ABSTRACT

Title of dissertation: DATA VISUALIZATION OF ASYMMETRIC DATA USING SAMMON MAPPING AND APPLICATIONS OF SELF-ORGANIZING MAPS

Haiyan Li, Doctor of Philosophy, 2005

Dissertation directed by: Professor Bruce L. Golden
Department of Decision and Information Technologies

Data visualization can be used to detect hidden structures and patterns in data sets that are found in data mining applications. However, although efficient data visualization algorithms to handle data sets with asymmetric proximities have been proposed, we develop an improved algorithm in this dissertation.

In the first part of the proposal, we develop a modified Sammon mapping approach that uses the upper triangular part and the lower triangular part of an asymmetric distance matrix simultaneously. Our proposed approach is applied to two asymmetric data sets: an American college selection data set, and a Canadian college selection data set which contains rank information. When compared to other approaches that are used in practice, our modified approach generates visual maps that have smaller distance errors and provide more reasonable representations of the data sets.

In data visualization, self-organizing maps (SOM) have been used to cluster points. In the second part of the proposal, we assess the performance of several software implementations of SOM-based methods. Viscovery SOMine is found to be helpful in determining the number of clusters and recovering the cluster structure of data sets. A genocide and politicide data set is analyzed using Viscovery SOMine, followed by another analysis on the public and private college data sets with the goal to find out schools with best values.

DATA VISUALIZATION OF ASYMMETRIC DATA USING SAMMON MAPPING AND APPLICATIONS OF SELF-ORGANIZING MAPS

by

Haiyan Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor Bruce L. Golden, Chair
Professor Edward A. Wasil
Assistant Professor Paul F. Zantek
Assistant Professor Steven Gabriel
Professor Ali Haghani

ACKNOWLEDGEMENTS

The work with this dissertation has been extensive and trying, but in the first place exciting, instructive, and fun. Without help, support, and encouragement from several persons, I would never have been able to finish this work.

First of all, I would like to thank my advisor, Dr. Bruce Golden, for his generous time and commitment. It is not often that one has an advisor that always contributes the time for listening to the little problems and roadblocks that unavoidably crop up in the course of performing research. Throughout my doctoral work he encouraged me to develop independent thinking and research skills. He continually stimulated my analytical thinking. His technical and editorial advices were essential to the completion of this dissertation proposal and have taught me innumerable lessons and insights on the workings of academic research in general.

I am also appreciative and thankful for the support and patience of Dr. Edward Wasil, who has read through my draft copies, listened to my complaining and encouraged me every step of the way. His insightful comments were crucial for editing the drafts into the dissertation proposal. His timeliness in returning the drafts did not go unnoticed.

I am indebted to Dr. Paul Zantek who has been a source of support throughout my graduate studies. His data mining class has been instrumental in shaping my understanding of clustering and data visualization. I am very grateful to Dr. Steven

Gabriel for the many valuable comments that helped shape this work. I would like to express my appreciation to Dr. Ali Haghani for gladly serving on my dissertation committee.

Thanks also to all my colleagues at the Department of Decision and Information Technologies for providing a good working atmosphere, especially Stacy Calo for assisting me with collecting survey and data entry during my late months of pregnancy.

At last, to my family thanks for supporting me with your love and understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this chapter, we provide background information on data visualization and describe the motivation and objectives of our research.

1.1 Motivation and the Problem of Interest

Data visualization is the process of "representing data as a visual image" (Latham, 1995) in which an image is created using a combination of points, lines, coordinate systems, numbers, symbols, words, shadings, and colors to represent different measured quantities (Tufte, 1983).

Data visualization is often used to make apparent any pattern in a data set that is large in size or dimensionality. For example, analyzing increasing amounts of data, such as customer data, to discover hidden patterns is a major problem facing businesses and organizations today. Visualization, together with other data mining techniques such as clustering and classification, can be employed to generate a data map that serves as a guide and provides the user with insights, i.e., detecting customer purchase patterns. The ability to show patterns attracts decision makers to use data visualization as a tool to get a

better understanding of the data set and then make better decisions. For example, consider the problem facing each high-school senior: selecting an undergraduate American university or college to pursue a bachelor's degree. Students can consult rankings in popular publications and reference books such as *The Fiske Guide*, which provides information on 300 universities and colleges. The information contained in these publications and books is not easy to assimilate. Condon et al. (2002) built a model of American universities that enables a student to visualize the data. Information that cannot be revealed in lists and tables, such as similarities between the universities, can be directly quantified according to the distances between universities on visual maps. These maps can assist students in identifying similar universities to consider.

As data visualization receives more and more attention, a variety of methods for visualization have been proposed including self-organizing maps (SOM), multi-dimensional scaling (MDS), and Sammon mapping (SM). These methods have been widely used in data visualization, as they are easy to implement and have modest computational requirements. Our proposed work is closely related to the Sammon mapping method, which we will describe in later chapters. SOM, MDS, and SM usually deal with the problems that have symmetric structures, for example, symmetric distance matrices. These methods take the table format of data sets as input, where rows are observations and columns are attributes. MDS and SM also accept a distance matrix as input but they require a symmetric distance matrix.

A proximity matrix or a similarity matrix contains the pairwise distances or similarities between all pairs of data items in a data set. In a proximity matrix, the distances are usually assumed to be symmetric. However, in practice, there are many interesting problems in which asymmetric proximities arise, especially in marketing or human behavior surveys. Asymmetric proximity is one type of proximity in which the pairwise distances are not symmetric. For example, in terms of teaching quality, the president of university *A* thinks that university *B* is the most competitive rival, while the president of university *B* thinks that the closest competitor is university *C* and not university *A*, and the president of university *C* thinks that university *A* is the closest competitor. The corresponding proximity matrix is shown in Table 1.1. If school *j* is a competitor of school *i*, then $d_{ij} = 1$, otherwise, $d_{ij} = 0$. Clearly, the matrix is asymmetric, i.e. entry $(i, j) \neq$ entry (j, i) for some i and j, $i \neq j$.

Since visualization methods normally work on symmetric cases, a question that needs to be answered is how to visualize asymmetric cases such as the matrix in Table 1.1. A visualization method that can handle asymmetry may need to be developed, or some modification may need to be made to an existing visualization method. A simple modification could be made by averaging off-diagonal entries to create a symmetric matrix. However, replacing asymmetric distances with an averaged distance alters the structure of the data set in a way that may result in a less accurate representation.

In order to maintain the asymmetric information of the data set, Merino and Munoz (2001) developed asymmetry coefficients. Mathematically, they defined an asymmetry coefficient of a data observation as a summation of the standardized similarities of the data observation to all of the other data observations (details such as

Competitor / School	A	B	C
A	--	1	0
B	0	--	1
C	1	0	--

Table 1.1 Asymmetric proximity matrix of schools A, B, and C.

transforming distances to similarities are described in the next chapter). Merino and Munoz incorporated asymmetry coefficients into the MDS and SM objective functions and corresponding search procedures. Their approach dealt with the symmetrized distance matrix of data observations. Based on their computational results, Merino and Munoz observed that data observations with large asymmetry coefficients were more influential in determining the structure of a map. However, data observations that are similar to many other data observations are usually dominant in forming the structure of a visual map, regardless of whether asymmetry coefficients are introduced into the MDS or SM methods. When most of the data observations have similar asymmetry coefficients, there is little or no impact the coefficients have on influencing the structure of a map.

In order to visualize asymmetric cases, we can examine the upper triangular part and the lower triangular part of the matrix simultaneously rather than study the symmetrized distance matrix derived by averaging corresponding upper and lower entries in the asymmetric distance matrix, or arithmetically calculate asymmetric coefficients. We expect that the visual maps generated by taking into account the upper triangular part and the lower triangular part of the matrix simultaneously are a better representation of the asymmetric data, and hopefully produce more insight into the data sets. For example,

we apply our approach to a small 3 × 3 asymmetric distance matrix given in Table 1.2. We use the GRG Solver, an optimization software developed by Frontline Systems (www.solver. com), to solve the small problem in our proposed approach and in one of the traditional approaches (i.e., the standard Sammon mapping approach which takes the average values of asymmetric parts as input) for comparison. Our approach seeks to optimize a Sammon mapping objective function directly on the whole asymmetric distance matrix. The common approach aims to solve the same optimization problem on the symmetrized distance matrix of the asymmetric data. The results show that our proposed approach is better than the one in the common approach (detailed discussions are given in Chapter 3) even in this tiny data set with only three data points. In addition, an asymmetric case with ranking information will be analyzed to extend our research on visualization of asymmetric problems.

Once an algorithm is implemented, it is necessary to check that if the algorithm works effectively as compared to other algorithms. Meanwhile, it is helpful to see if the algorithm is reliable on different data sets. We will apply our algorithm to two data sets: American college selection data and Canadian college data. The proximity matrix of each data set is asymmetric. The American college selection data are gathered from the *Fiske Guide* (1999). Canadian college data are collected from a survey that we conducted. Detailed information about each data set will be given in later chapters.

Another part of our study focuses on applications of data visualization. Clustering is such an application of data visualization that is widely used to detect hidden patterns

Asymm	A	B	C
A	0	1	3
B	2	0	2
C	3	4	0

Table 1.2 Example of a 3 × 3 asymmetric distance matrix.

that cannot be observed directly from enormous amounts of data. Several SOM-based software implementations for clustering are available either commercially or free of charge, such as Som_Pak (1997) and Viscovery (2002). However, there is little information indicating which implementation works better in practice. In our work, we will evaluate the performance of four software implementations of SOM-based clustering methods. Based on our evaluation, we will analyze several applications, such as clustering a state murder data set (Harff, 2003) and finding out colleges with best values (2003, 2004).

1.2 Summary of Objectives

A summary of the research work that has been done is asfollows.

Firstly, we developed a new visualization method that uses data with asymmetric proximities.

Second, we implemented our visualization method using a gradient descent algorithm.

Third, we applied our method to two data sets: American college selection and Canadian college selection and compared our results with the results generated by the standard method and the Merino's method.

Fourthly, we assessed four SOM-based clustering methods and analyzed clustering applications: state-sponsored data and college data with the best values.

1.3 Dissertation Organization

In Chapter 2, an overview of the literature on data visualization, including multidimensional scaling, Sammon mapping, and self-organizing maps, is provided. We define asymmetric distances and discuss several techniques that are used to handle an asymmetric distance matrix for visualization.

In Chapter 3, our modified Sammon mapping algorithm and its implementation are described and evaluated.

In Chapter 4, our modified Sammon mapping method is applied to American college selection data. We give the construction procedures of the data set. We discuss the results and compare our results to results generated by the commonly used standard SM method and Merino's method.

In Chapter 5, problems with ranking information are introduced. An example is given to show the process of constructing a distance matrix that incorporates ranking information. Canadian college selection data is analyzed by using our modified SM method. We discuss our preliminary results and provide insights gained from our work.

In Chapter 6, the performances of four SOM-based clustering software implementations are evaluated. We analyze an application (i.e., state-sponsored data) using Viscovery.

In Chapter 7, we apply Viscovery to public and private college data to find out colleges with best values. The visual maps of the public college data and the private college data are given and the results are discussed.

In Chapter 8, we summarize our research and point out our future work.

