (about 80%) and in all cases higher than the score for cluster 3. It can be seen that the difference between a “bad” cluster and a “good” one is not so relevant, but this is due to the fact that all clinics have a very high overall score. So that, a difference about 10% is important in this case. Hence the cluster 3 was labeled as bad and the rest as good. In order to confirm this

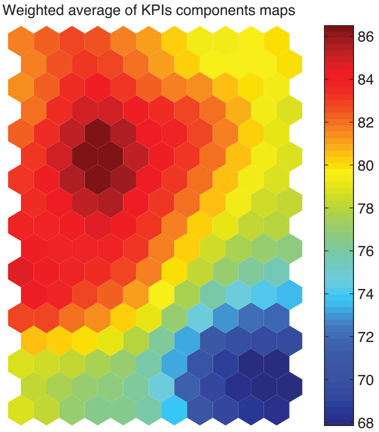


Figure 3.14: *Average of KPIs component planes of Portugal (Figure 3.11) according to the weight established by FME.*

decision it was considered to average all the component planes (KPIs) of the Turkey SOM (Figure 3.11), according to the weight established by *FME*, as in the case of Italy. The result is shown in Figure 3.14.

When comparing the different areas delimited by the clusters with the KPIs component planes, the following conclusions can be reached:

- **Cluster 1.** The clinics in this area are ISO compliant (KPIs 24 and 25 with score 100); this area presents the lowest values of the KPIs 15, 16, 17 and 22 (*Training hours*, *Treatment growth*, *Patient growth*, *Accidents to employee*).
- **Cluster 2.** This is the biggest cluster; the clinics that remain in this area also are ISO compliant (KPIs 24 and 25 with score 100) but they present the lowest values in the KPIs 1, 3, 4, 6 (*High flux dialysis*, *eKt/V*, *Hemoglobin*, *Treatment adequacy*) that correspond to patients perspective, and 20 and 21 (*Personnel costs* and *Other costs*).
- **Cluster 3.** Lower right corner; this is the area described as the worst among the obtained clusters. The clinics in this area are not ISO compliant and they present the lowest values in the KPIs 5, 9 and 11 (*Vascular access*, *Patients at risk of HepB infection*, *Turnover of personnel*). They present medium values in the KPIs 17 and 19 (*Patient growth* and *Scheduling efficiency*).

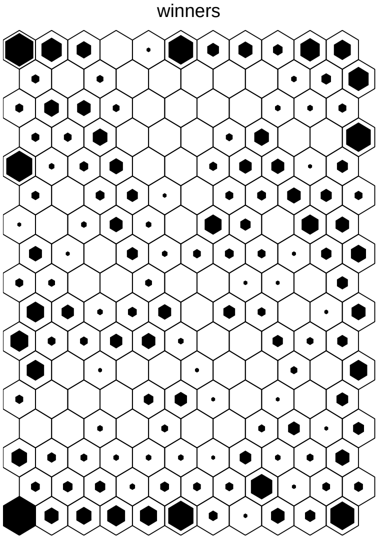


Figure 3.15: *SOM grid of Turkey showing the winner neurons after training.*

- **Cluster 4.** Area that contains clinics that are ISO compliant. If one pays attention to each KPI independently, this cluster present low values in KPIs 11, 17, 19 and 27 (*Turnover of personnel*, *Patient growth*, *Scheduling efficiency* and *Contaminated waste*); but if this cluster is seen globally, the last KPI presents really high values.

- **Cluster 5.** Lower left corner; the clinics in this area are not ISO compliant; if this area is seen in each KPI independently it presents the higher value in the 2nd and 19th KPIs, but this corner also presents the lowest values in the KPI 3.

The different records (in total 1,242) are distributed homogeneously, as shown in the winners map (Figure 3.15).

Parallel coordinates method is used again to represent the prototypes of the clusters, see Figure 3.16. Figure 3.16 represents clearly the difference among the clusters prototypes for the several KPIs. It can be observed that there are some KPIs that do not present differences among the 5 clusters: KPIs 1, 2, 3, 4 (related to patients perspective); and 20 and 22 (*Personnel costs* and *Accidents to employees*). So these

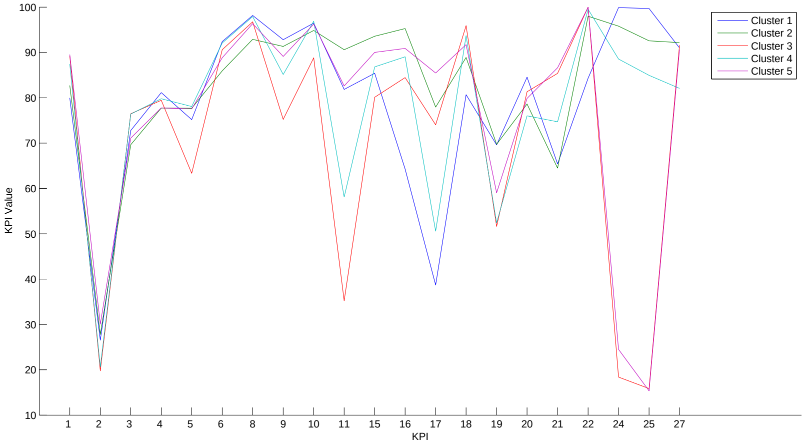


Figure 3.16: *Parallel coordinates plot of Turkey centroids.*

KPIs have not too much relevance in the transitions among clusters. However, there are others KPIs that present important differences among the clusters. These KPIs show a relevant role: for example the KPIs 11 and 17 (*Turnover personnel* and *Patient growth*). Moreover there are groups of KPIs that are similar in some clusters and different in others (for example the KPIs from 21th to 25th).

At this point, the transitions have been reduced to 5 zones (the clusters represented in Figure 3.13). As in the case of Italy, a transition matrix in percentage (Eq. 3.4) has been proposed with the objective of obtaining behavioral patterns of transitions. Due to the fact that the group labeled as bad is the third one, it must be paid attention to:

1. Third row (except the term belonging to the diagonal). This refers to the transitions from the area labeled as bad to an area labeled as good. This means good temporal behavior because the analyzed clinic moves to a better area.

2. Third column (except the term of the diagonal). This refers to the worst temporal behavior because the analyzed clinics move from a good area to a bad area.

$$\begin{bmatrix} 17.63 & 3.46 & 0.08 & 0.81 & 0.00 \\ 2.74 & 25.68 & 0.00 & 2.17 & 0.16 \\ 0.32 & 0.00 & 10.23 & 0.24 & 1.05 \\ 1.77 & 1.77 & 0.24 & 10.87 & 0.00 \\ 0.08 & 0.97 & 1.29 & 0.56 & 17.87 \end{bmatrix} \tag{3.4}$$

The following conclusions can be extracted from this matrix (Eq (3.4)):

- The higher percentages are found in the main diagonal which entail that, usually, there are not transitions among the different clusters.
- The biggest number of transitions between two clusters is produced in the transitions $1 \rightarrow 2$, $2 \rightarrow 1$, $2 \rightarrow 4$.
- The worst transitions (third column) present a low percentage; for this group of transitions, the cluster 5 have most chances of moving towards the third (bad) one.

The probability of transition (Markov chain) among clusters have been computed:

$$\begin{bmatrix} 0.80 & 0.16 & 0.00 & 0.04 & 0.00 \\ 0.09 & 0.84 & 0.00 & 0.07 & 0.01 \\ 0.03 & 0.00 & 0.86 & 0.02 & 0.09 \\ 0.12 & 0.12 & 0.02 & 0.74 & 0.00 \\ 0.00 & 0.05 & 0.06 & 0.03 & 0.86 \end{bmatrix} \tag{3.5}$$

The probabilities of the transition can be analyzed in the Figures 3.17, 3.18, 3.19, 3.20, 3.21. Each one of these figures represents the probability for a clinic of remaining in the current cluster (which starts with probability equal to 1) or change to another one as a function of the time since the beginning of the study.

Figure 3.17 shows the probability of remaining in cluster 3 or changing from the third cluster to another one (For this reason, the initial probability for this cluster is the unit). In this figure it can be also observed that there is an increasing probability

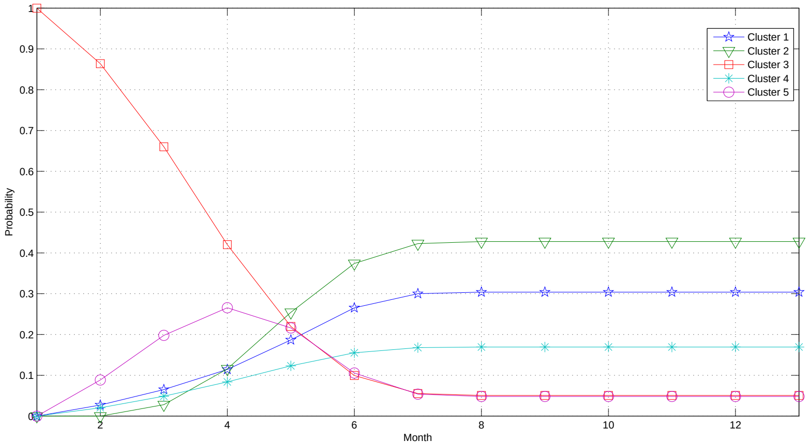


Figure 3.17: *Probabilities of changing from the third cluster to the rest.*

of changing from cluster 3 to another clusters starting from the 5th month; but in the 6th month is when there is an important difference.

It is also interesting to observe the probability of changing from one of the clusters labeled as good to the bad cluster (3). This fact can be observed by looking at Figures 3.18, 3.19, 3.20, 3.21. Noticed that in all these cases there is a greater probability of changing to a “good” cluster (1, 2, or 4) than changing to a “bad” cluster (3). In all the cases, the greatest probability corresponds to the transition to the cluster 2 starting from the 5th month of the study.

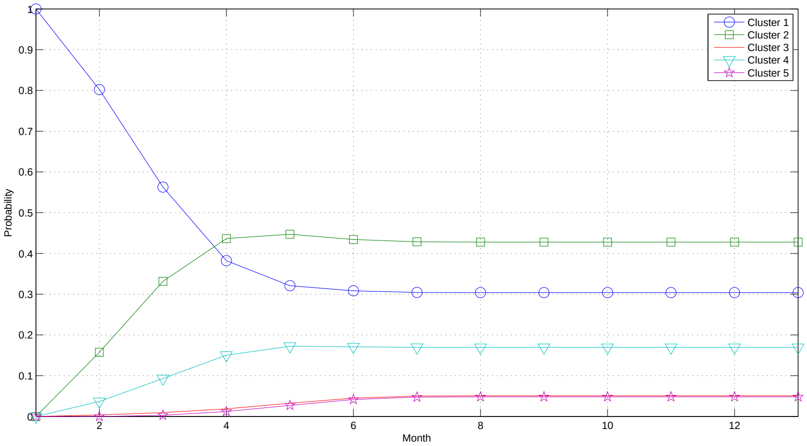


Figure 3.18: *Probabilities of changing from the first cluster to the rest.*

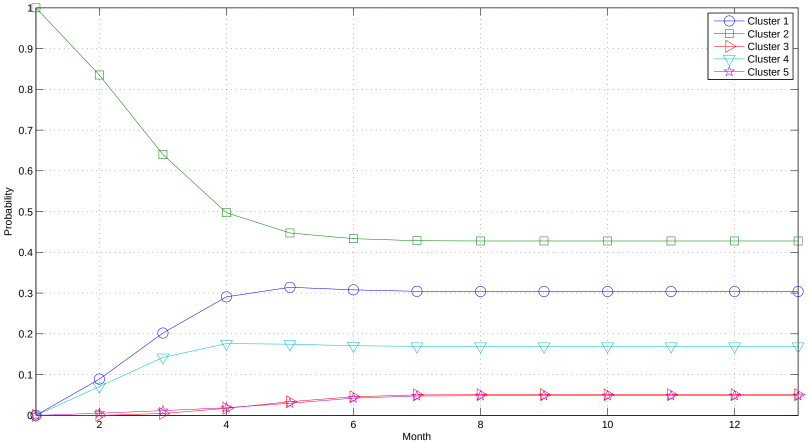


Figure 3.19: *Probabilities of changing from the second cluster to the rest.*

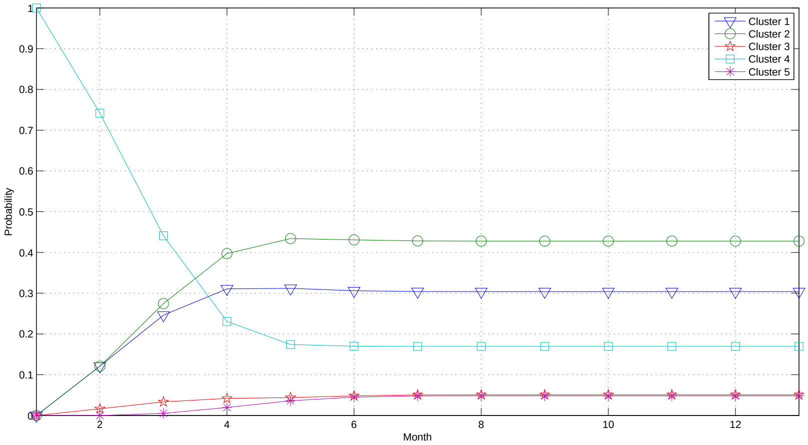


Figure 3.20: *Probabilities of changing from the fourth cluster to the rest.*

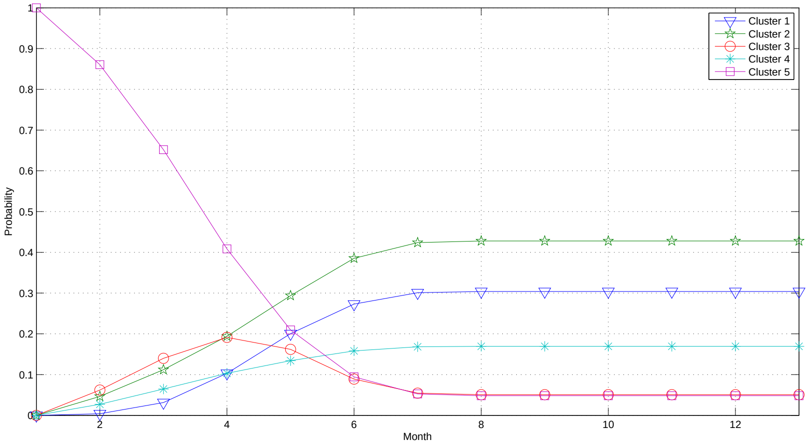


Figure 3.21: *Probabilities of changing from the fifth cluster to the rest.*

Analysis of Portugal KPIs

As in the previous cases, the main objective of the study is to analyze the data corresponding to this country using SOM techniques, in order to extract conclusions about KPIs and clinics. The same procedure as in the cases of Turkey and Italy was taking into account.

Following the same procedure that in previous countries, KPIs with a very low standard deviation were rejected. Obviously, those KPIs with empty values were also rejected. So in this country KPIs number 1 (*High Flux Dialysis*), 7 (*Reporting Compliance*), 12 (*Absenteeism*), 13 (*Overtime*), 14 (*Employee Satisfaction Survey*), 16 (*Treatment growth*), 19 (*Scheduling Efficiency*), 23 (*Patient education and support program*), 28 (*Electricity consumption*), and 29 (*Water consumption*) were not taken into account for training. Thus, the SOM was trained with the remaining KPIs, which yielded the component planes shown in Figure 3.22.

In the component planes, the following facts were drawn:

- Constant KPIs: In the map, the 10th KPI (*Seroconversion HepB-C*) presents a constant value, while KPIs 11 (*Turnover personnel*) and 20 (*Personnel costs*) show only small variations.

- Considering the SOM from a global point of view, KPI number 9 (*Patients at risk for HepB infection*) is the one with the lowest values. Therefore, certain actions must be considered for this area of business since it has low values for all clinics (whole map).

- KPIs 24 and 25 (related to ISO certification) are correlated, as happened in case of Turkey.

- Looking at special behaviors in KPIs, it is found that KPIs 17 (*Patient growth*) and 18 (*New patient inflow*) show the same spatial pattern: areas with lowest values are the same for both KPI.

- As in the case of Italy and Turkey, it can be seen that when the KPI 6 (*Treatment adequacy*) decreases its value, the same thing happen for the KPI 3 (eKt/V) (left side in both component planes). As previously discussed, this fact is because *eKt/V*, also defined as *dialysis dose*, is one of the main parameters used for treatment efficacy evaluation. In addition, the same applies to the KPI 4

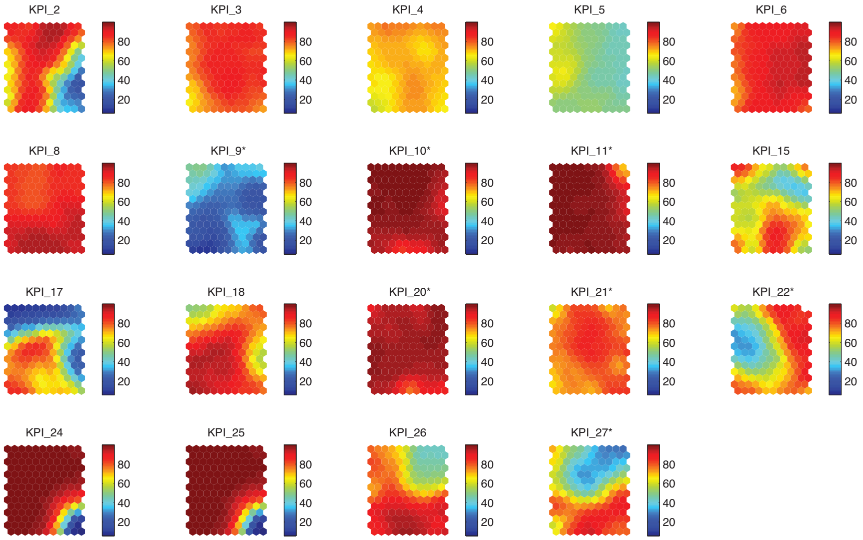


Figure 3.22: *Component planes for the SOM analysis obtained for Portugal after training the model without the KPIs (1, 7, 12, 13, 14, 16, 19, 23, 28 and 29). Notice that color scales are normalized to the same interval, where 100 corresponds to the largest value over all weight vectors, and 0 to the smallest one; therefore, the maximum value for one KPI might not correspond to dark red for all planes. Also notice that for KPIs marked with a * (also in Table 3.1), red regions correspond to good scores which in turn are obtained for low raw values on those parameters: for instance, a red region in the HepB infection risk component (KPI 9) does not mean that the risk is high, but rather that the performance regarding such aspect is a good one - that is, the infection risk is low.*

(*Hemoglobin*) as in the two previous cases. This is due to the fact that good hemoglobin levels in patients are also indicative of treatment adequacy.

- High rate of new patient income (KPI 18) corresponds to a higher risk of hepatitis B infection (KPI 9; note that high risk is indicated by low KPI scores, blue colors) as in the case of Italy. These partial inverse correlations find a possible explanation in that when new patients are accepted in a *FME* clinic to start the therapy they might display conditions that require time to be corrected.

Analysis of the Portuguese clinics evolution

As it has been made in previous countries, the prototype corresponding to every clinic in the different moments has been determined in the map and plotted it as a trajectory. This will bring information about the time evolution of the clinic. Figure 3.23 shows some of the trajectories obtained for several clinics.

As happened in the previous cases, there are many different paths and it is hard to find out information. Consequently, it was applied the *k-means* algorithm in order to determine areas (clusters) in the map. Clusters yielded by this algorithm are shown in Figure 3.24. The map of winner neurons (Figure 3.25) shows that data is spread out in the map homogeneously.

When comparing the different areas obtained by the clustering with the KPIs component planes, one can conclude that:

- **Cluster 1.** Lower right corner. It corresponds with the area of the map in which the clinical behavior is farther from the desired one; this cluster contains the clinics with worst values of KPIs 24 and 25 (clinics that are not ISO compliant), and low values of KPIs 2, 9, 15 and 17 (*HDF online dialysis*, *Patients at risk for HepB infection*, *Training hours* and *Patient growth*).

- **Cluster 2.** This group brings together clinics compliant and not compliant with ISO (that is why KPIs 24 and 25 component planes are not only in 0% or 100%).This area groups also the lowest values for KPIs 2, 9, 17 and 18 (*HDF online dialysis*, *Patients at risk for HepB infection*, *Patient growth*, and *New patient inflow*).

- **Cluster 3.** Upper right corner. It contains clinics that are ISO compliant

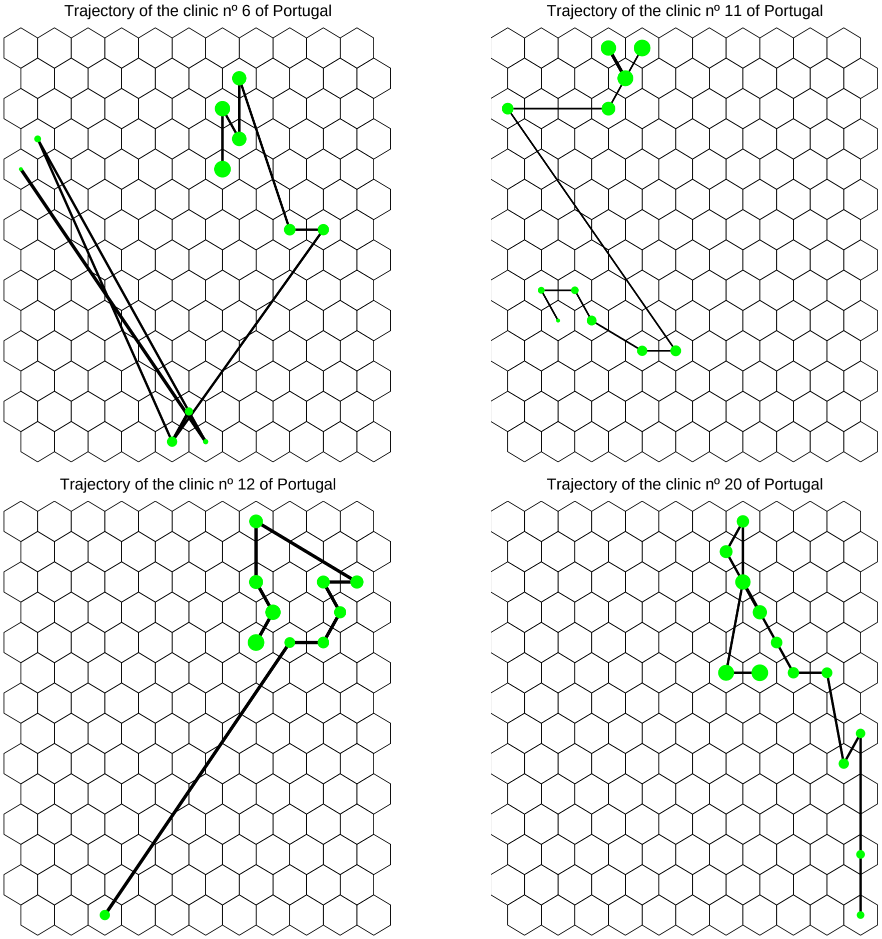


Figure 3.23: *Trajectories of several clinics of Portugal throughout the time in the SOM grid. Each trajectory starts at the smallest green point and finishes at the biggest one.*

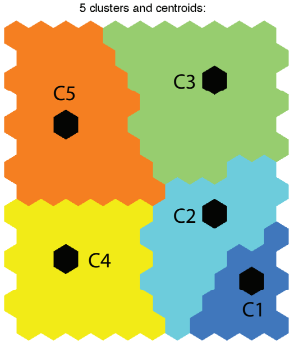


Figure 3.24: *Clustering obtained after applying k-means algorithm in the SOM map for Portugal clinics. Black hexagons represent the clusters centroids or prototypes.*

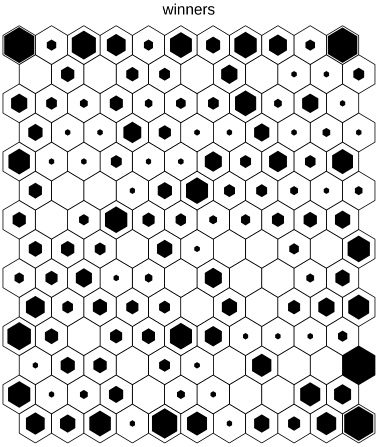


Figure 3.25: *SOM grid showing the winner neurons after training for Portugal clinics.*

(highest values for KPI 24 and 25) but with lowest values for KPIs 5, 9, 17, 26, and 27 (*Vascular access*, *Patients at risk for HepB infection*, *Patient growth*, *Compliance program* and *Contaminated waste*).

- **Cluster 4.** Lower left corner. It groups clinics that are ISO compliant but have the lowest values for KPIs 3, 4, 5, 6 and 9 (*eKt/V*, *Hemoglobin*, *Vascular access*, *Treatment adequacy* and *Patients at risk for HepB infection*). Additionally, KPI 21 (*Other costs*) shows lower values for these clinics.

- **Cluster 5.** Upper left corner. It contains clinics that are ISO compliant with high values for all KPIs except for KPIs 6, 18, 21 and 22 (*Treatment adequacy*, *New patient inflow*, *Other costs* and *Accidents to employees*), that do not show their highest values.

Therefore, cluster 1 is labeled as a “bad” cluster, and the rest as “good” ones. This fact is also reflected by computing the global score of each cluster as in previous cases (Table 3.4). According to this table, the lowest score is obtained for cluster 1 (71.14%). Moreover the score for the others clusters are quite similar among them (they are in the same range) but higher than the one obtained for cluster 1. It can be seen that the difference between a “bad” cluster and a “good” one is not so relevant, but this is due to the fact that all clinics have a very high score, so a difference about 10% makes the point in this case. This labeling was supported by the average of KPIs components map, represented in Figure 3.27.

As in the case of Turkey and Italy, Figure 3.26 represents the prototype centroids by the parallel coordinates method. This figure shows a great similarity among centroids, except for KPIs 2, 24 and 25 (*HDF online dialysis*, *ISO 9001* and *ISO 14001*), which are the KPIs that define the difference among “bad” cluster (1) and the others, as seen in Figure 3.22.

Table 3.4: *Prototype global scores for Portugal.*

Portugal Centroids Score				
C1	C2	C3	C4	C5
71.14	78.58	76.14	80.66	77.77

As it was made in previous countries, the transition matrix (in percentage) for

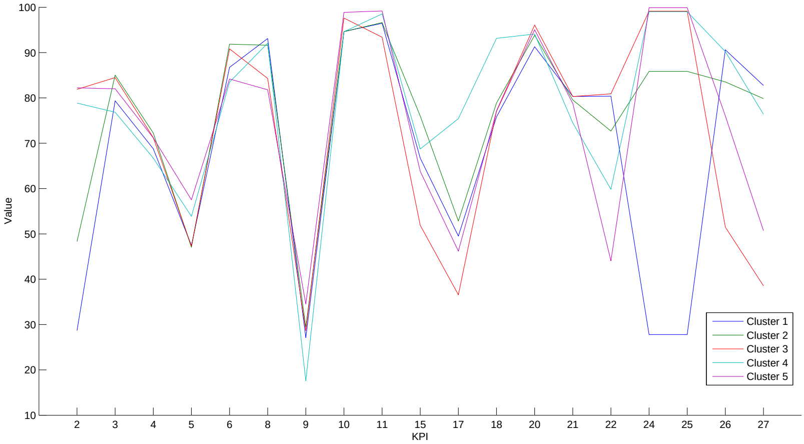


Figure 3.26: *Parallel coordinates plot of Portugal centroids.*

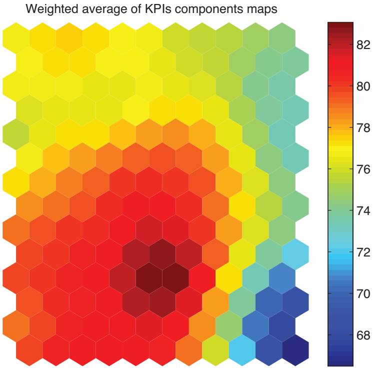


Figure 3.27: *Average of KPIs component planes of Portugal (Figure 3.22) according to the weight established by Fresenius.*

every clinic (33 clinics in total) was calculated, yielding the following results (Eq. 3.6):

$$\begin{bmatrix} 13.36 & 0.67 & 0.22 & 0.34 & 0.22 \\ 0.45 & 9.20 & 1.12 & 0.34 & 0.34 \\ 0.11 & 0.22 & 24.47 & 0.00 & 0.56 \\ 0.11 & 0.56 & 0.67 & 22.11 & 1.57 \\ 0.00 & 0.11 & 0.45 & 0.67 & 22.11 \end{bmatrix} \tag{3.6}$$

Due to the fact that Cluster 1 is labeled as “bad”, it must be paid attention to:

1. First row (except diagonal term). It represents transitions from an area labeled as bad to an area considered as good. This means a good temporal evolution, because the clinic moves to a better area.

2. First column (except diagonal term). This is the worst temporal evolution, because the clinic moves from a good area to a bad one.

The data of the transition matrix shows that:

- Highest percentages are located in the main diagonal, which means that, usually, there are not transitions between clusters. This conclusion is the same that in previous countries.

- Highest number of transitions among clusters is located in transitions from cluster 4 to 5 and from cluster 2 to 3.

- Worst behaviors (first column) have very low values; cluster 2 is the most probable cluster to downgrade to cluster 1.

The transition probability (Markov chain) among two states (namely, among two clusters) was calculated from the transitions matrix, yielding the matrix shown in 3.7.

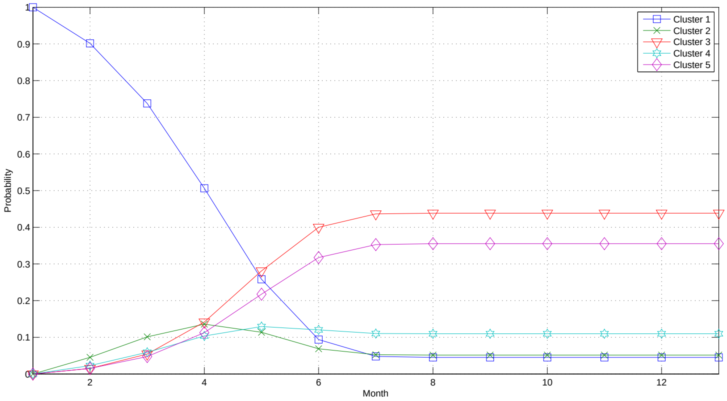


Figure 3.28: *Probabilities of changing from the first cluster to the rest.*

$$\begin{bmatrix} 0.90 & 0.04 & 0.02 & 0.02 & 0.02 \\ 0.04 & 0.80 & 0.10 & 0.03 & 0.03 \\ 0.00 & 0.01 & 0.96 & 0.00 & 0.02 \\ 0.00 & 0.02 & 0.03 & 0.88 & 0.06 \\ 0.00 & 0.00 & 0.02 & 0.02 & 0.95 \end{bmatrix} \tag{3.7}$$

Using the previous matrix (Eq. 3.7), it is possible to calculate the probability of changing from one cluster to another during the course of time, i.e. the temporal evolution of the clinics. Figures 3.28, 3.29, 3.30, 3.31, and 3.32 describe these probabilities starting from clusters 1, 2, 3, 4, and 5, respectively. As it can be seen, the highest probability of changing corresponds to moving to cluster 3 (labeled as "good").

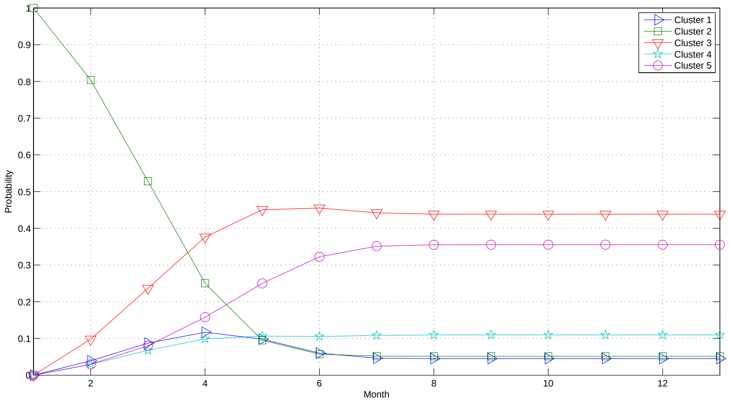


Figure 3.29: *Probabilities of changing from the second cluster to the rest.*

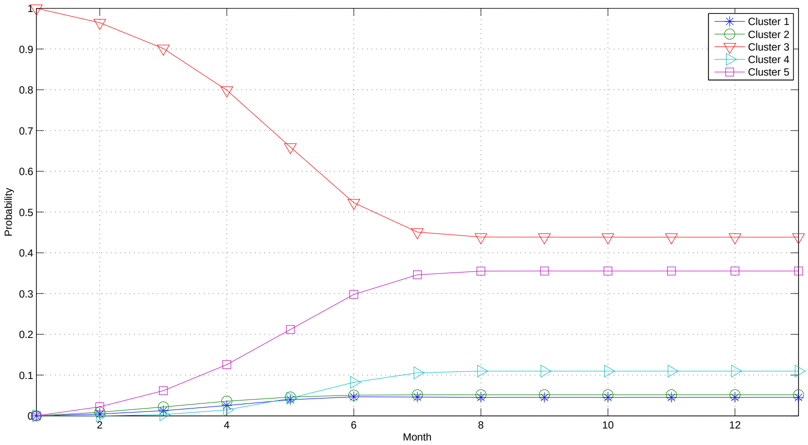


Figure 3.30: *Probabilities of changing from the third cluster to the rest.*

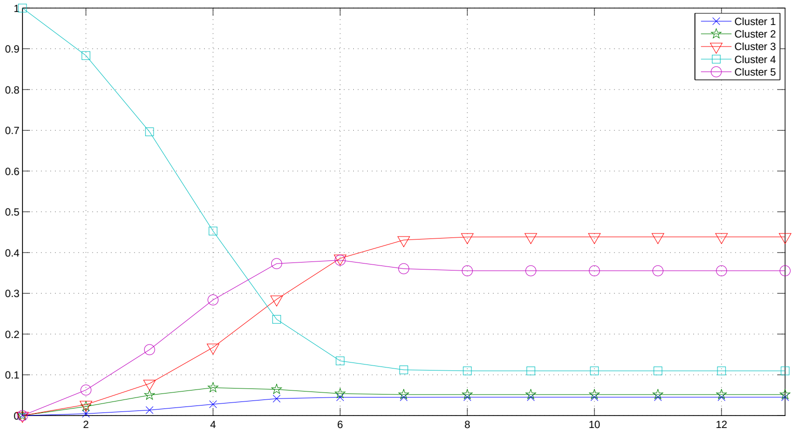


Figure 3.31: *Probabilities of changing from the fourth cluster to the rest.*

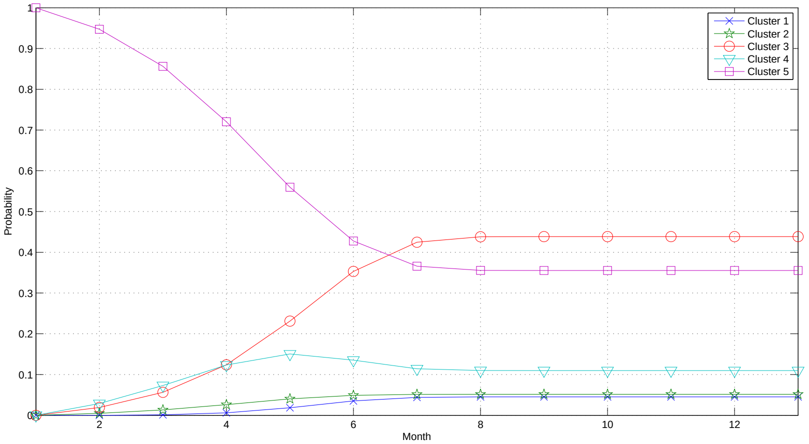


Figure 3.32: *Probabilities of changing from the fifth cluster to the rest.*

3.1.6 Conclusion

In general, in the light of a continuous improvement policy, the SOM analysis proposed here naturally lends itself to effectively identify both areas of excellence and aspects that might need improvements, also suggesting possible strategies for intervention. While this work can be considered a preliminary study for this specific data domain, the presented results provided evidences that support, and indeed encourage, the adoption of SOM analysis as a standard analytic tool for clinic performance monitoring.

The BSC has been successfully adopted for several years by *FME* to monitor its clinic efficiency. At present, the BSC analysis monthly provides an evaluation of financial and operational parameters with reference to specific perspectives and KPIs. This approach gives a punctual depiction of every single care unit's performance for the considered month.

Presented results have shown that SOMs may allow a comparative analysis among *FME* KPIs and perspectives over a large time lapse (not only month by month, as the currently available reports on the BSC allow) highlighting relationships that could not easily be inferred before. Moreover, it was possible to track clinic improvements (or declines), and to predict the probability of this changes so suggesting future interventions for business policy corrections. The extra value of SOMs in this context lies in that identified correlations among a set of KPIs can suggest a potential causal link that was not apparent to the management at first (as standard reports do not support this kind of inference). For instance, a report showing an increase in accidents to the personnel would suggest the need for corrective actions, but it would not provide any more clues for a focused intervention. On the other hand, when considering the component planes for the *Training Hours* and *Accidents to Employees* KPIs, no direct correspondence emerged between the number of training hours and the capability of the personnel of avoiding injurious events. This result has practical implications as it suggests that merely increasing the number of training hours might not be enough to improve safety. Rather, either the key factor for the high rate of accidents lies in the contents of the training program, or it lies in some totally different aspect: thus, SOMs, narrowing the spectrum of possible interventions, provide more clues to guide the management strategy for corrective actions. Hence, what emerged from this particular case study is largely beyond a simple theoretical speculation: indeed, SOMs concretely demonstrated to be a powerful integrative method to the traditional interpretation of an active BSC.

Moreover, the applicability of the SOMs does not exhaust in the specific case study above presented; on the contrary, given their flexibility, a wider utilization of the SOMs is not excluded also for the study of data coming from the BSCs of totally unrelated areas of business. Hence, the main goal of this work was not to provide unquestionable truths over a single case study; on the contrary, sharing the results it is aimed to propose the SOMs as a powerful tool to extrapolate cues for further considerations and analysis on general enterprise performances.

Moreover, an additional clustering step on the SOM vectors was also performed, as this provides a higher-level description of characteristic features of larger groups of KPI records, with respect to the more fine-grained information encoded by single neurons. The weight vectors of the SOM were automatically grouped based on their similarity by employing the k-means algorithm (Forgy, 1965). An analysis of cluster transitions was done based on such record trajectories: the aim was to infer from the observed, past records which behavior can be reasonably expected from clinics in future months by means the cluster trajectories. Transition probability (Markov chains) between states was calculated, so that such prediction might then be used to take corrective actions to make sure that clinics always maintain a high standard of performance.

3.2 Analysis of Patients Satisfaction Surveys (PSS) using Self-Organizing Maps

3.2.1 Introduction

Nowadays, continuous improvement in quality and safety of the medical treatment has become a central topic in dialysis. Under the hypothesis that a more effective delivery of care might lead to an improved dialysis patient's outcome, continuous quality improvement has to be applied to the delivered dialysis treatments (products and services) and to the other related services.

The evaluation of perception of care as patients' satisfaction has over the years become an important indicator for assessing health care quality (Fitzpatrick, 1991).

Patients are satisfied with a medical treatment when: they achieve a level of rehabilitation as similar as possible to healthy people; treatment is sustainable and not invasive; treatment is provided by professional and caring people; clinical staff is compassionate and promptly available; location of treatment is comfortable; and privacy is respected.

Patients who are generally more satisfied with the provided services will be more likely to accept their illness and thus might express a better quality of life (Kimmel, 2000a; Kimmel et al., 1998). Therefore, the goal of increasing patients' satisfaction is an indirect way of improving their quality of life.

Fresenius Medical Care (FME) as a global provider of dialysis services has a strong interest in monitoring patient satisfaction.

Using the same approach of the previously published manuscript (Kirchgessner et al., 2006), FME developed a specific questionnaire named Patient Satisfaction Programme (PSP) for the hemodialysis patients treated in the *NephroCare* clinics of 19 different countries.

The information provided by the PSP questionnaire was first analyzed using standard statistics. However, being on a high level of satisfaction, the option to detect opportunities to realize even marginal improvements was very limited. Searching for an alternative approach to analyze data, Self-Organizing Maps (SOMs) can be a more appropriate tool to find and visualize patterns (relevant profiles) in multi-dimensional data sets as the data set dealing with patients satisfaction is (Kohonen, 2001). As mentioned in Chapter 2, SOMs produce low dimensional maps (usually in two dimensions) having preserved the topological relationships of the original high dimensional data spaces. This means that two patients who are represented close to each other in SOM are actually similar patients in the original data space whereas patients who are mapped far away on the SOM representation are actually different in the original data space. Therefore, the analysis of the maps provides a direct interpretation of the original data since it is easy to find profiles of patients.

The suitability of SOM to deal with information extracted from satisfaction surveys has already been shown in previous research works in other fields. Lansiluoto evaluated users' satisfaction in macro-environmental analysis with an SOM model (Lansiluoto, 2007). He found that SOMs outperformed current methods in accuracy, content, ease of interpretation and format. Lee and Park utilized an integrated use of business intelligence tools, SOMs and decision trees for improving the Customer Relationship Management based on a customer satisfaction survey (Lee and Park, 2005). Garavaglia proposed to use SOMs to evaluate Customer Satisfaction with health care plans (Garavaglia, 2000). SOMs provided a useful visualization of the complexity of responses and highlighted specific areas for health plan quality and service improvements that might be missed if only simple average total scores were considered.

Therefore, on the basis of the reported positive experiences, it appears that SOMs are a suitable data analysis tool that provides a very straightforward interpretation of the achieved results while carrying out a non-linear accurate modeling of the data.

The aim of this study is to test and validate a methodology for the detection of a residual area of low satisfaction in dialysis patients.

3.2.2 PSP questionnaire

The adopted questionnaire was developed and validated using a methodology described by (Kirchgessner et al., 2006).

The questionnaire was distributed in 2008–2009 to hemodialysis patients treated in 335 centers belonging to the *NephroCare* clinics network located in 19 countries, namely Italy, South Africa, Poland, Hungary, Romania, Slovakia, France, Spain, Hungary, Czech Republic, Turkey, Germany, United Kingdom, Ireland, Slovenia, Russia, Serbia, Bosnia-Herzegovina and Portugal.

Respondents were asked to fill in the questionnaire in an anonymous form and to return it via a collection box located in each unit.

The PSP questionnaire contained 79 close ended questions that covered various aspects of the delivered care. The questions were grouped into eight blocks of statements:

- **Topics' Discussion:** General information and explanation of dialysis treatment given to patients before starting the treatment itself.

- **Dialysis Unit:** Dialysis center characteristics (comfort of the different rooms, bed or chairs used during treatment; cleanliness of unit; privacy; services provided).

- **Dialysis Arrangements:** Instructions and information given to patients by staff; dialysis procedures; safety arrangements.
- **About Nurses:** Attention that patients receive from nurses in terms of availability, approachability, courtesy, timeliness, confidence, competence, communication, respect.
- **About Doctors:** Attention that patients receive from doctors in terms of availability, approachability, courtesy, timeliness, confidence, competence, communication, respect.
- **Other Staff Members:** Relationships and communication between patient and other staff members.
- **Overall Rating:** Patient's general satisfaction with doctors, nurses, dialysis unit and transport service.
- **Overall Rating Yes/No:** Patient's measure of satisfaction based on whether he/she would recommend the dialysis unit to a friend in the same situation, and comparison with other units in which the patient has received treatment.

Table 3.5 shows the level of satisfaction (average) of the eight blocks of statements. The scale goes from -3 (strongly disagree) to +3 (strongly agree). In addition to these blocks, the parameters *Age* and *Gender* (*Male* = 1; *Female* = 2) were also collected.

Table 3.5: *Level of satisfaction (average) of the eight different blocks of statements. The scale goes from -3 (strongly disagree) to +3 (strongly agree).*

Domain	Level of satisfaction (average)
Topic's Discussion	1.33
Dialysis Unit	1.81
Dialysis Arrangements	1.94
About Nurses	2.11
About Doctors	2.16
Other Staff Members	1.86
Overall Rating	2.26
Overall Rateing YES/NO	2.75

Six blocks of statements (*Dialysis Unit*, *Dialysis Arrangement*, *About Nurses*, *About Doctors*, *Other Staff* and *Overall Rating*) required a response choice based on a seven point Likert scale (strongly disagree, disagree, slightly disagree, neither agree nor disagree, slightly agree, agree, strongly agree) and a last option (does not apply) (Likert, 1932). At the end of each block of statements, patients could also add individual personal comments.

It was recoded the original coding of the Likert scale from 1 to 7 into a new one from -3 (strongly disagree) to +3 (strongly agree). For each block of statements, it was used the mean value of the opinions given by the patients. Unanswered questions were not taken into account.

The items of the two remaining blocks of statements *Topics' Discussion* and *Overall Rating Yes/No* were based on a Yes/No answer. For the sake of using the same coding and meaning for all the satisfaction features, it was recoded the answer "NO" as -3 and the answer "YES" as +3.

3.2.3 Methodology

As mentioned in the introduction, this study deals with the use of SOMs as a visualization tool in order to extract information from the PSP questionnaire that may not be detected by other classical analyses.

For the training, it was followed the same procedure as in Section 3.1.4. Different options of the tuning parameters of the SOM algorithm were tested, combining all the possibilities (Weight Initialization, Neighborhood Function and Type of Training). Moreover, the random initialization was fulfilled 100 times for each combination of parameters. Regarding the size of the map, the total training time, the learning rate and the neighborhood radius, the default number of these parameters that SOM toolbox documentation considers the most appropriate was selected again.

After a preliminary analysis of the studied population as a whole, it is possible to select some sub-populations of specific interest (e.g. the group of patients with the lowest level of satisfaction) and then run a new SOM only on this subset of patients.

The procedure to produce the new SOM is the following:

1. The first step is to define the areas of interest in which a new analysis will be carried out. In the present study, since the main interest of the analysis is to understand how to improve the level of satisfaction, the relevant areas are those

including patients with the lowest level of satisfaction (e.g. component plane *Overall Rating i*, level of satisfaction equal or lower than 1). However, other areas can be significant as well.

2. Once the areas are identified, the related group of patients that are present in these areas is selected.

3. A new SOM is run for each new subset of patients. It should be emphasized that according to the previous procedure, Figure 3.34 is not a magnification of Figure 3.33 (see next section) but represents a new analysis of the area with critical information, involving a new SOM only for the chosen set of patients.

3.2.4 Results

A total of 10,632 haemodialysis patients from 335 units completed the questionnaire. The mean age of the survey respondents was 63.05 ± 14.93 years and 56.69% were males. The overall response rate was of 66%. The overall level of satisfaction was of 1.99 (in the range from -3 to +3) ranging from 1.33 in the block c "*Topics Discussion*" to 2.75 in the block j "*Overall Rating Y/N*".

The analysis of Figure 3.33 (blocks c–j) shows that on average the patients are very satisfied with all the issues studied in the survey. However, it is useful to analyze the different profiles that appear in the data set. Five significant areas have been highlighted with numbers from 1 to 5 in the map of winner neurons (Figure 3.33, block k). The same areas are also highlighted within a square in the blocks a to j of Figure 3.33, using the same color code as the one used in the map of winner neurons.

Areas 1 and 2 have many associated patients, thus showing that these areas are relevant. Both areas, representing almost 50% of the patients, include patients who are around 60 years old (women are mapped into area 2 and men into area 1). They are very satisfied with all the issues of the treatment showing a mean value of overall rating close to 3, which is the maximum possible value.

Area 3 is the part of the map which is most significantly different from the rest of the map. It covers those patients whose opinion is much worse than the average profile. These patients are around 60 years old, the ratio of men and women is balanced, and their opinion is quite negative in all the blocks of questions, with exception of Topics' Discussion (block c), in which the average value of the satisfaction survey is slightly

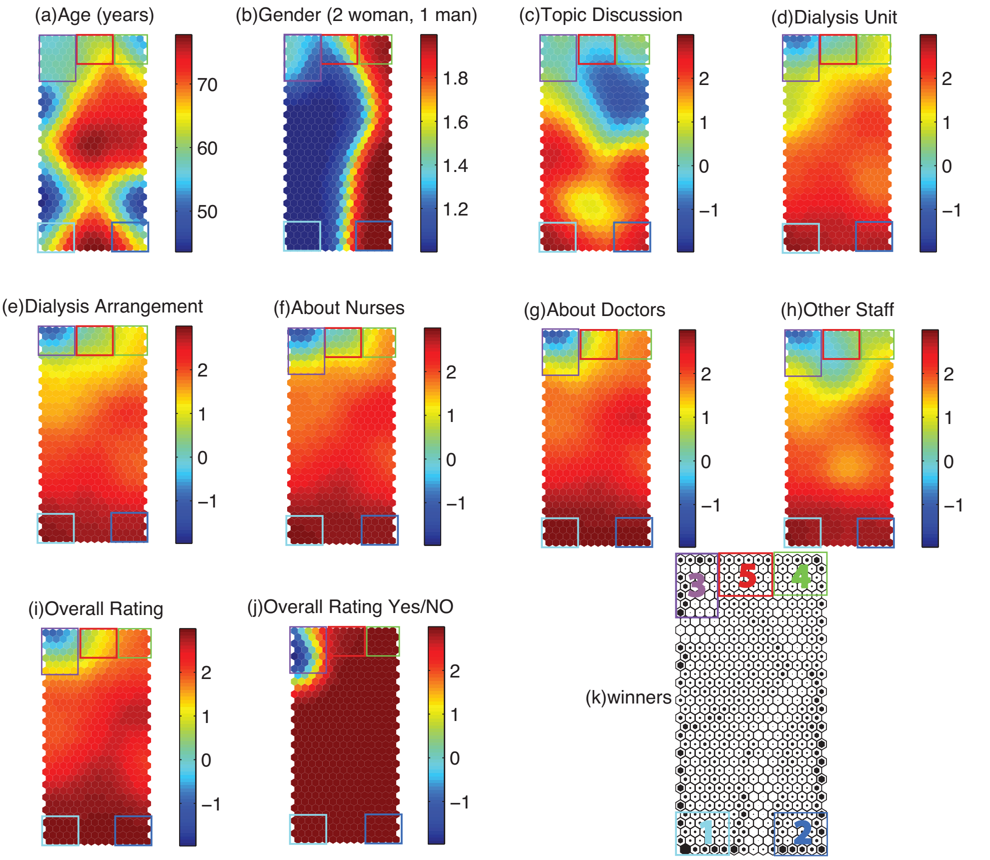


Figure 3.33: *Maps obtained using the complete data set. The lower-right map labeled with numbers (k) represents the winner neurons, and thus provides information about the number of patients represented by each neuron (the blacker the neuron, the higher the number of associated patients). The other maps (called component planes: blocks a–j) show projections corresponding to the different input features (Age, Gender and each one of the eight blocks of statements in the questionnaire).*

positive (>0).

Area 4 includes female middle-aged patients (in their late fifties). They have a very positive opinion of *Doctors* (level of satisfaction around 2) and *Nurses* (level of satisfaction between 1.5 and 2), but they are not so satisfied with *Other Staff* (level of satisfaction between 0 and 1), nor with other issues, such as *Topics' Discussion* (level of satisfaction slightly higher than 0), *Dialysis Unit* (level of satisfaction around

1) and *Dialysis Arrangement* (level of satisfaction slightly higher than 1). It should be emphasized that their opinions about the overall treatment (*Overall Rating* and *Overall Rating Y/N*) (level of satisfaction around 2 and 3, respectively) is also very positive, and therefore, their opinion can be considered as satisfactory, in general.

Finally, area 5 represents a striking group of patients, mainly males aged 60–65 years. Among the specific blocks of questions, they only show a fairly good level of satisfaction with *Doctors* (block g), with a level of satisfaction between 1 and 2, but their levels of satisfaction is much lower with the rest of staff (blocks f and h), *Topics' Discussion* (block c) with a level of satisfaction lower than 0, *Dialysis Unit* (block d) with level of satisfaction between 0 and 1, and *Dialysis Arrangement* (block e), that is also between 0 and 1. However, the *Overall Rating* is positive, since block i shows values between 1 and 2, and block j values between 2 and 3.

Due to the relevance of area 3 of block k in Figure 3.33 (outside this area, patients are basically satisfied with the treatment), Figure 3.34 shows another map in which only the patterns mapped in that negative area have been used to train another SOM in order to find relevant profiles that can help to improve the level of satisfaction of this set of patients. It should be emphasized that the analysis of this area representing patients with low levels of satisfaction (mean value of *Overall Rating* [block i] lower than 1), can help identify the profiles of those patients, and in turn, suggest actions to increase their levels of satisfaction.

The four corners of the blocks of Figure 3.34 are the most relevant profiles, and are labeled as areas A, B, C and D in the map of winner neurons (block k). Those areas are also highlighted in the other blocks with colored squares similarly to the map of winner neurons (block k).

The vast majority of women whose PSP is negative are represented in area A. They are between 50 and 60 years old, and although their opinions are quite neutral in almost all the blocks of statements, their responses to *Overall Rating Y/N* (block j) is strikingly low (level of satisfaction around the minimum value, i.e. -3).

Area B is relatively similar to the previous one, but in this case the patients' profile corresponds with men in their late fifties. The opinions about *Doctors* (block g) with a level of satisfaction higher than 1 and *Other Staff* (block h, level of satisfaction around 1) is relatively good, the satisfaction related to *Overall Rating* (block i, level of satisfaction between 0 and 1) can be considered as acceptable, but nevertheless,

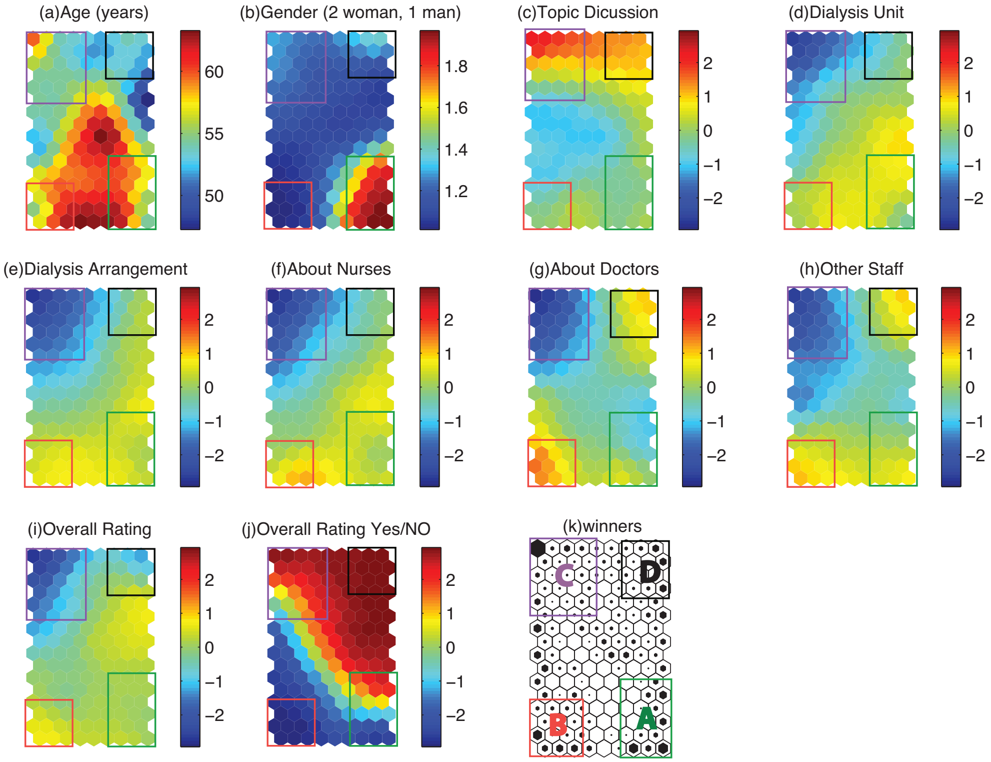


Figure 3.34: *Maps obtained using patients mapped from area 3 of the map shown in Figure 3.33. Component Planes (blocks a–j) and map of winner neurons (k) are shown.*

the *Overall Rating Y/N* (block *j*) shows levels of satisfaction much lower than desired (close to -3).

Area C represents men in the middle fifties. Their level of satisfaction is very low (level of satisfaction lower than -2) in all the specific blocks of statements except *Topics' Discussion* (block *c*), in which the level of satisfaction is around 2. *Overall Rating Y/N* (block *j*) also shows a high level of satisfaction thus suggesting a close relationship between an overall satisfaction and the knowledge of the treatment and implications of dialysis.

Finally, regarding area D, patients mapped into this corner correspond mainly with men in their early fifties. Their level of satisfaction is quite neutral for most of the blocks of statements but it is higher in the statements related to *Topics' Discussion* (block c), *Doctors* (block g), and *Other Staff* (block h), that show a level of satisfaction higher than 1. It possibly makes the satisfaction related to the block *Overall Rating Y/N* (block j) quite good as well (level of satisfaction close to the maximum), showing hence a behavior almost opposite to that found in area B.

3.2.5 Conclusions

The results of this Patient Satisfaction Surveys show that, in general, the level of patient satisfaction with the service provided is rather high. This makes research into opportunities for improvement quite difficult, and traditional analysis usually stops here. However, with the help of the so-called SOMs it is possible to nevertheless identify areas of potential improvement for specific patient groups. Clearly, in area 1 (Figure 3.33) this is difficult since this population of patients (mainly males aged between 60 and 70 years) shows a high level of satisfaction in all aspects addressed in the survey (Table 3.5). The same is true for area 2 which includes women of the same age group. However, area 3 which is populated by both males and females of a younger age (around 60 years) includes the most unsatisfied participants in the survey. In principle, this group is not satisfied with the organization, with the service provided by the staff and with the quality of follow-up provided by physicians. The only aspect they are relatively satisfied with is dealing with the level and quality of information provided to them. Those patients are usually empowered to understand their renal replacement treatment. They are usually even well informed regarding complex therapeutic options (such as convective treatment) and they are trained to follow the appropriate diet regimen for a dialysis patient. However, despite the fact that they are relatively young and have a good understanding of the implications of being on dialysis, they are quite discontent. This behavior has to do with the lower probability they have (or they presume to have) of receiving a kidney transplant, since they are also probably well aware of the current situation regarding scarcity of kidneys donors in most of the countries. They are less likely to accept that they will be on dialysis for the rest of their life, and they see their dependency on dialysis as depressing (Finkelstein and Finkelstein, 2000; Kimmel, 2000b; Kimmel et al., 2000). Patients located in area 5 of the map also show an interesting profile: mainly men aged between 60 and 65 years, they are not satisfied with the level of information they receive and they are only partially satisfied with the organization and the setting of the clinic, as well as with the level of service delivered by nurses and doctors. However, the overall rating is positive, and this suggests that their opinion about doctors plays a key role in their overall satisfaction with the service, much more so than other aspects. These patients are likely to have a lower understanding of their current situation and they definitively need more support. Ronco and Marcelli addressed the importance of time (t) dedicated by physicians (MD) to patients (Pt): ($MD \cdot t/Pt$), playing around the very well known concept of dialysis dose (Kt/V) (Ronco and Marcelli, 1999). Obviously, all patients would benefit from more attention and from more dedication in terms of time by the medical staff, but it is likely that this specific group can benefit even more than others, resulting in a lower grade of depression and a higher acceptance of the chronic disease. Area 4 is similar, with the only difference being a higher satisfaction with the nurses' service. One can conclude that this last group of patients, mainly women aged around 60 years, have a low understanding of their disease and of the technology used for treating them, but that they trust in the people taking care of them, transferring full responsibility to them.

This analysis does not claim to be complete, but it shows how a different presentation of the results can significantly improve the level of insight into an apparently well satisfied patient cohort. While traditional analyses provide only an average view of reality, this new non-canonical approach allows segmentation of the patients in order to detect their not-so-obvious needs. Segmentation of the patient populations is a well established marketing tool but something relatively new in medicine, specifically in the field of chronic diseases. Traditional marketing tools aim at achieving higher levels of customer satisfaction, and sophisticated models have been available for many years that help better understand customers' needs (Rese, 2003). By using the SOM representation, it was also able to portray the complexity of patients' needs and to identify niches of dissatisfaction. The particular advantage of using SOM lies in its ability to further analyze the high levels of overall satisfaction achieved in these kinds of surveys: the vast majority of patients are very satisfied with all the issues analyzed in the survey, as mentioned previously.

Moreover, compared with classical clustering techniques that can find typical profiles but are associated with complex presentation of the results, SOM is also able to find similar behaviors (typical profiles are represented in the same area of the map), and simultaneously depict the results in an easily interpreted 2- dimensional map. Furthermore, SOM modeling is non-linear whereas most of classical methods can only find linear relationships.

In addition to the component planes, i.e. the map representation in which high density data patterns are grouped together, another representation shows the number of patients associated with each area of the map, i.e. showing the relevance of the different areas of the map in terms of the number of patients represented in each area. Moreover, as has been done in this work, it is possible to carry out a magnification of some parts of the map, thus obtaining a hierarchical structure of maps. In this particular analysis, the experiment focused on patients whose levels of satisfaction were low in order to find different profiles within them. Useful conclusions could be extracted from this approach. These conclusions can help identify areas that should be changed or analyzed in order to increase the degree of patient satisfaction.

3.3 Visual Data Mining with Self-Organizing Maps for Ventricular Fibrillation Analysis

3.3.1 Introduction

Ventricular Fibrillation (VF) is a cardiac arrhythmia caused by a disorganized electrical activity of the heart (Moe et al., 1964) causing collapse and unconsciousness, following a serious risk of death unless an appropriate recovering therapy is applied (typically, a high voltage defibrillation shock) (Beck et al., 1947). Clinical and experimental studies have demonstrated that the success of defibrillation is inversely related to the time interval between the beginning of the VF episode and the application of the electrical shock (Yakaitis et al., 1980; Capucci et al., 2001; White et al., 1996). For this reason, the development of early VF detection algorithms for monitoring systems and automatic external defibrillators (AED) is being deeply studied. However, in general, these methods do not provide insight into the problem. For this reason, this work proposes a new methodology in order to obtain visual information about this problem.

Detection algorithms analyze the surface electrocardiogram (ECG), providing a fast and accurate diagnosis of VF in order to reduce the reaction time of specialist in case of monitory systems or even supply the appropriate therapy without the need of qualified personnel as in the case of Automatic External Defibrillators (AED) (Faddy, 2006). Non invasive detection of VF is typically based on extracting parameters from the ECG signal in different representations such as time, frequency, and time-frequency domains. Time-domain methods analyze the morphology of the ECG to discriminate VF rhythms (Chen et al., 1987; Clayton et al., 1993; Chen et al., 1996; Zhang et al., 1999). Frequency-domain measurements are motivated by experimental studies supporting that VF is not a chaotic and disorganized pathology and a certain degree of spatio-temporal organization exists during VF (Davidenko et al., 1992; Clayton et al., 1995; Jalife et al., 1998). Spectral description of the ECG has revealed important differences between normal and fibrillatory rhythms (Clayton et al., 1995; Herschleb et al., 1979; Murray et al., 1985). In this context, relevant parameters of the ECG spectrum have been used for developing VF detectors (M. E. Nygards and J. Hulting, 1978; Barro et al., 1989; Nolle et al., 1989). Concerning time-frequency domain, useful information can be extracted given the non-stationary nature of the VF signal, obtaining the continuous temporal evolution of frequency values in the ECG signal and thus, being able to detect changes leading to pathologic rhythms. Algorithms based on time-frequency distributions have also been proposed to detect VF episodes (Afonso and Tompkins, 1995; Clayton and Murray, 1998; Rosado et al., 1999), and different signal processing techniques are applied by different authors to detect VF accurately (Amann et al., 2007; Bai and Wang, 2011; Li et al., 2012; Zhang et al., 2011). An in-depth review of different time, frequency, wavelet and other VF detection methods is described in (Amann et al., 2005).

The combination of ECG parameters in different domains has been suggested as a useful approach to improve detection efficiency. In (Clayton et al., 1994; Neurauter et al., 2007; Pardey, 2007), a set of temporal and spectral features was used as input variables of a neural network, exhibiting better performance than other previously proposed methods. Following this approach, other statistical learning algorithms such as clustering methods (Jekova and Mitev, 2002), support vector machines (SVM) (Übeyli, 2008) or Data Mining procedures (Rosado-Muñoz et al., 2002) have been explored to enhance detection capabilities. The approach presented in this work makes use of Self-Organizing Maps, using also a set of temporal and spectral features as input variables, to obtain visual information about the faced problem. This study proposes the use of a supervised SOM to obtain visual information about four important groups of patients: *VF* (Ventricular Fibrillation), *VT* (Ventricular Tachycardia), *HP* (Healthy Patients) and *AHR* (Anomalous Heart Rates and Noise). Moreover, SOM is used to extract knowledge about the variable values and the profile for each group of patients, assisting in gaining a deeper understanding of this clinical problem.

3.3.2 Data set

This section details the characteristics of the dataset used in this study and the parameters extracted from the ECG signals.

ECG data

ECG signals were collected from the AHA Arrhythmia Database[2] (8,200 series) and the MIT-BIH Malignant Ventricular Arrhythmia database[3] where a single ECG channel is used. A total of 29 patient recordings were analyzed, each containing an average of 30 minutes of continuous ECG, from which approximately 100 minutes corresponded to VF. For each record, non-overlapping 128 sample length segments sampled at 125Hz were used, giving a 1.024s of continuous time window segment for the analysis. In total 57,908 observations were obtained. Before parameter calculation, a general signal pre-processing was done, firstly subtracting the mean ECG signal value, and secondly, low-pass filtering with cutoff frequency of 40Hz to remove the electrical network interference and other high frequency components not relevant for the analysis.

Fibrillatory rhythms are characterized by an absence of regularity in the ECG signal due to the creation of multiple independent re-entry activation circuits in the heart tissue, avoiding the correct transmission of the sinus activation pulse. This disorganized contraction of the ventricle fails to effectively eject blood from the ventricle, provoking a heart collapse. Figures 3.35 and 3.36 show the time, frequency and time-frequency results for an analyzed ECG window segment in a normal sinus and a fibrillatory rhythm, respectively.

Time-frequency domain parameters

Each window segment was processed to obtain a set of temporal (t), spectral (f) and time-frequency (tf) domain parameters. Before parameter calculation in the tf distribution, a denoising is done, consisting on removing all components less than 10% of the maximum spectral density value. These low value components are formed by noise or small interference terms (time-frequency distributions generate the so-called

[2] http://ecri.org (American Heart Association ECG Database)

[3] http://physionet.org

interference terms) (Cohen, 1995) and they do not add useful information.

Parameters are obtained from the Pseudo Wigner-Ville (PWV) time-frequency distribution as previously described in (Rosado et al., 1999; Rosado-Muñoz et al., 2002), only two out of twenty seven parameters come from the temporal signal, the rest are obtained from the PWV distribution.

In order to characterize and differentiate fibrillatory episodes from other cardiac rhythms, two spectral bands of interest were defined (Herschleb et al., 1979) in the tf domain. Since most of the energy components of fibrillatory episodes reside in the low frequencies band, a low frequency band (2-14Hz) called BALO was defined. A high frequency band (BAHI, 14-28Hz) was also considered, mainly containing energy components for non-VF rhythms. Based on the PWV distribution and the defined frequency bands, a number of temporal, spectral, and time-frequency parameters have been obtained (see Table 3.6). The chosen parameters provide different information about power spectral distribution along time, duration of significant frequency bands (those containing sinus or fibrillatory rhythms), and in general, all information leading to different measures from ECG rhythms that could provide useful information depending on the patient's pathology. The selection of relevant time-frequency parameters is very important for a sucessful further analysis and was carefully chosen after data analysis. A detailed description of the parameters can be found in (Rosado-Muñoz et al., 2002; Atienza et al., 2006; Alonso-Atienza et al., 2012).

Parameterization of ECG signal segments results in an input data set to the SOM consisting of $L = 57,908$ observations each containing 27 features. Each observation was labeled into four groups according to different rhythms, which appeared with different prior probabilities: *HP* ($p_1 = 40.25\%$), for healthy patients; *VT* ($p_2 = 8.84\%$), for ventricular tachycardia (VT) including their variants (regular VT, polymorphic VT or "torsade de pointes"); *VF* ($p_3 = 10.66\%$), for VF signal and flutter; and *AHR* ($p_4 = 40.25\%$), comprising the rest of cardiac rhythms.

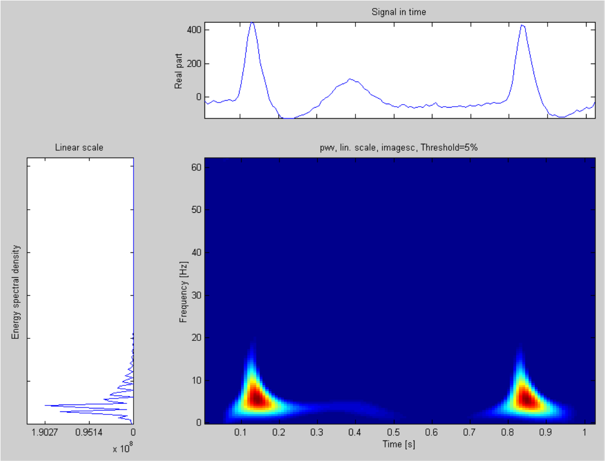


Figure 3.35: *Normal sinus rhythm in the surface ECG. Temporal signal (top) and its associated frequency (left) and PWV time-frequency representations.*

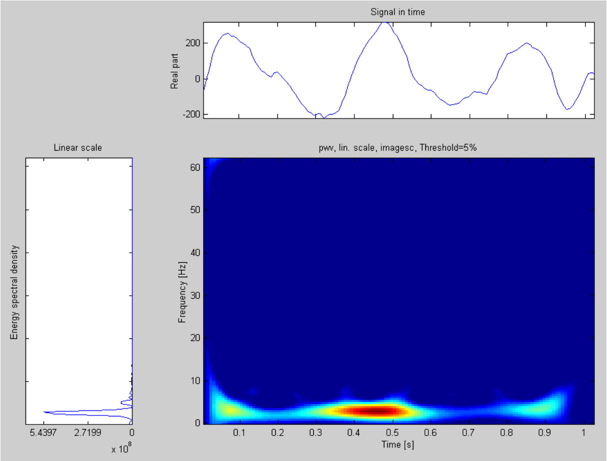


Figure 3.36: *Ventricular fibrillation rhythm in the surface ECG. Temporal signal (top) and its associated frequency (left) and PWV time-frequency representations.*

Table 3.6: *Obtained time-frequency parameters (mean ± std), where the "t" and "tf" in the "Domain" column refers to temporal and time-frequency domains.*

Variable	**Domain**	**HP**	**AHR**	**VT**	**VF**
`VR`	t	$(8.2 \pm 6.7) \cdot 10^{+0}$	$(6.0 \pm 5.0) \cdot 10^{+0}$	$(1.6 \pm 3.4) \cdot 10^{+0}$	$(1.5 \pm 1.1) \cdot 10^{+0}$
`ratiovar`	t	$(1.6 \pm 0.5) \cdot 10^{+0}$	$(1.8 \pm 0.5) \cdot 10^{+0}$	$(2.5 \pm 0.6) \cdot 10^{+0}$	$(2.7 \pm 0.4) \cdot 10^{+0}$
`pmxfrec`	f	$(5.5 \pm 3.2) \cdot 10^{+0}$	$(4.0 \pm 2.5) \cdot 10^{+0}$	$(2.8 \pm 2.0) \cdot 10^{+0}$	$(2.6 \pm 1.2) \cdot 10^{+0}$
`maximfrec`	f	$(2.2 \pm 0.8) \cdot 10^{+1}$	$(2.0 \pm 0.7) \cdot 10^{+1}$	$(1.5 \pm 0.8) \cdot 10^{+1}$	$(1.4 \pm 0.5) \cdot 10^{+1}$
`minimfrec`	f	$(7.3 \pm 4.9) \cdot 10^{-1}$	$(6.3 \pm 3.8) \cdot 10^{-1}$	$(6.4 \pm 3.5) \cdot 10^{-1}$	$(6.9 \pm 3.6) \cdot 10^{-1}$
`tsnz`	tf	$(1.1 \pm 0.6) \cdot 10^{+3}$	$(1.1 \pm 0.6) \cdot 10^{+3}$	$(1.6 \pm 0.5) \cdot 10^{+3}$	$(1.5 \pm 0.4) \cdot 10^{+3}$
`tsnzl`	f	$(6.4 \pm 3.1) \cdot 10^{+2}$	$(6.8 \pm 3.0) \cdot 10^{+2}$	$(1.2 \pm 3.1) \cdot 10^{+2}$	$(1.2 \pm 3.0) \cdot 10^{+2}$
`qtl`	f	$(0.6 \pm 1.0) \cdot 10^{-1}$	$(6.5 \pm 1.0) \cdot 10^{-1}$	$(7.7 \pm 1.1) \cdot 10^{-1}$	$(8.1 \pm 1.1) \cdot 10^{-1}$
`tsnzh`	f	$(2.0 \pm 2.3) \cdot 10^{+2}$	$(1.8 \pm 2.2) \cdot 10^{+2}$	$(1.5 \pm 2.1) \cdot 10^{+2}$	$(1.2 \pm 1.7) \cdot 10^{+2}$
`qth`	f	$(1.8 \pm 1.0) \cdot 10^{-1}$	$(1.5 \pm 0.9) \cdot 10^{-1}$	$(0.8 \pm 0.9) \cdot 10^{-1}$	$(0.6 \pm 0.7) \cdot 10^{-1}$
`mdl8`	t	$(9.1 \pm 4.1) \cdot 10^{+1}$	$(8.6 \pm 3.8) \cdot 10^{+1}$	$(6.8 \pm 3.5) \cdot 10^{+1}$	$(6.1 \pm 2.4) \cdot 10^{+1}$
`vdl8`	t	$(9.7 \pm 4.2) \cdot 10^{+1}$	$(8.7 \pm 3.8) \cdot 10^{+1}$	$(4.9 \pm 2.8) \cdot 10^{+1}$	$(4.5 \pm 2.0) \cdot 10^{+1}$
`te`	tf	$(0.6 \pm 1.0) \cdot 10^{+9}$	$(0.2 \pm 5.1) \cdot 10^{+10}$	$(0.1 \pm 2.0) \cdot 10^{+11}$	$(1.2 \pm 1.9) \cdot 10^{+9}$
`tel`	f	$(4.8 \pm 7.0) \cdot 10^{+8}$	$(0.1 \pm 2.6) \cdot 10^{+10}$	$(0.7 \pm 9.3) \cdot 10^{+10}$	$(1.1 \pm 1.5) \cdot 10^{+9}$
`qtel`	f	$(7.1 \pm 1.1) \cdot 10^{-1}$	$(7.3 \pm 1.1) \cdot 10^{-1}$	$(8.3 \pm 1.0) \cdot 10^{-1}$	$(0.9 \pm 1.0) \cdot 10^{-1}$
`teh`	f	$(0.8 \pm 1.2) \cdot 10^{+8}$	$(0.4 \pm 18.) \cdot 10^{+9}$	$(0.3 \pm 7.3) \cdot 10^{+10}$	$(0.3 \pm 1.2) \cdot 10^{+8}$
`qteh`	f	$(1.7 \pm 1.2) \cdot 10^{-1}$	$(1.1 \pm 0.8) \cdot 10^{-1}$	$(0.5 \pm 0.7) \cdot 10^{-1}$	$(0.3 \pm 0.5) \cdot 10^{-1}$
`ct8`	t	$(3.7 \pm 1.6) \cdot 10^{+0}$	$(3.9 \pm 1.5) \cdot 10^{+0}$	$(6.3 \pm 1.3) \cdot 10^{+0}$	$(6.2 \pm 1.3) \cdot 10^{+0}$
`tmy`	tf	$(1.5 \pm 0.7) \cdot 10^{+2}$	$(1.5 \pm 0.6) \cdot 10^{+2}$	$(2.9 \pm 1.2) \cdot 10^{+2}$	$(2.7 \pm 1.3) \cdot 10^{+3}$
`curve`	f	$(1.4 \pm 1.7) \cdot 10^{-1}$	$(1.7 \pm 1.7) \cdot 10^{-1}$	$(-1.0 \pm 2.8) \cdot 10^{-1}$	$(-1.8 \pm 3.0) \cdot 10^{-1}$
`nareas`	tf	$(1.4 \pm 0.7) \cdot 10^{+0}$	$(1.4 \pm 0.9) \cdot 10^{+0}$	$(2.0 \pm 0.9) \cdot 10^{+0}$	$(1.8 \pm 0.8) \cdot 10^{+0}$
`lfrec`	f	$(9.9 \pm 4.5) \cdot 10^{+0}$	$(8.0 \pm 3.1) \cdot 10^{+0}$	$(6.1 \pm 4.2) \cdot 10^{+0}$	$(5.0 \pm 1.5) \cdot 10^{+0}$
`maxfrec`	f	$(1.3 \pm 0.5) \cdot 10^{+1}$	$(1.0 \pm 0.4) \cdot 10^{+1}$	$(0.8 \pm 0.5) \cdot 10^{+1}$	$(0.7 \pm 0.2) \cdot 10^{+1}$
`minfrec`	f	$(2.6 \pm 1.6) \cdot 10^{+0}$	$(2.2 \pm 1.4) \cdot 10^{+0}$	$(1.9 \pm 0.9) \cdot 10^{+0}$	$(2.0 \pm 0.8) \cdot 10^{+0}$
`ltmp`	t	$(1.5 \pm 1.1) \cdot 10^{+1}$	$(1.7 \pm 1.3) \cdot 10^{+1}$	$(3.4 \pm 2.1) \cdot 10^{+1}$	$(3.5 \pm 2.2) \cdot 10^{+1}$
`dispersion`	tf	$(2.1 \pm 4.6) \cdot 10^{+0}$	$(1.9 \pm 4.6) \cdot 10^{+0}$	$(5.9 \pm 7.7) \cdot 10^{+0}$	$(5.8 \pm 7.8) \cdot 10^{+0}$
`area`	tf	$(1.3 \pm 1.1) \cdot 10^{+2}$	$(1.3 \pm 1.0) \cdot 10^{+2}$	$(1.9 \pm 1.4) \cdot 10^{+2}$	$(1.7 \pm 1.1) \cdot 10^{+2}$

3.3.3 Methodology

This section presents the methodology followed in this study. In order to explore different possibilities and results, three analysis were performed. This section is divided into three subsections which detail the different experiments carried out and their results.

In the present study, a supervised SOM was used because a supervised problem is faced, that is, each pattern is associated with a class, so it is essential to obtain results according to that information. Therefore, it is interesting to use a supervised SOM that provides information about the class in the training. The supervised SOM algorithm creates, initializes and trains a supervised SOM. It constructs the training data by adding M (number of classes) columns to the original data based on the class information. Therefore, the dimension of vectors after the process is $N+M$ (dimension of the input vectors + number of different classes). In each input vector, one of the new components has value '1' (if the input vector belongs to the class corresponding with the new component), and others '0' (if it does not belong to this class). After this, the classical approach presented in Chapter 2 is carried out. Then, the class of each map unit is determined by taking the maximum over these added components, and a label is given accordingly. Finally, the extra components are removed. In this work, instead of labeling a map with the labels provided by the algorithm, a colored "Hits map" is presented.