Chapter 2

Literature Review

Data visualization is one technique that can help researchers and business decision makers discover patterns in a data set. Data visualization has received lots of attention in the literature. Many methods (e.g., multidimensional scaling and Sammon mapping) and systems have been proposed and implemented in research and business areas like biomedical science, marketing, and financial services.

In this chapter, we provide an overview of the relevant literature in data visualization. First, we survey the papers that pertain to data visualization. Second, we examine papers on multidimensional scaling, Sammon mapping, and self-organizing maps. Third, we discuss several approaches dealing with asymmetric proximity data.

2.1 Data Visualization

There are many visualization methods that have been proposed for illustrating structures and multivariate relationships among data items. These methods can be classified into two categories: linear visualization methods and nonlinear visualization methods. In this section, we will discuss some of these methods.

Principle component analysis (PCA) (Hotelling, 1933) is a standard method in data analysis. Principle components are a set of variables that define a projection that encapsulates the maximum amount of variation in a dataset and is orthogonal (and therefore uncorrelated) to the previous principle component of the same (see Figure 2.1). Projection pursuit (Friedman, 1987) tries to show the best visual representation that reveals as much of the non-normally distributed structure of the data set as possible. A neural network implementation of this method is given by Fyfe and Baddeley (1995).

PCA cannot take into account nonlinear structures and projection pursuit cannot project the nonlinear structures onto a low-dimensional display if the data set has many dimensions and is highly nonlinear. Several approaches have been proposed to project nonlinear, high-dimensional structures onto a low-dimensional space. The most common methods allocate each individual data point onto a lower dimensional display and then optimize the display so that the distances between the points are as close as possible to the original distances. These methods differ in the selection of the objective function and the optimization approach.

Multidimensional scaling (MDS) refers to a group of methods that use proximities among data points to produce a representation of the data set (Kruskal, 1964a; Shepard, 1962). The representation consists of a geometric configuration of the points on a map

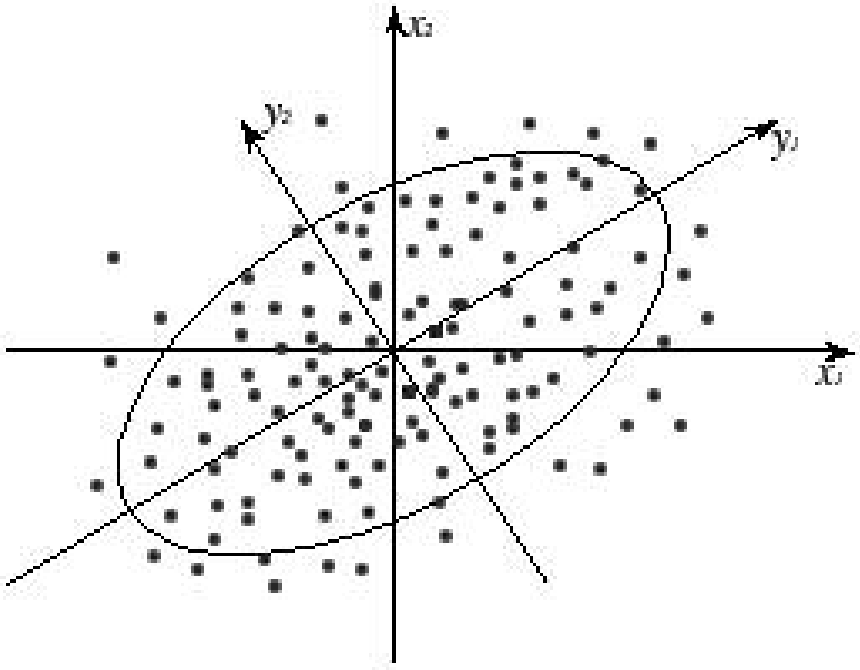

Figure 2.1 Example using principle components analysis.

where each point corresponds to one of the data items. MDS is widely used in behavioral, economic, and social sciences to analyze the pairwise proximities of the data points (e.g., similarity of brands in a market survey). MDS is discussed in greater detail in a later section.

Another nonlinear visualization method is Sammon mapping (Sammon, 1969). Sammon mapping (SM) is closely related to MDS. It tries to optimize an objective function in order to preserve the relative pairwise distance between data points. Details on MDS and Sammon mapping are provided in Sections 2.2 and 2.3.

Principal curves are a nonlinear generalization of PCA that projects a data set onto a nonlinear manifold after a linear manifold of the data set has been generated using PCA. Here, a manifold (or surface) refers to a topological space on which every point

has a neighborhood sharing some essential features of the data set (i.e., a sphere is a manifold). Principal curves were first defined as self-consistentsmooth curves (Hastie and Stuetzle, 1989) that pass through the middle of a *d*-dimensional probability distribution or data cloud (see Figure 2.2). Extensions of principal curves use multidimensional base functions to construct a nonlinear manifold, such as adaptive principal surfaces (LeBlanc and Tibshirani, 1994). Another popular approach is to use variants of the self-organizing map (SOM) algorithm (e.g., apply self-organizing maps to extract principal curves and extended principal curves from data (Der et al., 1998)). The self-organizing map is an efficient tool for the visualization of high-dimensional data sets (Vesanto, 1999).

Nonlinear visualization methods are computationally very intensive for large data sets. The triangulation method (Lee et al., 1977) can be used to reduce the computational complexity. An important property of the triangulation method is that the distances to its nearest two neighbors can be preserved exactly when inserting a new data item. Using the triangulation method, data items will be projected onto a map one by one and the nearest neighbor distances can always be preserved, that is, the generated map is based on a subset of distances in the original space. The projection process is, thus, substantially faster than nonlinear visualization methods. However, the generated map from the triangulation method may not be as accurate as the map from nonlinear visualization methods since the projection preserves only a part of distances.

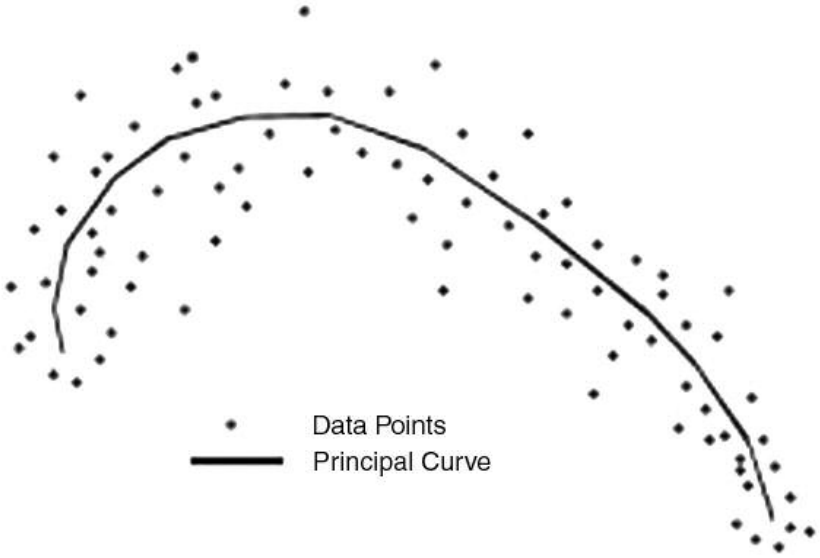

Figure 2.2 Illustration of a principal curve.

2.2 Multidimensional Scaling

2.2.1 Overview

As we mentioned in Section 2.1, multidimensional scaling is a collection of visualization methods that project proximity data onto lower dimensional space. In general, the goal of multidimensional scaling is to provide a visual representation of proximities in a set of investigated objects. Proximity data are always represented as distances. For example, in Table 2.1, each entry represents the pairwise distance between two buildings.

To better illustrate the idea of MDS, consider an example that visualizes the locations of 10 buildings. Given the symmetric matrix of distances among 10 buildings (see Table 2.1), MDS produces the map given in Figure 2.3.

	A	B	C	D	E	F	G	H	I	J
A	0	14.5	14	1	12.5	17.5	12	11.5	16.5	12
B	14.5	0	1	14.5	7	9.5	5.5	5	9.5	11.5
C	14	1	0	14	6	8.5	4.5	4	8.5	11.5
D	1	14.5	14	0	12.5	17.5	12	11.5	16.5	12
E	12.5	7	6	12.5	0	9.5	3	2	9.5	10.5
F	17.5	9.5	8.5	17.5	9.5	0	7.5	7.5	1.5	11
G	12	5.5	4.5	12	3	7.5	0	1	7.5	11
H	11.5	5	4	11.5	2	7.5	1	0	7.5	11
I	16.5	9.5	8.5	16.5	9.5	1.5	7.5	7.5	0	10.5
J	12	11.5	11.5	12	10.5	11	11	11	10.5	0

Table 2.1 Matrix of distances among 10 buildings.

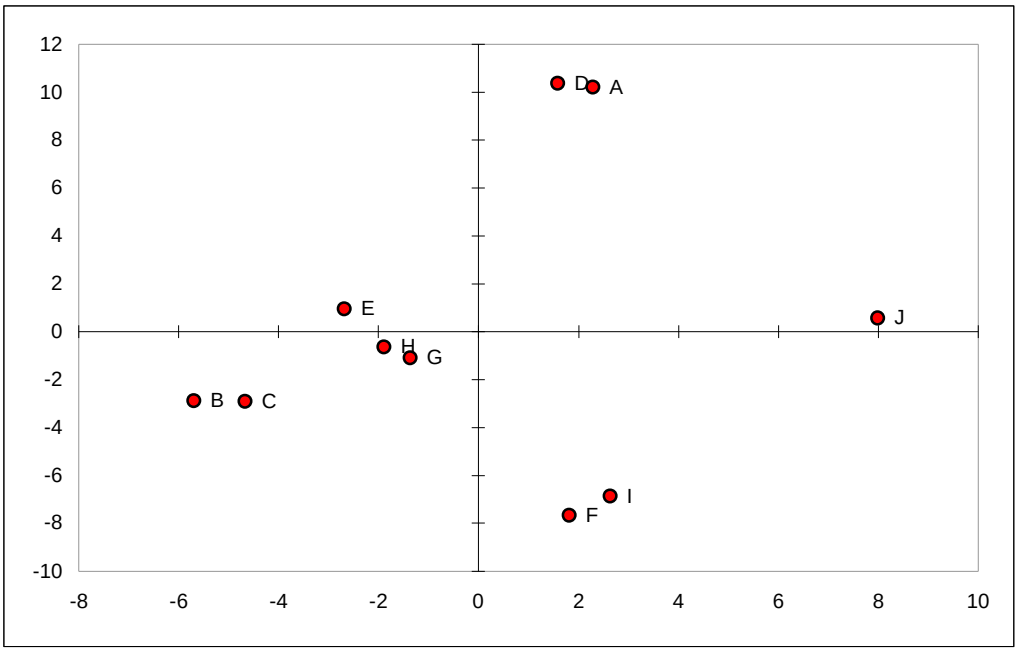

Figure 2.3 Example of multidimensional scaling.

In this example, the relationship between the original distances among data points and resulting distances shown on the map is positive, that is, the smaller the original distance, the closer the resulting distance between points, and vice versa. If the original proximity data had been represented as similarities, the relationship would have been negative which means the smaller the similarity between two data items, the farther apart in the map they would be. In our study, proximity data are represented as distances among data items.