The first problem found in this work is the large number of variables as discussed in Section 3.3.2, some of which may be redundant. Therefore, a feature selection must be carried out. In a previous work (Atienza et al., 2006), it is proposed the use of nonparametric bootstrap resampling technique using the same type of data as in the present work to provide a criterion for feature selection (11 features were selected). After selecting the variables pointed out in (Atienza et al., 2006), the training of several SOM was carried out. For the training, it was followed the same procedure as in the previous sections so that different options of the tuning parameters of the SOM algorithm were tested, combining all the possibilities, which provided 4008 different maps. Finally, it was selected the SOM that showed the minimum topographic error (Kiviluoto, 1996), which measures the topology preservation between the original space and the final space. To summarize, Figure 3.37 shows the methodology carried out in this study. The first step after recording the ECG is the segmentation of

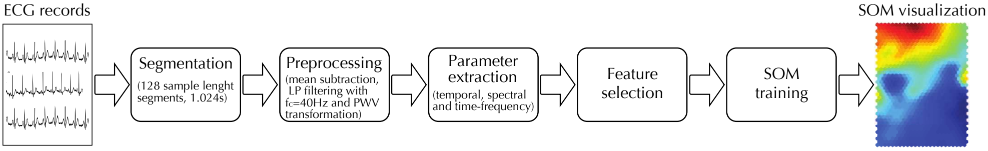

ECG records
Segmentation
(128 sample lenght segments, 1.024s)
Preprocessing
(mean subtraction, LP filtering with fc=40Hz and PWV transformation)
Parameter extraction
(temporal, spectral and time-frequency)
Feature selection
SOM training
SOM visualization

Figure 3.37: *Methodology carried out in this study.*

the signals, giving a 1.024s of continuous time window segment for the analysis. In total 57,908 observations were obtained. After segmentation, a pre-processing of each observation is done, which consists of mean subtraction and low-pass filtering. Afterwards, the parameter extraction is done by processing each window segment to obtain a set of temporal (t), spectral (f) and time-frequency (tf) domain parameters. After that, a feature selection is carried out, obtaining the final data set to train the SOM. This set consists of a matrix with 57,908 observations (rows) and 11 variables (columns). The last steps are the SOM training, and visualization by means of the component planes (before that, the hits map will be presented in order to compare the different patterns associated to each class by the distribution of their "hits" on the map).

3.3.4 Results

Supervised SOM training

This section presents the results obtained with the data set put forward in Section 3.3.2. Herein, the four pathology groups (*VF*, *VT*, *HP* and *AHR*) are included separately in the training. Once the map training is finished, the visualization of the two-dimensional map provides qualitative information about the input variables relationships for the data set used to train the map. Before visualizing the SOM component planes, the "Hits map", presented in Section 2.3.2, is represented in order to get spatial information about the classes in the map (Figure 3.38), that is, where each patient (corresponding to each class) is placed in the map.

The "Hits map" shows pathology groups with different colors (corresponding to each class) instead of directly labeling the map with the labels provided by the algorithm, as mentioned previously. This map provides more information than simply labeling because in each neuron (each hexagon on the map grid) we have information

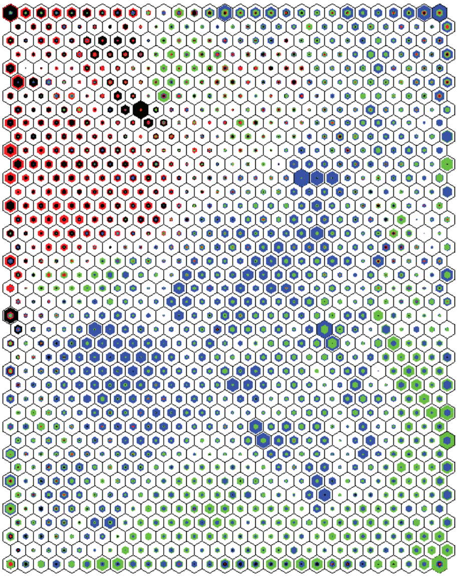


Figure 3.38: *"Hits map" obtained from the training with four groups of patients. VF is represented in red, VT in black, HP in green and AHR in blue.*

about all classes presented in a neuron instead of representing only the predominant class by a color.

Figure 3.38 shows that the classes labeled as *VF* (patients who suffer from ventricular fibrillation, represented in red) and *VT* (patients who suffer from Ventricular Tachycardia, represented in black) are very similar since they are completely overlapped. This is due to the fact that, in many cases, *VT* is an early stage of *VF* and thus, the pathologies can be considered as very similar in case of pathology profiling although the clinical recovering therapy for VT and VF is not the same. However, VT could be considered as the beginning of VF and thus, it was decided to join both classes in one to be able to extract knowledge, and to visually identify which variables are important to obtain differences between healthy patients and patients with Ventricular Fibrillation or Ventricular Tachycardia.

Supervised SOM training merging *VF* and *VT* classes

In this section, the results corresponding to the map using supervised training for three groups of patients are presented. These three groups correspond to: patients with Ventricular Fibrillation or Ventricular Tachycardia (*VFVT* as the merging of *VF* and *VT* classes), healthy patients (*HP*) and anomalous heart rates and noise (*AHR*).

Figure 3.39 shows the "Hits map" obtained from the training with these three groups of patients. It shows that *AHR* class (which included both anomalous heart rates and noise, represented in blue) spreads over the *HP* class (represented in green), which is not critical. That is, it would be a problem that other heart rates or noise would be overlapped with Ventricular Fibrillation or Ventricular Tachycardia because, in this case, it would not be possible to profile the diseases of interest (VF and VT). In this work, it is of great interest profiling patients with Ventricular Fibrillation or Ventricular Tachycardia pathology versus healthy patients using visual information of the variables in order to observe differences between both groups. Therefore, the *AHR* class is not determinative to profile healthy patients versus those suffering *VF* or *VT*. Moreover, to carry out the study in ideal conditions (without taking into account noise, only to obtain visual information between the differences in healthy and non-healthy patients) is of interest in order to profile patients suffering from *VF* and *VT* versus healthy patients. Due to these facts, further analysis is done drawing the patterns belonging to this class.

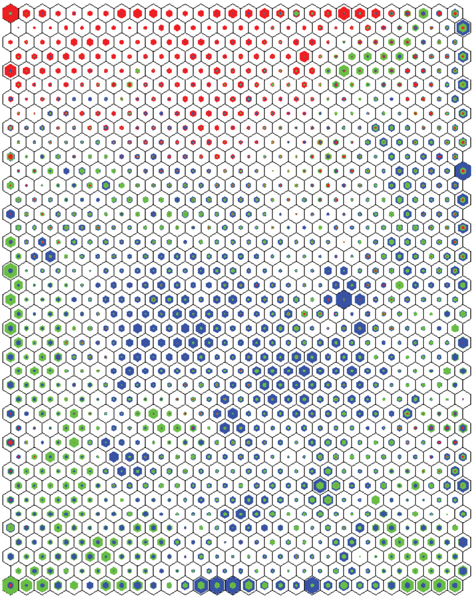


Figure 3.39: *"Hits map" obtained from the training with three groups of patients. VFVT is represented in red, HP in green and AHR in blue.*

Supervised SOM training merging *VF* and *VT* classes without *AHR* class

This section presents the results corresponding to the map using supervised training with two groups of patients (ideal condition), as mentioned previously. These two groups correspond to: patients who suffer Ventricular Fibrillation or Ventricular Tachycardia (*VFVT*) and healthy patients (*HP*).

Figure 3.40 represents "Hits map" obtained from the training with the two above-mentioned groups of patients. Figure 3.40 shows that *HP* class (represented in green) and *VFVT* class (represented in red) are clearly distinguishable. In general terms, the patients who suffer from some Ventricular Fibrillation or Ventricular Tachycardia (*VFVT*) are located at the bottom of the map, whereas the healthy patients (*HP*) are located at the top of the map. Therefore, there exist differences in the behavioral profile of each group of patients. To analyze the profiles of each group of patients, the component planes obtained after training the SOM (Figure 3.41) must be examined.

Figure 3.41 shows that there are three important variables when differentiating between healthy patients and those suffering from Ventricular Fibrillation or Ventricular Tachycardia. These variables are `qtel`, `ct8` and `curve`, described as:

- `qtel`: Percentage of the total spectral density contained in the BALO band. In case of a *VFVT*, the main spectral density is located in the BALO band.

- `ct8`: The time axis of the PWV distribution is divided into eight window subsegments. Then, for every subsegment, the energy in the BALO band is measured. The `ct8` corresponds to the number of subsegments that contain at least half of the spectral density if the total density of the band would be equally distributed along the time axis.

- `curve`: A vector containing the number of non-zero terms at every frequency bin of spectral resolution in the BALO and BAHI bands is computed and the curvature of the parabolic approximation of the vector is obtained. This value gives information about distribution of frequency terms along time. In case of *HP*, frequency distribution is not regular due to the QRS existence (high frequency terms are dominant) and the curvature is higher, which is contrary to the existence of an *VFVT* rhythm where frequency distribution is spread along all analyzed frequencies and the parabolic approximation curvature is lower.

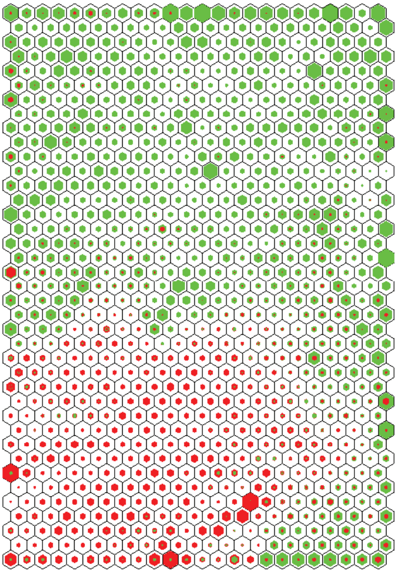


Figure 3.40: *"Hits map" obtained from the training with two groups of patients. VFVT is represented in red and HP in green.*

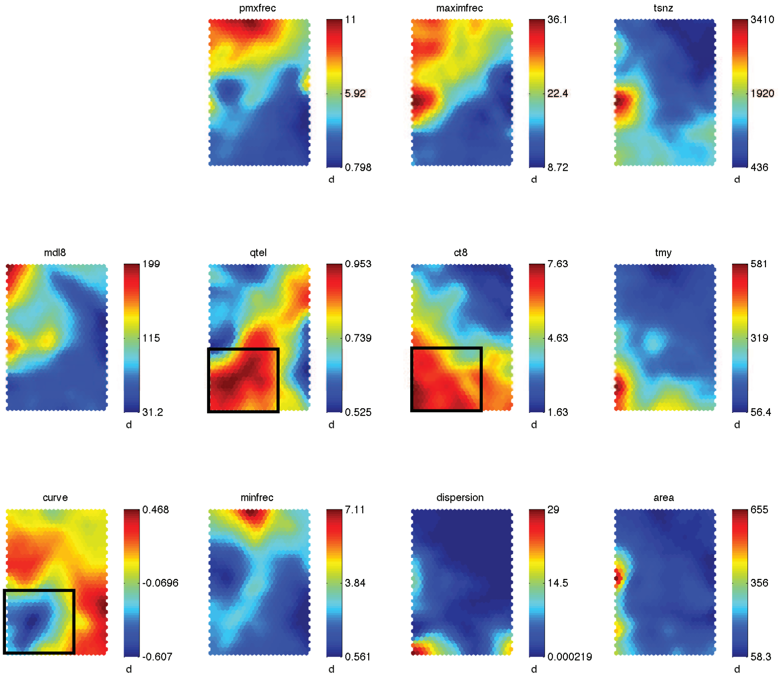


Figure 3.41: *"Component planes" obtained from the training with two groups of patients (VFVT and HP).*

These features are relevant due to the fact that the values of these variables undergo significant changes in the area where the patients suffering from Ventricular Fibrillation or Ventricular Tachycardia are located (bottom of the map) with respect to the area in which healthy patients are situated (top of the map). In case of patients who suffer from *VF* or *VT* the `qtel` and `ct8` variables take higher values (see areas with frame in Figure 3.41), whereas the `curve` variable takes lower values. This is contrary to healthy patients, showing lower values in the rest of the map to the variables `qtel` and `ct8` (except in the upper right area of `qtel` map) and higher values in the rest of map for the `curve` variable. The importance of these variables is due to the fact that *VF* and *VT* pathologies are very irregular both in time and frequency. Concerning `qtel`, the non-existence of a front wave in the heart avoids the blood being pumped from the heart (QRS absence) in case of *VF* and all spectral density components are mainly located in the BALO band. A similar reason arises in case of `ct8` due to the fact that distribution of density components along time is more regular in *VF* than in case of ECG existence (healthy patients) where specific time instants concentrate most of the spectral density. Thus, eventhough the variables use both domains, `qtel` and `ct8` provide relevant information related to frequency domain and time domain respectively. Finally, `curve` provides a combined information for both domains, showing that, due to special characteristics of *VF*, an adequate discrimination algorithm requires the usage of different domains.

Other training results obtained when trained all classes separately and merging *VF* and *VT* classes showed the same behavior with regard to the `qtel`, `ct8` and `curve` variables, being the most important to separate between patients who suffer from *VF* or *VT* and healthy patients.

3.3.5 Conclusions

This study proposes the use of a supervised Self-Organizing Map (SOM) to extract qualitative information about how the input variables are related to each other about four groups of patients: *VF*, *VT*, *HP* and *AHR*. In order to address the problem, three different trainings were carried out. Firstly, a supervised training considering all classes. It was noted that *VF* and *VT* classes were very similar (they were placed in the same location of the map) since, in many cases, *VT* is an early stage of *VF* and thus, the pathologies can be considered as very similar in case of pathology profiling.

For this reason, it was decided to merge these two classes since this fact does not entail an extremely relevant fact for visual analysis of the problem.

Finally, after noting that the AHR class did not provide relevant information to the problem, and with the aim of carrying out a final study only with the classes corresponding to patients who suffer Ventricular Fibrillation or Ventricular Tachycardia ($VFVT$) and healthy patients (HP), it was decided to draw the patterns corresponding to this class. Thus, the analysis was targeted to visual distinction between healthy patients and those suffering from Ventricular Fibrillation or Ventricular Tachycardia using the component planes obtained after training the SOM. It was observed a clear visual separation, resulting that the most relevant variables were `qtel`, `ct8` and `curve`. This analysis also showed that it was possible to perform a profile of patients suffering from Ventricular Fibrillation or Ventricular Tachycardia and other corresponding to healthy patients.

3.4 Visual Data Mining in physiotherapy field using Self-Organizing Maps

3.4.1 Introduction

Clinical data provide information that enables us to establish new and better diagnoses and treatments for certain pathologies. In physical therapy the analysis of clinical data is of particular significance because of its wide range of research options in relation to patients, pathologies and their treatment, and the important number of variables that can influence the evolution of an injury and its recovery. A complete and accurate analysis of the data can contribute to the development of more effective therapies and treatments for the patient. It should be noted that data in the clinical area (and specifically in physiotherapy) have a set of special characteristics compared to other kinds of data (DeMets et al., 2006), namely:

1. The human body and its interaction with its environment is one of the most complex systems that exist. Therefore, it is logical to consider these relationships might be non-linear, that is, an increase of one cause may not lead to a proportional increase of its effects.

2. There are many variables that define the evolution of an injury.
3. The patients' data collection sheets of a particular pathology or disease may be incomplete or contain errors of measurement.
4. The clinical data increase gradually over time, so the best models to apply are those that can take into account new data reliably.

These characteristics entail that the use of classical statistical models (i.e. multivariate regression or logistic regression) might be unsuitable given that these models do not highlight the subjectivity and the noise that, in many cases, affect these data. An alternative for the knowledge extraction from data is Visual Data Mining. In this case, a multidimensional visualization of the variables on the whole is considered (Chen et al., 2008a). In this way, the clinical specialist could extract his own conclusions with no need of learning the underlying of the models that the data specialist develops. As an example, if the results of a logistic regression are exposed, it is necessary to know what is understood by confidence intervals for the parameters, which are the initial hypotheses of the model as well as the interpretation of the model output. A visual approximation to the data analysis avoids all these problems since the clinical specialist observes the different behaviors that include the data in a direct way.

3.4.2 Case study

Anterior Cruciate Ligament injury (ACL) is the most frequent lesion in the knee joint (Ageberg, 2002) and the most of torn ligaments occurs during the participation in sports activities (Gotlin and Huie, 2000). The main function of the anterior cruciate ligament is to avoid the anterior displacement of the tibia on the femur. Likewise, it limits the tibial rotation and hyperextension, being considered as the first stabilizer of the knee joint in the sagittal (Imran and O'Connor, 1998). The injury risk of the ACL increases with high momentum strength that is generated when the corporal movements locate the knee joint in varus or valgus (Lloyd et al., 2005). Nevertheless, the movements that entail tibial rotation are which cause about 70% of the torn ACL.

As consequence of the torn ACL, it is produced a mechanical insufficiency that is manifested with synovial changes and arthrokinetics restrictions. A functional insufficiency also appears due to the affectation of the neuromuscular, and postural control of the propioception, and the strength of the musculature that surrounds the articulation. All the foregoing, causes static and dynamic instability that produces alterations in the movement patterns due to the deficient behavior of the implied mechanisms, as well as due to the fear associated with an aggravation of the lesion. In short, symptoms with alterations at a biomechanical level are produced. Among the different surgical techniques, most authors consider the intra-articular reconstruction techniques, which consist in the replacement of the injured ACL (Matsumoto and Seedhom, 1994), as the most successful for avoiding the pivot and restoring the biomechanical normality of the knee. Nowadays, the most used intra-articular reparation procedures are the autografts and allografts. At this moment, the most used plasties are patellar tendon and ischiotibial. In this study the semitendinosus tendon graft was used. After surgery, the subject must undergo a period of rehabilitation. This period is considered as important as the surgery or even more (Ménétrey et al., 2008). Thus, in order to facilitate the functional recovery of the affected knee, it is crucial a monitoring, a control and an evaluation of the patient. Accordingly, it is of vital importance to evaluate the strength levels and muscular measurements. Thus, the aim of the present work is to evaluate the efficiency of a rehabilitation protocol after an ACL reconstruction beside an ischiotibial tendon autograft. With this aim it was studied the difference between post and pre-surgery of the thigh contour at 5 cm representing the volume of the vastus medialis muscle, at 10 cm representing the vastus lateralis muscle, at 20 cm representing the rectus femoris. Also the two-legs jump in the take off moment and the routing of the knee joint at the flexion and extension were studied. The goal of the present study is to check if the analysis of these variables make possible to know if the recovery process has satisfied its final aim. Together with the measurements of the thigh contour and the muscle strength, SOM analysis also included the age, weight and height of each patient. Table 3.7 shows the mean and standard deviation of the employed variables.

Table 3.7: *Mean and standard deviation of the variables used in the SOM analysis.*

Variables	**Mean**	**Standard deviation**
Age (years)	28.05	8.96
Weight (Kg)	76.16	9.00
Height (cm)	174.90	7.90
Measurement5 (cm)	2.7×10^{-3}	73.8×10^{-3}
Measurement10 (cm)	9.7×10^{-3}	46.8×10^{-3}
Measurement20 (cm)	0.19	45.9×10^{-3}
Strength_Z	9.0×10^{-3}	97.8×10^{-3}
Strength_ischio	-0.81	1.27
Strength_quadriceps	-0.11	0.54

3.4.3 Methodology

The SOM algorithm basically depends on three parameters: kind of initialization (random or linear), neighborhood function (Gaussian, Cut Gaussian, Bubble, Epanechnikov) and the kind of training (batch or sequential) as mentioned previously. For obtaining the best SOM, the same procedure as in the previous studies was carried out. It consists in a sweep of parameters in order to train the maps with all the possible combinations. For the case of random initialization, 100 different random initializations were carried out for every combination of the other parameters (neighborhood function and kind of training). The best network was selected considering the best as the minimum topologic error.

3.4.4 Results

Once the best map has been selected according to the minimum topologic error, the winners' map was represented, as shown in Figure 3.42. In this representation the number of patients that represent each neuron is shown. The hexagons totally filled black represent 3 patients, the medium filled represent 2 patients and the less filled represent 1 patient. This figure must be used with the component planes in order to establish how many patients follow a particular behavior.

The map has been split into 6 different zones as Figure 3.42 shows. These zones were chosen according to which zones of the map represented an interesting or par-

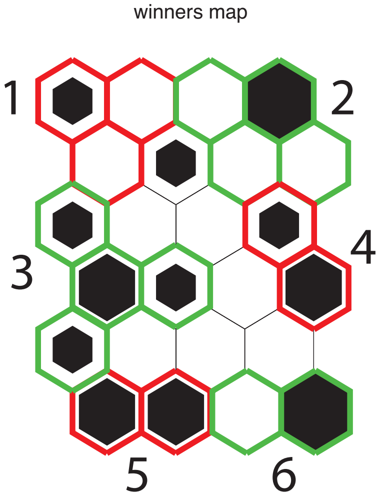


Figure 3.42: *"Winners map" of the obtained SOM.*

ticular behavior to study.

The component planes obtained after training the algorithm is shown in Figure 3.43. In general terms it can be seen that the variables *measurement5* and *strength_ischio* are highly correlated, because they have a very similar behavior since the upper left corner shows low values and the rest of the map shows high values. It is to note that, in fact, the high values in *strength_ischio* are negatives (see color bar), so in this strength there was no recovery for any of the patients in the study, but it can be affirmed that the decrement is higher or lower depending on the *measurement5*. However, the variable *measurement5* indicates the recovery of all the patients in the study except one patient located in the upper-left corner (see figure 3.42). Along with this, it can be observed that these two variables are inversely related to the variable *measurement10*.

As follows, the relationships among all the variables in each of the selected areas of study are explained.

- **Zone 1**: In this zone there is only one patient. It has been selected as a relevant zone because this pattern is far away from the others (it is an outlier) and it represents an abnormal or strange behavior. It is a medium age patient, medium weight and low height, as it can be seen on the component planes. Moreover,

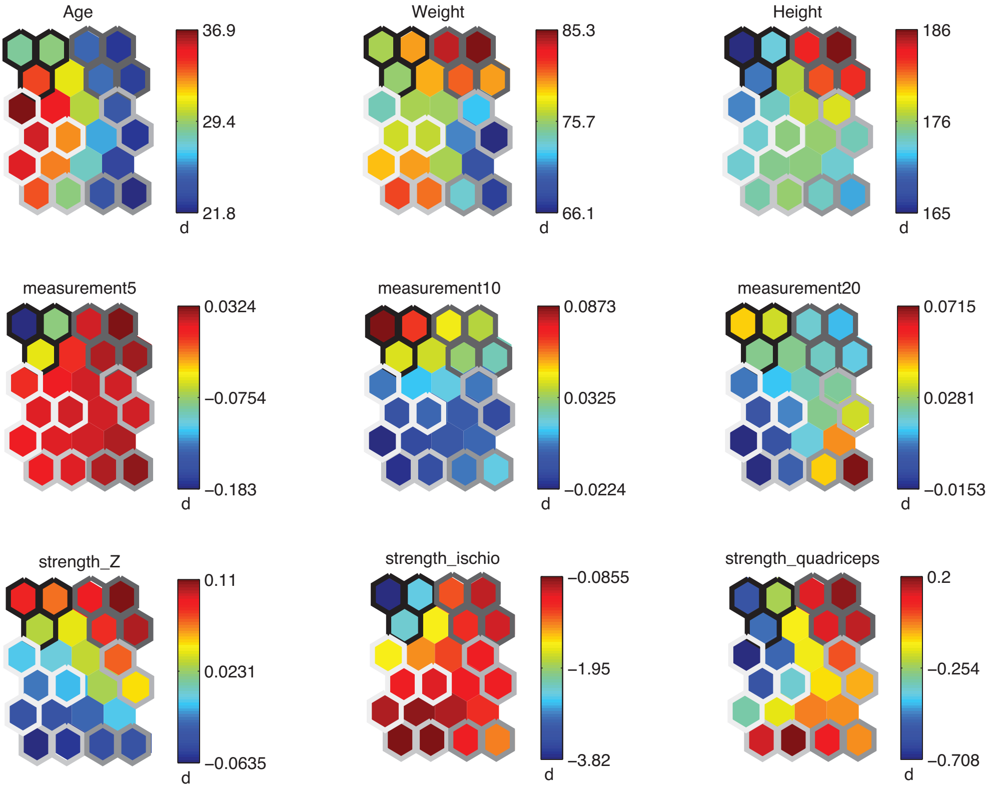


Figure 3.43: *Component planes obtained with SOM algorithm for patients after rehabilitation protocol.*

it can be observed that this patient has been recovered in measurements 10 and 20, in the former in a greater extent, and that measurement 5 has not been recovered. Regarding forces it can be observed that this patient recovers *strength_Z* but the same does not happen in *ischio* and *quadriceps*.

- **Zone 2**: In this zone, young patients and with heavy weight and height are found. It can be seen that these patients have been recovered in every measurement, being the 20th measurement to a lesser extent, in fact it has not incremented with respect to its initial value, it is about the same (note that the increment between the initial and final instants, that is what the SOM really represents, is about 0). Regarding the forces, it should be noted that in all of them, except in *strength_ischio*, there is a noticeable increment, so not only strength has been recovered but also it has been augmented. This group of patients represents a good enough group, the best of the study, because they have reached an excellent recovery regarding strengths and measurements, except in *strength_ischio*, that is not totally recovered, but it is near.

- **Zone 3**: This zone represents the older patients, medium-high weight and medium-low height. It can be observed in Figure 3.43 that in this zone of the map only the measurement 5 is recovered; the rest of the measurements have not been recovered or this has not been significant. Regarding the forces, in all of them the increment between the initial and final instants is negative, so there is no recovery in any case. This group of patients is not very desirable because they do not have a good recovery in general terms. This fact could be closely linked to the fact that these are the most aged patients in the study.

- **Zone 4**: This zone applies to young patients, low weight and medium-high height. In this zone it can be seen that there is a recovery of the measurements 5 and 20, while the measurement 10 has not been recovered or it has a negligible recovery. Regarding forces, there is a recovery in *strengthZ* and *strength_quadriceps*; and as in all the case studies, *strength_ischio* has not been recovered although this zone of the map represents one of the best zones of all the map in this type of force. Definitely, the recovery of the patients group belonging to this zone is positive given that there is an increment in two of their forces and measurements.

- **Zone 5**: In this zone, patients of medium-high age, high weight and medium height are allocated. These patients have a very good recovery of the measurement 5 while the same does not happen for the other measurements, in which the worst values of the map are found. Regarding forces, the only one recovered is *strength_quadriceps* although *strength_ischio*, that is not recovered, shows the best values of the map. In general these patients do not have a good recovery.

- **Zone 6**: In this zone are represented the youngest patients with low weight and height. It can be observed that in these patients the measurements 5 and 20 are considerably recovered while the measurement 10 is more or less the same. Although they can recover or remain equal, which is positive, it can be observed that only the *strength_quadriceps* has augmented.

3.4.5 Conclusions

In this study, a Visual Data Mining application is presented in the physical therapy field, which supposes a new approach in the knowledge extraction on this kind of data. With this approximation the clinic expert does not need the data specialist in order to interpret those models. According with the presented visualization, the clinical specialist is able to extract the data trends. In the case of thigh muscle contours, there were significant negative changes (decrease of the contour) on the vastus lateralis between pre- and post-surgery, but there was a final improvement of the overall thigh muscle contours at six months, due to the fact that a proper rehabilitation program was applied.

3.5 Use of SOMs for footwear comfort evaluation

3.5.1 Introduction

The footwear industry is, and has been, one of the main economic engines of some Spanish regions, such as some Southern regions in the province of Alicante. In the context of crisis times, as the period 2008-2013, specially stressed in Europe, the EU envisages to address this problematic situation pursuing the aim of promoting growth based on innovation and competitive and sustainable economy.

In this context of crisis and change, transformations needed in the global footwear industry are also being produced, which will configure a new competitive map and will test the strategic decisions of the companies in a particular difficult environment. With respect to Spain, footwear was affected by a reduction in consumption and a decrease in employment in 2009 (FundacionIndustrias, 2009). Therefore, in order to maintain levels of competitiveness and market share, the Spanish footwear industry must make the difference with respect to its main competitors. For this purpose, on one hand, it is found the product quality that has always been outstanding in the Spanish footwear, and in the other hand, the commitment to innovation and technology to produce better shoes and anticipate the future to maintain a privileged position in the coming years. This work is part of this second fact since it proposes the use of Visual Data Mining techniques, in particular the Self-Organizing Maps (SOMs), for evaluating data about comfort in footwear provided by *Instituto Tecnológico del Calzado y Conexas (INESCOP)*. For this purpose, the variables that may play a relevant role in this framework, and the crucial relationships between them, are studied for the comfort of a given shoe. This study tries to find the way of jointly represent valuations for different areas of the foot, with different variables (related to physical characteristics of the testers and characteristics or physical measures of foot-footwear) to see if there is a difference between buying/not buying groups.

The comfort in footwear is essential when walking because the foot is one of the structures of the human body that supports more weight when walking and further is the main shock absorber on the floor. Consumers are demanding ever higher levels of comfort and functionality on shoes. Shoe companies are aware of this and are investing a lot of efforts and resources to reach these levels, but it is a difficult task due to geometric differences presented between feet of the same size, and the differences between the designs built on the same last. The knowledge extraction about comfort in footwear can involve major improvements to both the user and footwear companies, and especially for users with some kind of foot problem as in the case of diabetic foot.

Comfort and functionality in footwear are the result of a complex interaction between the human body nature of the various elements of the footwear such as the shape, the properties of its components, materials and design of these components. Currently, the design of a shoe can be evaluated using two types of analyses: subjective and objective. Subjective analyses are based on comfort surveys conducted to users with standard feet, in size and shape, while walking with a given shoe. Objective analyses are based on measurements of biomechanical variables during the use of footwear under real or simulated conditions. Both analyses provide useful information for evaluating the comfort and functionality of the shoe; however, they do not guarantee comfort for any user and moreover, the equipment, instrumentation and personnel needed for its realization are costly in time and money. In this study, an approach based on Computational Intelligence is presented as an alternative to those procedures; it is based on a methodology to reduce the number of analyses needed to evaluate the footwear comfort.

3.5.2 Data set

The database shows a set of tests performed by 173 testers (of which 103 are women and 70 men) during the period between May 16th, 2010 to June 22nd of the same year. Analyzing the number of records, there are 1,624 of which 43% correspond with male testers and 57% with female ones. The database contains different variables corresponding to the tester, to the characteristics of foot-footwear, to the tester valuations and one dichotomous variable that indicates a hypothetical purchase. This last variable was not considered for training the SOM in order to avoid biasing the model.

As follows, the variables used in the SOM training are described and classified in the above-mentioned groups:

1. **Physical characteristics of the testers:**
    - Age.
    - Weight.
    - Height.
    - Gender.
2. **Characteristics of foot-footwear (percentage of difference between the last and the foot for different zones):**
    - Requested size on both feet.
    - Difference between the requested size and foot size (in percentage).
    - Projected width. Difference between the last and the foot (in percentage).
    - Heel width. Difference between the last and the foot (in percentage).

- Standard Ball girth. Difference between the last and the foot (in percentage).
- Oblique toe girth. Difference between the last and the foot (in percentage).
- Oblique toe width. Difference between the last and the foot (in percentage).
- Perpendicular toe width. Difference between the last and the foot (in percentage).

3. **Tester valuations on the different areas of the foot:**
   - Length. Valuation.
   - Projected width. Valuation.
   - Heel width. Valuation.
   - Medium instep girth. Valuation.
   - Anatomical Ball or joint girth. Valuation
   - Standard toe width. Valuation.

3.5.3 Results

This section presents the results obtained with the data set put forward in the previous section.

Figure 3.44 shows the projection of the above mentioned variables by means of the component planes of the trained SOM. In addition to the component planes shown in Figure 3.44, the winners map (Figure 3.45) must be analyzed. As mentioned in the Chapter 2, the number of patterns that are assigned to each neuron is proportional to the area that is filled in that neuron, thus, the larger the area, the larger the number of patterns assigned.

Figure 3.45 represents the “density” of data in the different areas of the map. Other interesting information is to know the behavior (buying or not buying) in different areas of the map (or group of neurons), as in the case study presented in Section 3.3 about ventricular fibrillation analysis, where different groups of patients were represented with different colors (corresponding to each class). This information is provided in Figure 3.46. The difference is that in the present study it has

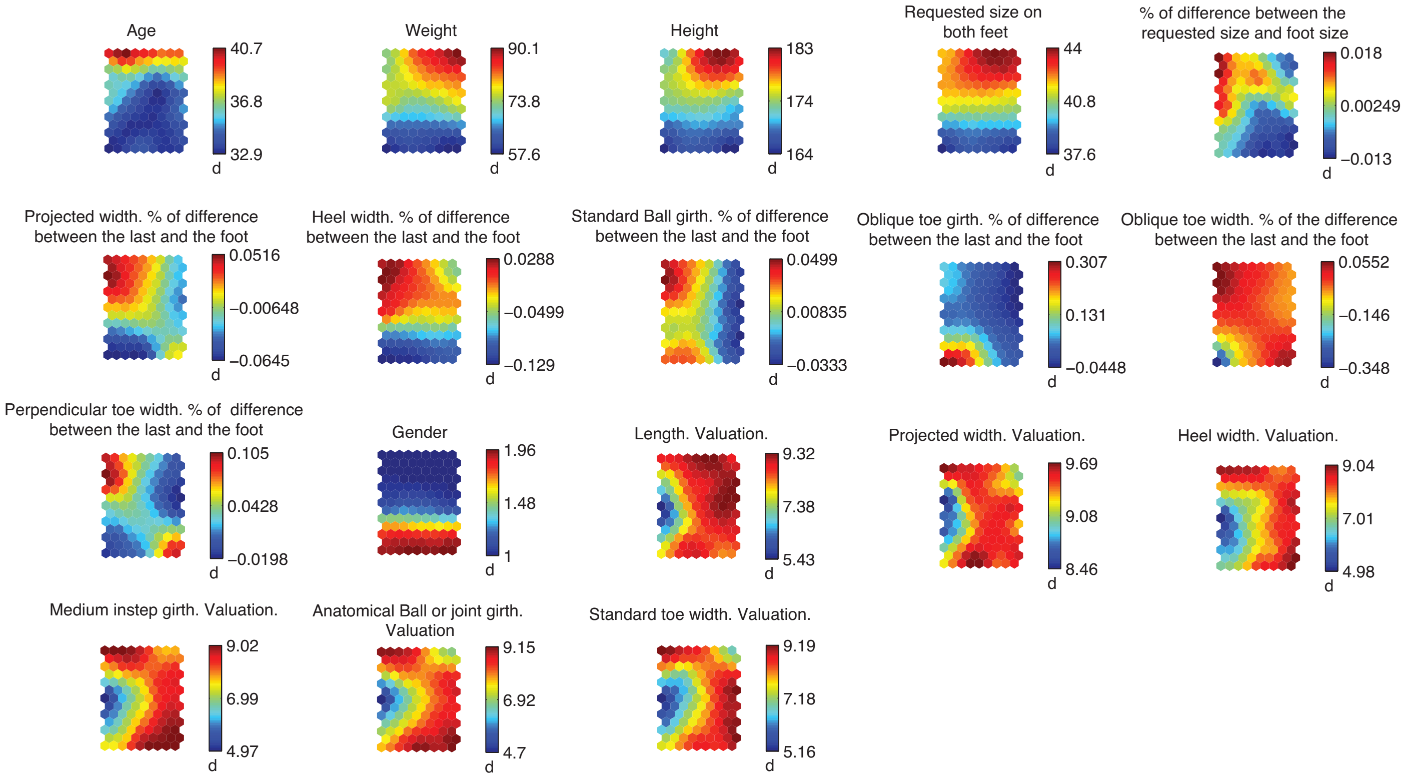


Figure 3.44: *Component planes obtained using after training the SOM with the complete data set.*

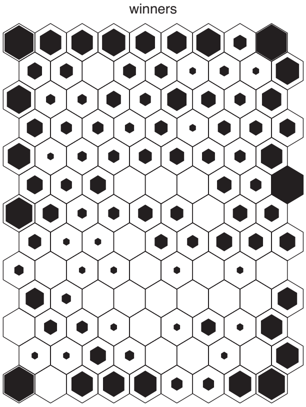


Figure 3.45: *Winners map obtained from the SOM training with the complete data set.*

been carried out an unsupervised problem, that is, information about the class (the dichotomous variable) is not included in the model training. This map (Figure 3.46) provides more information than simply labeling the majority class in a neuron because in each neuron (each hexagon on the map grid) we have information about all classes instead of representing only the predominant class by a color.

Paying attention to Figure 3.44 the following conclusions can be drawn:

- Firstly, some evident hypotheses are confirmed. Paying attention to *height* and *size* variables, it can be observed a clear correlation: the larger the height of the tester, the bigger the requested size. For example, in the upper right corner the tallest people are found, who requested largest sizes. On the other hand, in the bottom of the map, the smallest people are found, who ordered smallest sizes.

- Another obvious conclusion is the gender dependence with size. Looking at the *gender* variable (coded as 1 man and 2 women) it can be observed that women requested smaller sizes than men when comparing the component plane corresponding to *gender* with the component plane corresponding to *Requested size on both feet* (see the bottom part of both component planes).

- Regarding the physical characteristics of the testers, a couple of aspects attract attention: If compared the *height* and *weight* variables, the testers show to be well “proportioned” (they present normal body mass index), that is, the

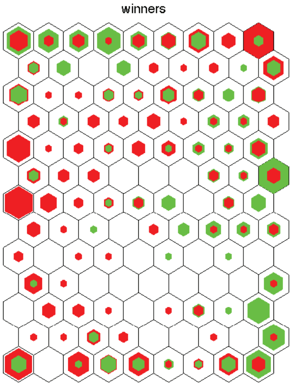


Figure 3.46: *Winners map obtained from the SOM training with the complete data set. Red color represents not buying behavior and green color represent buying behavior.*

heaviest testers are also the tallest ones whereas the shortest ones present the lowest weight. Moreover, another curious issue is that gender and age are clearly demarcated, that is, elderly testers are men and the younger ones are women. Looking at the component planes corresponding to *age* and *gender*, one can check that elderly male testers are located in the upper part. It is striking that there are no groups of testers with other physical characteristics (overweight) or elderly women.

- Paying attention to the different valuations, it can be observed that they are quite similar since if component planes corresponding to such valuations are compared, spatial areas that are most negative (blue color) are the same in all of them. This fact draws an important conclusion: taking all measures to obtain the valuation of the testers may not be necessary. If we try to clarify a bit more the resemblance between valuations, it can be observed that valuations corresponding to *projected width* and *toe width* are almost identical. Moreover, valuations corresponding to *heel width* and *medium instep girth* are quite similar.

- Another conclusion is that women are the testers who present the most negative difference between the last and the foot (in percentage) in heel width. There is a group of men that follow this trend, but to a lesser extent (upper right area of the map), which are also who present the most negative percentage of the difference between the requested size and the most negative percentage of the difference between the last and the foot of the projected width; these group corresponds to the tallest and heaviest testers of this study.

- The testers located in the upper right corner, mentioned in the previous item, make valuations decrease in that area. It can be observed that the variables corresponding to the valuations (except *Length valuation*) take values slightly lower in this area (except the middle left area, where the worst scores are found as discussed previously). The fact that the valuations decrease when the difference in percentage of the above-mentioned variables is low occurs only for men since, in the women case, low percentage differences in these variables do not lead to a decrease in their valuations. Therefore, it can be drawn as a conclusion that a tighter shoe has a negative influence on a man while this fact has no negative influence in a woman valuation.

- It is worth mentioning that the difference between *oblique toe girth* and *oblique toe width* (percentage difference between the last and the foot) is totally opposite. The *oblique toe girth* is positive only in a group of women; thus, in the majority of cases it is negative.

- Focusing on variables corresponding to characteristics of foot-footwear (percentage of difference between the last and the foot for different zones), they do not follow the same trend, that is, they differ for the several areas of the food. However, as mentioned before, the valuations for the different areas of the food are very similar. This fact reveals that the variables corresponding to characteristics of foot-footwear do not match with their respective valuations, which seems to be confuse since, a priori, the valuations on different foot areas should be related with the characteristics of foot-footwear in the same area. This may be due to the fact that the testers perform an overall assessment or overall rating (positive or negative for all the valuations) regardless of whether, in some areas of the foot, the last does not fit perfectly.

- It can be said that in the area where valuations are the worst ones (left side in Figure 3.44), the rest of variables (characteristics of foot-footwear) present values in the whole range (high, medium and low). This means that the variables corresponding to characteristics of foot-footwear are not related to valuations, that is, there is no variable or a group of them that are relevant or that directly influence on valuations. This only occurs for the variables *Oblique toe girth. Difference between the last and the foot (in percentage)* and *Oblique toe width. Difference between the last and the foot (in percentage)*, which present opposite behaviors. In the first case, it can be observed that for that area, this variable always presents low values, and in the second case, the variable only takes high values. That is, a low value for the first variable or a high value for the second, entails a poor valuation. However, this statement is not entirely conclusive because there are areas in such component planes that present low and high values respectively, and this fact does not lead to poor valuations.

- The component planes corresponding to *Projected width* and *Perpendicular toe width* are segmented in a similar manner. These component planes, are also similar to the *Standard Ball girth* one, except for a group of women who present higher values (bottom left corner). These variables are somehow related because all of them are related with the same thing or they measure different aspects the same target: the foot width.
- Relating the component planes (Figure 3.44) with the behavior when the tester bought (Figure 3.46), it can be observed that the group of testers with the worst valuations (left side in the middle of the map) corresponds to not buying. However note that a high valuation does not assure always a purchase. This is because the testers were encouraged to buy only in those cases in which the feeling of comfort was most excellent.

In the previous component planes (Figure 3.44), the scales of the variables were shown independently according to its own range. To verify differences between variables corresponding with tester valuations, the component planes were generated again taking the same color scale for this group of variables (Figure 3.47). The aforementioned conclusions are maintained (note that it is the same representation but the color bar scales of the valuations were combined to compare such variables). However, a new conclusion can be drawn: the *Projected width* valuation has much less variation than other valuations. In general, the value is quite high.

Due to the fact that testers may have different behavior when buying a shoe or ordering the size according to gender, which could directly affect to the other variables, it was decided to train a SOM for men and another one for women. Figures 3.48, 3.49 and 3.50 depict the maps (component planes, winners map and labels) corresponding

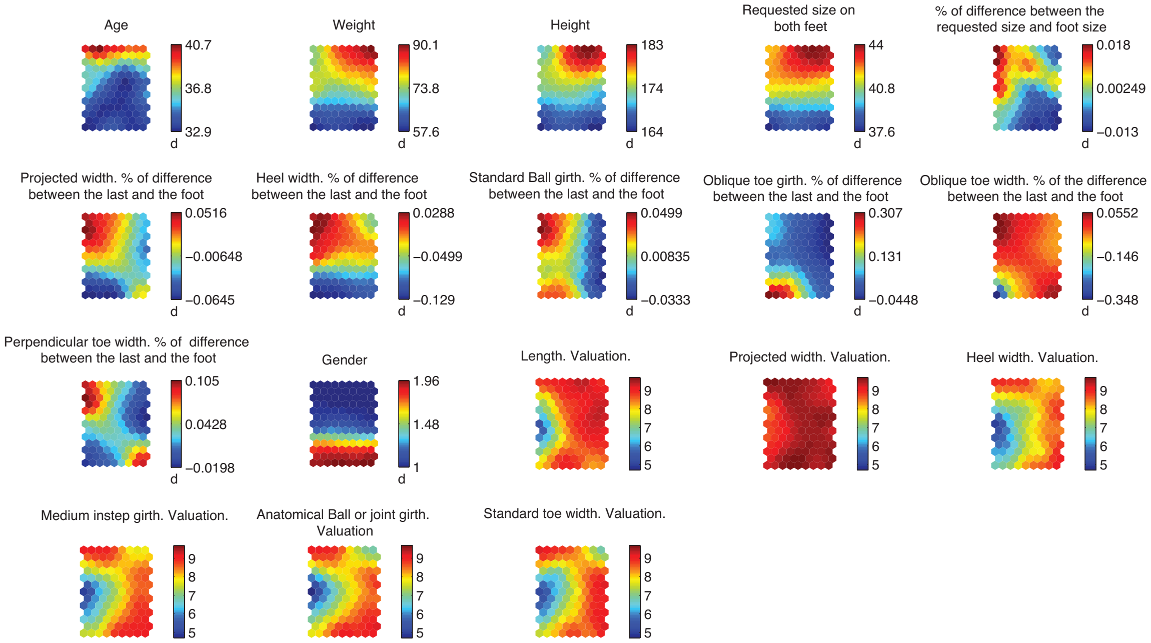


Figure 3.47: *Component planes obtained using after training the SOM with the complete data set with the same color scale for all component planes.*

to male testers.

Focusing on the male testers the next conclusions can be drawn:

- The variables *Standard Ball girth*, *Oblique toe girth*, *Oblique toe width* and *Perpendicular toe width* are correlated. This fact does not exist when the tester is a woman as discussed below. Given this correlation, taking only one of the four measures can be proposed when the tester is male.
- Regarding the variable *Difference between the requested size and foot size (in percentage)*, it can be observed that the testers who have a small value (tight fit of the footwear) correspond to those taking a large size (top of the map), which in turn correspond, logically, with the tallest and heaviest testers.
- About valuations, it is worth noting the similarity between *Heel width* and *Medium instep girth*. Moreover, *Length valuation* is also very similar to the previous ones, but in less extent. On the other hand, *Anatomical Ball or joint girth* and *Standard toe width* are also very similar between them.
- About the relationship between the valuations and the purchase, note that some correlation shows up; it is observed that the worst valuations, bottom right corner, have associated a group of people who do not buy (Figure 3.50). Also appears another curious group of men who do not buy (upper right corner). The valuations corresponding to *Anatomical Ball or joint girth*, *Projected width* and *Standard toe width*, which present medium and bad valuations, lead the tester not to buy. In this corner, the variables corresponding to *Difference between the requested size and foot size* and the other *differences between last and foot* also take low values. This fact again confirms the hypothesis that a tight shoe influences negatively the valuations of male testers.
- Another striking area to analyze is that located in the upper left corner. Paying attention to variables corresponding to valuations *Heel width*, *Medium instep girth*, *Anatomical Ball or joint girth*, *Standard toe width* it can be observed that there is a small patch of a lighter shade than the rest of the map (without taking into account the lower right corner which presents the lowest values). This means that the valuation decreases in this area, not presenting the maximum (about 9) as in almost the entire map. This entails that there exists testers that rejected the purchase of the footwear as shown in the same area of Figure

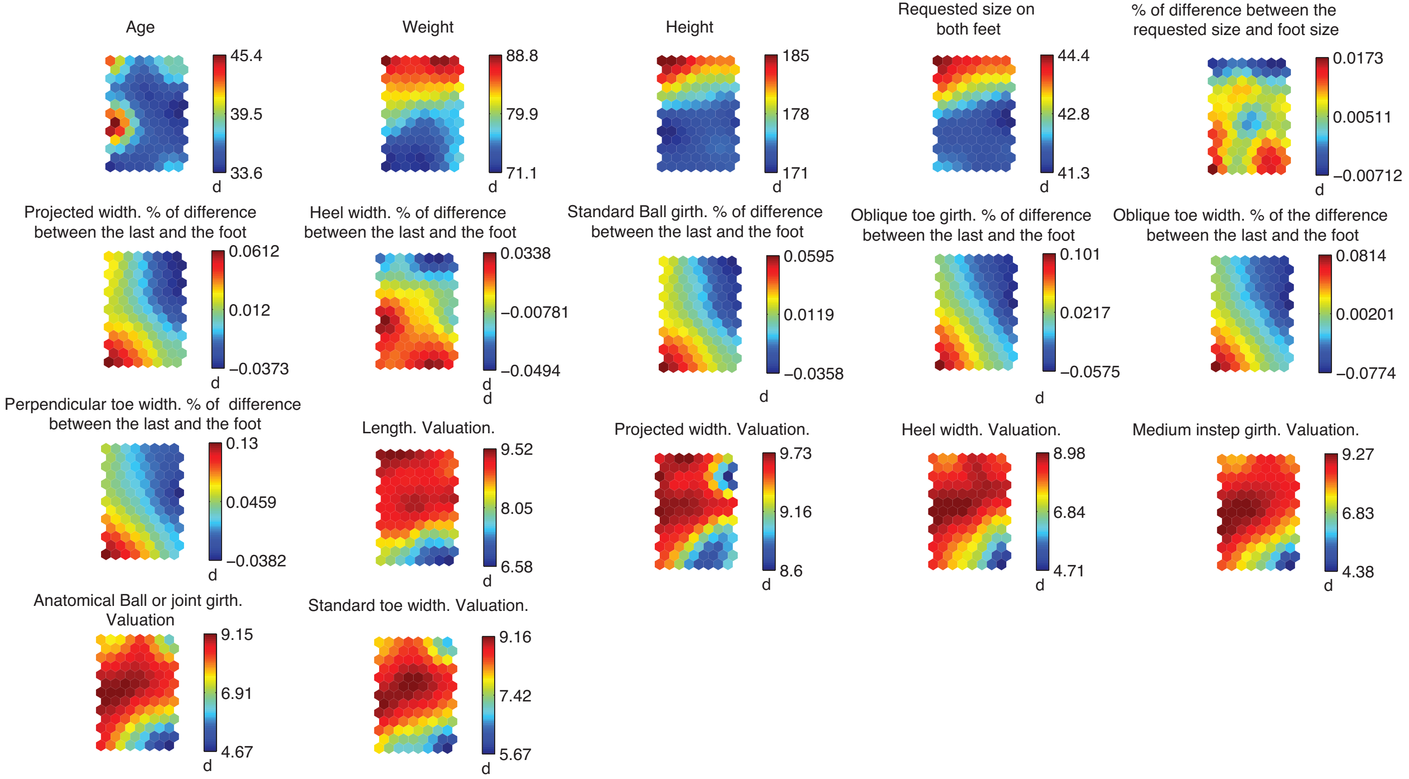


Figure 3.48: *Component planes obtained using after training the SOM considering only male testers.*

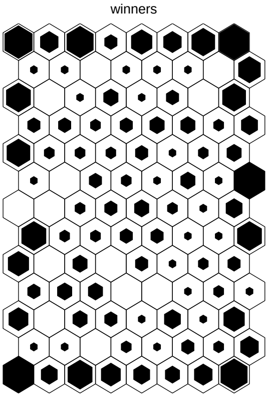


Figure 3.49: *Winners map obtained from the SOM training considering only male testers.*

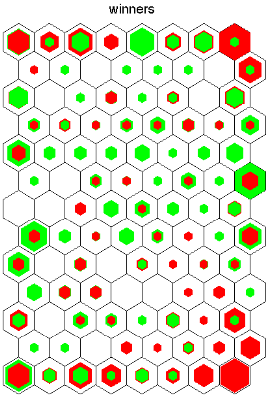


Figure 3.50: *Winners map obtained from the SOM training considering only male testers. Red color represents not buying behavior and green color represent buying behavior.*

3.50. Analyzing the area previously discussed in this figure, it can be seen that there are patterns as much from purchase as from the non purchase. This is because the intermediate values do not lead to a defined behavior of buying or not buying.

As follows, the analysis of female testers is presented. The maps corresponding to this analysis are depicted in Figures 3.51 3.52 and 3.53.

Focusing on the female testers the next conclusions can be drawn:

- Comparing these component planes with the corresponding with male testers, the first difference appears in the behavior between the *Requested size on both feet* and *Difference between the requested size and foot size*: there is no strong correlation between these two features. Furthermore, observing the variable *Difference between the requested size and foot size* it can be checked that women often ask for a tighter shoe. Paying attention to the color bars of these variables for both component planes (men and women), it can be checked that intermediate values on this color bars (green color) correspond to negative values for women and positive ones for men so that women have negative values around the entire map corresponding to such variable.

- Regarding the variables corresponding to the valuations, note that they are not always correlated, that is, they do not follow the same trend, contrary to the case of male testers training. This fact may indicate that women are more sensitive to different parts of the foot than men (men have a similar opinion, or global idea, about the different space areas of the footwear). However, there exists a resemblance between different pairs of valuations: *Projected width* and *Length*, *Heel width* and *Anatomical Ball or joint girth* and, finally, *Medium instep girth* and *Standard toe width*.

- There exists similarity between pairs of variables corresponding to the characteristics of foot-footwear, particularly between the percentage of difference between the last and the foot of *Standard Ball girth* and *Oblique toe girth*; and *Projected width* and *Perpendicular toe width*.

- The map of labels representing "purchase/non purchase" behavior (Figure 3.53) is divided by the secondary diagonal (from lower left to upper right corner). The

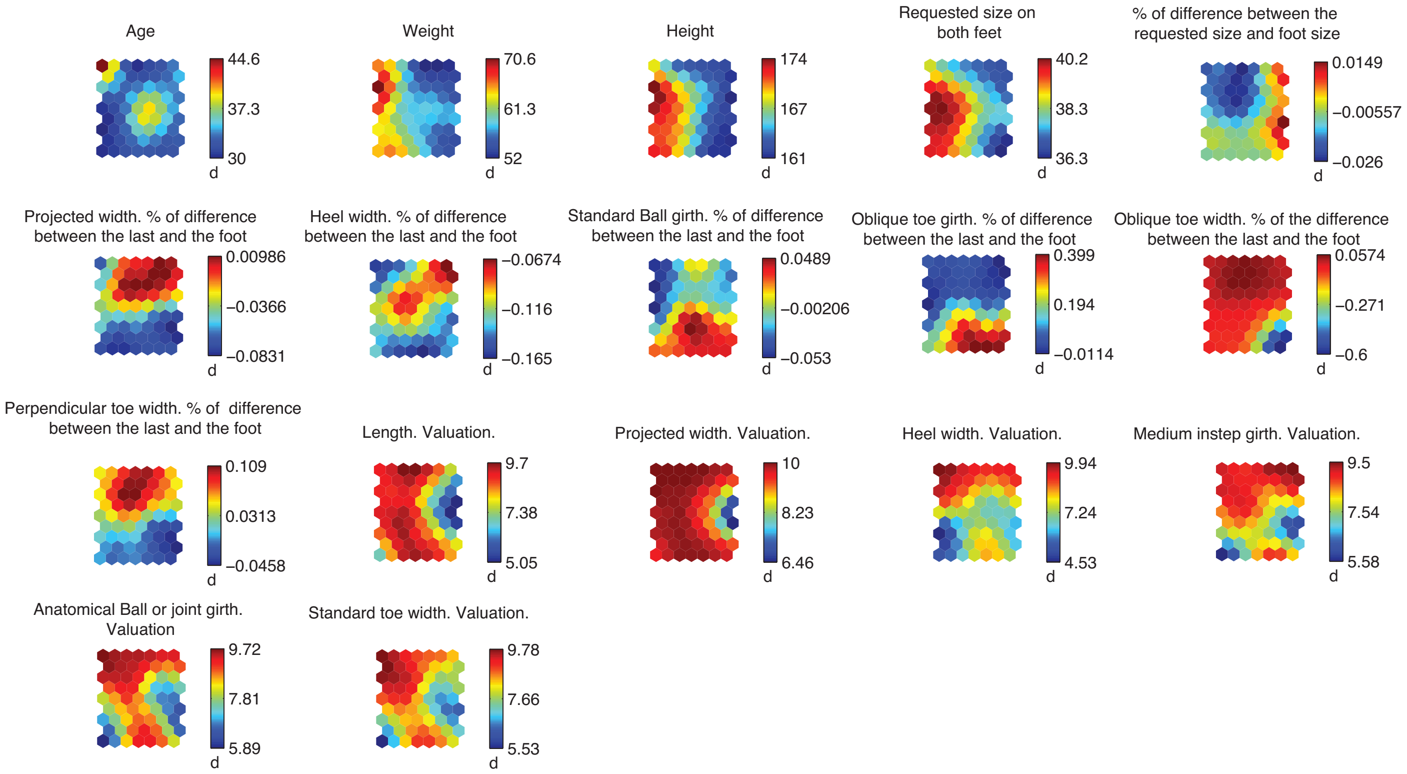


Figure 3.51: *Component planes obtained using after training the SOM considering only female testers.*

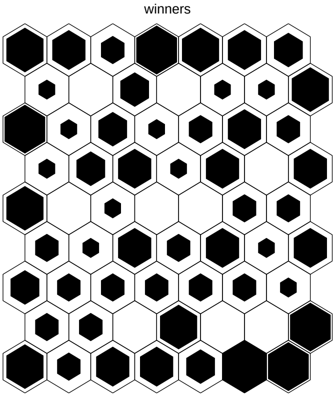


Figure 3.52: *Winners map obtained from the SOM training considering only female testers.*

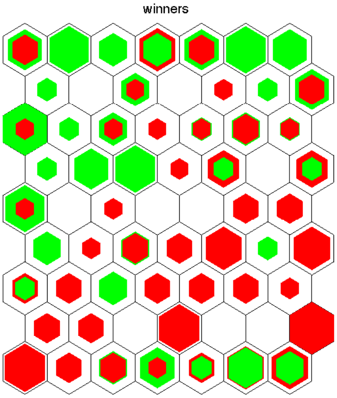


Figure 3.53: *Winners map obtained from the SOM training considering only female testers. Red color represents not buying behavior and green color represent buying behavior.*

bottom part has a majority of “non purchase” patterns, whereas the upper part reveals a majority of “purchase” behavior. The lower left corner represents testers that do not purchase the footwear. This is due to the fact that the lowest valuations (except in *Projected width*) are found in this area. There is another area of low valuation for all variables (right side of the map at middle height); here also a majority of testers who do not purchase the footwear are found.

3.5.4 Conclusions

The comfort in footwear is essential when walking because the foot is one of the structures of the human body that supports more weight when walking and further is the main shock absorber on the floor. Moreover, it is one of the most important factors, together with the aesthetic, when buying footwear. Due to this fact, this work is of great importance. Herein, it has been studied and analyzed data about comfort in footwear provided by *Instituto Tecnológico del Calzado y Conexas* (INESCOP) by means of the use of Visual Data Mining techniques, in particular the Self-Organizing Maps (SOMs). The study included different variables classified into three groups of measures related to physical characteristics of the testers, characteristics of foot-footwear (which represented the percentage of difference between the last and the foot for different zones) and tester valuations on the different areas of the foot. It has been studied which factors can be decisive when buying footwear, revealing interesting hidden relationships and patterns. Important conclusions were drawn from this study, but the most remarkable ones are detailed as follows. For example, as global conclusion it was proved that taking all measures to obtain the valuation of the testers may not be necessary, specially for men, whose valuation were almost equal in all the foot areas. Perhaps, this was due to the fact that men testers perform an overall assessment (positive or negative for all the valuations) regardless of whether, in some areas of the foot, the last does not fit perfectly. Another of the most important conclusions is that there is a different behavior between men and women in terms of valuations and when buying footwear. For example, women are the testers that presented the most negative difference between the last and the foot (in percentage) in heel width. This means that they usually ask tighter shoes than they really need. Moreover, a tighter shoe has a negative influence on a man while this fact has no negative influence in a woman valuation. Another difference between men and women behavior is that women are more sensitive to different parts of the foot than men (men have a similar opinion, or global idea, about the different space areas of the footwear).

Chapter 4

SonS and MDSonS: New Visualization Tools for Data Mining Techniques

Abstract

Clustering techniques and classification trees are two of the main techniques used in Data Mining but, at present, there is still a lack of visualization methods for these tools. Many graphs associated with clustering, also with hierarchical clustering, do not give any information about the values of the centroids' attributes and the relationships among them. In classification trees, graphical procedures can also be developed in order to help simplify their interpretation and to obtain a better understanding, but more visualization methods to support this tool are needed. This chapter presents a novel visualization technique called Sectors on Sectors (SonS)*, and an extended version called* Multidimensional Sectors on Sectors (MDSonS)*, for improving the interpretation of several Data Mining algorithms. These methods are applied for visualizing the results of: a) hierarchical clustering, which makes possible to extract all the existing relationships among centroids' attributes at any hierarchy level; b) Growing Hierarchical Self-Organizing Maps (GHSOM), a variant of the well-known Self-Organizing Maps (SOM), by means of which is possible to visualize, simultaneously,* *the data information at each hierarchy level compactly and extract relationships among variables; c) classification trees, in which the SonS is used for representing the input data information for each class presented in each terminal node of a classification tree providing extra information for a better understanding of the problem. These methods are tested by means of several data sets (real and synthetic). Achieved results show the suitability and usefulness of the proposed approaches.*

4.1 Theorical bases

This section discusses the main theoretical aspects of several algorithms used in this chapter, named hierarchical clustering algorithms, Growing Hierarchical Self-Organizing Maps (GHSOMs) and classification trees. Firstly, the notion of clustering and concepts related to it are introduced. Once explained the different types of existing clustering algorithms the section focuses on describing in detail the GHSOM algorithm, a variant of the SOM for hierarchical data sets. Afterwards, the main aspects related to classification trees are introduced. Herein it is explained the operation of these algorithms in order to understand deeply how they work, their possibilities and how they can be interpreted with the proposed visualization methods.

4.1.1 Clustering algorithms

Introduction

In clustering methods the focus of interest is turned to the unsupervised case, where class labeling of the training patterns is not available. Thus, the major concern becomes to "reveal" the organization of patterns into *"sensible"* clusters (groups), which will allow to discover similarities and differences among patterns and to derive useful conclusions about them (Theodoridis and Koutroumbas, 2008). Clustering may be also found under the name of unsupervised learning in pattern recognition.

A clustering procedure is normally dependent on several parameters in general terms. The first is the *proximity measure*. This is a measure that quantifies how "similar" or "dissimilar" two feature vectors are. It is natural to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others (Hastie et al., 2009). This must be taken care of during preprocessing. Another parameter is the *clustering criterion*. This depends on the interpretation the expert gives to the term “sensible”, based on the type of clusters that are expected to underlie the data set. For example, a compact cluster of feature vectors in the N-dimensional space, may be sensible according to one criterion, whereas an elongated cluster may be sensible according to another one. The clustering criterion may be expressed via a cost function or some other types of rules (Hastie et al., 2009). Finally, the last choice is based on the *clustering algorithms* to be used. Having adopted a proximity measure and a clustering criterion, this step refers to the choice of a specific algorithmic scheme that unravels the clustering structure of the data set.

As one may have already suspected, different choices of features, proximity measures, clustering criteria and clustering algorithms may lead to totally different clustering results. Hence, subjectivity is a reality in cluster analysis (Feldman and Sanger, 2007). To demonstrate this, let us consider the following example. Consider Figure 4.1. How many “sensible” ways of clustering can be obtained for these points? The most “logical” answer seems to be two. The first clustering contains four clusters (surrounded by dashed lines). The second clustering contains two clusters (surrounded by solid lines). Which clustering is “correct”? It seems that there is no definite answer. Both clusterings are valid. The best thing to do might be to consult with an expert and let the expert decide about the most sensible one. Thus, the final answer to these questions will be influenced by the knowledge of the expert. The rest of the section is devoted to presenting some basic concepts and definitions related to clustering.

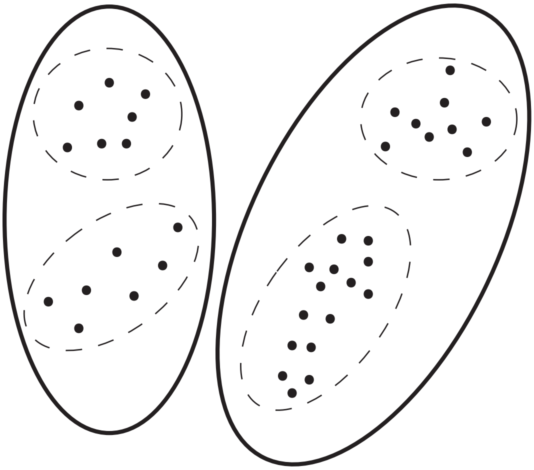


Figure 4.1: *A coarse clustering of the data results in two clusters (solid line), whereas a finer one results in four clusters (dashed line).*

Definitions of clustering

Cluster analysis is a collection of techniques for creating groups of objects (Berthold and J.Hand, 2002). The groups that are created are called clusters. The individuals within a cluster are similar in some sense.

In (Everitt et al., 2009), the vectors are viewed as points in the N-dimensional space and the clusters are described as "continuous regions of this space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points." Clusters described in this way are sometimes referred to as *natural clusters*. This definition is closer to the visual perception of clusters in the two- and three-dimensional spaces. As follows are given some definitions for "clustering", which, although they may not be universal, they give an idea of what clustering is. Let X be the data set, that is,

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \tag{4.1}$$

We define as an *m-clustering* of X, $\Re$, the partition of X into m sets (clusters), $C_1, \ldots, C_m$, so that the following three conditions are met:

- $C_i \neq \varnothing, i = 1 \ldots m$
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \varnothing, i \neq j; i, j = 1 \ldots m$

In addition, the vectors contained in a cluster C_i are "more similar" to each other and "less similar" to the feature vectors of the other clusters. Quantifying the terms "similar" and "dissimilar" depends very much on the types of clusters involved. For example, other measures (measuring similarity) are required for compact clusters, others for elongated clusters and different ones for shell-shaped clusters. Note that, under the preceding definitions of clustering, known as *classical* or *hard clustering*, each vector belongs to a single cluster. In *fuzzy clustering* (Miyamoto et al., 2008) (also referred to as *soft clustering*), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

Categories of clustering algorithms

Clustering algorithms may be viewed as schemes that provide us with sensible clusterings by considering only a small fraction of the set containing all possible partitions of X. The result depends on the specific algorithm and the criteria used. Thus, a clustering algorithm is a learning procedure that tries to identify the specific characteristics of the clusters underlying the data set. Clustering algorithms may be divided into the following major categories as pointed out in (Theodoridis and Koutroumbas, 2008).

- **Sequential algorithms.** These algorithms produce a single clustering. They are quite straightforward and fast methods. In most of them, all the feature vectors are presented to the algorithm once or a few times. The final result is, usually, dependent on the order in which the vectors are presented to the algorithm. These schemes tend to produce compact and hyperspherically, or hyperellipsoidally, shaped clusters, depending on the distance metric used.
- **Hierarchical clustering algorithms.** These schemes are further divided into:

  - *Aglomerative algorithms.* These algorithms produce a sequence of clusterings of decreasing number of clusters, m , at each step. The clustering produced at each step results from the previous one by merging two clusters into one. This category is detailed in the next subsection since it refers to the algorithm used to evaluate the performance of the two proposed visualization methods.
  - *Divisive algorithms.* These algorithms act in the opposite direction; that is, they produce a sequence of clusterings of increasing number of clusters at each step. The clustering produced at each step results from the previous one by splitting a single cluster into two.

- **Clustering algorithms based on cost function optimization.** This category contains algorithms in which "sensible" is quantified by a cost function, J. Usually, the number of clusters m is kept fixed. These algorithms use differential calculus concepts and produce successive clusterings while trying to optimize J. They terminate when a local optimum of J is determined or when a maximum number of iterations are reached. Algorithms of this category are also called iterative function optimization schemes.

Hierarchical clustering

Hierarchical clustering algorithms are of a different philosophy from the sequential algorithms. Specifically, instead of producing a single clustering in a single step they produce a hierarchy of clusterings (Han and Kamber, 2001).

In hierarchical clustering, m different partitions of the input data are generated into clusters, where m is the number of objects in the input data. One of these partitions corresponds to a single cluster made up of all m objects of the input data, while at the opposite extreme there is a partition corresponding to m clusters, each made up of just one object. Between these extremes there is a partition with 2 clusters, one with 3 clusters, and so on up to a partition with $m-1$ clusters. The key characteristic of these partitions, which makes them hierarchical, is that the partition with r clusters can be used to produce the partition with $r-1$ clusters by merging two clusters, and it can also be used to produce the partition with $r+1$ clusters by splitting a cluster into two.

As it is known both the top layer and the bottom layer of the hierarchy, there are two natural approaches to finding the intervening layers. We could start with m clusters, each containing one individual, and merge a pair of clusters to get $m-1$ clusters, and continue successively merging pairs of clusters; this approach is called agglomerative clustering (introduced in previous section). Alternatively, we could start with a single cluster, split it into two, then split one of the new clusters to give a total of three clusters, and so on; this approach is called divisive clustering, as mentioned in the previous section. Figure 4.2 shows an example of how the hierarchical clustering makes the merging (in case of agglomerative technique) or the partition (in case of divisive technique) for a data set of five patterns.

As pointed out in (Berthold and J.Hand, 2002), agglomerative clustering has been preferred traditionally, because the number of partitions considered in building the hierarchy is much smaller than for divisive clustering; the number of partitions considered is cubic in m for agglomerative, but exponential in m for divisive.

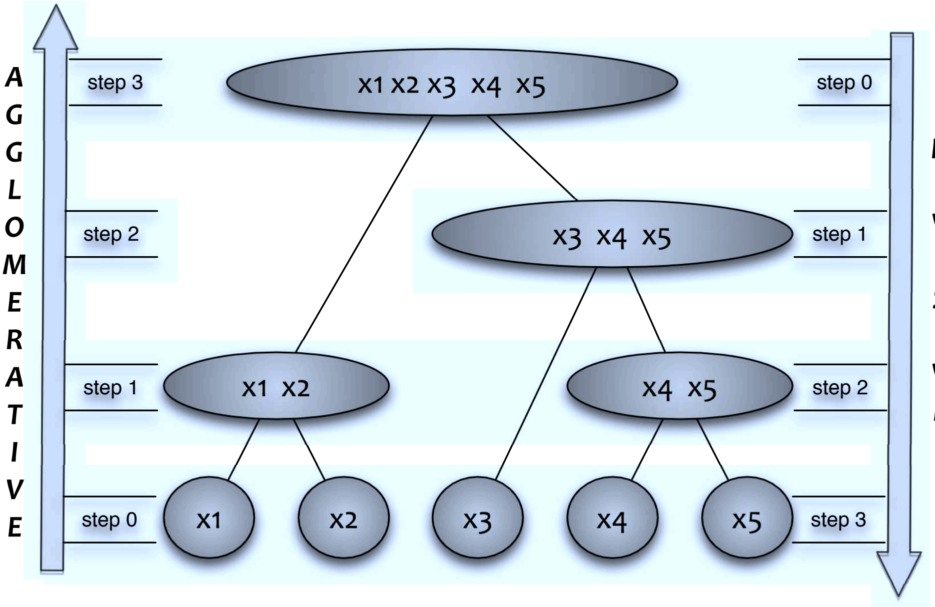


Figure 4.2: *Example diagram showing how the hierarchical clustering works for a data set of five patterns. For agglomerative technique see the diagram from bottom to top, and for divisive technique see it from top to bottom.*

The results of a hierarchical cluster analysis are almost always presented as a dendrogram. A dendrogram is an effective means of representing the sequence of clusterings produced by an agglomerative and divisive algorithm (Theodoridis and Koutroumbas, 2008). Cutting the dendrogram at a specific level results in a clustering.

A proximity dendrogram takes into account the level of proximity where two clusters are merged for the first time. This tool may be used as an indicator of the natural, or forced, formation of clusters at any level. That is, it may provide a clue about the best clustering for a given data set. An example of such a dendrogram, for agglomerative technique, is given in Figure 4.3.

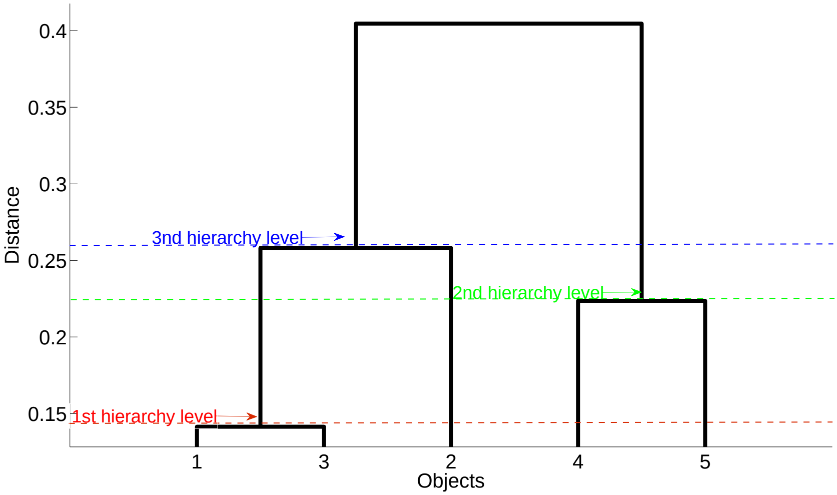


Figure 4.3: *Example dendrogram for hierarchical clustering algorithm that groups a set of five patterns.*

As it can be seen in Figure 4.3, three hierarchy levels are marked. In the initial step, the 5 objects represent 5 clusters. In the fist hierarchy level, 4 clusters are obtained, $C_1 = \{x_1, x_3\}$, $C_2 = \{x_2\}$, $C_3 = \{x_4\}$, $C_4 = \{x_5\}$; in the second hierarchy level 3 clusters, $C_1 = \{x_1, x_3\}$, $C_2 = \{x_2\}$, $C_3 = \{x_4, x_5\}$; and finally, in the third hierarchy level 2 clusters are found, $C_1 = \{x_1, x_3, x_2\}$, $C_2 = \{x_4, x_5\}$. In the last step, there is a single cluster containing all the objects.

Determining the best clustering within a given hierarchy is a very important task. This is equivalent to the identification of the number of clusters that best fits the data. An intuitive approach is to search in the proximity dendrogram for clusters that have a large *lifetime*. The lifetime of a cluster is defined as the absolute value of the difference between the proximity level at which it is created and the proximity level at which it is absorbed into a larger cluster (Theodoridis and Koutroumbas, 2008).

4.1.2 Growing Hierarchical SOM (GHSOM)

As mentioned in Chapter 2, the SOM is a popular visualization tool that provides qualitative information about how the input variables are related to each other given a data set used to train the map. Despite the popularity of SOM, at least two limitations have to be noted, which are related, on the one hand, to the static architecture of this model, as well as, on the other hand, to the limited capabilities for the representation of hierarchical relations of the data (Dittenbach et al., 2000).

Hierarchical models can provide more information from a data set than non-hierarchical models. SOM has been developed in several ways in order to set it within hierarchical frameworks. The key idea of *hierarchical feature maps* proposed in (Miikkulainen, 1990) is to use a hierarchical setup of multiple layers where each layer consists of a number of independent SOMs. Another variant of SOM which solve the above-mentioned limitations is the *Growing Hierarchical Self-Organizing Map* (GHSOM) (Dittenbach et al., 2000, 2002). The GHSOM is proposed as an extension to the SOM (Kohonen, 1982, 2001) and HSOM (Luttrell, 1989) with these two issues in mind:

- SOM has a fixed network architecture i.e. the number of units to use as well as the layout of the units has to be determined before training, as mentioned in Chapter 2.
- Input data that are hierarchical in nature should be represented in a hierarchical manner for clarity of representation.

GHSOM uses a hierarchical structure of multiple layers where each layer consists of a number of independent SOMs (Dittenbach et al., 2002). Only one SOM is used at the first layer of the hierarchy. For every unit in this map a SOM might be added to the next layer of the hierarchy. This principle is repeated with the third and any further layers of the GHSOM. In order to avoid SOM fixed size in terms of the number of units an incrementally growing version of SOM is used, similar to the *Growing Grid* (Fritzke, 1995).

The GHSOM will grow in two dimensions: in width (by increasing the size of each SOM) and in depth (by increasing the levels of the hierarchy) as shown in Figure 4.4.

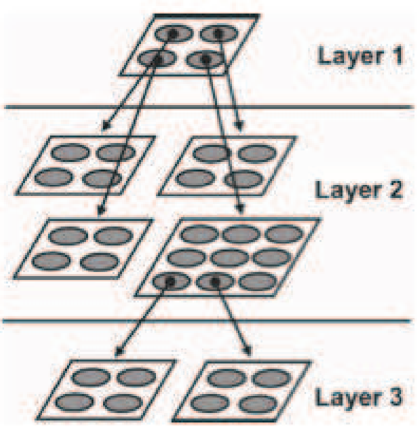
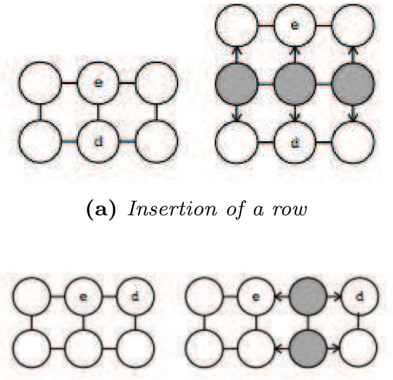


Figure 4.4: *GHSOM reflecting the hierarchical structure of the input data (Dittenbach et al., 2000).*

For growing in width, each SOM will attempt to modify its layout and increase its total number of units systematically so that each unit is not covering too large an input space. The training proceeds as follows:

1. The weights of each unit are initialized with random values.
2. The standard SOM training algorithm is applied.
3. The unit with the largest deviation between its weight vector and the input vectors that represents is chosen as the *error unit.*
4. A row or a column is inserted between the error unit and the most dissimilar neighbour unit in terms of input space.
5. Steps 2-4 are repeated until the mean quantization error (MQE) reaches a given threshold, a fraction of the average quantification error of unit i, in the proceeding layer of the hierarachy.



(a) *Insertion of a row*

(b) *Insertion of a column*

Figure 4.5: *Insertion of units: A row (a) or a column (b) of units (shaded gray) is inserted in between error unit e and the neighboring unit d with the largest distance between its model vector and the model vector of e in the Euclidean space.*

Figure 4.5 shows a graphical representation of the insertion process of the realization of a growing SOM, with the newly inserted units being depicted as shaded circles. The arrows point to the respective neighboring units used for model vector initialization.

As for deepening the hierarchy of the GHSOM, the general idea is to keep checking whether the lowest level SOMs have achieved enough coverage for the underlying input data. The details are as follows:

1. *Check the average quantification error of each unit to ensure it is above certain given threshold:* it indicates the desired granularity level of a data representation as a fraction of the initial quantization error at layer 0.

2. *Assign a SOM layer to each unit with an average quantification error greater than the given threshold, and train SOM with input vectors mapped to this unit.*

GHSOM provides a convenient way to self-organize inherently hierarchical data into layers and it gives users the ability to choose the granularity of the representation at the different levels of the hierarchy. Moreover, the GHSOM algorithm will automatically determine the architecture of the SOMs at different levels. This is an improvement over the Growing Grid as well as HSOM.

The drawbacks of this model include the strong dependency of the results on a number of parameters that are not automatically tuned. High thresholds usually result in a flat GHSOM with large individual SOMs, whereas low thresholds result in a deep hierarchy with small maps (Dittenbach et al., 2002).

4.1.3 Classification trees

Introduction

Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many applications. A decision tree classifier is a simple and widely used classification technique.

Tree models, also known as Classification And Regression Tree (CART), begin by producing a classification of observations into groups and then obtaining a score for each group (Breiman et al., 1984; Hastie et al., 2009). Tree models are usually divided into regression trees, when the response variable is continuous, and classification trees, when the response variable is quantitative discrete or qualitative (categorical)(Alpaydin, 2010). Tree models can be defined as a recursive procedure, through which a set of n statistical units are progressively divided into groups, according to a division rule that aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. At each step of the procedure, a division rule is specified by the choice of an explanatory variable to split and the choice of a splitting rule for the variable, which establishes how to partition the observations (Corporation, 1999).