2.2.2 Problem definition and stress function

From a slightly more technical point of view, for a set of observed distances between every pair of N items, multidimensional scaling methods aim to find a visual representation of the items in lower dimensional space such that the resulting distances among items match the original distances as closely as possible. The metric version of MDS aims to find configurations for data items where the resulting distances are as close as possible to the original distances of data items. Nonmetric MDS tries to keep the rank orders of the distances among data items as close as possible to the original rank orders. We consider only nonmetric MDS in this dissertation.

The mathematical definition of MDS now follows. Given N items and a corresponding distance matrix where entry d_{ij} is the pairwise distance between data items i and j, MDS seeks to find vector configurations $x_i^* = [x_{ik}^* : k = 1,..., p]$ and $x_j^* = [x_{jk}^* : k = 1,..., p]$ for data items i and j ($i, j = 1,\ldots,N, i \neq j$) in a p-dimensional space ($p \leq N - 1$), such that the Euclidian distance between x_i^* and x_j^*

$$d_{ij}^{*} = \left\| x_i^{*} - x_j^{*} \right\| = \sqrt{\sum_{k=1}^{p} (x_{ik}^{*} - x_{jk}^{*})^2} \;, \quad \forall_{i<j} d_{ij}^{*} \approx d_{ij} \;.$$

approximates the corresponding distance d_{ij}, for all pairs of data items *i* and *j*.

Since the proximity matrix is assumed to be symmetric, it is sufficient to take into account each pair of data items *i* and *j* just once. However, it may not be possible to perfectly represent the original distances (Johnson and Wichern, 1998) in a given lower dimensional space. Therefore, a numerical measure is needed to indicate the closeness and the measure is called a stress function.

Kruskal's stress (Kruskal, 1964a), known as Stress formula 1 or Stress 1 for short, measures the extent to which a representation deviates from a perfect match and is defined as

$$\frac{\sum_{i<j} (d_{ij} - d_{ij}^{*})^2}{\sum_{i<j} d_{ij}^2} \;.$$

If the stress is zero, then the resulting pairwise distances are exactly the same as the pairwise distances in the original distance matrix. However, in order to be useful, it is not necessary to require a zero stress value as long as a certain amount of distortion is tolerable. Kruskal (1964a) provides guidelines for the amount of stress to tolerate (see Table 2.2).

Multidimensional scaling has been applied in many areas--the literature is vast and dispersed over many periodicals and books. We will not attempt to give an overview of the developments to date, and refer the reader elsewhere for details (Borg and Groenen, 1997; Cox and Cox, 1994).

Stress	Goodness of fit
0.2	Poor
0.1	Fair
0.05	Good
0.025	Excellent
0	Perfect

Table 2.2 Stress guidelines suggested by Kruskal (1964a).

2.3 Sammon Mapping

2.3.1 Overview

Frequently, a Sammon map is used for data exploration. A practical area of Sammon mapping is in the visualization of protein structures based on measures of similarity between molecules. For example, Sammon maps have been used to analyze protein sequence relationships (Agrafiotis, 1997).

Like other MDS visualization methods, SM deals with proximity data. We are given distances between every pair of data items (we possibly have no direct access to any high-dimensional data but we do have access to a measure of distances between every two data items). SM tries to reconstruct the original data based solely on the given distance matrix. For example, given the distances between two buildings, SM can be used to construct a map for the coordinates of the building themselves. A demonstration of SM is presented in Figure 2.4 using the distance matrix shown in Table 2.1.

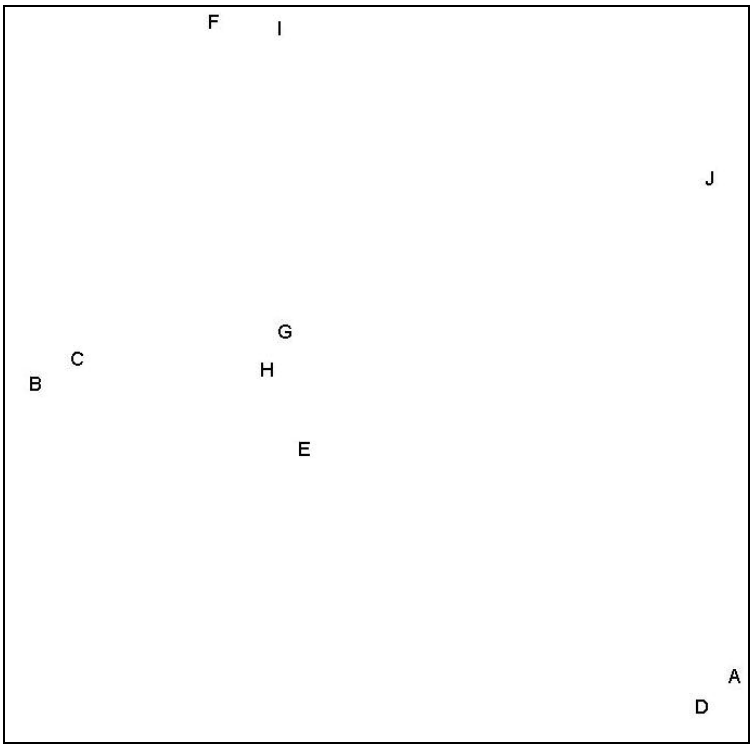

Figure 2.4 Sammon map of the 10 building.

2.3.2 SM algorithm

SM is an unsupervised nonlinear method that tries to preserve relative distances (Lerner et al., 1998). Here "unsupervised" means no targeted information or outcome to predict. To preserve the inherent structure, the algorithm that generates a Sammon map employs a nonlinear transformation of the observed distances among data items when mapping data items from a high-dimensional space onto a low-dimensional space.

If we denote the distance (usually Euclidean distance) between two data items *i* and *j*, $i \neq j$, in the original dimensional space by d_{ij} and the distance in the required projected space by $d_{ij}^{'}$, the error function of SM is given by

$$E = \frac{1}{\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}} \sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{(d_{ij} - d_{ij}^{'})^2}{d_{ij}} \quad .$$

In the error function *E*, the smaller the error value, the better the map we get. However, in practice, we are often unlikely to obtain perfect maps especially when the data set is large and in high-dimensional space. Therefore, approximate preservation is the likely result when we project high-dimensional data onto a low-dimensional plane.

The error function of SM is similar to Kruskal's S stress function (see Section 2.2.2) except that each squared difference of distances is divided by the corresponding input distance rather than the squared input distance. In other words, the only difference between SM and MDS is that the errors in distance preservation are normalized with the distance in the original space. Because of the normalization, SM places greater emphasis on smaller distances rather than on larger distances; this is different from Kruskal's MDS which treats small and large distances roughly the same.

2.3.3 Implementation

SM can also be viewed as an optimization problem and its error function can be minimized using several available techniques. Sammon solved the minimization problem using steepest gradient descent that is also referred to as pseudo-Newton minimization (Becker and Le Cun, 1989). This optimization procedure can be achieved by iteratively using the following rule

$$x_{ik}^{'}(t+1) = x_{ik}^{'}(t) - \alpha \frac{\frac{\partial E(t)}{\partial x_{ik}^{'}(t)}}{\left| \frac{\partial^2 E(t)}{\partial x_{ik}^{'}(t)^2} \right|} \quad .$$

Note that x'_{ik} is the k^{th} coordinate of the position of data item *i* in the required projected space, and α is called "magic factor" which controls the step size of updating coordinates.

We point out that Sammon (1969) suggested a value of 0.3 to 0.4 for α. However, since α is experimentally determined, the suggested value may not be optimal for all problems. Therefore, multiple experiments are necessary in order to find an appropriate value of α to minimize *E*.

Because a second derivative is used in the denominator, the update rule can be unstable at some points where the second derivative is very small. To avoid the instability, an alternative minimization technique, called normal gradient descent, has been used, where

$$x'_{ik}(t+1) = x'_{ik}(t) - \alpha \frac{\partial E(t)}{\partial x'_{ik}(t)} \quad .$$

When employing the gradient descent procedure to search for a minimum error value, a local minimum in the error surface could be obtained. Therefore, several experiments with different random initializations may be necessary. Another problem is the computational requirement of SM is $O(n^2)$. The pairwise distances and the derivatives have to be computed each iteration. Therefore, as the number of data items increases, the computational time increases dramatically. To lower the computational overhead, the Hamming metric has been used as a distance measure instead of the Euclidean metric (White, 1972). This showed some improvement in the computational efficiency, but the resulting maps could be distorted when the input space is the Euclidean space (Chien, 1978) and the interpretation becomes more complex.

So far, all of the problems that we have discussed are assumed to have symmetric proximity data, that is, the distance between data items i and j denoted by d_{ij} is exactly same as the distance between data items j and i denoted by d_{ji}. However, in practice, there are lots of interesting problems that have asymmetric proximity data, which means that the distance d_{ij} may be different from the distance d_{ji}.

To our knowledge, there are few papers that discuss the visualization of asymmetric proximity data sets. One of our major research goals is to develop a method that visualizes asymmetric proximity data sets. Our objective function and search procedure for implementing SM with asymmetric data are discussed in Chapter 3.

2.4 Self-organizing Maps

2.4.1 Overview

The self-organizing map (SOM) invented by Kohonen in the early 1980s is a type of neural network based on the idea of self-organized or unsupervised learning (Kohonen, 1995). This means that the algorithm has no targeted information or outcome to predict. Consequently, SOMs are ideal for clustering where no requirement of output fields is defined. However, SOM can also be employed to visualize high-dimensional data items (Flexer, 1999).

Being a stable and flexible method for clustering and visualization, SOM has been used for a wide range of purposes, ranging from controlling industry processes to analyzing gene data (Kaski et al., 2001). Many applications of SOM are given in the survey by Oja et al. (2003).

An illustration of the mechanism of SOM is shown in Figure 2.5. The 3×3 map consists of two layers: the input layer, and the output layer, which is often referred to as the output map. All the input neurons are connected to all the output neurons, and weights are assigned to each connection (not shown).

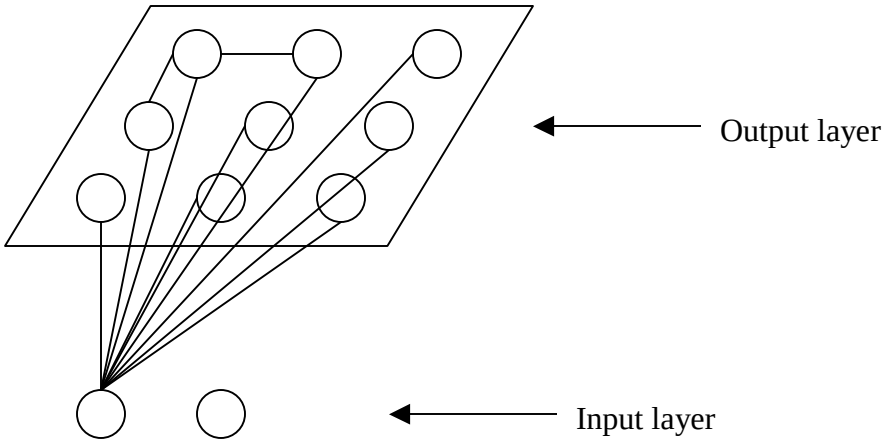

Figure 2.5 Illustration of a self-organizing map.

2.4.2 SOM algorithm

The self-organizing map trains by locating data items, one by one, to the output map. When an input data item is presented to an output neuron, its characteristics are compared with those of all output neurons, which are given initial weights. The neuron with the characteristics that are most similar to that of the input data item is chosen to represent the input data item; at the same time, the surrounding neurons of this chosen neuron are adjusted to be more similar to the chosen neuron in order to attract input data items similar to the mapped data item. This chosen neuron has a better chance, as compared to other neurons, of representing input data items that have similar patterns, and its neighbors are gradually able to represent similar input data items.

Each input data item is attracted to one and only one neuron, while each neuron can attract one or more data items. Each neuron i has a reference vector $m_i = [m_{i1}, \ldots, m_{in}]$, which is used to represent an input data item, where n is the dimensionality of the

input space. When locating an input data item *x* on the output map, the neuron *w* is the winner selected based on the minimum Euclidean distance, that is,

$$w = \min_i \{ \|x - m_i\| \}, \text{ where } \|.\| \text{ is Euclidean distance.}$$

During the training process, the winning neuron, also called the best matching unit, and its neighbors are allowed to modify their reference vectors as close to the current input data item as possible. The general form of the modification is given by

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

where $\alpha(t)$ is the learning rate that controls the training speed and $h_{wi}(t)$ is the neighborhood function centered on the winning neuron *w* (this function indicates the radius of neighborhood set) and $x(t)$ is the input at time *t*. Initially, the neighborhood function is set to a large value; this value decreases monotonically with time, as does the learning rate.

2.4.3 SOM software packages

In this section, we focus on two SOM software packages that are publicly available: SOM_Pak and Viscovery SOMine.

SOM_Pak was developed at the Helsinki University of Technology. It is a command line program and the interface is not user friendly; for example, there is no simple option for executing repetitive commands and the user has to type in each command. SOM_Pak can be downloaded for free; see the Web site at http://www.cis.hut.fi/research/som_pak/. A screenshot of SOM_Pak is shown in Figure 2.6.

Viscovery SOMine was developed by Eudaptics (www.eudaptics.com). An important advantage of SOMine is that it allows the user to visualize and analyze a data

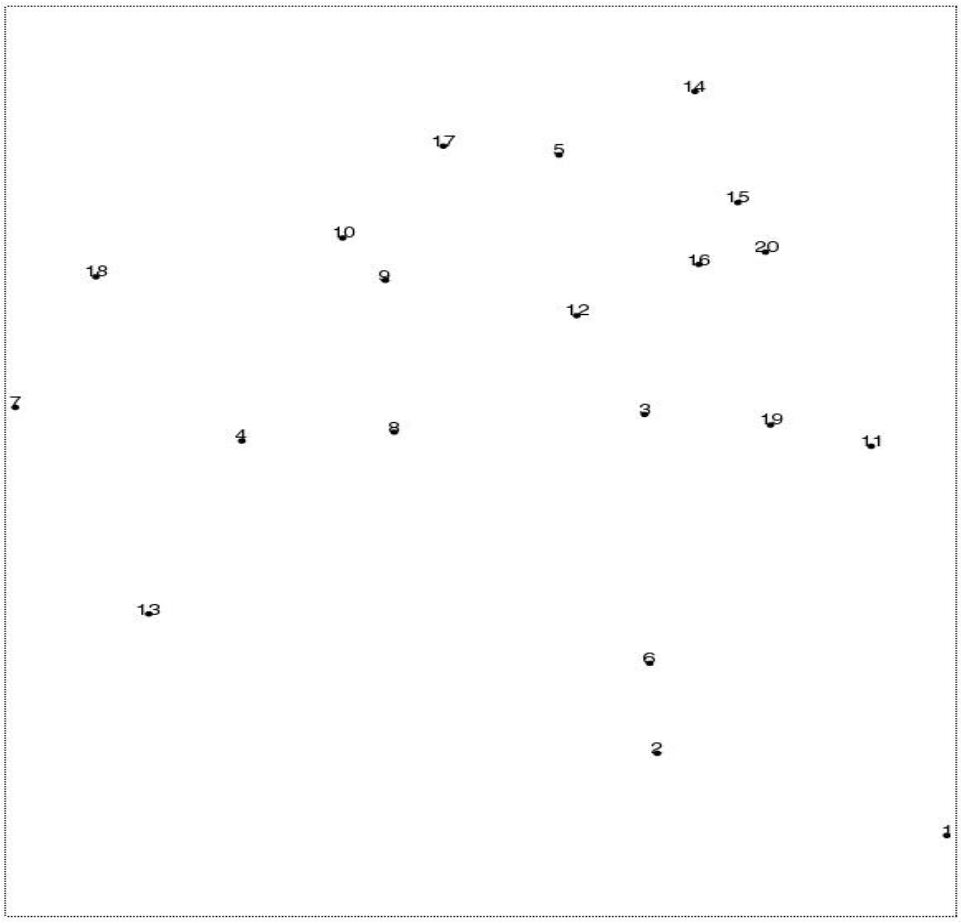

Figure 2.6 Sammon map of a distance matrix of 20 data items produced by SOM_Pak.

set without any prior statistical knowledge of the data set. The software provides suggestions as to which data items should be grouped together. The user can use or modify several parameters to control data processing. A screenshot of a Viscovery SOMine map is shown in Figure 2.7.

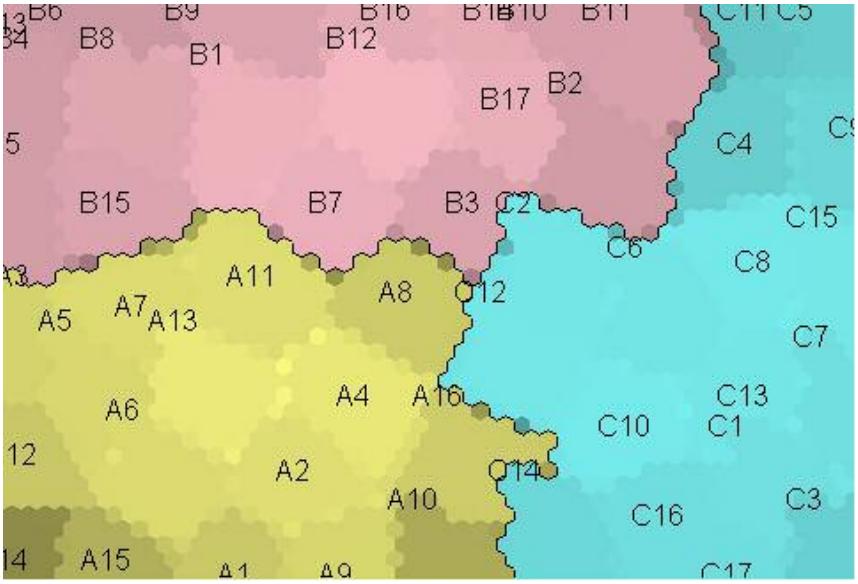

Figure 2.7 Screenshot of Viscovery. Three groups (As, Bs, and Cs) are shown in the map; the dark lines separate groups.

2.5 Asymmetric Proximity

2.5.1 Overview

Proximity refers to the similarity (sometimes also refers to dissimilarity or distance) between two data items. If the proximity between two data items is measured in Euclidean space (i.e., $d_{ij} = d_{ji} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$), then the proximity is symmetric. However, the proximity between data items or objects might not be symmetric. When objects are compared from different perspectives, for example, object *a* is said similar to object *b* in terms of color closeness, while object *b* is said dissimilar to object *a* because of their different shapes, the proximity relationship between these two objects is asymmetric, i.e., $d_{ab} \neq d_{ba}$.

Asymmetric proximity data arises in a number of diverse research areas such as marketing, psychology, sociology, and information retrieval. Asymmetric proximity data can usually be found in frequency matrices. For example, brand switching data has been utilized in marketing to examine the structure of competition within a particular product class. An example of car-switching data is given in Table 2.3. The rows represent cars last owned and the columns reveal cars currently owned. Out of all customers, 40 who last owned a Ford switch to a Honda. Other examples of asymmetric proximity data are migration rates between countries, frequencies of journal citations, word relationships in text documents, etc.

In the past, especially from the late 1970s to the early 1990s, analysis of asymmetry in proximity data had been one of the most provocative research topics in psychological research areas, in contrast to traditional MDS approaches. Researchers realized the psychological relevance of asymmetry in proximity data.

	BMW	Ford	Honda	Toyata	GM
BMW	180	40	20	0	10
Ford	20	343	40	30	70
Honda	10	20	120	10	20
Toyata	30	20	30	70	10
GM	10	20	0	20	250

Table 2.3 An example of car switching data.

Not surprisingly, many approaches for asymmetric proximity data were proposed in the psychological area. Tversky (1977) initially challenged the validity of the traditional spatial model (i.e., MDS model) regarding the observed violation of symmetry, minimality, and triangle inequality conditions of the metric axioms in actual data. Krumhansl (1978) also discussed the problems of traditional spatial model and proposed her distance-density model as an alternative to Tversky's (1977) feature-matching model. Other researchers proposed many different treatments of asymmetric proximity data. However, there has not been found any model that is not only mathematically sound but also practically applicable and easily interpretable.

2.5.2 Mathematical modeling for asymmetric proximity data

As many relationships are intrinsically asymmetric (Tversky, 1977) and increasing attention has been paid to asymmetric proximity data, a number of approaches have been developed to analyze asymmetry in proximity data. Most of these can be classified into three categories.

Since traditional approaches represent relationships between data items symmetrically, asymmetry is discarded as noise with respect to the symmetric part of the proximity data. The symmetric part is extracted and the data are symmetrized. One common approach to do this is to average corresponding off-diagonal entries (Kruskal, 1964b), i.e., substitute d_{ij} by d_{ij}', where $d_{ij}' = \frac{d_{ij} + d_{ji}}{2}$, and then apply the symmetric model, i.e., traditional MDS. For example, Tversky and Hutchinson (1986) analyzed 39 asymmetric proximity data by averaging. Another way of symmetrization is proposed by Levin and Brown (1979) who derived row multiplicative constants from two least square procedures to scale rows or columns of the asymmetric matrix to maximize symmetry. However, symmetrization of asymmetric proximity data may ignore some important information brought by asymmetry, and the symmetric solution found in the dimensional spaces does not depict anything about the asymmetry.

Approaches in the second category aim to capture the information of asymmetry in addition to the symmetric structure of the data. All major models in this category involve a symmetric component and an asymmetric component. Krumhansl (1978) specified a distance-density model in which object *A* and object *B* are represented in projected low dimensional space, the similarity between *A* and *B* can be interpreted not only by the interpoint distance but also the density of points in the surrounding configuration. In other words, asymmetries are accounted for through points around *A* and *B*. Saito (1986) developed an MDS approach in which estimated constants considered as density constants are added to the symmetric configuration in relation to the distance-density model. Different from the distance-density model, Constantine and Gower proposed an approach partitioning an asymmetric matrix into two matrices: *S*

(symmetric) and N (skew-symmetric; i.e., $n_{ij} = -n_{ji}$). A singular value decomposition of N was performed to obtain a least-squares fit to be plotted in low dimensional space (i.e., in two dimensions) and an interpretation of asymmetry was provided. Weeks and Bentler (1982) specified a model in which similarity is represented as distance, and traditional additive constants are combined to reflect asymmetries. Description of other models for asymmetric proximity data can be found in the paper of Zielman and Heiser (1996) in which most of mentioned models for asymmetric proximity can be decomposed into a symmetric part and a completely asymmetric part (i.e., skew-symmetry). These models are mathematically elegant and the resulting dimensional configurations are interpretable. However, these models assume an underlying symmetric component of the data but the assumption might not fit every case. Besides, software for computing the model parameters is not available (Zeilman and Heiser, 1996).