The main result of a tree model is a final partition of the observations. To achieve this, it is necessary to specify stopping criteria for the division process. The output of the analysis is usually represented as a tree. This implies that the partition performed at a certain level is influenced by the previous choices. The two main aspects are the division criteria and the methods employed to reduce the dimension of the tree (pruning). The ideal final tree configuration is both parsimonious and accurate. The first property implies that the tree has a small number of leaves, so that the predictive rule can be easily interpreted. The second property implies a large number of leaves that are maximally pure. The final choice is bound to be a compromise between the two opposing strategies. The results of a tree model can be very sensitive to the choice of a stopping rule. First the tree is built to its greatest size. This might be the tree with the greatest number of leaves, or the tree in which every node contains only one observation or observations all with the same outcome value or level. Then the tree is "trimmed" or "pruned" according to a cost-complexity criterion. Moreover, CART can easily handle both numerical and categorical variables. Among other advantages of CART method is its robustness to outliers. An important practical property of CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent variables.

Once the decision tree has been constructed, classifying a test record is straightforward. Starting from the root node, it is applied the test condition to the record and follow the appropriate branch based on the outcome of the test. It then lead us either to another internal node, for which a new test condition is applied, or to a leaf node. When we reach the leaf node, the class label associated with the leaf node is then assigned to the record (Tan et al., 2006).

Classification trees theory

Classification trees are used when for each observation of learning sample we know the class in advance. Classes in learning sample may be provided by the user or calculated in accordance with some exogenous rule.

Let t_p be a parent node and t_l,t_r respectively left and right child nodes of parent node t_p. Consider the learning sample with variable matrix X with M number of variables x_j and N observations. Let class vector Y consist of N observations with total amount of K classes.

Classification tree is built in accordance with *splitting rule*, that is, the rule that performs the splitting of learning sample into smaller parts. We already know that each time data have to be divided into two parts with maximum homogeneity (Figure 4.6).

CART methodology consists of tree parts:

1. **Construction of *maximum tree*.** This part is most time consuming. Building the *maximum tree* implies splitting the learning sample up to last observations, i.e. when terminal nodes contain observations only of one class.
2. **Choice of the right tree size.** Maximum trees may turn out to be of very

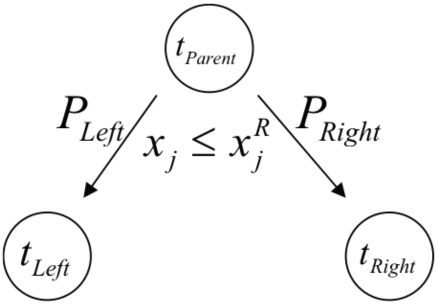


Figure 4.6: *Splitting algorithm of CART, where t_p, t_l, t_r are parent, left and right nodes respectively; x_j is variable j; and x_j^R is the best splitting value of variable x_j.*

high complexity and consist of hundreds of levels. Therefore, they have to be optimized before being used for classification of new data. Tree optimization implies choosing the right size of tree - cutting off insignificant nodes and even sub-trees. Two pruning algorithms can be used in practice: optimization by number of points in each node and cross-validation (Patil et al., 2010).

3. **Classification of new data using constructed tree.** As the classification tree is constructed, it can be used for classification of new data. The output of this stage is an assigned class or response value to each of the new observations. By set of questions in the tree, each of the new observations will get to one of the terminal nodes of the tree. A new observation is assigned with the *dominating class/response value* of terminal node, where this observation belongs to. *Dominating class* is the class, that has the largest amount of observations in the current node. For example, the node with 5 observations of class 1, two observations of class 2 and 0 observations of class 3, will have class 1 as a dominating class.

4.2 Visualization methods for Data Mining techniques

As mentioned in Chapter 1 there is a number of well-known techniques for visualizing data, such as x-y plots, line plots, and histograms. These techniques are useful for data exploration but are usually limited to relatively small and low dimensional data sets. In the last decades, a large number of novel information visualization techniques have been developed, allowing visualizations of multidimensional data sets. Nice overviews of the most visual methods can be found in (Card et al., 1999; Spence, 2001; Ware, 2012).

Also many visualization techniques have been developed to support specific Data Mining tasks, such as classification and clustering. On one hand, in classification, the most popular approaches are algorithms that construct decision trees. Since most algorithms do not provide information about the distribution of the input data, it is often difficult to understand and optimize the decision model. There exist some tools for these tasks, as the decision tree visualizer in *SGIs MineSet system*® that shows an overview of the decision tree together with important parameters such as the attribute value distributions. The system allows an interactive selection of the attributes shown and helps the user understand the decision tree. A more sophisticated approach which also helps in decision tree construction is visual classification as proposed in (Ankerst et al., 2000). The basic idea is to show each attribute value by a colored pixel and arrange them in bars. The pixels of each attribute bar are sorted separately and the attribute with the purest value distribution is selected as the split attribute of the decision tree. These methods help to optimize the model generation and the classification process, but they do not help to extract knowledge, or to obtain a better understanding, about a classification tree nor to gain information about the result of the carried out classification.

On the other hand, results from partitioning cluster analysis can be visualized by projecting the data into a two-dimensional space. Cluster membership is usually represented by different colors and glyphs, or by dividing clusters into several panels of a trellis display (Chen et al., 2008a). In addition, silhouette plots (Rousseeuw, 1987) provide a popular tool for diagnosing the quality of a partition. Sometimes, the high dimensional data sets involve some level of hierarchical structure making difficult the use of the same visualization tools (Chen et al., 2008a; Theodoridis and Koutroumbas, 2008). Regarding hierarchical clustering, it is difficult to find methods for visualizing their results. Hierarchical cluster analysis is almost always accompanied by a dendrogram. Cutting the dendrogram at a specific level results in a clustering, as explained in Section 4.1.1. Another visualization tool is the so-called treemap (Shneiderman, 1991). A treemap works by dividing the display area into a nested sequence of rectangles whose areas correspond to an attribute of the dataset. In some works, treemaps are used to visualize hierarchical clustering (Thomas and Tajudin, 2006; Makanju et al., 2008; McConnell et al., 2002; Baehrecke et al., 2004). Also, other popular tools as convex cluster hulls or silhouettes are specific to clustering (Chen et al., 2008a). Despite the fact that the dendrogram is an excellent tool to determine the number of clusters in a given hierarchical data set, treemaps are very useful when visualizing the hierarchy, and convex cluster hulls and silhouettes give information about how the centroids partition the input space, and how well each object lies within its cluster, respectively; nevertheless, these techniques do not provide any information about the values of the attributes in each cluster centroid and the relationships among them. This drawback is solved by the visualization methods proposed in this chapter, which are also able to visualize hierarchical structures.

In spite of the lack of methods for hierarchical clustering visualization, there are techniques for defining hierarchical information structures, that is, structured information, previously stored, in a hierarchical way, e.g., the file system on a computer, the organization of employees, Internet addressing, library cataloging, etc. These techniques are based on hierarchical visualization, but they do not use clustering algorithms since the hierarchy and the clustering is known a priori. Some of these techniques are the classic tree drawing algorithm for ordered binary trees (Reingold and Tilford, 1981), Cheops (Beaudoin et al., 1996), Hierarchical Edge Bundles (Holten, 2006), or Reconfigurable Disc Tree (RDT) (Jeong and Pang, 1998). There are also some software or Grafical User Interfaces as Hyperbolic Browser (Andrews and Kasanicka, 2007), Information Slices (Andrews and Heidegger, 1998), Magic Eye View (Kreuseler and Schumann, 1999), Cone Trees (Robertson et al., 1991), Information Pyramids (Andrews, 2002) and 3D Hyperbolic Browser (Munzner, 1997), among others. The point is that these techniques are used when the hierarchy is very deep. Thus, they aim to represent correctly the hierarchy given by the structured information already stored, not to extract information about the relationship among the attributes, which is the goal of the approaches presented is this chapter.

The rest of this chapter is organized as follows. The details of the proposed methods are described in Section 4.3. The data sets used to validate the proposed methods are described in Section 4.4. In Section 3.3.3, the proposed methods are applied in several Data Mining techniques, such as hierarchical clustering, Growing Hierarchical Self-Organizing Maps (GHSOM) and classification trees, for visualizing the achieved results in the mentioned data sets. Finally, Section 4.6 summarizes the conclusions of the proposed visualization techniques.

4.3 Proposed visualization methods

This section is devoted to a detailed description of both graphics produced by the *SonS* and *MDSonS* visualization methods in order to show the differences between both techniques. They are explained in detail to understand how they are interpreted when applied to a dataset. Moreover, it should by noted that *SonS* method has been developed as a interactive tool, where the user controls that visualization interacting with such interface in a way that reveals new information as the user explores the piece. For further information about this software tool, the reader is referred to Appendix A.

4.3.1 Sectors on Sectors (SonS)

Sectors on Sectors (SonS) is a visualization method that extracts visual information of data groups by representing the number of instances in each group, the value of the centroids of these groups of data and the existing relationships among the several groups and variables. This method is based on the well-known pie chart visualization. Each cluster, or group, is represented by a slice of a circle (pie sectors). The arc length of each pie sector is proportional to the number of patterns included in each cluster. By means of new divisions in each pie sector and a color bar with the same number of labels as attributes, the existing relationships among centroids' attributes of the different clusters can be inferred. Figure 4.8 represents the three steps followed to create the *SonS* visualization method; which are stated as follows[1]:

1. **Division of one circle on several sectors depending on the number of clusters:** First of all the circle is divided into several pie segments or sectors corresponding to each cluster. The arc length of each sector is proportional to the number of patterns included in each cluster. The number of patterns belonging to each cluster is shown within parentheses. In this way, the significance of each cluster is easily recognizable (Figure 4.8, left).

[1] This steps, or procedure, are extensible to other hierarchies in the case of hierarchical clustering or GHSOM.

2. **Division of the pie sectors depending on the number and the value of attributes:** After the first step, each sector is divided into as many subsectors as variables presented in the problem. The inner part corresponds to the first variable, and going outwards, the next variables are appearing. Each one of these parts vary its radius. This radius corresponds to the relative value of each variable, with respect to the sum of all of them[2]. That is, let X be a centroid corresponding to one cluster, so that,

$$X = \{x_1, x_2, \dots, x_N\} \tag{4.2}$$

Then, the radius of each subsector (corresponding to each centroid attribute) is calculated as follows:

$$r_i = \frac{|x_i|}{\sum_{k=1}^{N} |x_k|}, i = 1 \dots N \tag{4.3}$$

In this way the bigger the radius corresponding to each variable, the higher the weight of the variable and therefore, the more relevant the feature. This is a good method to identify the relevance of each variable within each cluster in a straightforward way (Figure 4.8, middle). Figure 4.7 depicts an example of how the method computes the radii of the different subsectors, corresponding to each attribute, in each cluster in order to represent the relevance of the features in each one of them.

In this example, one centroid is shown with the values $[25, -5, -12]$. After applying Eq. (4.3) to this vector, the relevance of each attribute is obtained. Notice that the relevance for the first attribute (inner subsector) is 0.59, 0.12 for the second one (subsector in the middle) and 0.29 for the third one (outer sector) and that the sum of all these "transformed" attributes is equal to 1. Notice that this example only presents 3 variables or attributes, so this is not a high-dimensional example. The intention of this example is to explain a simple

[2] Each variable is standardized to zero mean and unit variance before applying the clustering algorithm in order to avoid a biased model. Moreover, the standardization makes that the relevance of each variable (represented by the size of the radius) is independent of its range. The use of standardized variables, guarantees that the radius is the relevance of the variable within the cluster.

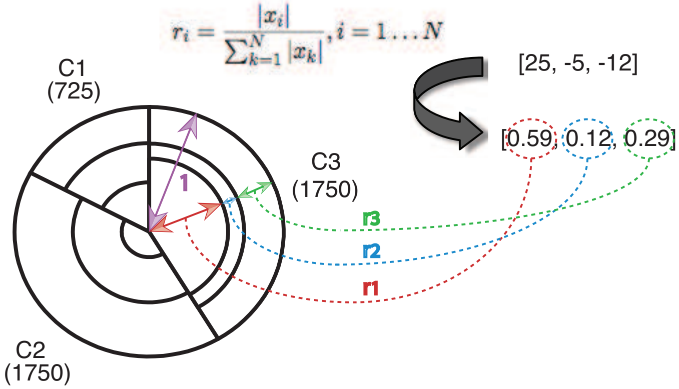


Figure 4.7: *Example of how the method computes the radii of the different subsectors in order to represent the relevance of the features in each cluster. After applying Eq. (4.3) to the vector* $[25, -5, -12]$*, the relevance of each attribute is obtained. The sum of all these "transformed" attributes is equal to 1.*

case for a better understanding. In the Section 4.5, high-dimensional data sets are applied to this method.

3. **Color coding for identifying the real value of features:** Attached to the graph, there is a color bar with the same number of column labels as variables (each column label for each variable). The mean value of the variables of each cluster (normally, the centroid) is codified by means of colors[3]. The value of the color for the first feature (inner subsector) is given by the first column label, the second feature by the second column label and so on. In this way, it is possible to know the exact value of each variable for each cluster centroid (Figure 4.8, right).

[3] This is automatically extensible to other measures such as the median, which is a much more adequate prototype measure in presence of outliers, for instance. Therefore, *SonS* is not restricted to the use of a particular prototype measure.

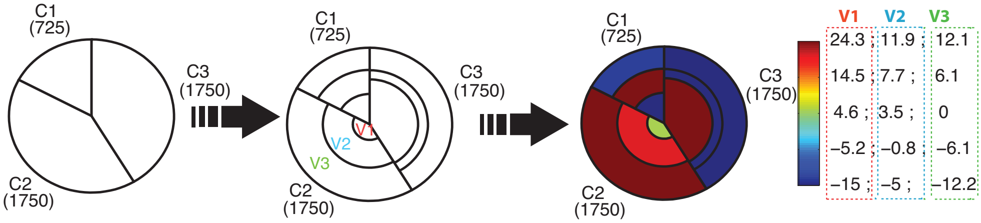


Figure 4.8: *The three steps followed to create the SonS visualization method. From left to right: 1) producing as many sectors as clusters, 2) splitting each sector according to the attributes and 3) color coding to identify real values.*

4.3.2 Multidimensional Sectors on Sectors (MDSonS)

The method proposed in this section is an improvement of the SonS visualization technique, called *Multidimensional Sectors on Sectors (MDSonS)*, which makes possible to visualize the distances between clusters. The visualization method is different from that proposed in the previous section due to the need of accommodating the information provided by Multidimensional Scaling (MDS) to the new visualization. Multidimensional Scaling (MDS) technique (Borg and Groenen, 2005) is used for representing the distances between clusters. Therefore, this technique is put forward below.

Multidimensional Scaling

Multidimensional Scaling (MDS) is a technique for the analysis of similarity, or dissimilarity, on a set of patterns. Such data may be inter-correlations of test items, ratings of similarity on political candidates, trade indices for a set of countries, etc. MDS attempts to model such data as distances among points in a geometric space. The main reason for doing this is that one wants a graphical display of the structure of the data, one that is much easier to understand than an array of numbers and, moreover, one that displays the essential information in the data, smoothing out noise. The graphical display of the correlations provided by MDS enables the data analyst to literally “look” at the data, and to explore their structure visually. This often shows regularities that remain hidden when studying arrays of numbers.

There are numerous varieties of MDS. Some facets for distinguishing among them are the particular type of geometry into which one wants to map the data, the mapping function, the algorithms used to find an optimal data representation, the treatment of statistical error in the models, or the possibility to represent not just one but several similarity matrices at the same time. Other facets relate to the different purposes for which MDS has been used, to various ways of looking at or "interpreting" an MDS representation.

The data to be analyzed is a collection of I objects on which a distance function $\delta_{i,j}$ is defined, where $\delta_{i,j}$ is the distance between i-th and j-th objects. These distances are the entries of the dissimilarity matrix:

$$\Delta = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix} \tag{4.4}$$

The goal of MDS is, given Δ, to find I vectors $x_1, \ldots, x_I \in \mathbb{R}^N$ such that:

$$\|x_i - x_j\| \approx \delta_{i,j} \text{ for all } i, j \in I \tag{4.5}$$

where $\|\cdot\|$ is a vector norm. In classical MDS, this norm is the Euclidean distance, but more generally it may be an arbitrary distance function.

In other words, MDS attempts to find an embedding from the I objects into $\mathbb{R}^N$ such that distances are preserved. If the dimension N is chosen to be 2 or 3, the vectors x_i can be plotted in order to obtain a visualization of the similarities between the I objects. Note that the vectors x_i are not unique: using the Euclidean distance, they may be arbitrarily translated and rotated since these transformations do not change the pairwise distances $\|x_i - x_j\|$.

There are various approaches to determining the vectors x_i. Usually, MDS is formulated as an optimization problem where $(x_1, \ldots, x_I)$ is found as a minimizer of some cost function, for example,

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2 \tag{4.6}$$

A solution may then be found by numerical optimization techniques.

MDSonS Methodology

MDSonS is also based on the well-known pie chart visualization. In this method, each group of data or cluster is represented by a circle. The area of each circle is proportional to the number of patterns included in each group and the distance among circles is proportional to the distance among clusters. By means of each slice of a circle (pie sectors) and a color bar with the same number of labels as attributes, the existing relationships among centroids' attributes at any hierarchy level can be extracted.

Once applied the Data Mining technique (hierarchical clustering, GHSOM or classification trees), and obtained the groups of data, the visualization graph is produced in three steps (see Figure 4.9) described as follows [4]:

1. **Representation of the different clusters and their size:** First of all, as many circles as clusters are drawn. The area of each circle is proportional to the number of patterns included in each cluster and the distance among circles is proportional to the distance among clusters' centroids. The distances among centroids are computed by MDS, which produces a representation of the similarity (or dissimilarity) between pairs of objects in a multidimensional space as distances between points of a low-dimensional space (Borg and Groenen, 2005), as mentioned in the previous section. The number of patterns belonging to each cluster is shown within parentheses. In this way it is easily recognizable the significance of each cluster and the distance among them (Figure 4.9, top left).

2. **Division of the circles depending on the number and the value of attributes:** Once the data is divided into clusters and after knowing the size

[4] This steps, or procedure, are extensible to other hierarchies in the case of hierarchical clustering or GHSOM, as in the *SonS* case.

of each one of them, the value of the attributes (or features) for each cluster centroid is analyzed. For this task, each circle corresponding with each cluster, is divided into several sectors, which correspond to each variable. The first variable is the one that starts with a vertical line in the top middle of the circle, and the rest of the variables is appearing sequentially counter clockwise. The arc length of each sector corresponds to the relative value of each variable, respect to the sum of all of them[5]. In this way, the bigger the arc length of a given variable, the more relevant the variable. With this method the relevance of each variable within each cluster, or within all of them, can be identified in a straightforward way. (Figure 4.9, top right)

3. **Color coding for identifying the real value of features:** Attached to the graph, there is a color bar with the same number of columns labels as variables (Figure 4.9, bottom). This step is exactly the same as the one presented in *SonS* method.

As mentioned previously, in the case of hierarchical clustering, and GHSOM algorithm, the description for the first hierarchy level can be extended to the rest of levels. For instance, in the data set analyzed in Figure 4.19, it can be observed that from each circle (corresponding to each cluster) in level 1, a new graph with new values of cluster centroids emerges, which corresponds with the second hierarchy.

The main advantage of the proposed visualization technique is that it is possible to observe relationships among different variables in the same cluster and relationships among the same variables in different clusters, in the different levels of the hierarchy; but specially, what is remarkable in comparison to *SonS* is the representation of the distances among clusters centroids.

[5] As in *SonS*, each variable is standardized to zero mean and unit variance before applying the Data Mining algorithm in order to avoid a biased model.

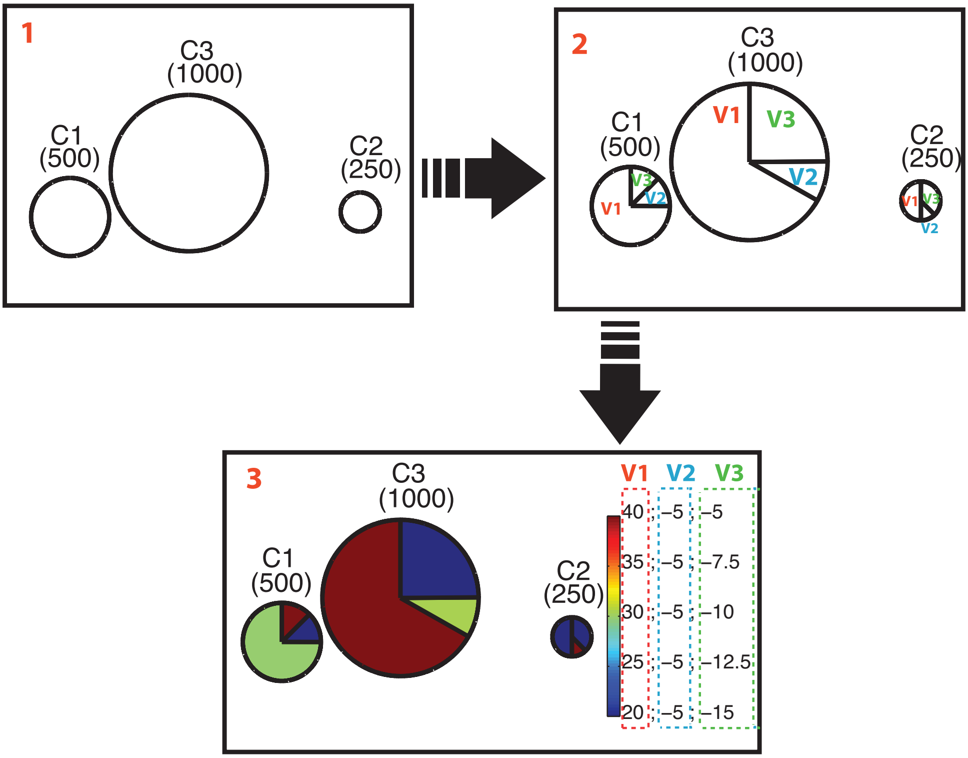


Figure 4.9: *The three steps followed to create the MDSonS visualization method.*

4.4 Data sets

This section explains the different data sets used to show the performance of the proposed visualization methods. All these data sets were not applied to each of the case studies presented in this chapter (*SonS* applied to hierarchical clustering, *MDSonS* applied to hierarchical clustering, *SonS* applied to GHSOM algorithm and *SonS* applied to classification trees). Nevertheless, One of the data sets (*German elections* data set, see section 4.4.2) was used to test the visualization results of hierarchical clustering, provided by *SonS* and *MDSonS* techniques, to compare the visualizations produced by both methods. However, for the other cases, in which it was attempted to visualize the results produced by other data mining techniques with the *SonS* method, other data sets were used to demonstrate the utility of such method in different scenarios.

4.4.1 Synthetic data set

The first data set is a synthetic data set created to show the performance of the proposed visualization method. The data consist of three clouds of points defined by X, Y, and Z coordinates, as shown in Figure 4.10. These three clouds of points can be divided into nine, three new clouds for each one of them; thus being a hierarchical structure.

Another variant of this data set was also taken into account (Figure 4.11). In this case, the cloud of points corresponding to cluster B was slightly displaced to the left with regard to the previous case. Notice that, while in the first case (Figure 4.10) the distances between cluster B and the rest were practically the same, in the second case (Figure 4.11) the cluster B was significantly closer to cluster A. In this way, the distances between clusters were different; the goal of this variant of the data set is to show the capabilities of *MDSonS* to represent distances among clusters.

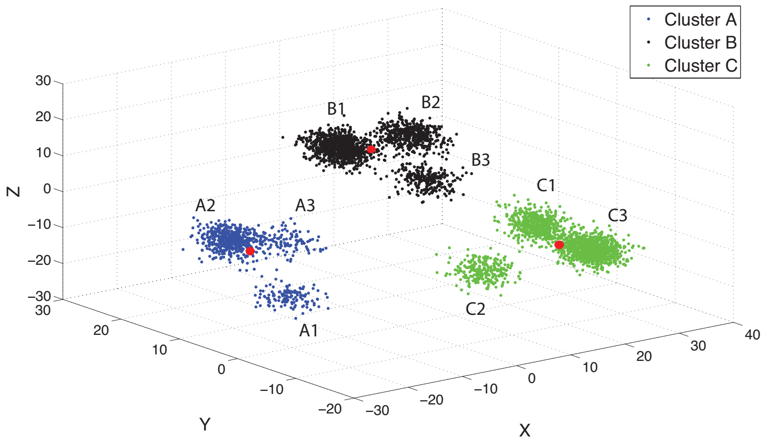


Figure 4.10: *Representation of the first synthetic data set variant. The points corresponding to each cluster after the first level (three clusters: A, B, C) are shown in different colors. Their centroids are represented with red dots. Subclusters at the second level of hierarchy are indicated with a number (1, 2, 3) after the corresponding letter (A, B, C).*

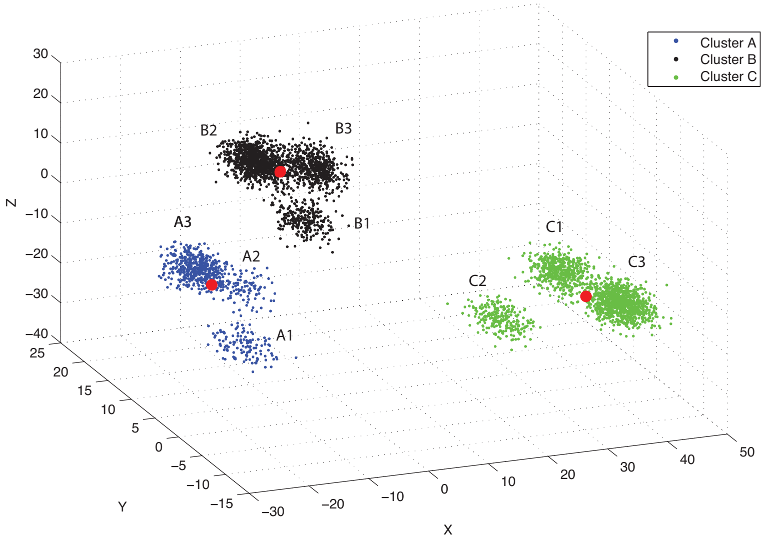


Figure 4.11: *Representation of the second synthetic data set variant. The points corresponding to each cluster after the first level (three clusters: A, B, C) are shown in different colors. Their centroids are represented with red dots. Subclusters at the second level of hierarchy are indicated with a number (1, 2, 3) after the corresponding letter (A, B, C).*

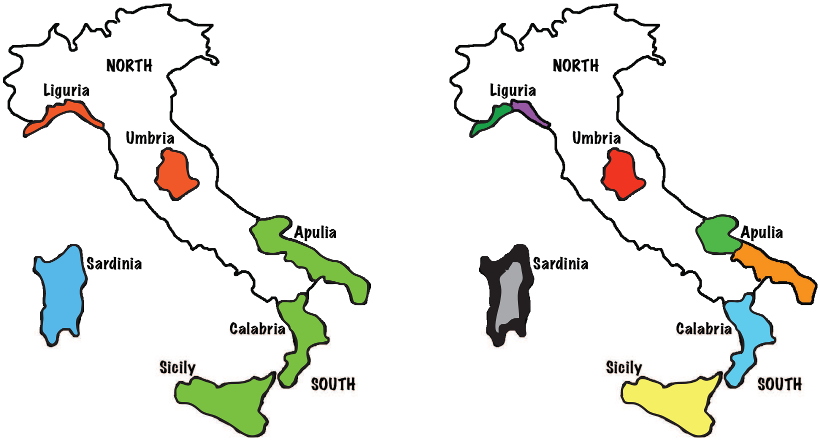
4.4.2 *German elections* data set

As a real example, a data set of the German parliamentary elections of September 18, 2005 was used. The data, extracted from package *flexclust* of "R" software[6], consist of the proportions of "second votes" obtained by the five parties that got elected to the first chamber of the German parliament for each of the 299 electoral districts. The "second votes" are actually more important than the "first votes" because they control the number of seats each party has in parliament. It should be emphasized that the proportions do not sum the unity because parties that did not get elected into parliament were omitted from the data set. Before election day, the German government comprised a coalition of Social Democrats (SPD) and the Green Party (GRUENE); their main opposition consisted of the conservative party (Christian Democrats, UNION) and the Liberal Party (FDP). The latter two intended to form a coalition after the election if they gained a joint majority, so the two major "sides" during the campaign were SPD+GRUENE versus UNION+FDP. In addition, a new "left-leaning party" (LINKE) canvassed for the first time; this new party contained the descendents of the Communist Party of the former East Germany and some left-wing separatists from the SPD in the former West Germany. This real example has been chosen to show the performance of the presented methods due to the qualitative conclusions that can be drawn from this data set.

4.4.3 *Italian olive oil* data set

This data set contains information about the percentage composition of fatty acids found in the lipid fraction of Italian olive oils (Forina and Tiscornia, 1983). The data set consists of 572 samples and 10 variables. The training variables are eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) in % × 100 (per 10 thousand). The other two variables contain information about the classes. There are two kinds of classes: super-classes that correspond with three regions of Italy: North, South, and the island of Sardinia (see Figure 4.12a); and sub-classes corresponding with nine collection areas: three from the Northern region (Umbria, East and West Liguria), four from the South (North and South Apulia, Calabria, and Sicily), and two from the island of Sardinia (inland and coastal Sardinia)(see Figure 4.12b). The data set arises from a study to determine the authentic-

[6] http://cran.R-project.org. (*Last checked November 2013*)



(a) *"Super-classes" or regions of Italy*

(b) *"Sub-classes" or collections areas of Italy*

Figure 4.12: *Regions and collections areas of Italy corresponding to the two kind of classes of the Italian olive oil data set.*

ity of an olive oil. The goal is to distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids. As in the case of *German elections* data set, this real example has been chosen to show the performance of the presented methods due to the qualitative conclusions that can be drawn from this data set.

4.4.4 *Iris flower* data set

The *"Iris flower data set"*[7] contains 3 classes of 50 instances each, where each class refers to a type of iris plant (*Setosa, Versicolor, Virginica*). One class is linearly separable from the other two; the latter are not linearly separable from each other. The input variables are *sepal length, sepal width, petal length and petal width.*

[7] http://archive.ics.uci.edu/ml/datasets/Iris. (*Last checked November 2013*)

4.5 Results

In this section the *SonS* and *MSonS* visualization methods are applied to different Data Mining techniques in order to evaluate their performance using the data sets described in Section 4.4.

4.5.1 *SonS* applied to hierarchical clustering

In this section *SonS* method is applied to visualize hierarchical clustering. The use of the method, described in Section 4.3.1, can be extended to every hierarchical level found in hierarchical clustering techniques. For the second hierarchy level, for instance, Figure 4.14 shows that for each sector (in level 1) emerge a new pie chart with new values of cluster centroids. If in a given data set more than two levels are present, a new pie will emerge from the sectors of the previous level as occurs when having two levels. It should be emphasized that not always a new pie emerges from all the sectors but it depends on the selected hierarchy. This method is highly recommended since it provides a compact visualization of each cluster making possible to observe the information of several hierarchy levels simultaneously; thus it makes possible to extract information in the different levels of hierarchy.

Example 1. Synthetic data set

This case of study, like the one presented in Section 4.5.2, does not represent a high-dimensional example. However, they are studied with the aim of understanding the proposed methods and how to interpret the information provided by such methods. Thus, these examples are useful to prove that the information provided by the proposed methods and the information underlying the data set (Figures 4.10 and 4.11) matches.

Figure 4.13 shows the dendrogram corresponding with the data set shown in Figure 4.10, which makes possible to extract the number of clusters visually. In particular, if the dendrogram is analyzed when the distance is 1.75 (higher dashed line), three clusters will be obtained (level 1) which are represented in green, red and blue colors. The second clustering level corresponds to a distance around 0.75 (lower dashed line), in which each former cluster is now divided into 3 new clusters represented in different

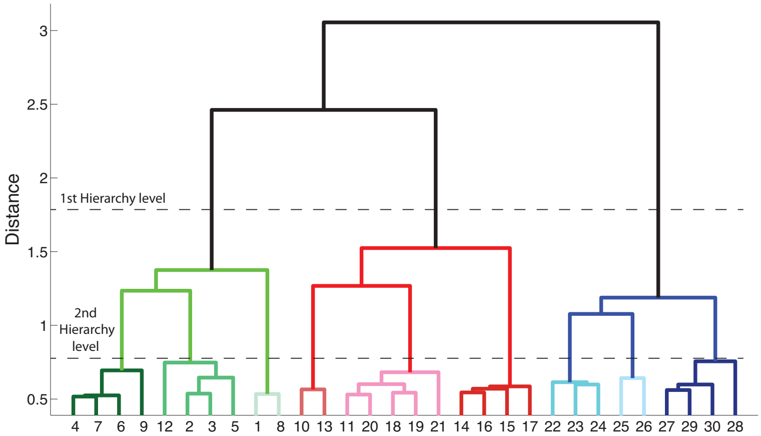


Figure 4.13: *Dendrogram corresponding to the clustering of the first variant of synthetic data set. Higher dashed line represents the distance of the first hierarchical level. Lower dashed line represents the distance of the second hierarchical level.*

shades of the same color that its parent.

Once the hierarchy is determined and the clusters formed in each level, the clustering is represented with the *SonS* visualization method (Figure 4.14).

The sector corresponding to cluster A, in the first level (center pie chart), has similar radii for the different variables. That means that the centroids' attributes have approximately the same relevance after clustering (in cluster A).

Also, in the first hierarchy level and regarding cluster C, the most relevant variable is the first one (X coordinate) since it has the largest radius (see inner subsector), and that matches the information shown in Figure 4.10; where, for this cluster, the first coordinate shows values around 24 whereas the others are about -5 and -12 respectively. Analyzing relationships among the same features in different clusters, in the first hierarchy level; it can be observed that the last feature (coordinate Z) shows a value of approximately 12 in cluster B and around -12 in cluster C (see in both cases the last column of labels). Due to this fact, as it can be observed in Figure 4.10, the cloud of points corresponding to cluster B is higher than the cloud of points corresponding to cluster C (variable Z corresponds to the height).

Summarizing, after applying the proposed method to a synthetic dataset, it can be seen that, although the dendrogram is an excellent tool to determine the number of

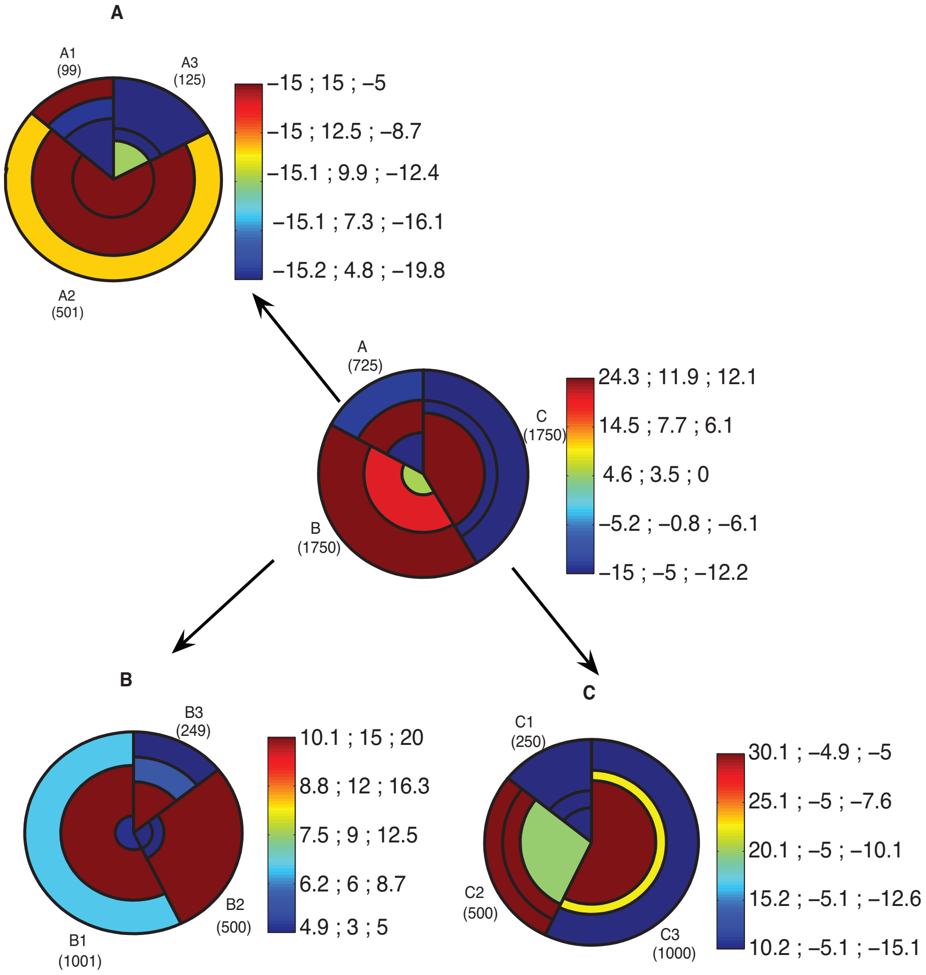


Figure 4.14: *SonS visualization method for the first variant of the synthetic data set.*

clusters in a given hierarchical data set, *SonS* provides an additional value by making possible to visualize relationships between centroids' attributes of all clusters at any hierarchical level.

Example 2. German elections

Figure 4.15 shows the dendrogram obtained for the *German elections* data set. Again, the number of clusters and the hierachy can be visually stablished by the interpretation of the expert. Analyzing the dendrogram when the distance is 0.2 (higher dashed line), four different clusters (level 1) are obtained, which are represented in red, black, blue and green colors. The second level can be obtained cutting when the distance is around 0.14 (lower dashed line). In this way the red and blue clusters are now divided into two new ones represented in different shades of the same color that its parents.

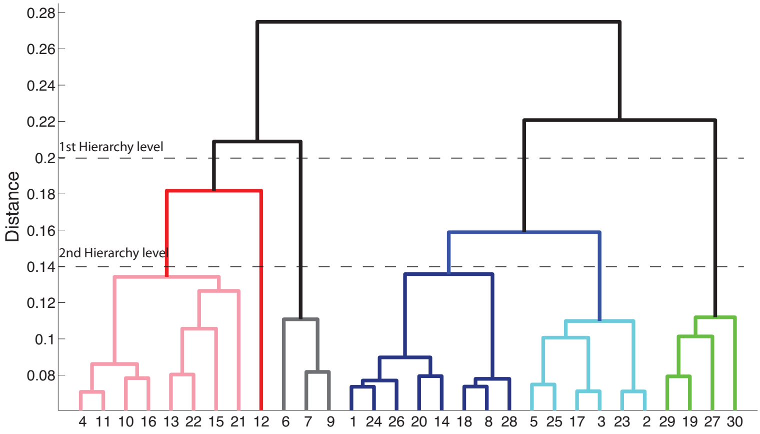


Figure 4.15: *Dendrogram corresponding with the clustering of the German elections data set. Higher dashed line represents the distance of the first hierarchical level. Lower dashed line represents the distance of the second hierarchical level.*

Each electoral district belongs to one of the 16 German federal states. After carrying out the clustering, the state corresponding with each pattern of the different clusters was analyzed. Therefore, the most predominant states can be found in order to check if each cluster corresponds with different German areas. The conclusion is that the 4 clusters (first level in the hierarchy) correspond with 4 different regions, namely, West Germany (without Saarland), East Germany (without Berlin and with-

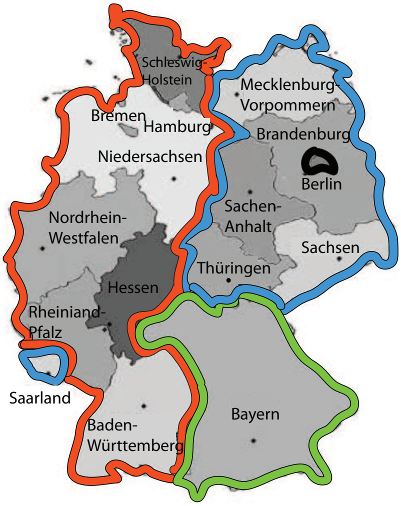


Figure 4.16: *German map with the 16 different German federal states. The 4 regions corresponding with the clustering in the first hierarchy level are shown in colors.*

out Bayern) together with Saarland, Bayern and finally, Berlin represented in Figure 4.16 in red, blue, green and black, respectively.