Another category of approaches for scaling asymmetric proximities is a graph-theoretic representation of asymmetric proximity data (Cunningham, 1978; Hutchingson, 1989; Klauer, 1989). In these models, asymmetries are represented as directed distances. These models do not require an underlying symmetric relationship. The differences between these graphic models are the representation type, for example, Cunningham (1978) employed directed trees as representations of proximities, and Hutchingson (1989) used networks to represent proximities data. One disadvantage of a graph representation is that the representation is limited to small data sets. Graph representation of large data sets seems messy because of a large number of arcs.

2.5.3 Other approaches of analyzing asymmetric proximity data

In addition to the above three categories of approaches of modeling asymmetric proximity data, there are other approaches sharing some features of the second and the third categories (Rodgers and Thompson, 1992; Merino and Munoz, 2001). Rodgers and Thompson proposed an approach in which asymmetric proximity data have been preprocessed using the idea of seriation before applying MDS to the data. Seriation is a procedure that orders data items on a continuum in order to maximize the sum of the elements above (or below) the main diagonal (Baker and Hubert, 1977). Rodgers and Thompson used the seriation algorithm and ordered the data items according to number of dominances over other data items. They defined dominance as: if $s_{ij} > s_{ji}$, then i dominates j. The data items that consistently dominate other data items are placed lower in the ordering, i.e., the data item that dominates all other data items is placed on the bottom row in the below diagonal triangular. MDS is fit to the ordered above diagonal or below diagonal triangular that explicitly exhibits the dominance relationships of data items and the resulting MDS configuration contains a directed distance.

Merino and Munoz (2001) introduced another approach to scaling asymmetry proximity data in which asymmetry coefficients derived from the skew-symmetry matrix are attached to MDS objective functions and MDS scaling was performed on the symmetrized proximity data. An asymmetry coefficient is defined as a summation of the standardized similarities of a data item to all other data items. Intuitively, data items that are similar to more data items will have larger asymmetry coefficients. In some sense, asymmetry coefficients convey the dominance information of data items, and data items with large asymmetry coefficients determine the structure of the configuration.

Empirically, it shows that the most dominant data items have a tendency to be placed in the center of a configuration, and this phenomenon may be interpreted as indirect confirmation of the usefulness of the Euclidean space, which recognized the central role of the dominant data items (Tversky and Hutchinson, 1986).

The approaches proposed by Rodgers & Thompson (1992), and Merino & Munoz (2001) are more of an exploratory data analysis and less specific asymmetric models than the methods described above. The approach by Rodgers & Thompson is flexible and tractable, i.e., the ordered triangular of data items, and substantive information is more accessible to data analysts. Merino and Munoz's approach suggests that it might be worth incorporating asymmetry coefficients to improve the visual configuration (Merino and Munoz, 2001). However, both approaches have some disadvantages. Since Rodgers and Thompson only considered the maximized triangular part (i.e., below the diagonal), some important information contained in the other triangular part would be lost, especially if the other triangular part contained many non-zero values. In Merino and Munoz's approach, the definition of asymmetry coefficient might not be suitable in every case, for example, if most of the data items have approximately equal asymmetry coefficients, i.e., defined as sum of row (or column) similarities, then the defined coefficients may not reveal dominant information. In this case, it might be better to define the coefficient as some function of dominance to reveal the centrality status. Details on centrality are in Tversky and Hutchinson's paper (1986).

From a visualization perspective, the visual configurations generated from the upper triangular and lower triangular parts of an asymmetric proximity matrix are implicitly different. For example, there are two SM maps generated by using the upper

triangular part (see Figure 2.8) and the lower triangular part (see Figure 2.9), respectively, of the asymmetric distance matrix given in Table 2.4. It is not convincing that maps generated from the upper triangular part are better than maps generated from the lower part, or vice versa. In addition, if there are no substantive reasons to assume that the underlying relationships between data items are symmetric, a natural way to visualize asymmetric proximity data is to simultaneously take into account the asymmetric parts, i.e., the upper triangular part and the lower triangular part of the proximity matrix. To assess the quality of maps, we can quantify the rank preservation by comparing the generated map results with the order relationships of the original asymmetric dataset. Chapter 3 discusses our proposed modified SM method and measures for quality assessment.

	B1	A2	A3	B4	A5	A6	A7	A8	A9	C10
B1	0	9	8	1	5	15	4	3	13	4
A2	20	0	1	20	6	10	4	4	8	3
A3	20	1	0	20	5	9	3	3	7	3
B4	1	9	8	0	5	15	4	3	13	4
A5	20	8	7	20	0	14	3	2	12	1
A6	20	9	8	20	5	0	2	3	1	2
A7	20	7	6	20	3	13	0	1	11	2
A8	20	6	5	20	2	12	1	0	10	2
A9	20	11	10	20	7	2	4	5	0	1
C10	20	20	20	20	20	20	20	20	20	0

Table 2.4 Asymmetric distance matrix taken from the American college selection data set.

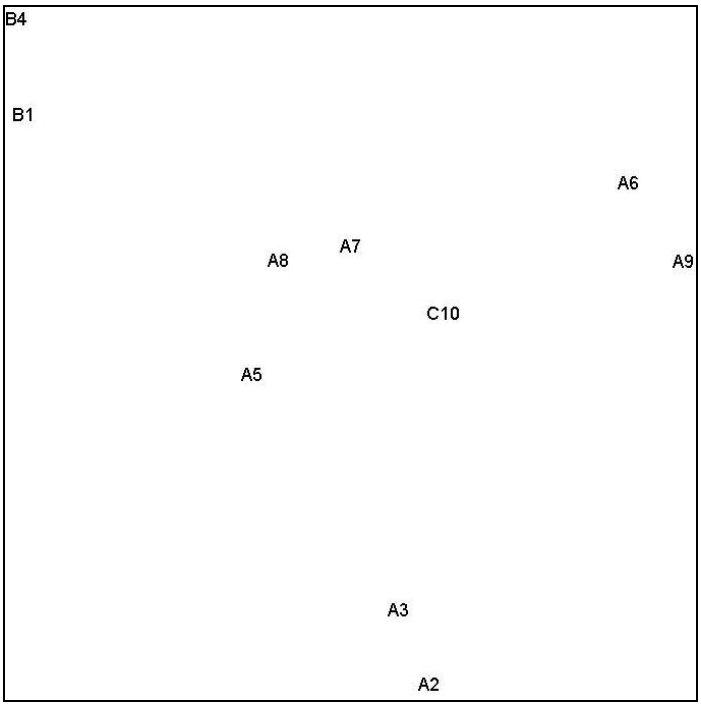

Figure 2.8 Sammon map of the upper triangular matrix.

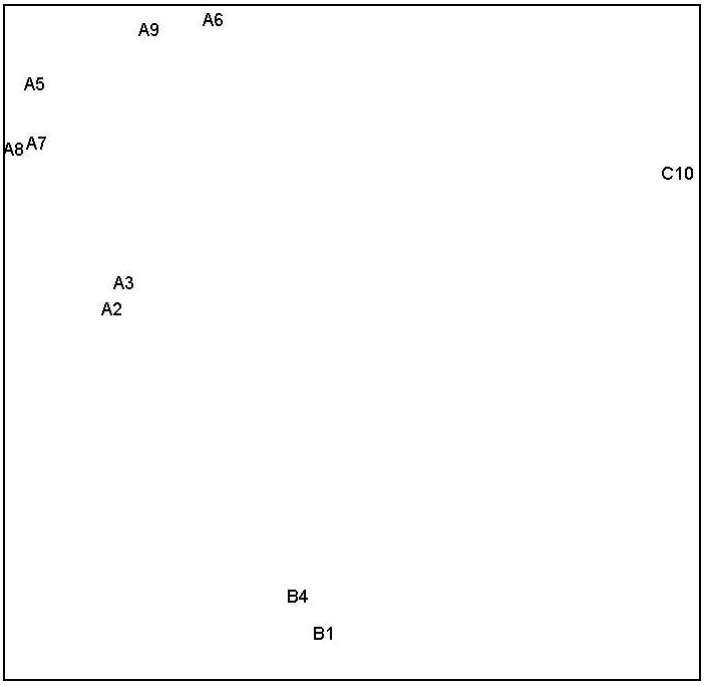

Figure 2.9 Sammon map of the lower triangular matrix.

Chapter 3

Constructing Sammon Maps from Asymmetric Data

Sammon maps are one of the most widely used tools in visualization and clustering. Sammon mapping (SM) projects high-dimensional data onto a 2-dimensional output map. A Sammon map is usually created for proximity data with a symmetric distance matrix. However, there are many applications (e.g., American college selection data) that have asymmetric distance matrices. In this chapter, we discuss the use of SM to visualize asymmetric proximity data sets. We describe the objective function, the associated update rule, and our implementation of SM for proximity data with an asymmetric distance matrix.

3.1 Modification of Sammon Mapping Method

SM tries to minimize the following objective function

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{(d_{ij} - d'_{ij})^2}{d_{ij}} \quad , \qquad (1)$$

where d_{ij} denotes the input distance (usually Euclidean distance) between data items *i* and *j*, $i \neq j$, $i, j = 1, ..., n$, in the original space. d'_{ij} denotes the output distance between *i* and *j* in the mapped D-dimensional space and $d'_{ij} = \sqrt{\sum_{k=1}^{D} [x_{ik} - x_{jk}]^2}$, where x_{ik} and x_{jk} are decision variables, $i = 1,\ldots, n-1$, $j = i+1,\ldots, n$, $k = 1,\ldots, D$. Note that only half of all entries in the distance matrix are taken into account because the distance matrix is assumed to be symmetric.

To minimize the objective function *E*, Sammon used the steepest gradient descent procedure to search for a minimum value of *E*. For convenience, the updating rule for his procedure is shown again as

$$x_{ik}(t+1) = x_{ik}(t) - \alpha \frac{\dfrac{\partial E(t)}{\partial x_{ik}(t)}}{\left|\dfrac{\partial^2 E(t)}{\partial x_{ik}(t)^2}\right|} \quad , \qquad (2)$$

where x_{ik} is the k^{th} coordinate of the position of *i* in the mapped space and α is the "magic factor" (Sammon, 1969). The magic factor is a parameter (Apostol and Szpankowski, 1999) that controls the step size for configuration update. Its value is determined experimentally. It is treated as a constant over all iterations.

The objective function (1) and the associated search procedure work well when the distance matrix is symmetric. However, problems arise when the distance matrix is

asymmetric. Since the objective function E assumes that the input distance matrix is symmetric, it is inappropriate to use an asymmetric distance matrix as the input. To overcome this problem, one simple technique is to symmetrize asymmetric distances by simply averaging, that is, replacing the entry d_{ij} with $(d_{ij} + d_{ji})/2$. Suppose that there are three data points *a*, *b*, *c* in an asymmetric distance matrix, and their pairwise distances are: in the upper triangular part, $d_{a,b} > d_{a,c}$, and in the lower triangular part, $d_{b,a} < d_{c,a}$. Therefore, it is uncertain that the distance between *a* and *b* is greater or less than the distance between a and c. Using symmetrized distances, the uncertainty of the order relationships can be resolved. However, Sammon maps generated from symmetrized distances will lose the asymmetry information (Merino and Munoz, 2001).

To generate maps that better represent and help visualize asymmetric proximity data sets, it is natural to consider the original asymmetric distance matrix instead of the symmetrized distance matrix in the objective function. We would like to account for the upper triangular portion and the lower triangular portion of the asymmetric distance matrix simultaneously in the optimization process.

The objective function that we propose has two parts denoted by U and L. The first part (U) takes into account the upper triangular part, while the second part (L) deals with the lower triangular part of the original asymmetric distance matrix. Our proposed objective function, denoted by E_1, is defined by

$$E_1 = \frac{1}{C}(U + L) \quad , \tag{3}$$

where

$$U = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(d_{ij} - d'_{ij})^2}{d_{ij}} \quad ,$$

$$L = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{(d_{ij} - d_{ij}^{'})^2}{d_{ij}} \quad ,$$

$$d_{ij}^{'} = \sqrt{\sum_{k=1}^{D} [x_{ik} - x_{jk}]^2}$$

and $$C = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} d_{ij} \quad ,$$

constrainted by $x_{ik}, x_{jk} \neq 0,\ i, j = 1,\ldots,n,\ k = 1,\ldots,D.$

x_{ik}, x_{jk} are decision variables and represent k^{th} coordinates of data items i and j in the mapped D-dimensional space.

By using U and L, we seek to obtain a configuration of data items such that the structures in the upper triangular part and the lower triangular part can be considered separately. We will use the steepest gradient method as the search procedure.

Let $E_1(t)$ denote the error value at t^{th} iteration and $U(t)$ and $L(t)$ denote the error values of the upper triangular part and the lower triangular part, respectively. Let $d_{ij}^{'}(t)$ denote the distance between i and j at the t^{th} iteration, that is,

$$d_{ij}^{'}(t) = \sqrt{\sum_{k=1}^{D} [x_{ik}(t) - x_{jk}(t)]^2}$$

and D denotes the dimensionality of the mapped space (usually D is 2). The new q^{th} coordinate of data item p at iteration t+1is given by

$$x_{pq}(t+1) = x_{pq}(t) - (MF) \frac{\partial E_1(t) / \partial x_{pq}(t)}{\left| \partial^2 E_1(t) / \partial x_{pq}(t)^2 \right|} \quad ,$$

where MF stands for “magic factor”, and the first derivative is

$$\frac{\partial E_1(t)}{\partial x_{pq}(t)} = \frac{1}{C} \frac{\partial U(t)}{\partial x_{pq}(t)} + \frac{1}{C} \frac{\partial L(t)}{\partial x_{pq}(t)} \qquad (4)$$

$$= \quad \frac{-2}{C}\sum_{j=1,\, j\neq p}^{n}\frac{(d_{pj}-d'_{pj})}{d_{pj}}\frac{(x_{pq}-x_{jq})}{d'_{pj}} \; +$$

$$\frac{-2}{C}\sum_{j=1,\, j\neq p}^{n}\frac{(d_{jp}-d'_{pj})}{d_{jp}}\frac{(x_{pq}-x_{jq})}{d'_{pj}} \quad .$$

The second derivative is

$$\frac{\partial^2 E_1(t)}{\partial x_{pq}(t)^2} = \frac{1}{C}\frac{\partial^2 U(t)}{\partial x_{pq}(t)^2} + \frac{1}{C}\frac{\partial^2 L(t)}{\partial x_{pq}(t)^2} \tag{5}$$

$$= \quad \frac{-2}{C}\sum_{j=1,\, j\neq p}^{n}\frac{1}{d_{pj}d'_{pj}}\left[\left(d_{pj}-d'_{pj}\right)-\frac{\left(x_{pq}-x_{jq}\right)^2}{d'_{pj}}\left(1+\frac{d_{pj}-d'_{pj}}{d'_{pj}}\right)\right] \; +$$

$$\frac{-2}{C}\sum_{j=1,\, j\neq p}^{n}\frac{1}{d_{jp}d'_{pj}}\left[\left(d_{jp}-d'_{pj}\right)-\frac{\left(x_{pq}-x_{jq}\right)^2}{d'_{pj}}\left(1+\frac{d_{jp}-d'_{pj}}{d'_{pj}}\right)\right] \quad .$$

Note that in the update rule, no two points are allowed to be identical. This prevents the partial derivatives from "blowing up."

The minimization problem in (3) is a nonlinear optimization problem and is non-convex (Klock and Buhmann, 1999). Therefore, we cannot guarantee finding the global minimum. The best we can do is to obtain a local minimum from each starting solution. We use the GRG software (2004) to solve the small asymmetric distance matrix given in Table 3.1 with our proposed objective function (3). Two sets of random starting configurations are used as initial coordinates for these three data points. Listed in Tables 3.2 and 3.3 are resulting configurations of three data points corresponding to two different sets of random starting configurations. The resulting objective function values are given in Tables 3.2 and 3.3. Meanwhile, we use GRG to solve the same asymmetric problem in the common approach that takes symmetrized distances as inputs. In Tables 3.4 and 3.5, we show the results of configurations and objective function values obtained

in the common approach. Since it is an optimization problem seeking a minimum value for the SM function, it is better to get smaller resulting objective function values. The shown

Asymmetric	A	B	C
A	0	1	3
B	2	0	2
C	3	4	0

Table 3.1 Asymmetric distance matrix of three data points.

Point	x_1-coordinate	x_2-coordinate
A	0.270706143	0.159258481
B	0.812893328	1.377376338
C	3.26577432	0.331200098
Objective Function Value	0.066666667	

Table 3.2 Results in our proposed approach with random starting point 1.

Point	x_1-coordinate	x_2-coordinate
A	3.481685023	4.95005622
B	2.987622292	6.188468288
C	5.448716802	7.215177319
Objective Function Value	0.066666667	

Table 3.3 Results in our proposed approach with random starting point 2.

Point	x_1-coordinate	x_2-coordinate
A	0.331987222	0
B	0.289836968	1.499407704
C	3.214502644	0.831327182
Objective Function Value	0.075000006	

Table 3.4 Results in the common approach with random starting point 1.

Point	x_1-coordinate	x_2-coordinate
A	3.505923937	4.963893963
B	2.008518266	5.052079549
C	2.927993608	7.907699703
Objective Function Value	0.075000012	

Table 3.5 Results in the common approach with random starting point 2.

objective function values generated by our approach are smaller than those generated by the common approach. This confirms our conjecture that our proposed approach performs better than the common approach in visualizing asymmetric problems, at least from the perspective of optimization.

3.2 Implementation of the Modified SM Method

In this section, we discuss the implementation procedures and provide small examples to illustrate them. We discuss problems that we encountered when implementing the modified Sammon mapping method. As illustrated in the previous section, GRG can be used to solve asymmetric problems. However, as the size of the asymmetric problem increases, it becomes burdensome to formulate the proposed objective function and the computational time increases significantly. A good alternative to GRG for solving asymmetric problems is C/C++. C/C++ is a commonly used software for coding. It is machine portable and requires only a small amount of changes to run on other computers. It is very fast, almost as fast as assembler. It allows structured

programming and is very flexible. It is suited to large and complex problems. Therefore, we implement the proposed modified SM method in C/C++.

We use random values as a starting configuration and employ the steepest gradient method as an optimization procedure to generate maps for an asymmetric distance matrix. If there is no improvement in the value of the objective function after a certain number of iterations, then the algorithm is considered converged and the resulting configuration is obtained to visualize the asymmetric distance matrix. Since the magic factor is experimentally determined, multiple experiments are necessary to find an appropriate value. We determined from the sensitivity analysis that the recommended magic factor of 0.4 is a good choice in our study for updating the configuration. The corresponding sensitivity analysis is discussed in the next chapter.

We illustrate our procedure with a small example. The data set denoted by 30A is an asymmetric distance matrix for 30 schools taken from the American college selection data set of 100 schools that was constructed using information provided in *The Fiske Guide* (Condon et al., 2002). This data set contains pairwise distances between each pair of 30 American colleges (see Table 3.6). For example, the distance between A5 and A19 is not symmetric, that is, $d_{A5,\ A19} = 11$ and $d_{A19,\ A5} = 6$. In this data set, our resulting Sammon map provides us with a visualization of the asymmetric data set (see Figure 3.1). In Table 3.7, we give the symmetrized distances of this data set, for example, the entry of $d_{A5,\ A19}$ equals the entry of $d_{A19,\ A5}$, which is the average value of these two entries in the asymmetric distance matrix. Figure 3.2 is a Sammon map generated by the standard Sammon map with the symmetrized distances. It is obvious that there are some differences between these two Sammon maps. For example, the distances between pairs