Saarland's behavior (located in the southwest of Germany, at the French border) may attract some attention because they voted in a similar way to the Eastern states. This is most likely due to the fact that Oskar Lafontaine, one of the two leaders of LINKE (which contained the descendents of the Communist Party of the former East Germany), is a former prime minister of Saarland as pointed out in (Kesselman et al., 2009). Another striking state is Berlin, which exhibits very diverse voting behavior and thus spreads over the rest of the clusters except some patterns, which form a different cluster because they are quite far away from others.

In Figure 4.17 the clustering solution for the *German elections* data set is represented with the *SonS* method. Unlike the representation provided by the dendrogram, this visualization is appropriate for this kind of data because a large radius in one subsector, corresponding to one variable (parties), means a large number of votes. In this way, it is easily recognizable which party has the strongest performance but also the exact value of the percentage support for each party looking at the color bar and its labels.

Focusing on the first level of hierarchy (top *SonS* graph), there are four different

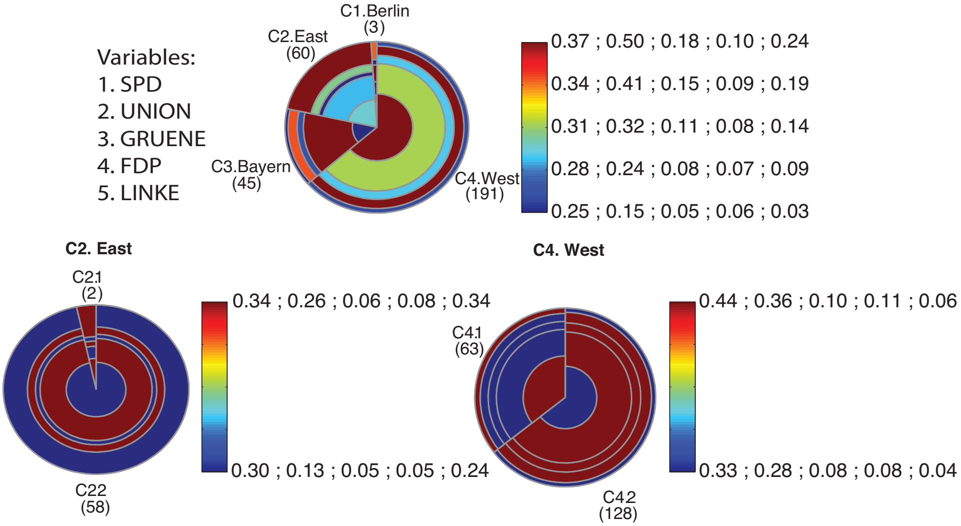


Figure 4.17: *SonS visualization method for German elections data set.*

clusters corresponding with the geographic areas marked with several colors in Figure 4.16. From now on, the red area will be called *"West"*, the blue one *"East"*, the green one *"Bayern"* and the black one *"Berlin"* as it is pointed out in Figure 4.17. The first cluster represented is the corresponding with *"Berlin"*. This cluster has only three patterns, the rest of patterns of Berlin (nine) spread over the rest of the clusters as mentioned previously. The second cluster corresponds to *"East"*, in which the parties with strongest performance are SPD (0.3), UNION (0.25) and LINKE (0.24). Notice that this is the only cluster where LINKE has an important relevance. This makes sense since LINKE party contained the descendents of the Communist Party of the former East Germany and some left-wing separatists from the SPD in the former West Germany. In the third cluster, corresponding to *"Bayern"*, the winner party is UNION. According to the fourth cluster *"West"* the two main parties are SPD and UNION having a little bit more support SPD than UNION (0.37 and 0.32, respectively) which are in opposite wings.

In the next hierarchy level, the cluster 2 (*"East"*) and cluster 4 (*"West"*) are divided into two new clusters each one of them (C2.1, C2.2, C4.1 and C4.2 respectively). Notice that the first new cluster extracted from the cluster corresponding to *"East"* (C2.1) has very similar values in its variables as the cluster *"Berlin"*. That is because the two patterns which are members of this new cluster actually belong to the cluster *“Berlin”*. The point is that in the first level, the clustering algorithm was not able to distinguish it because these two patterns were located between the clusters *“East”* and *“Berlin”*, but in the second level the difference is more evident; thus showing up the importance of the hierarchical approach. The second new cluster (C2.2) corresponds totally with *“East”* (blue area in Figure 4.16) where the parties with strongest performance are SPD, UNION and LINKE as mentioned before.

According to the division of cluster 4 (*“West”*), the two new formed clusters are similar to each other. In both clusters, the first and second features are the most relevant ones, whereas the last three have a lower significance. Although each variable presents a maximum value for one cluster and a minimum one for the other, actually the difference between the maximum and minimum values is not very significant (as shown in the color bar labels). Therefore, in deeper hierarchy levels, the clustering division is done depending on which party has the strongest performance. That is, the first cluster (C4.1) corresponds with the case when SPD is the most supported party (0.44), and the second cluster (C4.2) corresponds with the case when the party with the biggest support is UNION (0.36). The first new cluster (C4.1) corresponds with the south part of West Germany together with the Northern region Schleswig-Holstein, and the second new cluster (C4.2) corresponds with the north part of West Germany, except Schleswig-Holstein. The region Nordrhein-Westfalen has approximately the same number of patterns in each cluster.

The main conclusions and ideas extracted from this data set have been contrasted with (Leisch, 2009; Chen et al., 2008a). Moreover, new information and new ideas have been extracted by the proposed visualization method, which could not be obtained with other classical visualization tools.

4.5.2 *MDSonS* applied to in hierarchical clustering

In this section the *MDSonS* method is used to visualize hierarchical clustering, as in the previous section, in order to highlight the differences between *SonS* and *MDSonS*. The use of the method can be extended to every hierarchical level found in hierarchical clustering techniques, as occurred in the *SonS* case.

Example 1. Synthetic data set

Figure 4.18 shows the dendrogram corresponding with the second variant of the synthetic data set, which makes possible to extract the number of clusters visually. In particular, if the dendrogram is analyzed when the distance is about 1.5 (higher dashed line), the three clusters (level 1) will be obtained, which are represented in blue, green and red colors. The second clustering level corresponds to a distance around 0.75 (lower dashed line), in which each former cluster is now divided into three new clusters represented in different shades of the same color that its parent.

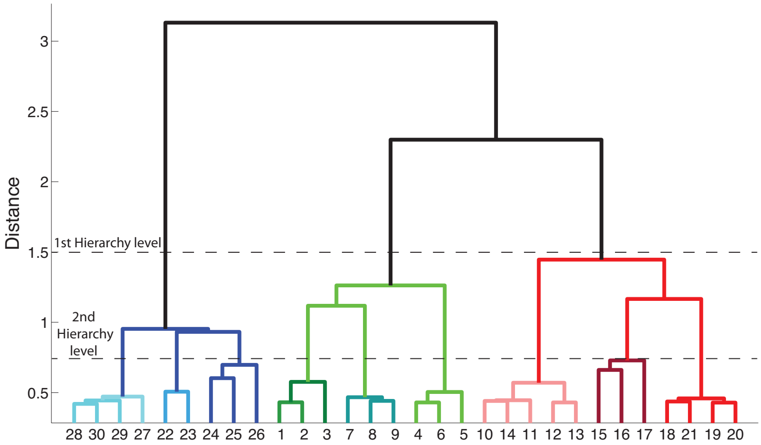


Figure 4.18: *Dendrogram corresponding to the clustering of the second variant of synthetic data set. Higher dashed line represents the distance of the first hierarchical level. Lower dashed line represents the distance of the second hierarchical level.*

Once the hierarchy is determined and the clusters formed in each level, the clustering is represented by means of *MDSonS* (Figure 4.19). Figure 4.19 shows the clustering for the two different hierarchies. The first one is shown in the centre of the figure without frame. From each one of the clusters in the first hierarchy level, three new circles emerge. Those circles enclosed in frames correspond to the second hierarchy level. Focusing on the first hierarchy, three different clusters appear in Figure 4.19, one small (cluster A) and two bigger clusters (clusters B and C). It can also be observed that there are two clusters (A and B) that are closer to each other than to the other one (cluster C); this conclusion matches the representation shown in Figure 4.11.

Focusing on each cluster, the different variables of cluster A (delimited by the

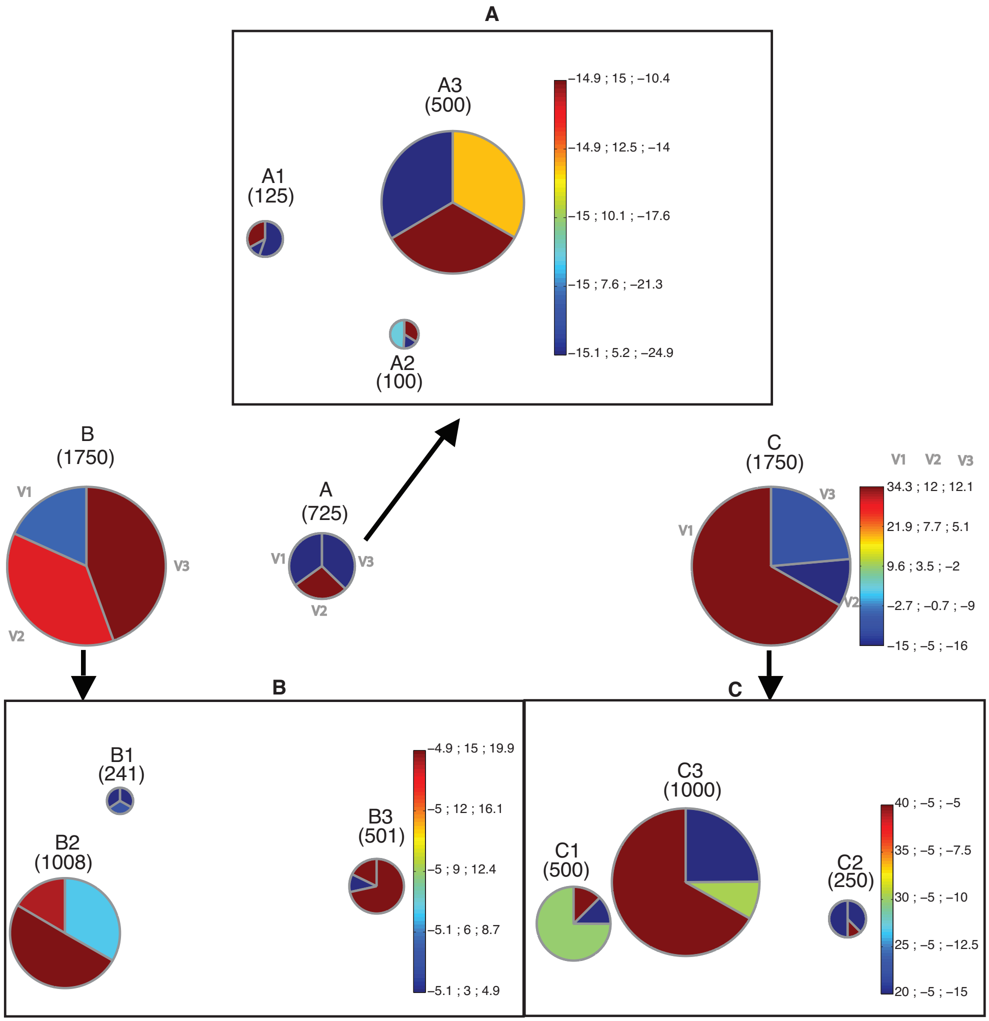


Figure 4.19: *MDSonS representation for the second variant of the synthetic data set.*

sectors) have a similar area. That means that the centroids' attributes have the same relevance after clustering. Exactly they take the values $[-15, 12, -16]$ as it can be observed with the color bar and checked in Figure 4.11. The cluster B has a similar area for the second and third variables (Y and Z coordinates) whereas the area of the sector corresponding to the first variable is smaller (less relevant). In particular, they take the values $[-5, 10, 12]$. However, in cluster C the area of the first sector (first feature) is the biggest one by far. As it can be observed with the color bar the exact values of the different features are $[34, -5, -12]$. All these conclusions agree the representation of Figure 4.11 regarding the relevance of the different variables to define each cluster.

Analyzing relationships among the same features in different clusters; in the first hierarchy level, it can be observed that the first feature (X coordinate) presents low values in cluster B and A (-5 and -15 respectively) whereas cluster C presents a high value (34), very different from the other two values. This is the reason why clusters A and B are far away from cluster C. However, in order to see which feature is the most relevant for distinguishing between cluster A and B, it can be seen that the last feature (Z coordinate) shows a value of approximately 12 in cluster B and about -16 in cluster A (see in both cases the last column of labels) whereas the other features are quite similar. Thus, as it can be observed in Figure 4.11, the cloud of points corresponding to cluster B is higher than the cloud of points corresponding to cluster A (variable Z corresponds to the height). Notice that, in fact, the third feature can distinguish not only cluster A and B, but can distinguish between cluster B and the other two (high values for cluster B and low values for the rest). In order to distinguish among subclusters, a similar procedure can be carried out for the next hierarchy level.

Example 2. German elections

Figure 4.20 shows the clustering achieved by the proposed method. As in *SonS* case, this visualization is appropriate for this kind of data because a long arc length, corresponding to each variable, means a large number of votes.

From Figure 4.20 similar conclusions to those seen in Section 4.5.1 can be extracted. Summarizing, regarding the cluster *"West"* the two main parties are SPD and UNION having a little bit more support SPD than UNION. In cluster *"Bayern"* the parties with the biggest support are SPD (0.25) and UNION (0.5) having

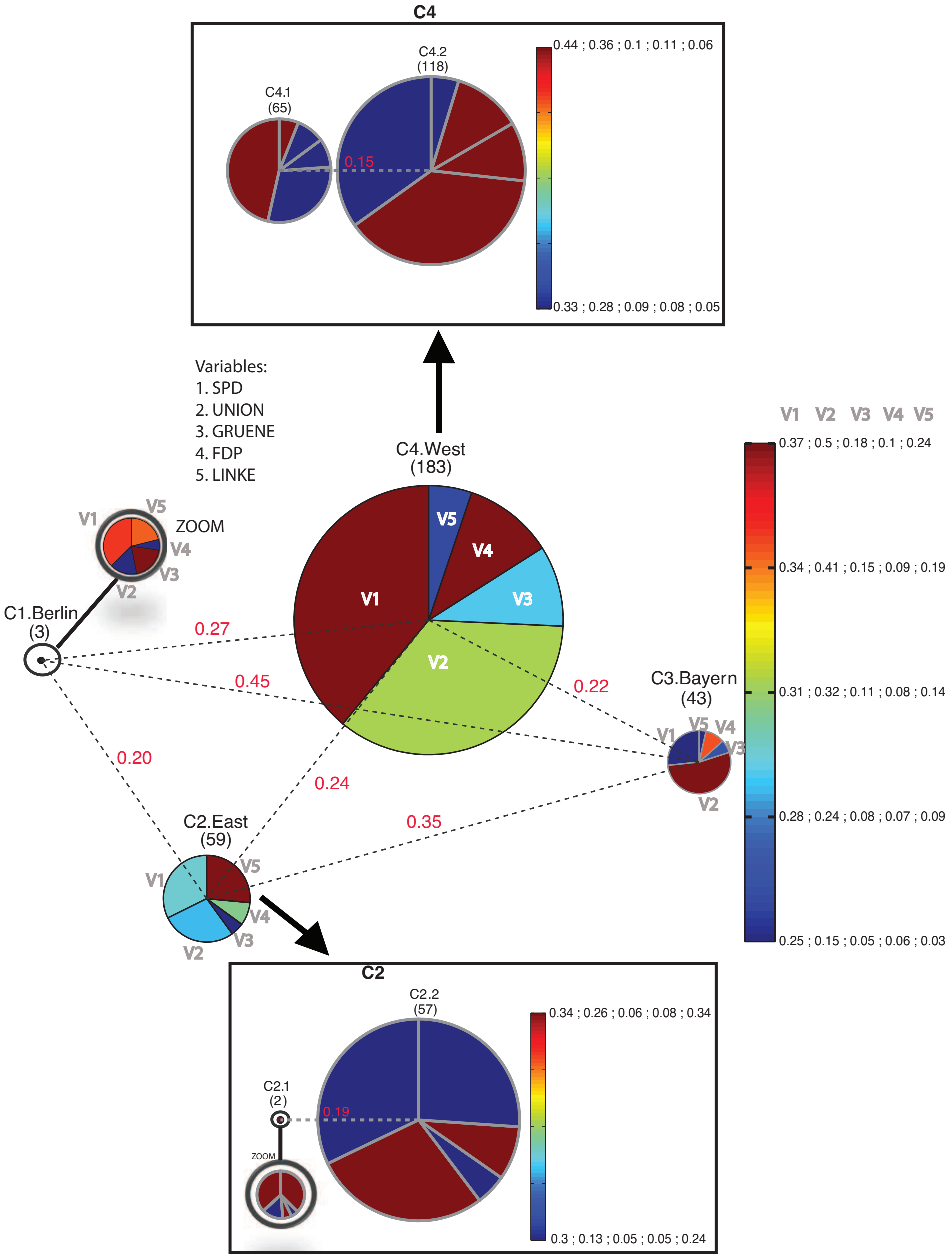


Figure 4.20: *MDSonS visualization method for German elections data set.*

more support the latter. The cluster *"Berlin"* has only three patterns, the rest of Berlin patterns (nine) spread over the rest of the clusters. Thus, these three patterns corresponding with cluster *"Berlin"* cannot be considered as a global behavioral pattern of Berlin. Actually, these patterns are related to the communist part of Berlin (East). Anyway, it should be pointed out that for these 3 patterns, the party with the strongest performance is SPD (0.35); and also LINKE has a significant performance (0.19) compared with the clusters commented previously (see Figure 4.20). Finally, in cluster *"East"*, the parties with strongest performance are SPD (0.3), UNION (0.25) and LINKE (0.24). As commented previously, LINKE has a significant performance compared with other clusters.

In addition to conclusions of the features within each cluster, information about the relationship of features in the different clusters can also be extracted. For example, SPD presents the biggest support in cluster *"West"* (0.37), UNION in cluster *"Bayern"* (0.5), GRUENE in cluster *"Berlin"* (0.18), FDP in cluster *"West"* (0.1) and LINKE in cluster *"East"*(0.24).

Moreover, new conclusions and ideas can be extracted when using *MDSonS* method, which are related to the information provided by the representation of the distances in this method. This information provides great utility to the method. It also helps to contrast hypotheses. For example, in *SonS*, it is known by intuition that the clusters *"Bayern"* and *"West"* are similar since these two clusters are the only ones that have the largest support in the first two parties, receiving the other three parties much lower support, as mentioned previously. However, representing the distances among clusters it can be proved that the most similar cluster to *"Bayern"*, is *"West"*. This assumption can be proved checking all the distances (red numbers) between the cluster *"Bayern"* and the other clusters since the minimum distance appears between the two mentioned clusters.

Moreover, it can also be proved the hypothesis, extracted from *SonS*, that clusters *"Berlin"* and *"East"* are very similar because these two clusters are the only ones where LINKE has an important relevance. This information, which is not available in the *SonS* method, provides an essential aid to the problem understanding. Actually, checking all the distances among all clusters it can be proved that these are the two most similar (closest) clusters among all of them . In addition, the result of the method is more intuitive to analyze, and since it presents information in a less compact design, it allows a neat representation of a larger number of variables.

In the next hierarchy level (pies enclosed in frames), also similar conclusions to those seen in Section 4.5.1 can be extracted. No more relevant information can be extracted in this hierarchical level apart from the distances between the clusters.

4.5.3 *SonS* applied to GHSOM algorithm

In this section *SonS* is also used for representing the results provided for GHSOM algorithm. The main advantage of GHSOM over hierarchical clustering is that in the former, the hierarchical structure is found "automatically" (actually tuning some parameters); and in the latter, the user should visually establish the data structure using a dendrogram. The main problem of GHSOM is that it is not possible to visualize simultaneously the data information in each level; that is, GHSOM visualization only makes possible to depict the component planes corresponding to the deepest SOMs so that the information about the component planes corresponding to the first SOMs in first or intermediate hierarchies is visually inaccessible. In this Section *SonS* visualization technique is used for visualizing the results obtained from the GHSOM algorithm to circumvent that drawback, since it allows a simultaneous and compact visualization of the different hierarchy levels. *SonS* also enables the extraction of knowledge in terms of relationships among variables. This is not possible using other classical visualizations. From now on, and for this particular application, the several sectors will correspond to neurons instead of clusters.

Italian olive oils

After training the data set explained in section 4.4.3 using the GHSOM algorithm, two hierarchy levels were produced. The first level started the training with four neurons (four sectors in Figure 4.21). After this, the predominant region for each neuron, in the first hierarchy level, was checked. For the first hierarchy level (top pie chart, Figure 4.21) there is one sector corresponding to the island of Sardinia, another to the South (specifically South Apulia), another to the North and, finally, another corresponding to the South again (specifically North Apulia, Calabria and Sicily). As it can be observed, the radius of the last variable, for some sectors, is very small. This fact makes difficult the visualization of the value in the mentioned variable. Because of this, a zoom of the image has been carried out. The labels of the color bars have been removed for a better visualization; only the color bars are shown in order to indicate a qualitative value (reddish colors correspond to high values and bluish colors to low values). Notice that although a given color may be very similar for two different variables, it does not mean a very similar value for the variables since each variable has its own range of values.

For distinguishing the oils from the different regions, in the first hierarchy level, the most important variable is the 8th (outer subsector in the circle, marked in green) because it takes a maximum value for one region and a minimum one for the others. If the mentioned variable is high, it involves that the oil belongs to the South and if it is low belongs to either the North or the Island of Sardinia. For distinguishing between North and Sardinia, the 5th and the 7th variables, marked in pink and yellow in Figure 4.21, play a relevant role (high values for Sardinia and low for the North). Summarizing, for the first hierarchical level, a number of rules can drawn from visual inspection of the generated graph:

- **NORTH** if V8, V5 and V7 $\downarrow\downarrow$
- **I. SARDINIA** if V8 $\downarrow\downarrow$; V5 and V7 $\uparrow\uparrow$
- **SOUTH** if V8 $\uparrow\uparrow$

In the next hierarchy level three new *SonS* graphs were found which emerged from the previous sectors corresponding to Island of Sardinia, North and finally the South, specifically the sector which represented North Apulia, Calabria and Sicily (Figure 4.21). In order to distinguish among the sub-classes in this hierarchy, a similar procedure can be carried out, which consists of checking which variables take maximum values for one region and minimum for others. Thus, an additional advantage of the *SonS* is that it enables to make a feature selection visually, since it is possible to separate the different classes using fewer features than the ones presented in the problem.

The pie which emerged from the sector “I. Sard” has four sectors, one corresponding to Coast Sardinia, other two to the Island of Sardinia and, finally, another corresponding to East Liguria. Although it might be expected to find only neurons (sectors) corresponding to Island of Sardinia there is also one belonging to the North (E.Liguria); this is because the sectors were labeled with the name of the region which had the biggest number of patterns in this neuron. Moreover the regions Island of Sardinia and the North are similar (they were only distinguished by means of two

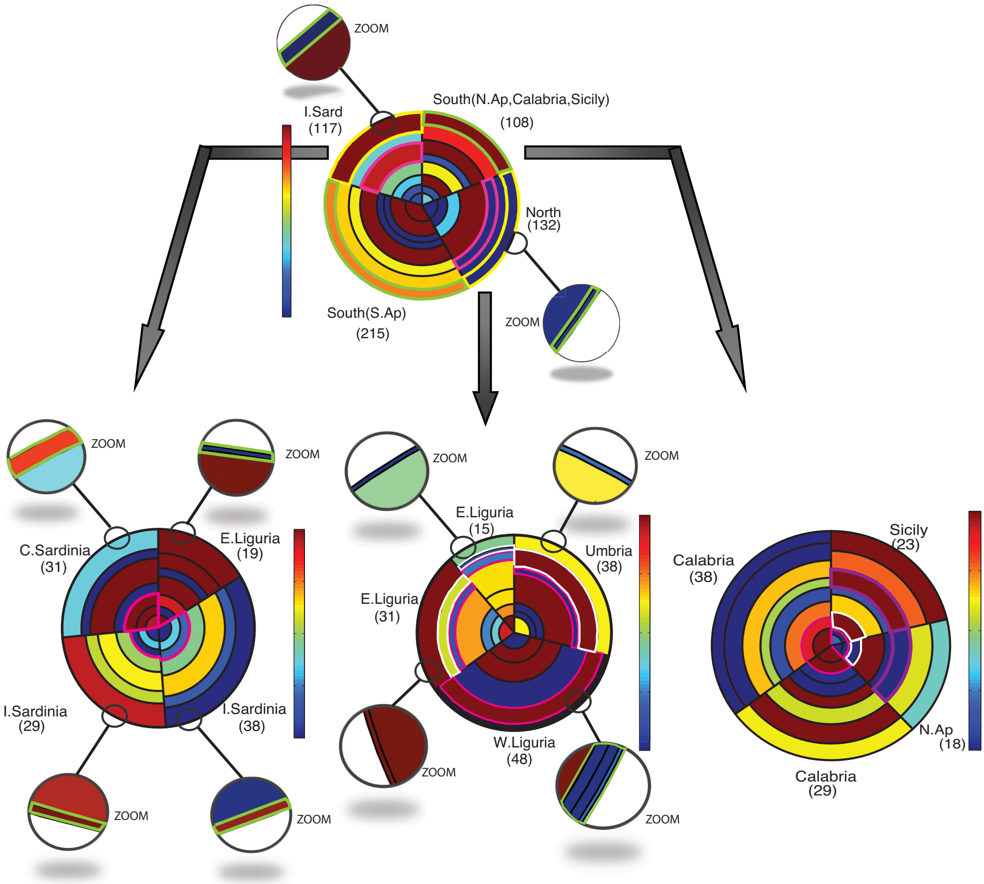


Figure 4.21: *SonS visualization method after training Italian olive oil data set with GHSOM algorithm.*

variables, the 5th and the 7th, in the previous hierarchy). However, in the left pie (second level of hierarchy) East Liguria is easily distinguishable from the other sectors by means of the 8th variable (outer subsector, marked with green), which takes low values (blue color) for East Liguria whereas for the rest of sectors (neurons) it presents high values (red color). In order to distinguish between the two "sub-classes" corresponding to the region Island of Sardinia (inland and coastal Sardinia) the variables that must be taken into account are the 1st, 2nd and 3rd (three inner subsectors marked with pink color in Figure 4.21); coastal Sardinia takes high values for these variables whereas inland Sardinia takes low values.

- **INLAND SARDINIA** if V8 ↑↑; V1, V2 and V3 ↓↓
- **COASTAL SARDINIA** if V8 ↑↑; V1, V2 and V3 ↑↑

Regarding the "sub-classes" of the North (central pie of second hierarchy level, Fig 4.21), there are again four sectors, two corresponding with East Liguria, one corresponding with West Liguria and other one with Umbria. One of the two neurons corresponding with East Liguria is basically formed by oils from this area but the other also contains a considerable number of patterns belonging to Umbria. Low values of the variables 6th, 7th and 8th (marked with green color in sector corresponding with West Liguria) makes possible to distinguish West Liguria from the rest of areas. Also the 5th variable (marked with pink color) must be taken into account because it presents the maximum value for West Liguria and low values for the rest of Northern areas. Now for distinguishing between the rest of oils from these areas (East Liguria and Umbria), the 6th variable, marked with white color, must be used (low or medium values for East Liguria and high values for Umbria). Summarizing, for the oils from North, we have that the oil will belong to:

- **W. LIGURIA** if V6, V7 and V8 ↓↓ or V5 ↑↑
- **E. LIGURIA** if V5 ↓↓ and V6 ↓↓
- **UMBRIA** if V5 ↓↓ and V6 ↑↑

The pie located in the right side in the second level of hierarchy (Figure 4.21) describes the areas from the South. The 1st and the 2nd features (marked with pink color) distinguish Calabria from the other Southern areas; the oils from Calabria present high values in these variables, whereas for the other two Southern areas (Sicily and North Apulia) they present minimum values. In order to distinguish between the oils from Sicily and North Apulia it must be taken into account the variables 3rd and 6th (marked with white and purple respectively). They present maximum values for Sicily and minimum for North Apulia. As it can be seen in this pie, the sector corresponding with Sicily only presents 23 patterns whereas it actually has 36. This is due to the fact that the rest of patterns spread over the rest of clusters; in particular, some of them where included in the sector corresponding with North Apulia, in one of sectors of Calabria (in the second hierarchy level, right pie) and one of the sectors (in the first hierarchy level) corresponding with the South (specifically to South Apulia). Summarizing, for the oils from South (specifically North Apulia, Calabria and Sicily), we have that the oil will belong to:

- **CALABRIA** if V1 and V2 $\uparrow\uparrow$
- **SICILY** if V1 and V2 $\downarrow\downarrow$; V3 and V6 $\uparrow\uparrow$
- **NORTH APULIA** if V1 and V2 $\downarrow\downarrow$; V3 and V6 $\downarrow\downarrow$

4.5.4 *SonS* applied to classification tree models

Classification tree analysis is one of the main techniques used in Data Mining (Berthold and J.Hand, 2002; Hastie et al., 2009), but there is still a lack of visualization methods to support this tool. Therefore, graphical procedures should be developed in order to improve the interpretation of the solutions provided by these models. The *Sectors on Sectors (SonS)* visualization method is used to visualize the input space in the terminal nodes of the classification tree. Once the classification tree is built, each one of the subsectors obtained by *SonS*, corresponding to each variable, vary its radius in order to represent the relevance of each variable in each cluster; but for the sake of simplicity in the visualization, this step has been omitted.

For classification problems, in which we focus on this section, the goal is to find a tree where the terminal tree nodes are relatively “pure” i.e., contain observations that (almost) all belong to the same category or class. However, not always the terminal nodes are pure. Because of this, a visualization tool is proposed in which it is possible to obtain visually the number of patterns belonging to each class presented in each terminal node as well as to extract the maximum information by means of representing the input data for each class presented in each terminal node. The proposed graphical procedure helps to simplify interpretation even for complex trees and helps to interpret the different data found in the terminal nodes.

Example 1: Iris flower data set

Figure 4.22 shows the classification tree obtained for the *"Iris flower data set"*. In each terminal node, the *SonS* graph has been drawn unless all the patterns included in the terminal node belong to the same class (as occurs in terminal node labeled as "*Setosa*").

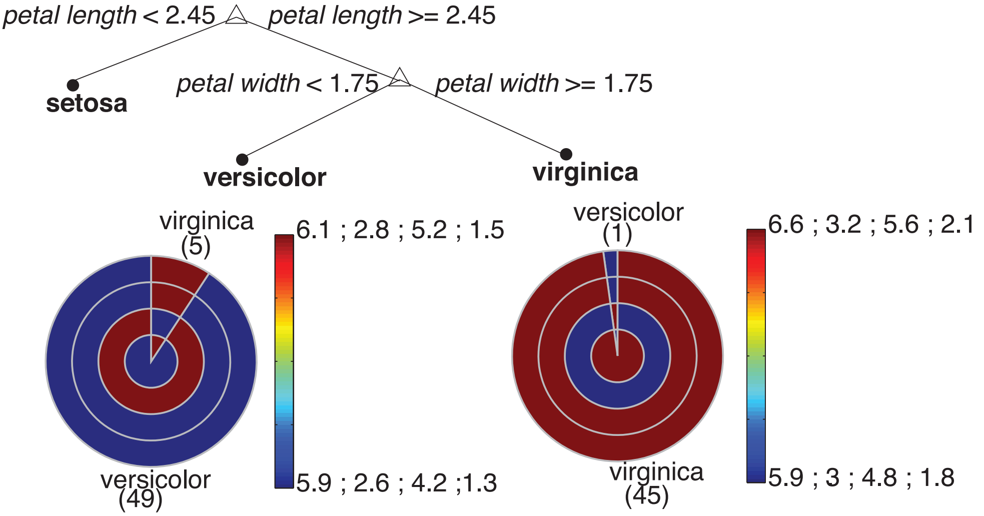


Figure 4.22: *Classification tree obtained for the "Iris flower data set" with the SonS graph in the terminal nodes.*

As shown in Figure 4.22, the most important variable to separate between *Setosa* class and others is the 3rd variable (*petal length*). If it takes a value lower than 2.45, it means that the input pattern will belong to *Setosa*, and it will belong to any of the other two classes otherwise. The classification tree indicates that in order to differentiate between *Versicolor* and *Virginica* classes, the last variable (*Petal Width*) must be taken into account. If this variable is lower than 1.75, the input pattern will belong to the *Versicolor* class; and if it is greater than or equal to 1.75 the pattern will belong to the *Virginica* class. However, as extracted from the *SonS* graph, in the terminal node corresponding with *Versicolor*, there are 5 patterns belonging to *Virginica* class. Looking at the last variable (outer subsector), which distinguish between the *Versicolor* and *Virginica* classes, along with the last column of the color bar, it can be seen that the sector corresponding to *Virginica* takes a value of 1.5, whereas the *Versicolor* class takes a value of 1.3. Thus, we could say that, *Versicolor* class corresponds with a value lower than 1.5 instead of 1.75, as classification tree indicates. In the terminal node corresponding to *Virginica*, it can be observed that just one pattern belonging to *Versicolor* class has been erroneously included.

Example 2: Italian olive oils data set

Figure 4.23 shows the classification tree obtained for the *"Italian olive oils data set"*. To extract the most significant conclusions, special attention will be paid to those terminal nodes where there is a considerable number of patterns erroneously included (more than 20%). Therefore, Figure 4.23, only shows the *SonS* graphs that follow this rule. The first *SonS* graph that attracts some attention is the corresponding to the first Calabria terminal node (1st chart starting from the right) because more than the 30% of the patterns are wrong. This chart has one sector corresponding to Calabria (9 patterns), another one corresponding to Sicily (3 patterns) and finally one corresponding to South Apulia (1 pattern). In order to distinguish among these groups of patterns, new decision rules must be established. For example, Calabria and Sicily are easily distinguishable by means of the 4th variable (marked with pink colour) because Calabria presents a maximum value (7352), indicated by deep red color, and Sicily presents a minimum value (7103), indicated by blue color. Notice that other variables also present maximum values in one of these regions, and minimum values for the other one, but the 4th variable presents the widest range (in relative values) between the maximum and minimum values. Therefore, the procedure to follow is to choose an intermediate value (7227.5) to separate between these two regions. Hence, the new rule is that if the 4th variable has a value lower than 7227.5, the patterns will belong to Sicily; and if it is greater than or equal to 7227.5, the patterns will belong to Calabria. For distinguishing South Apulia from others regions, a similar procedure can be followed, but since only one pattern is affected, an ad-hoc definition of a rule might be pointless.

Another terminal node that presents a large number of patterns erroneously included is the corresponding to the second Calabria terminal node (middle chart). In

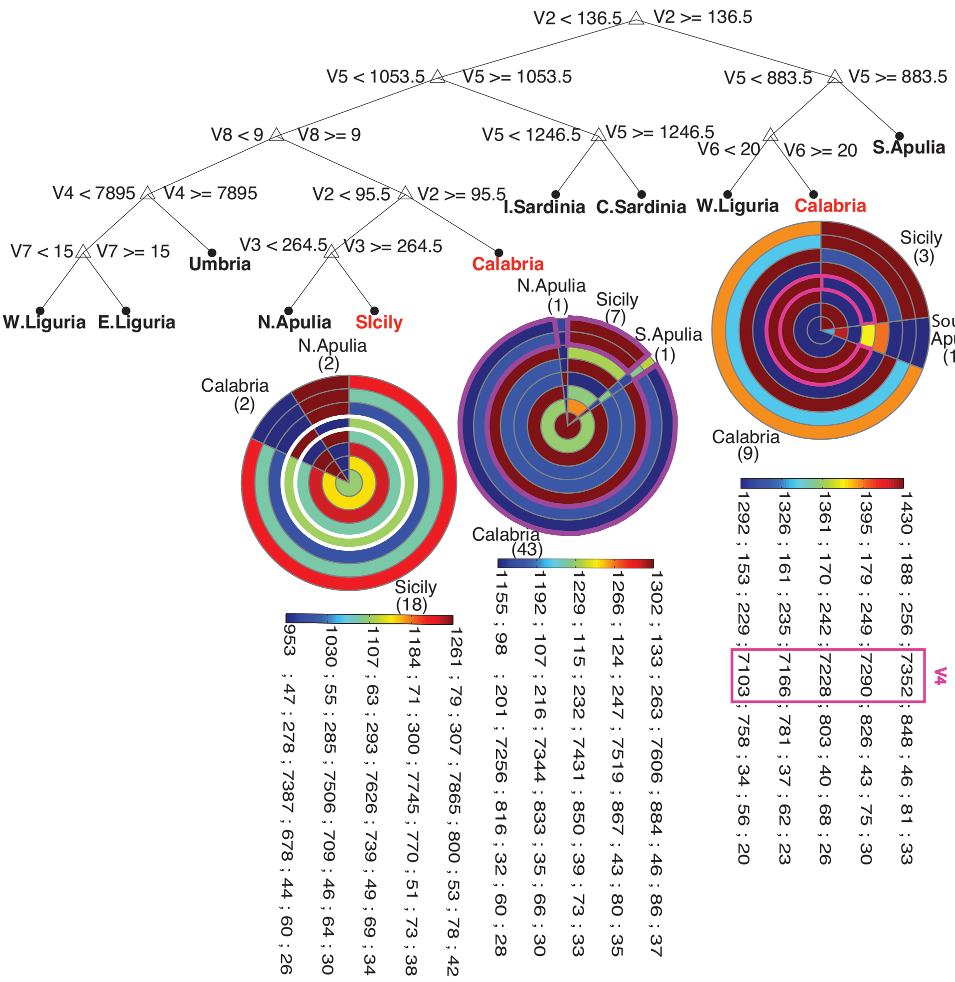


Figure 4.23: *Classification tree obtained for the "Italian olive oils" data set with the SonS graph in the terminal nodes.*

this case, low values of the 8th and 7th variables (marked with purple color) separate Calabria from Sicily.

The last terminal node to consider is that corresponding to Sicily (3rd chart starting from the right). In this case, the 5th variable (marked with white color) takes relevance in order to distinguish among the regions in this terminal node. Notice that, if this variable takes low values (blue) the olive oil will belong to North Apulia, if it takes intermediate values (green) the olive oil will belong to Sicily, and finally, if it takes high values (red) the olive oil will belong to Calabria. This is the only variable that distinguishes among the three classes included in this terminal node. It is worth mentioning that a deeper tree (which would have less generalization ability) could achieve to separate these classes. Our approach allows to extract, visually, this separation as well as gain knowledge about the problem while preserving the generalization capabilities of the tree. Anyway, in this example, the classification tree makes his role quite well because although it makes mistakes in particular final nodes, the patterns erroneously included actually belong to the same super-class; therefore, the *SonS* can be seen as an improvement or a fine-tuning.