	A5	A7	A8	A19	A20	A24	A26	A27	A29	A31	A34	A41	A45	A50	A53	A56	A58	A61	A62	A67	A73	A76	A77	A78	A86	A89	A91	A93	A94	A98
A5	0	3	2	11	11	15	6	4	9	7	7	8	14	8	3	5	11	11	1	4	10	7	11	9	8	12	5	6	8	6
A7	3	0	1	10	8	12	5	3	6	4	4	5	11	5	2	2	8	8	2	1	7	4	10	6	5	9	2	3	7	3
A8	2	1	0	9	9	13	4	2	7	5	5	6	12	6	1	3	9	9	1	2	8	5	9	7	6	10	3	4	6	4
A19	6	5	4	0	11	15	8	6	9	3	7	4	14	8	5	2	7	11	5	4	10	3	13	9	4	12	5	2	10	6
A20	6	5	4	13	0	15	8	6	9	3	7	4	14	8	5	2	7	11	5	4	10	3	13	9	4	12	5	2	10	6
A24	4	1	2	11	4	0	6	4	3	5	5	6	2	6	3	3	9	4	3	2	3	5	11	7	6	2	3	4	8	4
A26	4	2	2	5	10	9	0	3	8	4	6	2	11	7	2	2	5	10	3	3	9	3	5	8	5	11	4	3	2	5
A27	3	1	1	10	9	13	5	0	7	5	5	6	12	6	2	3	9	9	2	2	8	5	10	7	6	10	3	4	7	4
A29	3	4	3	12	3	6	7	5	0	4	2	7	5	3	4	5	10	2	2	5	2	6	12	4	4	3	1	5	9	2
A31	7	5	5	14	8	12	9	7	6	0	4	5	11	5	6	3	8	8	6	5	7	4	14	6	1	9	2	1	11	3
A34	5	3	4	13	4	8	8	6	2	4	0	8	7	1	5	5	11	4	4	4	3	7	13	2	3	5	1	5	10	2
A41	4	3	2	11	10	14	6	4	8	4	6	0	13	7	3	2	3	10	3	2	9	1	11	8	5	11	4	3	8	5
A45	5	2	3	12	4	1	7	5	3	6	5	7	0	6	4	4	10	4	4	3	3	6	12	7	7	2	4	5	9	5
A50	5	2	3	12	6	10	7	5	4	3	2	7	9	0	4	4	10	6	4	3	5	6	12	1	2	7	1	4	9	1
A53	3	1	1	8	9	12	3	1	7	5	5	5	12	6	0	3	8	9	2	2	8	5	8	7	6	10	3	4	5	4
A56	6	5	4	13	11	15	8	6	9	3	7	4	14	8	5	0	7	11	5	4	10	3	13	9	4	12	5	2	10	6
A58	4	3	2	11	10	14	6	4	8	4	6	2	13	7	3	2	0	10	3	2	9	3	11	8	5	11	4	3	8	5
A61	3	2	3	12	3	4	7	5	1	5	3	7	3	4	4	4	10	0	2	3	2	6	12	5	5	1	2	5	9	3
A62	1	2	1	10	10	14	5	3	8	6	6	7	13	7	2	4	10	10	0	3	9	6	10	8	7	11	4	5	7	5
A67	4	1	2	11	8	12	6	4	6	3	4	4	11	5	3	1	7	8	3	0	7	3	11	6	4	9	2	2	8	3
A73	4	5	4	13	1	7	8	6	2	4	4	5	6	5	5	3	8	3	3	5	0	4	13	6	5	4	3	3	10	4
A76	5	2	3	12	9	13	7	5	7	4	5	1	12	6	4	1	4	9	4	1	8	0	12	7	5	10	3	3	9	4
A77	8	5	6	2	11	15	10	8	9	5	7	6	14	8	7	4	9	11	7	6	10	5	0	9	6	12	5	4	12	6
A78	4	1	2	11	7	11	6	4	5	2	3	6	10	4	3	3	9	7	3	2	6	5	11	0	1	8	1	3	8	1
A86	7	4	5	14	7	11	9	7	5	1	3	6	10	4	6	4	9	7	6	5	6	5	14	5	0	8	1	2	11	2
A89	4	1	2	11	3	3	6	4	2	5	4	6	2	5	3	3	9	3	3	2	2	5	11	6	5	0	2	4	8	3
A91	6	3	4	13	6	10	8	6	4	3	2	8	9	3	5	5	11	6	5	4	5	7	13	4	3	7	0	4	10	1
A93	6	5	4	13	9	13	8	6	7	1	5	4	12	6	5	2	7	9	5	4	8	3	13	7	2	10	3	0	10	4
A94	6	4	4	3	11	7	2	5	10	6	8	4	9	9	4	4	7	11	5	5	10	5	3	10	7	9	6	5	0	7
A98	5	2	3	12	6	10	7	5	4	4	2	7	9	3	4	4	10	6	4	3	5	6	12	4	4	7	1	5	9	0

Table 3.6 Asymmetric distance matrix of 30 American colleges.

	A5	A7	A8	A19	A20	A24	A26	A27	A29	A31	A34	A41	A45	A50	A53	A56	A58	A61	A62	A67	A73	A76	A77	A78	A86	A89	A91	A93	A94	A98
A5	0	3	2	8.5	8.5	9.5	5	3.5	6	7	6	6	9.5	6.5	3	5.5	7.5	7	1	4	7	6	9.5	6.5	7.5	8	5.5	6	7	5.5
A7	3	0	1	7.5	6.5	6.5	3.5	2	5	4.5	3.5	4	6.5	3.5	1.5	3.5	5.5	5	2	1	6	3	7.5	3.5	4.5	5	2.5	4	5.5	2.5
A8	2	1	0	6.5	6.5	7.5	3	1.5	5	5	4.5	4	7.5	4.5	1	3.5	5.5	6	1	2	6	4	7.5	4.5	5.5	6	3.5	4	5	3.5
A19	8.5	7.5	6.5	0	12	13	6.5	8	11	8.5	10	7.5	13	10	6.5	7.5	9	12	7.5	7.5	12	6	7.5	10	9	12	9	7.5	6.5	9
A20	8.5	6.5	6.5	12	0	9.5	9	7.5	6	5.5	5.5	7	9	7	7	6.5	8.5	7	7.5	6	5.5	6	12	8	5.5	7.5	5.5	5.5	11	6
A24	9.5	6.5	7.5	13	9.5	0	7.5	8.5	4.5	8.5	6.5	10	1.5	8	7.5	9	12	4	8.5	7	5	9	13	9	8.5	2.5	6.5	8.5	7.5	7
A26	5	3.5	3	6.5	9	7.5	0	4	7.5	6.5	7	4	9	7	2.5	5	5.5	8.5	4	4.5	8.5	5	7.5	7	7	8.5	6	5.5	2	6
A27	3.5	2	1.5	8	7.5	8.5	4	0	6	6	5.5	5	8.5	5.5	1.5	4.5	6.5	7	2.5	3	7	5	9	5.5	6.5	7	4.5	5	6	4.5
A29	6	5	5	11	6	4.5	7.5	6	0	5	2	7.5	4	3.5	5.5	7	9	1.5	5	5.5	2	6.5	11	4.5	4.5	2.5	2.5	6	9.5	3
A31	7	4.5	5	8.5	5.5	8.5	6.5	6	5	0	4	4.5	8.5	5.5	5.5	3	6	6.5	6	4	5.5	4	9.5	4	1	7	2.5	1	8.5	3.5
A34	6	3.5	4.5	10	5.5	6.5	7	5.5	2	4	0	7	6	1.5	5	6	8.5	3.5	5	4	3.5	6	10	2.5	3	4.5	1.5	5	9	2
A41	6	4	4	7.5	7	10	4	5	7.5	4.5	7	0	10	7	4	3	2.5	8.5	5	3	7	1	8.5	7	5.5	8.5	6	3.5	6	6
A45	9.5	6.5	7.5	13	9	1.5	9	8.5	4	8.5	6	10	0	7.5	8	9	12	3.5	8.5	7	4.5	9	13	8.5	8.5	2	6.5	8.5	9	7
A50	6.5	3.5	4.5	10	7	8	7	5.5	3.5	4	1.5	7	7.5	0	5	6	8.5	5	5.5	4	5	6	10	2.5	3	6	2	5	9	2
A53	3	1.5	1	6.5	7	7.5	2.5	1.5	5.5	5.5	5	4	8	5	0	4	5.5	6.5	2	2.5	6.5	4.5	7.5	5	6	6.5	4	4.5	4.5	4
A56	5.5	3.5	3.5	7.5	6.5	9	5	4.5	7	3	6	3	9	6	4	0	4.5	7.5	4.5	2.5	6.5	2	8.5	6	4	7.5	5	2	7	5
A58	7.5	5.5	5.5	9	8.5	12	5.5	6.5	9	6	8.5	2.5	12	8.5	5.5	4.5	0	10	6.5	4.5	8.5	3.5	10	8.5	7	10	7.5	5	7.5	7.5
A61	7	5	6	12	7	4	8.5	7	1.5	6.5	3.5	8.5	3.5	5	6.5	7.5	10	0	6	3	6	5	12	6	6	2	4	7	10	4.5
A62	1	2	1	7.5	7.5	8.5	4	2.5	5	6	5	5	8.5	5.5	2	4.5	6.5	6	0	3	6	5	8.5	5.5	6.5	7	4.5	5	6	4.5
A67	4	1	2	7.5	6	7	4.5	3	5.5	4	4	3	7	4	2.5	2.5	4.5	5.5	3	0	6	2	8.5	4	4.5	5.5	3	3	6.5	3
A73	7	6	6	12	5.5	5	8.5	7	2	5.5	3.5	7	4.5	5	6.5	6.5	8.5	2.5	6	6	0	6	12	6	5.5	3	4	5.5	10	4.5
A76	6	3	4	7.5	6	9	5	5	6.5	4	6	1	9	6	4.5	2	3.5	7.5	5	2	6	0	8.5	6	5	7.5	5	3	7	5
A77	9.5	7.5	7.5	7.5	12	13	7.5	9	11	9.5	10	8.5	13	10	7.5	8.5	10	12	8.5	8.5	12	8.5	0	10	10	12	9	8.5	7.5	9
A78	6.5	3.5	4.5	10	8	9	7	5.5	4.5	4	2.5	7	8.5	2.5	5	6	8.5	6	5.5	4	6	6	10	0	3	7	2.5	5	9	2.5
A86	7.5	4.5	5.5	9	5.5	8.5	7	6.5	4.5	1	3	5.5	8.5	3	6	4	7	6	6.5	4.5	5.5	5	10	3	0	6.5	2	2	9	3
A89	8	5	6	12	7.5	2.5	8.5	7	2.5	7	4.5	8.5	2	6	6.5	7.5	10	2	7	5.5	3	7.5	12	7	6.5	0	4.5	7	8.5	5
A91	5.5	2.5	3.5	9	5.5	6.5	6	4.5	2.5	2.5	1.5	6	6.5	2	4	5	7.5	4	4.5	3	4	5	9	2.5	2	4.5	0	3.5	8	1
A93	6	4	4	7.5	5.5	8.5	5.5	5	6	1	5	3.5	8.5	5	4.5	2	5	7	5	3	5.5	3	8.5	5	2	7	3.5	0	7.5	4.5
A94	7	5.5	5	6.5	11	7.5	2	6	9.5	8.5	9	6	9	9	4.5	7	7.5	10	6	6.5	10	7	7.5	9	9	8.5	8	7.5	0	8
A98	5.5	2.5	3.5	9	6	7	6	4.5	3	3.5	2	6	7	2	4	5	7.5	4.5	4.5	3	4.5	5	9	2.5	3	5	1	4.5	8	0

Table 3.7 Symmetrized distance matrix of 30 American colleges.

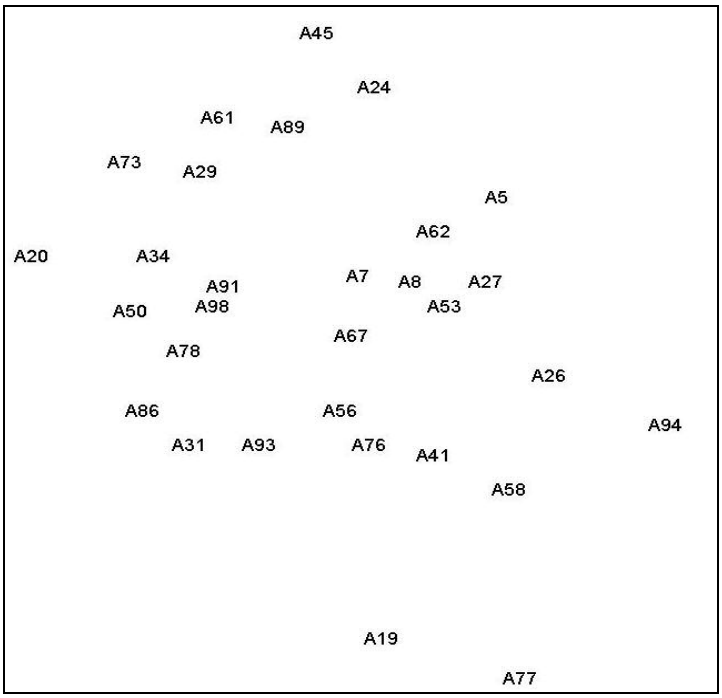

Figure 3.1 Sammon map of the asymmetric distance matrix for data set 30A generated by the modified SM method.