Another advantage of this method is that it could also be used to build shallow classification trees. That means that, it may be no longer necessary to produce very deep trees because the same conclusions can be extracted visually (starting in previous nodes). That is, if the nodes of the tree are removed at some level, it will be possible to establish the rules visually without needing to build deep trees. Moreover, the *SonS* graphs could be used in other nodes (not only in terminal nodes) in order to obtain visual information about how the classification tree evolves.

Another interesting use of the original *SonS* method, in classification trees, could be to carry out a clustering algorithm with the data included in each terminal node and visualize the result. In this case, visual information about the different clusters obtained in each terminal node would be extracted.

4.6 Conclusions

This chapter presents a novel visualization technique called *Sectors on Sectors (SonS)*, and a modified version called *Multidimensional Sectors on Sectors (MDSonS)*. The *MDSonS* method makes use of Multidimensional Scaling to solve a drawback of

SonS, namely, the lack of representing distances between pairs of clusters. It has been shown the performance of these visualization tools by means of real and synthetic data sets, demonstrating its applicability.

Firstly, *SonS* and *MDSonS* methods have shown to be very useful tools when visualizing hierarchical clustering since it is possible to infer relationships among features, clusters and different levels of the hierarchy. However, *MDSonS* entails a new improvement over the *Sectors on Sectors (SonS)* method, which consists of carrying out a Multidimensional Scaling (MDS) of the centroids; and drawing each pie chart, corresponding with each cluster, in the location provided by MDS. MDS provides centroids coordinates in 2D taking into account the distances among all clusters.

Secondly, *SonS* applied in Growing Hierarchical Self-Organizing Maps (GHSOM) has demonstrated to be a useful alternative visualization tool for this algorithm since it allows a simultaneous and compact visualization of the different hierarchy levels that is not provided by the classical GHSOM. It is also a useful tool when visualizing hierarchical data since it is possible to infer relationships among features, neurons and different levels of the hierarchy demonstrating its capacity for extracting information. This fact is complicated, or not possible, in the most of classical visualizations known by the author of this thesis.

Finally, *SonS* applied in classification trees helps to extract knowledge and to obtain a better understanding even for complex trees, since it represents the input data information, for the classes associated with each terminal node (although the approach can also be applied to non-terminal nodes), of a classification tree. This method is capable of providing visual information of the patterns belonging to a terminal node in the decision tree, so that it will be possible to extract information about the values of their variables and information about the patterns erroneously included. Therefore new decision rules can be established visually in order to distinguish them. As follows another advantages and uses of the *SonS* applied in classification trees are described:

- Another advantage of this method is that it could also be used to build shallow classification trees; deep trees might not be necessary because the same conclusions can be extracted visually (starting in previous nodes). That is, if the nodes of the tree are removed at some level, it will be possible to establish the rules visually without needing to build deep trees.

- As previously mentioned, the *SonS* graphs could be used in other nodes (not only in terminal nodes) in order to obtain visual information about how the classification tree evolves.
- Another interesting use of the original *SonS* method, in classification trees, could be to carry out a clustering algorithm with the data included in each terminal node and visualize the result. In this case, visual information about the different clusters obtained in each terminal node would be extracted.

As far as the author knows there are no previous works addressing the issue of hierarchical clustering visualization in terms of obtaining information about the values of clusters centroids'; and relationship with the hierarchical arrangement provided by the clustering algorithm. Nor there is literature about methods capable of providing visual information of the patterns belonging to a terminal node in the decision tree, to the author's knowledge. Therefore, the work represents a novelty in the field of data visualization and knowledge extraction since the performance of the presented visualization methods has been shown by means of different examples (synthetics and real) demonstrating its applicability in several Data Mining techniques.

Chapter 5

ManiSonS: A New use of *SonS* method for visualizing the results of Manifold Clustering

Abstract

Feature extraction is usually needed to interpret a given problem when dealing with the visualization of high-dimensional data sets. Hence, manifold learning together with the SonS *method is presented in this chapter as a solution when dealing with data sets with a very large number of variables. In this chapter, clustering algorithms are used after applying the manifold technique. Therefore, this chapter presents a new use of the* Sectors on Sectors (SonS) *visualization technique in order to show the results of the clustering carried out on the manifold. The proposed approach extracts knowledge about the manifold and makes possible to easily find the most important variables of the manifold in order to distinguish among the different clusters. The methodology is tested in one synthetic data set and one real data set.*

5.1 Introduction

High-dimensional datasets can be very difficult to visualize. While data in two or three dimensions can be plotted to show the inherent structure of the data in a very straightforward way, equivalent high-dimensional plots are much less intuitive. Sometimes, these datasets may contain several tens of variables, making the visualization complex and non-trivial. For example, extracting information from a SOM or any of the methods proposed in this thesis with about thirty variables can be an arduous task. To aid visualization of the structure of a dataset, the dimension must be reduced in some way. One approach to simplification is to assume that the data of interest lie on an embedded non-linear manifold within the higher-dimensional space. If the manifold is of a sufficient low dimension, the data can be visualized in the low dimensional space or in a high dimensional space more simple than the original one.

Many manifold learning methods have been developed in the last decade, and it has become a hot topic in the field of Data Mining (Lee and Verleysen, 2010). These dimension reduction methods can be approached from the point of view of either unsupervised learning or supervised learning. They can be divided into linear and non linear methods. Recent research has focused on nonlinear manifolds, and the long list of manifold learning algorithms provides sophisticated examples of dimension reduction (Lee and Verleysen, 2010; Tenenbaum et al., 2000b). In the context of machine learning, manifold methods may be viewed as a preliminary feature extraction step, after which pattern recognition algorithms are applied. Therefore, when dealing with data sets with very high-dimensionality, it may be needed to carry out a feature extraction that, being smaller in number, represents the problem similarly. This also occurs in visualization problems in which may be interesting reducing the number of variables to interpret the problem more clearly. Hence, manifold learning together with the *SonS* method is presented as a solution when dealing with data sets with a very large number of variables. In this chapter, clustering algorithms are used after applying the manifold technique, so that visualizing the clustering results after applying the manifold can be of great interest. Therefore, this chapter presents a new use, in manifold field, of the *Sectors on Sectors (SonS)* visualization technique, presented in Chapter 4, in order to show the results of the clustering carried out on the manifold. For that purpose, supervised approaches were used because this chapter deals with two classification problems where the information about the label of each pattern is available. In order to solve these problems, both linear and nonlinear methods have been used. The methods used to solve these problems were Linear Discriminant Analysis (LDA) (McLachlan, 2004), Neighborhood Components Analysis (NCA) (Goldberger et al., 2004) and Maximally Collapsing Metric Learning (MCML) (Globerson and Roweis, 2006). The next section describes the main theoretical aspects of these methods.

5.2 Supervised Manifolds

As mentioned in previous section, manifolds can be approached from the point of view of either unsupervised learning or supervised learning. In the next sections, the supervised approaches used in this study are explained.

5.2.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a method used to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements (Fisher, 1936; McLachlan, 2004). In the other two methods however, the dependent variable is a numerical quantity, while for LDA it is a categorical variable (i.e. the class label). LDA assumes that the independent variables are normally distributed. The difference between ANOVA and LDA is that the former uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable (Wetcher-Hendricks, 2011).

LDA is also closely related to principal component analysis (PCA) in that they both look for linear combinations of variables which best explain the data (Martinez and Kak, 2001). LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis (Perrière and Thioulouse, 2003; Abdi and Valentin, 2007).

5.2.2 Neighbourhood Components Analysis (NCA)

Neighbourhood Components Analysis (NCA) aims at "learning" a distance metric by finding a linear transformation of input data such that the average leave-one-out (LOO) classification performance (Lachenbruch, 1967) is maximized in the transformed space. Leave-one-out (LOO) classification tries to predict the class label of a single data point by consensus of its k-nearest neighbours with a given distance metric. Therefore, NCA can learn a low-dimensional linear embedding of labelled data for data visualisation. Unlike other methods, this classification model is non-parametric without any assumption on the shape of the class distributions or the boundaries between them.

The key insight to the algorithm is that a matrix A corresponding to the transformation can be found by defining a differentiable objective function for A, followed by use of an iterative solver such as conjugate gradient descent (Surhone et al., 2010).

The matrix A is defined by means of an objective function describing classification accuracy in the transformed space and try to determine A^* such that this objective function is maximized (Eq. 5.1).

$$A^* = \text{argmax}_A f(A) \tag{5.1}$$

5.2.3 Maximally Collapsing Metric Learning (MCML)

Maximally Collapsing Metric Learning algorithm (MCML), relies on the simple geometric intuition that if all points in the same class could be mapped into a single location in feature space and all points in other classes mapped to other locations, this would result in an ideal approximation of our equivalence relation (Globerson and Roweis, 2006). MCML algorithm approximates this scenario via a stochastic selection rule, as in Neighborhood Component Analysis (NCA). However, unlike NCA, the optimization problem is convex and thus MCML is completely specified by its objective function. Different initialization and optimization techniques may affect the speed of obtaining the solution but the final solution itself is unique. This method also approximates the local covariance structure of the data, as opposed to Linear Discriminant Analysis (LDA) methods which use only global covariance structure.

5.3 Results

5.3.1 Data sets

The first data set is a synthetic data set created to show the performance of the proposed visualization method. The data consists of three clouds of points defined by six coordinates. The first three coordinates contain the most relevant information about the three different clusters, while the remaining three provide irrelevant information or noise, that is, very small values that barely provide information. Table 5.1 shows the variation ranges of each one of the variables.

Table 5.1: *Information about the variable ranges of the synthetic data set.*

Coordinate	max.	min.	mean	σ
1st	-14.0083	-34.9812	-18.1518	6.8112
2nd	15.9955	0.0119	12.2174	5.6212
3rd	30.9962	10.0240	17.4124	6.2271
4th	0.0200	0.0100	0.0151	0.0029
5th	0.0200	0.0100	0.0150	0.0028
6th	0.0200	0.0101	0.0153	0.0028

Moreover, as a real example, the *seeds data set*[1] was used, which contains X-ray images of wheat. The examined group comprised kernels belonging to three different varieties of wheat: “Kama”, “Rosa” and “Canadian”, 70 elements each, randomly selected for the experiment. Studies were conducted using combine harvested wheat grain coming from experimental fields, explored at the *Institute of Agrophysics of the Polish Academy of Sciences* in Lublin. The data set will be used for clustering tasks. To construct the data, seven geometric parameters of wheat kernels were measured (Area, Perimeter, Compactness, Length of kernel, Width of kernel, Asymmetry coefficient, Length of kernel groove); all of these parameters were real-valued.

[1] http://archive.ics.uci.edu/ml/datasets/seeds. (*Last checked September 2013*)

5.3.2 Performance evaluation

Synthetic data set

As previously mentioned, three different manifolds have been used to tackle this problem (LDA, NCA and MCML) because they are supervised techniques, that is, they make possible to introduce labels in the learning procedure. Moreover, they support exact out-of-sample extension, that is, they learn an explicit function between the data space and the low-dimensional latent space, with the same number of patterns in both spaces. After applying the different dimensionality reduction techniques (the data was reduced to two dimensions), a clustering based on k-means algorithm was performed.

Since the three manifolds produced the same success rate (100%), after applying the clustering algorithm, only the results obtained by NCA are shown. Figure 5.1 shows the results provided by the *ManiSonS* visualization method after applying the clustering algorithm on the manifold.

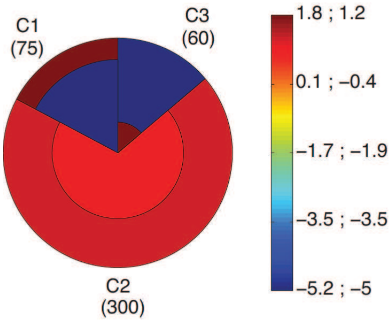


Figure 5.1: *ManiSonS visualization method applied to the synthetic data set.*

Figure 5.1 provides relevant information about the clustering carried out in the reduced space. For example, it can be seen which variables are important to separate between clusters. To this end, those variables that take high values for one cluster and low values for the other one, must be sought. For example, V1 (inner subsector) is the most important variable to separate clusters C1 and C2, since it takes high values for C2 and low values for C1. In order to separate between C1 and C3, the two variables of the manifold are useful. The input pattern will belong to C1 if V1 takes low values or also if V2 takes high values. Otherwise, it will belong to C3. To separate between C2 and C3, the most relevant variable becomes V2.

Seeds data set

In this section, the three different manifolds that were used in the previous section are used again to tackle this problem. The dimensionality reduction technique finally presented is LDA (again the data was reduced to two dimensions) since it provided the highest success classification rate (96.67%); NCA and MCML techniques provided a success classification rate of 85.24% and 91.90% respectively. Figure 5.2 shows the results provided by the *ManiSonS* visualization method after applying the clustering algorithm on LDA manifold.

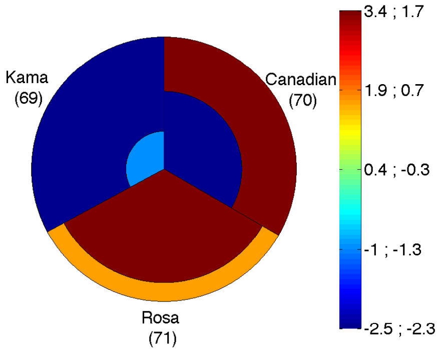


Figure 5.2: *ManiSonS visualization method applied to the seeds data set.*

As shown in Figure 5.2, it is possible to characterize each cluster by means of the variables of the manifold. For example,"Kama" cluster is characterized because it is the only cluster in which the 2nd variable (outer subsector) takes minimum values (dark blue). "Rosa" cluster is characterized because it is the only cluster in which 1st variable takes maximum values (dark red). The same occurs in "Canadian" cluster, but with the 2nd variable. Besides characterizing each cluster using the values that one variable in particular can take (or using the values of several variables in other possible problems) it can be determined which variables, or planes in the dimensional space of the manifold, are relevant to separate between clusters. For example, in order to distinguish between patterns belonging to "Kama" from "Rosa" variety, the 1st variable is the most relevant. If the 1st variable takes low values, the wheat will belong to “Kama” variety, and if it takes high values it will belong to “Rosa” variety. For distinguishing between “Kama” and “Canadian” variety, the 2nd variable takes the highest relevance. If this variable takes low values, the wheat will belong to “Kama” variety, while if it takes high values it will belong to “Canadian” one. Finally, for distinguish between “Rosa” and “Canadian” variety it should be checked the 1st variable. If it takes low values, the wheat will belong to “Canadian” variety, while if it takes high values the wheat will belong to “Rosa” one.

5.4 Conclusions

In this chapter, a method called *ManiSonS*, which is based on *SonS* method applied to manifold clustering, has been presented by means of two examples (one synthetic and one real), demonstrating its applicability in order to extract rules from the visualization of the manifold clustering. The proposed method has shown to be a very useful tool when visualizing the clustering carried out on a manifold since it is possible to infer relationships among features and clusters. Moreover, it makes possible to determine which variable, or planes in the dimensional space of the manifold, are relevant to separate between clusters. The proposed graphical procedure helps to extract knowledge and to obtain a better understanding about the results of the manifold.

This method can be used even with data sets with a large number of variables due to the fact that dimensionality reduction will make possible to represent the results of the clustering in the low-dimensional space without overloading the graph.

Chapter 6

Conclusions and future prospects

This dissertation has aimed to, on one hand, the use of the well-known Self-Organizing Maps (SOMs) to solve real problems in several different fields of research and, on the other hand, research new visualization approaches that make possible visualize the results of several Data Mining algorithms.

For the first goal, it was necessary the study and analysis of different visualization techniques. Finally, the choice of SOM was driven by two main advantages of this technique. First, SOMs enable to summarize large collections of complex data in a compact and easily interpretable graphical representation. It allows unanticipated relationships between variables to emerge freely. It must be noted that the SOM is by no means the only technique proposed in the literature for performing data visualization. However, the simplicity and intuitiveness of the SOM makes it preferable for the case studies presented in Chapter 3 since the ability to produce an easily-readable map is a key factor for supporting the use of SOMs in the problems addressed in this thesis.

For the second objective, it was necessary an exhaustive search of existing visualization methods to see which contributions existed in the literature about this area. Moreover, it was performed a comprehensive analysis of Data Mining techniques that could be susceptible to use data visualization to improve their interpretation and that of their results. In particular, as the first option, it was attempted to develop new visualization methods in the field of cluster analysis that, in addition to represent information about the centroids, also information about the relevance of the variables (information about the most relevant variables in each cluster and the most important ones to distinguish among clusters) was available. After verifying the usefulness of the proposed methods to be used in clustering methods, in particular hierarchical clustering, the use of these techniques was extended to other Data Mining techniques as Growing Hierarchical Self-Organizing Maps (GHSOM) and classification trees.

6.1 General summary

This section summarizes the contents of the thesis discussing briefly the work carried in each of the chapters. This is intended to provide the reader a general overview of the work done to contextualize the contributions of this thesis, which are presented in the next section.

Chapter 1 introduces a theoretical framework of the thesis. Moreover, a brief review about data visualization and Visual Data Mining is presented. In this chapter, the research problem, the motivations, and the goals of this research are also described. Chapter 2 discusses the main theoretical aspects of the Self-Organizing Maps (SOMs) since they are an important part of this thesis due to the fact that they were used to solve several real problems in different fields of research. Chapter 3 illustrates the usefulness of the SOMs on several real problems:

- The first problem addressed was about the study of *Balanced Scorecard* (BSC), which is a validated tool to monitor enterprise performances against specific objectives. Therein, it was proposed the SOMs as an innovative approach to extrapolate information from the BSC data and to present it in an easy-readable informative form. In this problem, it was provided evidence of the innovation offered by the SOMs: indeed, SOMs allowed a comparative analysis among *Fresenius Medical Care* KPIs (Key Performance Indicators) and perspectives over a large time lapse (not only month by month, as the currently available reports on the BSC allow) highlighting relationships that could not easily be inferred before; moreover, it was possible to track clinic improvements, and to predict the probability of these changes thus suggesting future interventions for business policy corrections.

- In the second problem, SOMs were used to evaluate Patients Satisfaction Surveys (PSS), which evaluation has become an important indicator for assessing health care quality. The aim of this work was to test and validate a methodology for identification of areas of potential improvement for specific patient groups. While traditional analyses provide only an average view of reality, this new non-canonical approach allowed segmentation of the patients in order to detect their not-so-obvious needs. By using the SOM map representation, it was also able to portray the complexity of patients' needs and to identify niches of dissatisfaction. The particular advantage of using SOMs lied in its ability to further analyze the high levels of overall satisfaction achieved in these kinds of surveys. The vast majority of patients were very satisfied with all the issues analyzed in the survey; many classical methods would have been biased by this fact and would have not been able to provide useful further analysis (almost all the data would have been grouped together since the results of the surveys were very similar). Moreover, compared with classical clustering techniques that can find typical profiles but are associated with complex presentation of the results, SOM was also able to find similar behaviors (typical profiles are represented in the same area of the map), and simultaneously depict the results in an easily interpreted 2- dimensional map.

- The third real problem was framed in the field of cardiology. This study proposed a new methodology in order to obtain visual information among four important groups of patients: *VF* (Ventricular Fibrillation), *VT* (Ventricular Tachycardia), *HP* (Healthy Patients) and *AHR* (other Anomalous Heart Rates and Noise) since methods used up to now do not provide insight into the problem (such methods only attempt to classify the different groups of patients). This analysis showed that it was possible to perform a profile of patients suffering from Ventricular Fibrillation or Ventricular Tachycardia and other corresponding to healthy patients.

- The fourth problem addressed the application of SOMs in a physiotherapy problem by means of the valuation analysis of the knee in athletes in the pre- and post-surgery of the anterior cruciate ligament, studying variables of force and measurements at different distances of the knee. The SOM was able to show that in the case of thigh muscle contours, there were significant negative changes (decrease of the contour) on the vastus lateralis between pre- and post-surgery, but there was a final improvement of the overall thigh muscle contours at six months, due to the fact that a proper rehabilitation program was applied.

- The last study proposed the use of SOMs for evaluating data about comfort in footwear provided by *Instituto Tecnológico del Calzado y Conexas (INESCOP)*. Here, it was studied which factors can be decisive when buying footwear, revealing interesting hidden relationships and patterns. Important conclusions were drawn from this study, but the most remarkable ones are detailed as follows. As global conclusion, it was proved that taking all measures to obtain the valuation of the testers may not be necessary, specially for men, whose valuation were almost equal in all the foot areas. Another of the most important conclusions is that there is a different behavior between men and women in terms of valuations and when buying footwear. Women usually ask tighter shoes than they really may need according to their physical measures. Moreover, a tighter shoe has a negative influence on a man while this fact has no negative influence in a woman valuation. Another difference between men and women behavior is that women are more sensitive to different parts of the foot than men (men have a similar opinion, or global idea, about the different space areas of the footwear).

Chapter 4 focused on presenting new visualization methods. This chapter presented a novel visualization technique called *Sectors on Sectors (SonS)*, and an extended version called *Multidimensional Sectors on Sectors (MDSonS)*, for improving the interpretation of several Data Mining algorithms. These methods were applied for visualizing the results of: a) hierarchical clustering, which made possible to extract all the existing relationships among centroids' attributes at any hierarchy level; b) Growing Hierarchical Self-Organizing Maps (GHSOM), a variant of the well-known Self-Organizing Maps (SOM), by means of which was possible to visualize, simultaneously, the data information at each hierarchy level compactly and extract relationships among variables; c) classification trees, in which the *SonS* was used for representing the input data information for each class presented in each terminal node of a classification tree providing extra information for a better understanding of the problem. These methods were tested by means of several data sets (real and synthetic). Achieved results showed the suitability and usefulness of the proposed approaches.

Chapter 5 presented manifold learning together with the *SonS* method as a solution when dealing with data sets with a large number of variables. This approach made possible to determine which variables, in the dimensional space of the manifold, were relevant to separate between clusters carried out on it. The proposed graphical procedure helps to extract knowledge and to obtain a better understanding about the results of the manifold. This method can be used even with data sets with a large number of variables due to the fact that dimensionality reduction will make possible to represent the results of the clustering in the low-dimensional space without overloading the graph. Finally, Appendix A presents an interactive software tool based on *SonS* technique, and programmed in *Processing*, to endow such method of an added value. This software tool allows the user to control that visualization, interacting with such interface in a way that reveals new information as the user explores the piece.

6.2 Summary of contributions

The major contributions of this thesis can be divided roughly into the objectives mentioned above, that is, the use of SOMs to solve real research problems and to create new visualization tools for high dimensional data sets. But in particular, the present work makes several contributions to the principles and practice of the visualization, which are described below:

- This thesis has contributed to the knowledge extraction into various research problems using data visualization, which would not have been possible using classical techniques. The conclusions drawn in each of them were novel and valuable, and they provided evidence of the innovation offered by the SOM to address these problems.

- Another contribution is a novel visualization technique, called *Sectors on Sectors (SonS)*, firstly focused on sequential and hierarchical clustering. This method aims to extracts visual information of data groups by representing the number of instances in each group, the value of the centroids of these clusters and the existing relationships among the several groups and variables.

- Another novel method, with the same aim as the *SonS*, which emerged as an improvement of the former. This method, called *Multidimensional Sectors on Sectors (MDSonS)* makes possible visualizing the distances among clusters. The visualization method is different from that proposed in *SonS* method due to the need of accommodating the information provided by Multidimensional Scaling (MDS) to the new visualization.

- Extension of one of the proposed methods to other Data Mining techniques. The *SonS* has been used to visualize the Growing Hierarchical Self-Organizing Maps (GHSOM) and classification trees results showing to be very useful. On one hand, *SonS* visualization applied in GHSOM has demonstrated to be a useful alternative visualization tool for this algorithm since it allows a simultaneous and compact visualization of the different hierarchy levels that is not provided by the classical GHSOM. The main problem of GHSOM is that it is not possible to visualize simultaneously the data information in each level; that is, GHSOM visualization only makes possible to depict the component planes corresponding to the deepest SOMs so that the information about the component planes corresponding to the first SOMs in first or intermediate hierarchies is visually inaccessible. *SonS* visualization technique was used to circumvent that drawback, since it allows a simultaneous and compact visualization of the different hierarchy levels, and it also enables the extraction of knowledge in terms of relationships among variables. This is not possible using other classical visualizations. On the other hand, *SonS* applied in classification tree models helps to extract knowledge and to obtain a better understanding even for complex trees, since it represents the input data information, for the classes associated with each terminal node (although the approach can also be applied to non-terminal nodes), of a classification tree. This method is capable of providing visual information of the patterns belonging to a terminal node in the decision tree, so that it will be possible to extract information about the values of their variables and information about the patterns erroneously included. Therefore new decision rules can be established visually in order to distinguish them. Another advantage of this method is that it could also be used to build shallow classification trees; deep trees might not be necessary because the same conclusions can be extracted visually (starting in previous nodes).

- Use of the *SonS* method with manifold learning. Thus, it is possible to visualize data sets with a large number of variables due to the fact that dimensionality reduction will make possible to represent the results of the clustering in the low-dimensional space without overloading the graph. Moreover, this method makes possible to obtain information about the manifold learning. It should be noted that this procedure could also be performed with the *MDSonS* method.

- A new interactive tool, programmed in *Processing* that represents the *SonS* method in which the user controls that visualization interacting with such interface in a way that reveals new information as the user explores the piece.

6.3 Strengths and weaknesses of the proposed visualization techniques

The different visualization techniques that are proposed in this thesis are distinct in some respect from each other. The main advantages and disadvantages of each technique are listed below.

The main advantage of the proposed visualization techniques, in clustering analysis field, is that they make possible to observe relationships among different variables in the same cluster's centroid and relationships among the same variables in different clusters's centroids, in the different levels of the hierarchy. As far as the author of this thesis knows there are no previous works addressing two issues of hierarchical clustering visualization: obtaining information about the values of clusters' centroids and, at the same time, discovering the hierarchical arrangement provided by the clustering algorithm. Regarding using the proposed methods in GHSOM and classification trees, they provide also new visual information that helps to the interpretation of the results of such algorithms. Again, there is not literature about methods capable of providing visual information of the patterns belonging to a terminal node in the decision tree, to the author's knowledge. Therefore, the work represents a novelty and an important advance in the field of data visualization and knowledge extraction since the performance of the presented visualization methods has been shown by means of different examples (synthetics and real) demonstrating its applicability in several Data Mining techniques. It is noteworthy that, although the *MDSonS* provides more information, the *SonS* has the benefit of a more compact visualization, which is a plus when dealing with hierarchical clustering, GHSOM or classificatin trees (a single pie chart represents the clustering performed in a whole hierarchy). It should be said that this may be a drawback when facing a problem with more variables, since being more compact, the visualization provided is overloaded and more difficult to interpret.

The weakness of the proposed methods is that when dealing with a large number of variables the graphs can be overloaded. Therefore, a new procedure that uses manifold learning was performed. In this way, it was able to reduce the number of variables so that the proposed methods continue to be useful when faced with problems of very high dimension (i.e twenty or thirty variables). It should be noted that this approach partially solves the problem as it is actually faced in a transformed space where the variables do not correspond with the originals. Another weakness is that the proposed methods are not very intuitive for an inexperienced user. To make the most of the methods, and use them effectively, it is necessary that the user have worked with them previously and to be familiarized with them. Even so, despite this shortcoming, the proposed methods are in great usefulness and it is convenient to learn how to interpret them because the benefits and contributions of using them are very significant.

6.4 Future prospects

Future work focuses on improving the proposed methods and trying to provide more information without overloading their visualization. One possibility lies in including information about preprocessing of input data (using statistics such as standard deviation or eigenvalue in principal component analysis). A possible way to include new information in the *MDSonS* method is to provide knowledge about the clusters shape. It may be of great interest knowing whether one is dealing with a spherically or ellipsoidally (more elongated) shaped cluster. For this purpose, the two principal components of each cluster could be computed (by means of a principal component analysis). After this, an ellipse could be painted using the two principal components.

Moreover, it would be interesting working on other clustering algorithms (using for example Mahalanobis metric). One possible approach could be develop the above mentioned concept in the "interactive *SonS*" presented in Appendix A, where by means of a selector or buttons the user could select the desired metric to show, and that the visualization was adapted to such selection at the precise moment in which the user interacts. Also, the ongoing work is on developing the *MDSonS* as interactive tool (as in *SonS* case).

Regarding the use of manifolds, for reducing the dimensionality of the data, along with the methods presented in this thesis, the research on new methods to automatically select the number of variables used in the reduced space (number of variables in the manifold) would be interesting. One possibility would be to define a measure that took into account, for example, the distance between input patterns and between clusters in the reduced space.

Finally, another possible future work for improving the proposed visualization method could be to come up with other strategies, apart from manifold learning, that might improve its performance when dealing with a large number of features.

6.5 Scientific publications achieved with this thesis

During the research stage developed in the doctoral studies related to this thesis, several scientific results have been produced. These results are reflected in a number of scientific papers and participation in conferences. The following sections list the scientific results related to this thesis.

6.5.1 Journal publications

- Cattinelli, I., Bolzoni, E., Barbieri, C., Mari, F., Martín-Guerrero, J.D., Soria-Olivas, E., **Martínez-Martínez, J. M.**, Gomez-Sanchis, J., Amato, C., Stopper, A. and Gatti, E. (2011) Use of Self-Organizing Maps for Balanced Scorecard analysis to monitor the performance of dialysis clinic chains. *Health Care Management Science*, 15(1):79:90.

- Martín-Guerrero, J.D., Marcelli, D., Soria-Olivas, E., Mari, F., **Martínez-Martínez, J. M.**, Soley-Bech, I., Martínez-Sober, M., Scatizzi, L., Gomez-Sanchis, J., Stopper, A., Serrano-lópez, A.J., and Gatti, E. (2012) Self-Organising Maps: A new way to screen the level of satisfaction of dialysis patients. *Expert Systems with Applications*, 15(1):79:90.

- Rosado-Muñoz, A., **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E. (2013). Visual data mining with self-organising maps for ventricular fibrillation analysis. *Computer Methods and Programs in Biomedicine*, 39(10):8793-8798.

- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D. and Gómez-Sanchis, J. (*submitted* in 2013). Multidimensional Sectors on Sectors (MDSonS): A New Hierarchical Clustering Visualization Tool. *Data & Knowledge Engineering.*

- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D. and Serrano-lópez, A.J. (*submitted* in 2013). A New Visualization Tool for Data Mining Techniques. *IEEE Transactions on Visualization and Computer Graphics.*

6.5.2 Book chapters

- Alakhdar, Y., **Martínez-Martínez, J.M.**, Guimerá-Tomás, J., Escandell-Montero, P. and Benítez, J. (2012). Visual Data Mining in Physiotherapy Using Self-Organizing Maps: A New Approximation to the Data Analysis. In *Medical Applications of Intelligent Data Analysis: Research Advancements*, 187-194. IGI Global.

- Escandell-Montero, P., Alakhdar, Y., Soria-Olivas, E., Benítez, J. and **Martínez-Martínez, J.M.** (*Accepted* in 2013). Artificial neural networks: a new analysis tool in physical therapy. In *Encyclopedia of Information Science and Technology, Third Edition.* IGI Global.

6.5.3 Conferences papers

- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D., Martínez-Sober M., and Gómez-Sanchis, J. (2011). Sectors on Sectors (SonS): A New Hierarchical Clustering Visualization Tool. *Computational Intelligence and Data Mining, 20011. CIDM '11. IEEE Symposium on*,304-309.

- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D., Gómez-Sanchis, J., and Vila-Francés, J.(2011). Growing Hierarchical Sectors on Sectors. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2011. ESANN '11*, 239-244.

- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D., Gómez-Sanchis, J., and Vila-Francés, J.(2012). Extended visualization method for classification trees. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2012. ESANN '12*, 197-202.
- **Martínez-Martínez, J.M.**, Escandell-Montero, P., Martín-Guerrero, J.D., Vila-Francés, J. and Soria-Olivas, E. (2013). ManiSonS: A New Visualization Tool for Manifold Clustering. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2013. ESANN '13*, 561-566.

6.5.4 Research projects

Although this work has been partially supported by *Fresenius Medical Care (FME)* in the framework of the project "*Intelligent analysis of Fresenius Medical Care data*", the contributions of this research have had an impact in the following projects:

- Intelligent analysis of Fresenius Medical Care data. *Fresenius Medical Care (FME)*. Duration: 2010-2012. Main researcher: José David Martín Guerrero.
- Análisis inteligente de datos para la evaluación de confort en el calzado. *Instituto Tecnológico del Calzado y Conexas (INESCOP)*. Duration: 2010-2011. Main researcher: José David Martín Guerrero.
- Contribuciones de la Inteligencia Computacional a la mejora y evaluación del confort del calzado (CIFOOTCOM). *AEUV - Accions especials de la Universitat de València*. Duration: 2012-2012. Main researcher: José David Martín Guerrero.
- Inteligencia Computacional aplicada a datos médicos de Nephrocare e-Services'.
- Intelligent analysis of Fresenius Medical Care data. *Fresenius Medical Care (FME), nephrocare e-Services*. Duration: 1/04/2013-1/10/2014. Main researcher: José David Martín Guerrero.
- Smart tools for the Prescription of orthopaedic Insoles and Footwear (SMART-PIF). *European Commission (VII Framework Program)*. Duration: 2013-2015. Main researcher: Joé David Martín Guerrero.

Appendix A

SonS method developed as an interactive tool

Abstract

Due to the importance of interactive visualization methods, where the user controls that visualization interacting with such interface in a way that reveals new information as the user explores the piece, it was decided to develop a software tool based on SonS method and programmed in Processing.

A.1 Introduction

As mentioned throughout this thesis, data visualization is of great importance because it helps us to extract underlying information in data that would otherwise not be possible. Visualizations encompass a wide and growing range of projects, reflecting creative ways of representing all sorts of data visually, with virtually no limit to what kind of information can be translated into an image. The designer of a visualization determines which visual element (color, shape, size, motion, and so forth) will represent individual data points. Images can be 2D or 3D, can be static or dynamic, and can allow user interaction with computers to create graphic illustrations of information so that the user can draw more effectively on the data and focus on what really interests him from the visualization. Interactive visualization is emerging as a vibrant new form of communication, providing compelling presentations that allow viewers to interact directly with information in order to construct their own understandings of it. Building on a long tradition of print-based information visualization, interactive visualization is interface-based products that utilize on-page navigation features (i.e. mouse-overs, clicks, drop-down menus, check boxes, etc.), allowing the user to interact with such interface in a way that reveals new information and/or data as the user explores the piece. It utilizes the technological capabilities of computers, the Internet, and computer graphics to marshal multifaceted information in the service of making a point visually. Interactive data visualization goes a step further, moving beyond the display of static graphics and spreadsheets to using computers and mobile devices to drill down into charts and graphs for more details, and interactively changing what data you see and how it is processed.

For all the above, and after seeing the potential of the new visualization methods proposed in Chapter 4 (endorsed by various scientific publications), it was decided to go a step further and develop an interactive tool to endow such method of an added value. Therefore, the present chapter presents a new interactive tool based on *SonS* method through the use of *Processing*[1].

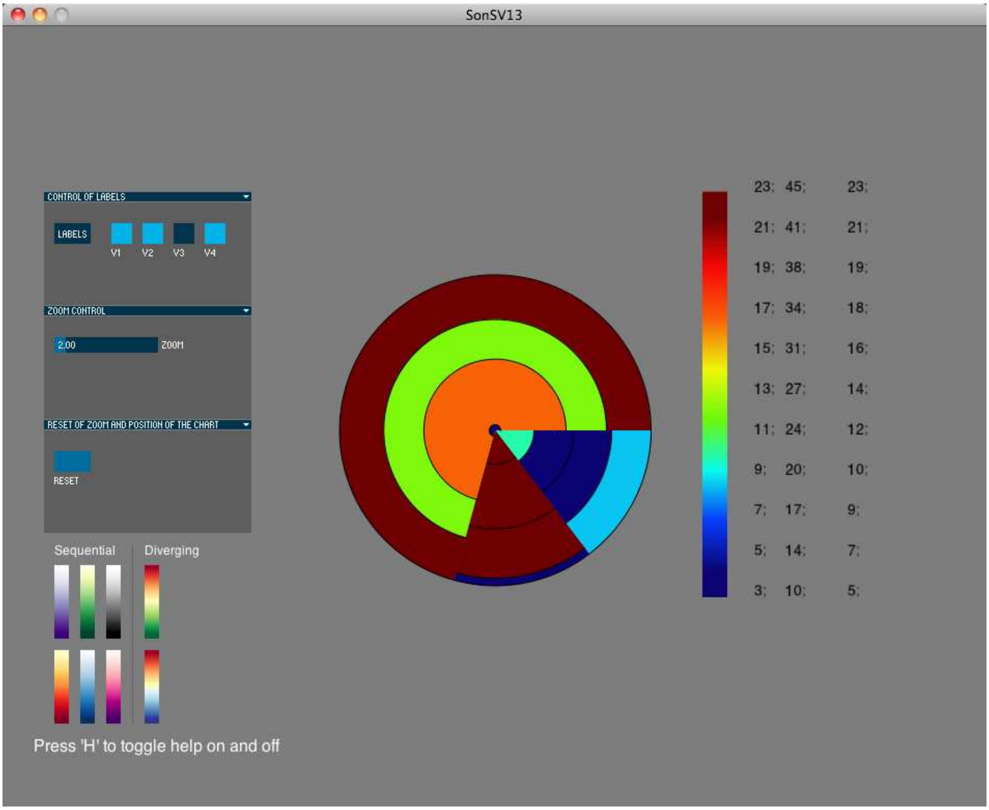
Processing is a programming environment to develop visually oriented applications with an emphasis on animation and providing users with instant feedback through interaction. Originally built as a domain-specific extension to Java targeted towards artists and designers, *Processing* has evolved into a full-blown design and prototyping tool used for large-scale installation work, motion graphics, and complex data visualization. *Processing* is based on Java and is open source.

Written in *Processing* language, the proposed tool aims to provide the user with a complete tool, supporting the control of interactive information visualization that can be tailored to the addressed problems.

[1] http://processing.org/

A.2 Developed interactive tool

One of the advantages of working with *Processing* is that the developed tools can be exported as applications compatible to run on Windows, Mac OS X, Linux, Android or embedded in web. Figure A.1 shows the application developed in *Processing*.

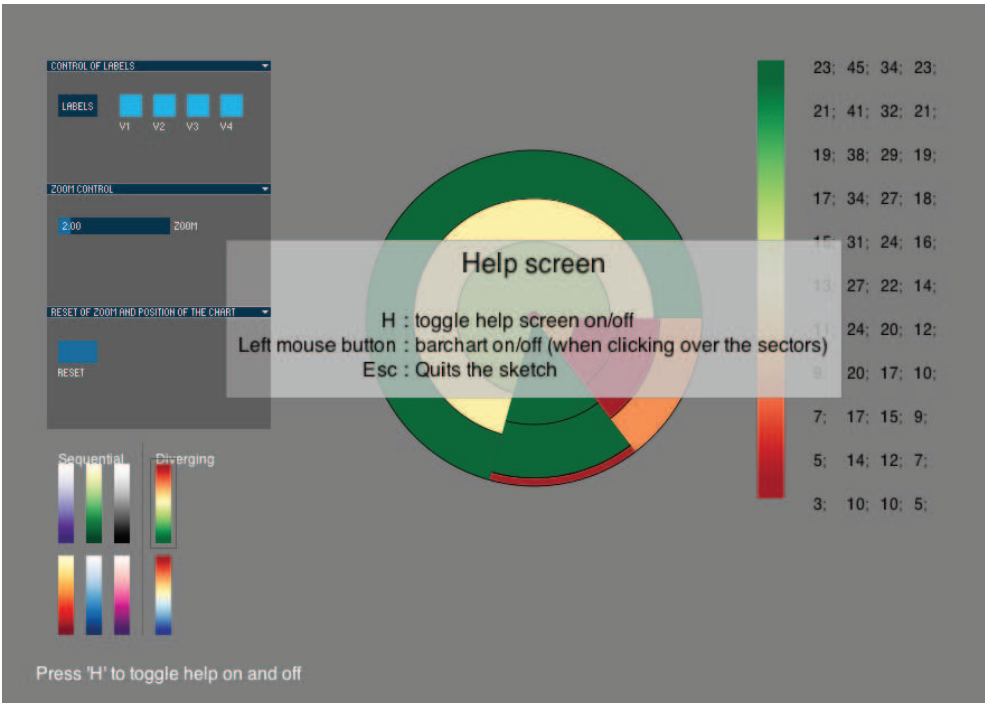


Press 'H' to toggle help on and off

Figure A.1: *Main window of the developed interactive tool.*

It is an easy to use and intuitive tool that by means of a brief help (when typing “H” the help of the method toggle on and off, see Figure A.2) the user can start interacting with visualization tool.

Such application contains several modules (on the left in Figure A.1) for the control of *SonS* graph, located on the right in the picture. These modules are endowed of functionalities so that the user can modify the visualization. These are some of the features added to the method:



Help screen

H : toggle help screen on/off
Left mouse button : barchart on/off (when clicking over the sectors)
Esc : Quits the sketch

Press 'H' to toggle help on and off

Figure A.2: *Help window superimposed on the main window of the developed interactive tool. The help window appears after typing "H".*

- **Control of labels.** This interactive menu makes possible to select the labels of variables that the user wants to display. In Figure A.1 , it is observed that in the module corresponding to "control of labels" there are four control buttons, three of which are activated/pressed (V1, V2 and V4), whereas V3 button is disabled. If the labels next to the color bar are observed, it can be seen that there is just a gap in the third column of labels, corresponding to the third variable, due to the deactivation of the V3 control button.

- **ZOOM control.** By means of a slider, it is possible to zoom in and zoom out the *SonS* graph in case of necessity (as for example occurred in Chapter 4, Section 4.5.3, where the radius of the last variable, for some sectors, was very small). Users can pan across the surface in two dimensions and zoom into objects of interest.

- **Reset.** This button makes possible to reset the zoom and position of the chart.

- **Selection of color scheme:** Clicking on the different "mini" color bars, it is possible to select the color scheme (or "color map") of the *SonS* graph. Two different legend types have been included, named *Sequential* and *Diverging*.

In addition to these interactive functionalities, new visualization features were added to provide more information to the method:

- **Visualization of the number of patterns included in each cluster.** When mouse is over one sector, the number of patterns included in the cluster corresponding to such sector is indicated (see Figure A.3).
- **Visualization of the classes of the patterns included in each cluster.** When left-clicking on a cluster, information about patterns including in such cluster is shown. This information, which is related to the classes of the problem addressed (if supervised problem), is reduced to a bar chart showing the number of patterns of each class existing in such cluster (see Figure A.4). In this way, one can get information on the number of patterns erroneously included. In Figure A.4, it can be seen that "Cluster 1" contains 12 patterns. In the bar chart corresponding with this cluster it is perceived that 8 of the 12 patterns corresponds to C1, 3 to C2 and 1 to C3.

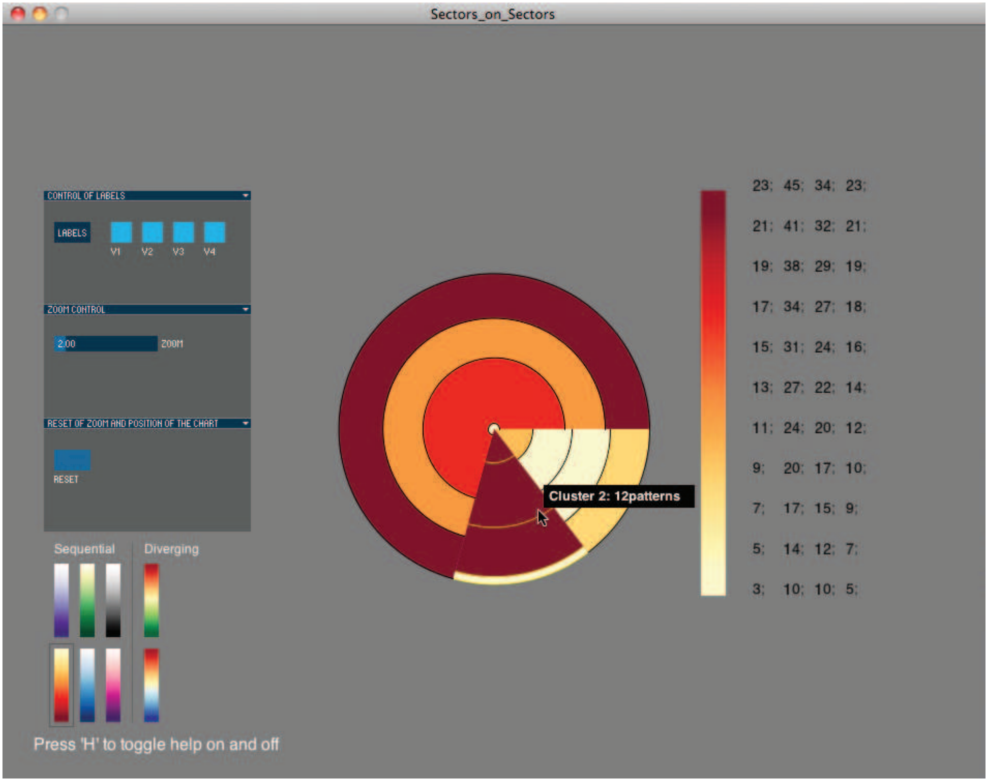
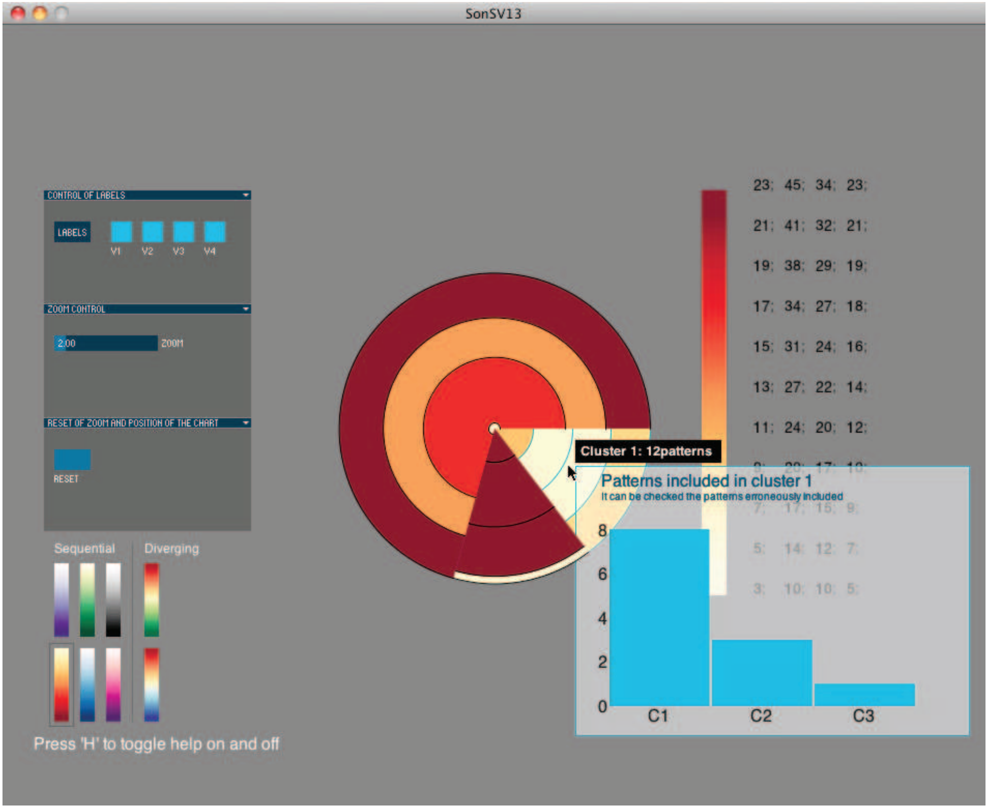


Figure A.3: *Snapshot of the application when mouse is over one sector. The number of patterns included in the cluster corresponding to such sector is indicated.*



Cluster 1: 12patterns

Patterns included in cluster 1

It can be checked the patterns erroneously included

Press 'H' to toggle help on and off

Figure A.4: *Snapshot of the application when left-clicking on a sector. Information about patterns including in such cluster is shown by means of a bar chart showing the number of patterns of each class existing in such cluster*

A.3 Conclusions

In this Appendix, an interactive software tool based on *SonS* technique, and programmed in *Processing* has been presented to endow such method of an added value. This software tool allows the user to control that visualization, interacting with such interface in a way that reveals new information as the user explores the piece. By means of several control buttons, and interacting with the mouse, the visualization becomes more intuitively and the user can gain more knowledge than in the non-interactive version. The functionalities added to this interactive version are: control of labels (it is possible to select the labels of variables that the user wants to display), zoom control (by means of a slider, it is possible to zoom in and zoom out the *SonS* graph), reset button (this button makes possible to reset the zoom and position of the chart), visualization of the number of patterns included in each cluster when mouse is over one sector and, finally, visualization of classes of the patterns included in each cluster when left-clicking on a sector.

Bibliography

Abdi, H. and Valentin, D. (2007). Discriminant correspondence analysis. In *Encyclopedia of measurement and statistics. Thousand Oaks: Sage.*, pages 270–275.

Afonso, V. X. and Tompkins, W. J. (1995). Detecting ventricular fibrillation. *IEEE Engineering in Medicine and Biology.*, 14(2):152–9.

Ageberg, E. (2002). Consequences of a ligament injury on neuromuscular function and relevance to rehabilitation using the anterior cruciate ligament-injured knee as model. *Journal of Electromyography and Kinesiology*, 12(3):205–212.

Alakhdar, Y., Martínez-Martínez, J. M., Guimerà-Tomás, J., Escandell-Montero, P., Benítez, J., and Soria-Olivas, E. (2012). Visual data mining in physiotherapy using self-organazing maps: A new approximation to the data analysis. In Magdalena, R., Soria, E., Guerrero, J., Gómez, J., and Serrano, A. J., editors, *Medical Application of Intelligent Data Analysis: Research Advancements*, chapter 12, pages 186–193. IGI Global.

Alonso-Atienza, F., Rojo-Álvarez, J. L., Rosado-Muñoz, A., Vinagre, J. J., García-Alberola, A., and Camps-Valls, G. (2012). Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. *Expert Systems with Applications*, 39(2):1956 – 1967.

Alpaydin, E. (2010). *Introduction to Machine Learning.* The MIT Press, 2nd edition.

Amann, A., Tratnig, R., and Unterkofler, K. (2005). Reliability of old and new ventricular fibrillation detection algorithms for automated external defibrillators. *BioMedical Engineering Online*, 4(60).

Amann, A., Tratnig, R., and Unterkofler, K. (2007). Detecting ventricular fibrillation by time-delay methods. *Biomedical Engineering, IEEE Transactions on*, 54(1):174 –177.

Andrews, K. (2002). Visual exploration of large hierarchies with information pyramids. *Information Visualisation, International Conference on*, 0:793.

Andrews, K. and Heidegger, H. (1998). Information slices: Visualising and exploring large hierarchies using cascading, semi-circular discs. In *Information Visualization, 1998. Proceedings. IEEE Symposium on.*

Andrews, K. and Kasanicka, J. (2007). A comparative study of four hierarchy browsers using the hierarchical visualisation testing environment (hvte). In *Information Visualization, 2007. IV '07. 11th International Conference*, pages 81 –86.

Ankerst, M., Ester, M., and Kriegel, H.-P. (2000). Towards an effective cooperation of the user and the computer for classification. In *International Conference on Knowledge Discovery & Data Mining (KDD-'2000)*, pages 179–188.

Atienza, F. A., Rojo-Álvarez, J. L., Camps-Valls, G., Muñoz, A. R., and García-Alberola, A. (2006). Bootstrap feature selection in support vector machines for ventricular fibrillation detection. In *European Symposium on Artificial Neural Networks, 2006. ESANN '06.*, pages 233–238.

Baehrecke, E., Dang, N., Babaria, K., and Shneiderman, B. (2004). Visualization and analysis of microarray and gene ontology data with treemaps. *BMC bioinformatics*, 5(1):84.

Bai, B. and Wang, Y. (2011). Ventricular fibrillation detection based on empirical mode decomposition. In *Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on*, pages 1 –4.

Ball, S. and Petsimeris, P. (2010). Mapping urban social divisions. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 11(2):20.

Bansal, K. L. and Sood, S. (2011). Data visualization a tool of data mining. *International Journal of Computer Science and Technology*, 2(3):197–198.

Barro, S., Ruiz, R., Cabello, D., and Mira, J. (1989). Algorithmic sequential decision making in the frequency domain for life threatening centricular arrhythmias and aimitative artifacts: a diagnostic system. *Journal of Biomedical Engineering.*, 11(4):320–8.

Basara, H. and Yuan, M. (2008). Community health assessment using self-organizing maps and geographic information systems. *International Journal of Health Geographics*, 7(67).

Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1994). Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry: Theory and Applications*, 4(5):235–282.

Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1999). *Graph Drawing*. Prentice Hall.

Beaudoin, L., Parent, M.-A., and Vroomen, L. (1996). Cheops: a compact explorer for complex hierarchies. In *Visualization '96. Proceedings.*, pages 87 –92.

Beck, C. S., Pritchard, W. H., Giles, W., and Mensah, G. (1947). Ventricular fibrillation of long duration abolished by electric shock. *The journal of the American Medical Association.*, 135:985–6.

Berthold, M. and J.Hand, D. (2002). *Intelligent Data Analysis*. Springer, 2nd edition.

Borg, I. and Groenen, P. J. (2005). *Modern Multidimensional Scaling. Theory and Applications*. Springer.

Bourne, M., Neely, A., Mills, J., and Platts, K. (2003). Implementing performance measurement systems: a literature review. *International Journal of Business Performance Management*, 5(1):1–24.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, 1st edition.

Buja, A. and Asimov, D. (1986). Grand tour methods: an outline. In *Proceedings of the Seventeenth Symposium on the interface of computer sciences and statistics on Computer science and statistics*, pages 63–67, New York, NY, USA. Elsevier North-Holland, Inc.

Capucci, A., Aschieri, D., Piepoli, M., Rosi, A., Arvedi, M., and Villani, G. (2001). Early defibrillation through first responders triples survivals rates from out of hospital cardiac arrest in an Italian community. *European Heart Journal.*, 22:242.

Card, S. K., Mackinlay, J. D., and Shneiderman, B., editors (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making unit. *European Journal of the Operational Research*, 2:429–44.

Chen, C.-h., Hardle, W., and Unwin, A. (2008a). *Handbook of Data Visualization (Springer Handbooks of Computational Statistics)*. Springer-Verlag TELOS, Santa Clara, CA, USA.

Chen, C.-h., Hardle, W., Unwin, A., and Inselberg, A. (2008b). Parallel coordinates: Visualization, exploration and classification of high-dimensional data. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pages 643–680. Springer Berlin Heidelberg.

Chen, S., Thakor, N., and Mower, M. (1987). Ventricular fibrillation detection by a regression test on the autocorrelation function. *Journal of Medical and Biolical Engineering Computing.*, 25(3):241–9.

Chen, S. W., Clarkson, P. M., and Fan, Q. (1996). A robust sequential detection algorithm for cardiac arrhythmia classification. *IEEE Transactions on Biomedical Engineering*, 43(11):1120–5.

Chernoff, H. (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, 68(342):361–368.

Clayton, R. H. and Murray, A. (1998). Comparison of techniques for time-frequency analysis of the ecg during human ventricular fibrillation. In *IEEE Proceeeding -Science, Measurements and Technoly.*, volume 145, pages 301–6.

Clayton, R. H., Murray, A., and Campbell, R. W. (1993). Comparison of four techniques for recognition of ventricular fibrillation from the surface ECG. *Medical and Biological Engineering and Computing.*, 31(2):111–7.

Clayton, R. H., Murray, A., and Campbell, R. W. (1994). Recognition of ventricular fibrillation using neural networks. *Medical and Biological Engineering and Computing.*, 32(2):217–20.

Clayton, R. H., Murray, A., and Campbell, R. W. (1995). Evidence for electrical organization during ventricular fibrillation in the human heart. *Journal of Cardiovascular Electrophysiology.*, 6(8):616–24.

Cohen, L. (1995). *Time Frequency Analysis*. Prentice Hall PTR, New Jersey, USA, 1st edition.

Corporation, T. C. (1999). *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation.

Cox, D. and Snell, E. (1981). *Applied Statistics: Principles and Examples*. Chapman and Hall/CRC Texts in Statistical Science Series. Chapman and Hall.

Daugirdas, T. (1993). Second generation logarithmic estimates of single-pool variable volume Kt/V: an analysis error. 4:1205–1213.

Davidenko, J. M., Pertsov, A. V., Salomonsz, R., Baxter, W., and Jalife, J. (1992). Stationary and drifting spiral waves of excitation in isolated cardiac muscle. *Nature*, 355(6358):349–51.

de Francisco, A. L., Kim, J., Anker, S. D., Belozeroff, V., Canaud, B., Chazot, C., Drüeke, T. B., Eckardt, K.-U., Floege, J., Kronenberg, F., Macdougall, I. C., Marcelli, D., Molemans, B., Passlick-Deetjen, J., Schernthaner, G., Stenvinkel, P., Wheeler, D. C., Fouqueray, B., and Aljama, P. (2010). An Epidemiological Study of Hemodialysis Patients Based on the European Fresenius Medical Care Hemodialysis Network: Results of the ARO Study. *Nephron. Clinical practice*, 118(2):c143–c154.

Deboeck, G. J. and Kohonen, T. K., editors (1998). *Visual Explorations in Finance*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition.

DeMets, D., Furberg, C., and Friedman, L. (2006). *Data Monitoring in Clinical Trials: A Case Studies Approach*. Springer.

Díaz, I., Cuadrado, A. A., Diez, A. B., Domínguez, M., Fuertes, J. J., and Prada, M. A. (2012). Supervision of industrial processes using self organizing maps. In Magdalena-Benedito, R., Martínez-Sober, M., Martínez-Martínez, J. M., Escandell-Montero, P., and Vila-Francés, J., editors, *Intelligent Data Analysis for Real-Life Applications: Theory and Practice*, chapter 11, pages 206–227. IGI Global.

Dittenbach, M., Merkl, D., and Rauber, A. (2000). The growing hierarchical self-organizing map. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 15 –19 vol.6.

Dittenbach, M., Rauber, A., and Merkl, D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*, 48(1):199–216(18).

Everitt, B., Landau, S., and Lee, M. (2009). *Cluster Analysis*. Wiley, 4th edition.

Faddy, S. C. (2006). Reconfirmation algorithms should be standard of care in automated external defibrillators. *Resuscitation*, 68(3):409–15.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fernández, C., Soria, E., Martín-Guerrero, J. D., and Serrano, A. J. (2006). Neural networks for animal science applications: Two case studies. *Expert Systems with Applications.*, 31(2):444–450.

Finkelstein, F. and Finkelstein, S. (2000). Depression in chronic dialysis patients: assessment and treatment. *Nephrology Dialysis Transplantation*, 15(12):1911–1913.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Fitzpatrick, R. (1991). Surveys of patients satisfaction: I-important general considerations. *BMJ*, 302(6781):887–889.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769.

Forina, M., A. C. L. S. and Tiscornia, E. (1983). *Classification of olive oils from their fatty acid composition*, pages 189–214. Food Research and Data Analysis. Applied Science Publishers, London.

Friedman, V. (2008). Data visualization and infographics. In *Graphics, Monday Inspiration, January 14th, 2008.*

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200.

Friendly, M. (2008). Milestones in the history of thematic cartography, statistical graphics, and data visualization. In *Seeing Science: Today*. American Association for the Advancement of Science.

Fritzke, B. (1995). Growing grid: a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters*, 2:9–13. 10.1007/BF02332159.

FundacionIndustrias (2009). Fundación de industrias de calzado español. anuario 2009. (spanish).

Garavaglia, S. (2000). Health care customer satisfaction survey analysis using self-organizing maps and ldquo;exponentially smeared rdquo; data vectors. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 4, pages 119–124 vol.4.

Globerson, A. and Roweis, S. (2006). Metric learning by collapsing classes. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 451–458. MIT Press, Cambridge, MA.

Goicoechea, M., Caramelo, C., Rodriguez, P., Verde, E., Gruss, E., Albalate, M., Ortiz, A., Casado, S., and Valderrábano, F. (2001). Role of type of vascular access in erythropoietin and intravenous iron requirements in haemodialysis. *Nephrology, Dialysis, Transplantation*, 16(11):2188–93.

Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. (2004). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press.

Gotch, F. A. and Sargent, J. A. (1985). A mechanistic analysis of the National Cooperative Dialysis Study (NCDS). *Kidney international*, 28(3):526–34.

Gotlin, R. and Huie, G. (2000). Anterior cruciate ligament injuries. operative and rehabilitative options. *Physical Medicine & Rehabilitation Clinics of North America*, 11(4):895–928.

Greenwood, R. N., Ronco, C., Gastaldon, F., Brendolan, A., Homel, P., Usvyat, L., Bruno, L., Carter, M., and Levin, N. W. (2003). Erythropoeitin dose variation in different facilities in different countries and its relationship to drug resistance. *Kidney international. Supplement*, 64(87):S78–86.

Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series In Data Management Systems. Elsevier Books, Oxford.

Harris, R. (1999). *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press.

Hartigan, J. and Kleiner, B. (1981). Mosaics for contingency tables. In *13th symposium on the Interface*, pages 268–273. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition.

Herschleb, J. N., Heethaar, R. M., de Tweel, I. V., Zimmerman, A. N. E., and Meijler, F. L. (1979). Signal analysis of ventricular fibrillation. In *IEEE Computers in Cardiology*, pages 49–54.

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika*, 51:11–26.

Holten, D. (2006). Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):741 –748.

Imran, A. and O'Connor, J. (1998). Control of knee stability after acl injury or repair: interaction between hamstrings contraction and tibial translation. *Clinical Biomechanics*, 13(3):153 – 162.

Inamdar, N., Kaplan, R. S., and Bower, M. (2002). Applying the Balanced Scorecard in healthcare provider organizations. *Journal of Healthcare Management*, 47(3):195–196.

Inselberg, A. and Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multidimensional geometry. In *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*, pages 361 –378.

Iosifescu, M. (2007). *Finite Markov Processes and Their Applications.* Dover books on mathematics. Dover Publications, Incorporated.

Jalife, J., Gray, R. A., Morley, G. E., and Davidenko, J. M. (1998). Evidence for electrical organization during ventricular fibrillation in the human heart. *Chaos*, 8(1):79–93.

Jekova, I. and Mitev, P. (2002). Detection of ventricular fibrillation and tachycardia from the surface ecg by a set of parameters acquired from four methods. *Physiological Measrement*, 23(4):629–634.

Jensen, R. and Shen, Q. (2008). *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches.* Wiley-IEEE Press.

Jeong, C.-S. and Pang, A. (1998). Reconfigurable disc trees for visualizing large hierarchical information space. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 19 –25, 149.

Kaplan, R. S. and Norton, D. P. (1992). The Balanced Scorecard – Measures that drive performance. *Harvard Business Review*, 70(1):71–79.

Kaplan, R. S. and Norton, D. P. (1996). *The Balanced Scorecard: Translating Strategy into action.* Harvard Business School Press, Boston, USA (MA).

Kaski, S. (1997). *Data Exploration Using Self-Organizing Maps.* PhD thesis. DTech Thesis, Helsinki University of Technology, Finland.

Keim, D. and Kriegel, H.-P. (1994). Visdb: database exploration using multidimensional visualization. *Computer Graphics and Applications, IEEE*, 14(5):40–49.

Keim, D. A. (2001). Visual Exploration of Large Data Sets. *Communications of the ACM (CACM)*, 44(8):38–44.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 08(1):1–8.

Kesselman, M., Krieger, J., and Joseph, W. (2009). *Introduction to Comparative Politics: Political Challenges and Changing Agendas.* Wadsworth.

Kimmel, P. L. (2000a). Just whose quality of life is it anyway? controversies and consistencies in measurements of quality of life. *Kidney International*, 57(S113–S120).

Kimmel, P. L. (2000b). Psychosocial factors in adult end-stage renal disease patients treated with hemodialysis: Correlates and outcomes. *American Journal of Kidney Diseases*, 35(4, Supplement):S132 – S140.

Kimmel, P. L., Peterson, R. A., Weihs, K. L., Simmens, S. J., Alleyne, S., Cruz, I., and Veis, J. (2000). Multiple measurements of depression predict mortality in a longitudinal study of chronic hemodialysis outpatients. *Kidney International*, 57(5):2093–2098.

Kimmel, P. L., Peterson, R. A., Weihs, K. L., Simmens, S. J., Alleyne, S., Cruz, I., and Veis, J. H. (1998). Psychosocial factors, behavioral compliance and survival in urban hemodialysis patients. *Kidney International*, 54:245–254.

Kirchgessner, J., Perera-Chang, M., Klinkner, G., Soley, I., Marcelli, D., Arkossy, O., Stopper, A., and Kimmel, P. L. (2006). Satisfaction with care in peritoneal dialysis patients. *Kidney International*, 70:1325–1331.

Kirk, A. (2012). *Data Visualization: A Successful Design Process*. Community experience distilled. Packt Publishing, Limited.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. In *Neural Networks, 1996., IEEE International Conference on*, volume 1, pages 294–299 vol.1.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69. 10.1007/BF00337288.

Kohonen, T. (1989). *Self-organization and associative memory: 3rd edition*. Springer-Verlag New York, Inc., New York, NY, USA.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer Berlin Heidelberg, 3rd edition.

Kohonen, T. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer series in information sciences. Springer-Verlag.

Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10):1358 –1384.

Kreuseler, M. and Schumann, H. (1999). Information visualization using a new focus+context technique in combination with dynamic clustering of information space. In *NPIVM '99: Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, pages 1–5, New York, NY, USA. ACM.

Lachenbruch, P. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23(4):639–645.

Lameire, N., Jager, K., Van Biesen, W., de Bacquer, D., and Vanholder, R. (2005). Chronic kidney disease: a European perspective. *Kidney international. Supplement*, 68(99):S30–8.

Lansiluoto, A. (2007). Suitability of self-organising maps for analysing a macro-environment – an empirical field survey. *International Journal of Business Information Systems*, 2(2):149–161.

Lee, J. and Verleysen, M. (2010). *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer.

Lee, J. H. and Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Expert System with Applications.*, 29(1):145–152.

Leisch, F. (2009). Neighborhood graphs, stripes and shadow plots for cluster visualization. *Statistics and Computing*.

Levkowitz, H. (1991). Color icons-merging color and texture perception for integrated visualization of multiple parameters. In *Visualization, 1991. Visualization '91, Proceedings., IEEE Conference on*, pages 164–170, 420.

Li, Y., Bisera, J., Weil, M., and Tang, W. (2012). An algorithm used for ventricular fibrillation detection without interrupting chest compression. *Biomedical Engineering, IEEE Transactions on*, 59(1):78 –86.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Lloyd, D. G., Buchanan, T. S., and Besier, T. F. (2005). Neuromuscular Biomechanical Modeling to Understand Knee Ligament Loading. *Medicine and Science in Sports and Exercise*, 37:1939–1947.

Lloyd-Williams, M. and Williams, T. (1996). A neural network approach to analyzing health care information. *Top Health Information Management*, 17(2):26–33.

Luttrell, S. (1989). Hierarchical self-organising networks. In *Artificial Neural Networks, 1989., First IEE International Conference on (Conf. Publ. No. 313)*, pages 2 –6.

M. E. Nygards and J. Hulting (1978). Recognition of ventricular fibrillation utilizing the power spectrum of the ecg. In *IEEE Computers in Cardiology*, pages 393–7.

Magdalena, R., Fernández, C., Martín, J. D., Soria, E., Martínez, M., Navarro, M. J., and Mata, C. (2009). Qualitative analysis of goat and sheep production data using self-organizing maps. *Expert Systems*, 26(2):191–201.

Makanju, A., Brooks, S., Zincir-Heywood, A., and Milios, E. (2008). Logview: Visualizing event log clusters. In *Privacy, Security and Trust, 2008. PST '08. Sixth Annual Conference on*, pages 99 –108.

Makinen, V.-P., Forsblom, C., Thorn, L. M., Waden, J., Gordin, D., Heikkila, O., Hietala, K., Kyllonen, L., Kyto, J., Rosengard-Barlund, M., Saraheimo, M., Tolonen, N., Parkkonen, M., Kaski, K., Ala-Korpela, M., and Groop, P.-H. (2008). Metabolic Phenotypes, Vascular Complications, and Premature Deaths in a Population of 4,197 Patients With Type 1 Diabetes. *Diabetes*, 57:2480–2487.

Marcelli, D., Kirchgessner, J., Amato, C., Steil, H., Mitteregger, A., Moscardò, V., Carioni, C., Orlandini, G., and Gatti, E. (2001). EuCliD (European Clinical Database): a database comparing different realities. *Journal of Nephroly.*, 14(Suppl 4):S94–100.

Marr, B. and Schiuma, G. (2003). Business performance measurement – past, present and future. *Management Decision*, 41(8):680–687.

Martinez, A. M. and Kak, A. (2001). PCA versus LDA. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):228–233.

Matsumoto, H. and Seedhom, B. (1994). Treatment of the pivot-shift intraarticular versus extraarticular or combined reconstruction procedures. a biomechanical study. *Clinical Orthopaedics and Related Research*, 299:298–304.

McConnell, P., Johnson, K., and Lin, S. (2002). Applications of Tree-Maps to hierarchical biological data. *Bioinformatics*, 18(9):1278.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Ménétrey, J., Duthon, V., Laumonier, T., and Fritschy, D. (2008). "biological failure" of the anterior cruciate ligament graft. *Knee Surgery, Sports Traumatology, Arthroscopy*, 16(3):224–231.

Miikkulainen, R. (1990). Script recognition with hierarchical feature maps. *Connection Science*, 2:83–101.

Miyamoto, S., Ichihashi, H., and Honda, K. (2008). *Algorithms for Fuzzy Clustering: Methods in C-Means Clustering with Applications*. Studies in Fuzziness and Soft Computing. U.S. Government Printing Office.

Moe, G. K., Abildskov, J. A., and Han, J. (1964). Factors responsible for the initiation and maintenance of ventricular fibrillation. In Surawicz, B. and Pellegrino, E., editors, *Sudden Cardiac Death*. New York: Grune and Stratton.

Montefiori, M. and Resta, M. (2008). A computational approach for the health care market. *Health Care Management Science*, 12(4):344–350.

Moustafa, R. E. and Hadi, A. S. (2009). Grand tour and the Andrews plot. *Wiley Interdisciplinary Reviews Computational Statistics*, 1(2):245–250.

Munzner, T. (1997). H3: laying out large directed graphs in 3d hyperbolic space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 2 –10.

Murray, A., Campbell, R. W. F., and Julian, D. G. (1985). Characteristics of the ventricular fibrillation waveform. In *IEEE Computers in Cardiology*, pages 275–8.

Nelson, D. W., Bellander, B.-M., MacCallum, R. M., Axelsson, J., Alm, M., Wallin, M., Weitzberg, E., and Rudehill, A. (2004). Cerebral microdialysis of patients with severe traumatic brain injury exhibits highly individualistic patterns as visualized by cluster analysis with self-organizing maps. *Critical Care Medicine*, 32(12):2428–2436.

Neurauter, A., Eftestol, T., Kramer-Johansen, J., Abella, B. S., Sunde, K., Wenzel, V., Lindner, K. H., Eilevstjonn, J., Myklebust, H., Steen, P. A., and Strohmenger, H.-U. (2007). Prediction of countershock success using single features from multiple ventricular fibrillation frequency bands and feature combinations using neural networks. *Resuscitation*, 73(2):253–263.

Nolle, F. M., Bowser, R. W., Badura, F. K., Catlett, J. M., Gudapati, R. R., Hee, T. T., Moos, A. N., and Sketch, M. H. S. (1989). Evalutation of frequency-domain algorithm to detect ventricular fibrillation in the surface electrocardiogram. In *IEEE Computers in Cardiology*, pages 337–40.

Oliveira, M. C. F. D. and Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 9:378–394.

Pardey, J. (2007). Detection of ventricular fibrillation by sequential hypothesis testing of binary sequences. In *IEEE Computers in Cardiology*, pages 573–6.

Patil, D. D., Wadhai, V., and Gokhale, J. (2010). Evaluation of decision tree pruning algorithms for complexity and classification accuracy. *International Journal of Computer Applications*, 11(2):23–30. Published By Foundation of Computer Science.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6):559–572.

Perrière, G. and Thioulouse, J. (2003). Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins. *Computer Methods and Programs in Biomedicine*, 70(2):99–105.

Pickett, R. and Grinstein, G. (1988). Iconographic displays for visualizing multidimensional data. In *Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on*, volume 1, pages 514–519.

Reingold, E. and Tilford, J. (1981). Tidier drawings of trees. *Software Engineering, IEEE Transactions on*, SE-7(2):223 – 228.

Rese, M. (2003). Relationship marketing and customer satisfaction: An information economics perspective. *Marketing Theory*, 3(1):97–117.

Resta, M. (2011). Assessing the Efficiency of Health Care Providers: A SOM Perspective. In Laaksonen, J. and Honkela, T., editors, *Advances in Self-Organizing Maps*, volume 6731 of *Lecture Notes in Computer Science*, pages 30–39. Springer.

Richard A. Becker, William S. Cleveland, M.-J. S. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Satatistics*, 5(2):123–155.

Robertson, G. G., Mackinlay, J. D., and Card, S. K. (1991). Cone trees: animated 3d visualizations of hierarchical information. In *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 189–194, New York, NY, USA. ACM.

Ronco, C. and Marcelli, D. (1999). Are better outcomes in hemodialysis in europe explained by superior esrd care? *Seminars in Dialysis*, 12(5):345–348.

Rosado, A., Serrano, A., Martínez, M., Soria, E., Calpe, J., and Bataller, M. (1999). Detailed study of time-frequency parameters for ventricular fibrillation detection. In *ESEM Europen Society for Engineering and Medicine*, pages 379–380.

Rosado-Muñoz, A., Camps-Valls, G., Guerrero-Martínez, J., Francés-Villora, J. V., Muñoz-Marí, J., and Serrano-López, A. J. (2002). Enhancing feature extraction for VF detection using data mining techniques. In *IEEE Computers in Cardiology*, volume 29, pages 209 – 212.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326.

Sammon Jr, J. W. (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Schieppati, A. and Remuzzi, G. (2005). Chronic renal diseases as a public health problem: epidemiology, social, and economic implications. *Kidney international. Supplement*, 68(98):S7–S10.

Seiford, L. M. (1996). Data envelopment analysis: the evolution of the state of the art (1978-1995). *Journal of Productivity Analysis*, 7(99-137).

Shneiderman, B. (1991). Tree visualization with tree-maps: A 2-d space-filling approach. *ACM Transactions on Graphics*, 11:92–99.

Simoff, S., Böhlen, M., and Mazeika, A. (2008). *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. LNCS sublibrary: Information systems and applications, incl. Internet/Web, and HCI. Springer.

Soria, E., Martín, J. D., Fernández, C., Magdalena, R., Serrano, A. J., and Moreno, Á. (2006). Qualitative modelling of time series using self-organizing maps: application to animal science. In *Proceedings of the 6th WSEAS international conference on Applied computer science*, ACS'06, pages 183–187, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Soukup, T. and Davidson, I. (2002). *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Spence, R. (2001). *Information visualization*. ACM Press Bks. Addison-Wesley.

Steil, H., Amato, C., Carioni, C., Kirchgessner, J., Marcelli, D., Mitteregger, A., Moscardo, V., Orlandini, G., and Gatti, E. (2004). EuCliD – a medical registry. *Methods of information in medicine*, 43(1):83–88.

Stevenson, K. B., Hannah, E. L., Lowder, C. a., Adcox, M. J., Davidson, R. L., Mallea, M. C., Narasimhan, N., and Wagnild, J. P. (2002). Epidemiology of hemodialysis vascular access infections from longitudinal infection surveillance data: predicting the impact of NKF-DOQI clinical practice guidelines for vascular access. *American journal of kidney diseases*, 39(3):549–555.

Stopper, A., Amato, C., Gioberge, S., Giordana, G., Marcelli, D., and Gatti, E. (2007). Managing complexity at dialysis service centers across Europe. *Blood purification*, 25(1):77–89.

Surhone, L., Timpledon, M., and Marseken, S. (2010). *Neighbourhood Components Analysis*. VDM Publishing.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Pearson international Edition. Pearson Addison Wesley.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000a). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000b). A global geometric framework for nonlinear dimensionality reduction. *Science*.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. Academic Press.

Thomas, J. J. and Tajudin, D. A. (2006). Visualizing the examination timetabling data using clustering method and treemaps. In *Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications*.

Tzanako, E. (2002). *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence*. Industrial Electronics. Taylor & Francis.

Übeyli, E. D. (2008). Usage of eigenvector methods in implementation of automated diagnostic systems for ecg beats. *Digital Signal Processing*, 18(1):33 – 48.

Udani, S., Lazich, I., and Bakris, G. L. (2011). Epidemiology of hypertensive kidney disease. *Nature reviews. Nephrology*, 7(1):11–21.

Vesanto, J., Alhoniemi, E., Himberg, J., Kiviluoto, K., and Parviainen, J. (1999). Self-Organizing Map for Data Mining in MATLAB: the SOM Toolbox. *Simulation News Europe*, (25):54.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (2000). SOM Toolbox for MATLAB 5. Technical report, Helsinki University of Technology.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210.

Ware, C. (2012). *Information Visualization: Perception for Design*. Interactive Technologies. Elsevier Science & Technology.

Wasse, H., Kutner, N., Zhang, R., and Huang, Y. (2007). Association of initial hemodialysis vascular access with patient-reported health status and quality of life. *Clinical journal of the American Society of Nephrology*, 2(4):708–714.

Wetcher-Hendricks, D. (2011). *Analyzing Quantitative Data: An Introduction for Social Researchers*. Wiley.

White, R., Asplin, B., Bugliosi, T., and Hankins, D. (1996). High discharge survival rate after out-of-hospital ventricular fibrillation with rapid defibrillation by police and paramedics. *Annals of Emergency Medicine.*, 28(5):480–5.

Wilkinson, L. (2006). Revising the pareto chart. *The American Statistician*, 60:332–334.

Yakaitis, R. W., Ewy, G. A., and Otto, C. W. (1980). Infuence of time and therapy on ventricular fibrillation in dogs. *Critical Care Medicine.*, 8(3):157–63.

Zelman, W. N., Pink, G. H., and Matthias, C. B. (2003). Use of the Balanced Scorecard in health care. *Journal of Health Care Finance*, 29(4):1–16.

Zhang, C., Zhao, J., Tian, J., Li, F., and Jia, H. (2011). Support vector machine for arrhythmia discrimination with TCI feature selection. In *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*, pages 111 –115.

Zhang, S., Zhang, C., and Wu, X. (2004). *Knowledge Discovery In Multiple Databases.* Advanced information and knowledge processing. Springer.

Zhang, X. S., Zhu, Y. S., Thakor, N. V., and Wang, Z. Z. (1999). Detecting ventricular tachycardia and fibrillation by complexity measure. *IEEE Transactions on Biomedical Engineering.*, 46(5):548–55.

Visual Data Mining: Real Applications and New Approaches.

José María Martínez Martínez, January 2014