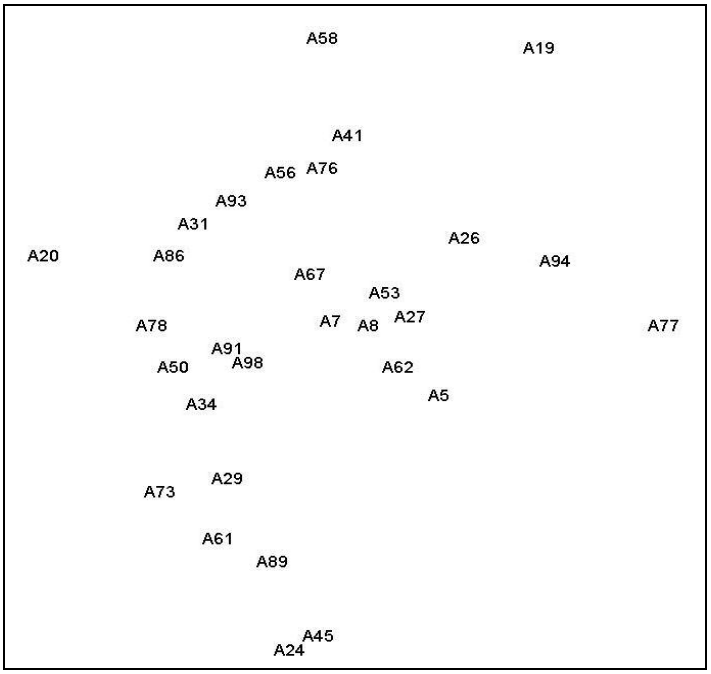

Figure 3.2 Sammon map of the symmetrized distance matrix for data set 30A generated by the standard SM method.

of the triplet A24, A45, and A89 are close to each other in the original asymmetric distance matrix, where this relationship ideally need be visualized as an equilateral triangular in the map. This relationship is better represented in the map given by the modified SM method than in the map given by the standard SM. In Chapter 4, we will compare Sammon maps generated by different SM methods and discuss the differences between them and give our tentative recommendation of choosing a better visualization map.

3.3 Performance Measurements of Sammon Maps

In order to assess the quality of maps produced from asymmetric distance data, we introduce two performance measures. The first is the distance error that measures the extent to which the pairwise distances projected on the map deviate from the original pairwise distances. The distance error function, denoted by *DE*, is defined by

$$DE = \sum_{i \neq j} \frac{(d_{ij} - d'_{ij})^2}{d_{ij}} \quad ,$$

where d_{ij} is the original distance and d'_{ij} is the projected distance. The smaller the value of *DE*, the better the map.

The second measure is an order preservation coefficient that indicates how well a map preserves the distance order relationships of the original asymmetric data. In Table 3.8, we give a small asymmetric distance matrix. The pairwise distance between two data items may be very different; for example, we see $d_{ab} = 1$ while $d_{ba} = 2$. Since that $d_{ab} < d_{ac}$ and $d_{ba} < d_{ca}$, the distance between *a* and *b* is less than the distance between *a* and *c*. If a Sammon map *A* preserves more order relationships for an asymmetric distance matrix than Sammon map *B*, then *A* is said to be more accurate than *B*. We are concerned about a map's ability to preserve order relationships. Our proposed order preservation measure, denoted by *OP*, is defined by,

$$OP = \frac{\text{number of relationships preserved}}{\text{number of relationships of the asymmetric data}} \quad .$$

Note that the order relationship can be uncertain. In Table 3.8, we see that $d_{ab} < d_{cd}$ and $d_{ba} > d_{dc}$. Therefore, it is incomparable if the distance between *a* and *b* is greater than the distance between *c* and *d* on the projected map.

	a	*b*	*c*	*d*
a	0	1	2	3
b	2	0	1	9
c	5	7	0	3
d	6	8	1	0

Table 3.8 Asymmetric distance matrix of four data points.

In Table 3.8, consider the four data items (a, b, c, d) and the twelve non-zero distance entries (six entries are in the upper triangular part of the matrix and six entries are in the lower triangular part of the matrix). The total number of relationships is 15, which is given by 6(6-1)/2. In Table 3.9, we provide all of the order relationships of the asymmetric distance matrix given in Table 3.8. If six order relationships are preserved in a Sammon map, the order preservation coefficient equals 40% (that is, 6/15).

If a Sammon map exhibits greater accuracy than another, then it should have a smaller distance error and a larger order preservation value. We use these two measures to assess the quality of Sammon maps in our applications.

$d_{ab} < d_{ac}$	and	$d_{ba} < d_{ca}$	Distance(a, b)	<	Distance(a, c)
$d_{ab} < d_{ad}$	and	$d_{ba} < d_{da}$	Distance(a, b)	<	Distance(a, d)
$d_{ab} = d_{bc}$	but	$d_{ba} < d_{cb}$	Distance(a, b)	≤	Distance(b, c)
$d_{ab} < d_{bd}$	and	$d_{ba} < d_{db}$	Distance(a, b)	<	Distance(b, d)
$d_{ab} < d_{cd}$	but	$d_{ba} > d_{dc}$	Distance(a, b)	?	Distance(c, d)
$d_{ac} < d_{ad}$	and	$d_{ca} < d_{da}$	Distance(a, c)	<	Distance(a, d)
$d_{ac} > d_{bc}$	but	$d_{ca} < d_{cb}$	Distance(a, c)	?	Distance(b, c)
$d_{ac} < d_{bd}$	and	$d_{ca} < d_{db}$	Distance(a, c)	<	Distance(b, d)
$d_{ac} < d_{cd}$	but	$d_{ca} > d_{dc}$	Distance(a, c)	?	Distance(c, d)
$d_{ad} > d_{bc}$	but	$d_{da} < d_{cb}$	Distance(a, d)	?	Distance(b, c)
$d_{ad} < d_{bd}$	and	$d_{da} < d_{db}$	Distance(a, d)	<	Distance(b, d)
$d_{ad} = d_{cd}$	but	$d_{da} > d_{dc}$	Distance(a, d)	≥	Distance(c, d)
$d_{bc} < d_{bd}$	and	$d_{cb} < d_{db}$	Distance(b, c)	<	Distance(b, d)
$d_{bc} < d_{cd}$	but	$d_{cb} > d_{dc}$	Distance(b, c)	?	Distance(c, d)
$d_{bd} > d_{cd}$	and	$d_{db} > d_{dc}$	Distance(b, d)	>	Distance(c, d)

Question mark (?) denotes that the relationship between the distances is incomparable.

Table 3.9 Order relationships.

Chapter 4

Visualizing American College Selection Data

4.1 Description of the Data Set

The Fiske Guide (Fiske, 1999) is a well-known publication that has been used for nearly 20 years to help students and parents select the right college. In the 2000 edition of *The Fiske Guide*, information on tuition cost, SAT scores, social life, and quality of life has been provided for over 300 colleges and universities in the United States. *The Fiske Guide* also lists a school's overlaps, that is, the major competitors to which applicants are also applying in greatest numbers. The overlaps provide students and parents with possible alternatives when selecting a school. For example, the overlaps of the University of Pennsylvania are Harvard, Princeton, Yale, Cornell, and Brown. Students who applied to the University of Pennsylvania also applied to those five schools. However, the overlaps of Harvard University -- Princeton, Yale, Stanford, MIT, and Brown -- do not include the University of Pennsylvania. The overlaps of two schools are not necessarily symmetric.

The American college selection data set is derived from the overlap data of 100 schools in *The Fiske Guide*. This data set was constructed by Condon et al. (2002). There were four steps involved in the construction process: building an adjacency matrix,

constructing a directed graph, computing distance measures, and modifying the distance matrix.

In the first step, Condon et al. created a 100×100 0-1 asymmetric adjacency matrix S for 100 schools, where entry $s_{ij} = 1$ (row i and column j) indicates that school j is an overlap of school i. For example, in Table 4.1, we show an 6×6 adjacency matrix for six universities (Brown, Cornell, Harvard, MIT, Penn, and Stanford). The entry in the Penn row and the Harvard column is 1, that is, Harvard is an overlap of Penn. In the second step, Condon et al. converted the adjacency matrix S to a directed graph with 100 nodes (one node for each school) and a directed arc for each non-zero s_{ij} entry, where i is the start node, j is the end node, and the directed arc connects i and j. In Figure 4.1, we converted the 6×6 adjacency matrix, which is shown in Table 4.1, into a directed graph with six nodes. We see that there is a directed arc starting at the Penn node and ending at the Harvard node.

In the third step, Condon et al. set the distance of an arc in the directed graph to one and computed the all-pairs shortest path distance matrix T, where each entry t_{ij} is calculated as the shortest distance from node i to node j. In the final step, the authors modified the distance matrix for disconnected nodes, that is, they set each entry t_{ij} for a disconnected pair to a value V that is greater than the longest distance in the matrix. It is necessary to carefully choose the value V -- a large value of V will push the connected points closer together so that it will be difficult to observe the inner relationships. A small value of V will result in merging disconnected points.

School	Brown	Cornell	Harvard	MIT	Penn	Stanford
Brown	0	1	1	0	0	1
Cornell	1	0	1	0	1	0
Harvard	1	0	0	1	0	1
MIT	0	1	1	0	0	1
Penn	1	1	1	0	0	0
Stanford	1	0	1	1	0	0

Table 4.1 Adjacency matrix for six schools.

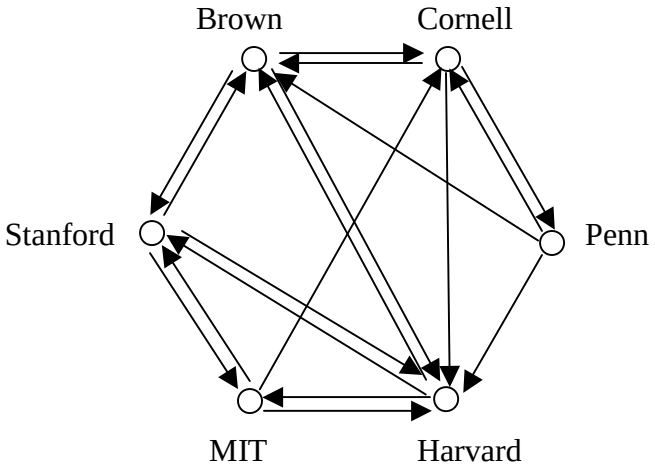

Figure 4.1 Directed graph generated from the adjacency matrix.

The American college selection data set generated by Condon et al. contains four groups of schools denoted by A, B, C, and D (see Table 4.2). There are 74 schools in A, 11 schools in B (these are schools from the southern United States), 8 schools in C (six schools are from the Ivy League), and 7 schools in D (all from California).

Key	School	State
A2	Arizona State University	AZ
A3	Arizona, University of	AZ
A5	Barnard College (Columbia University)	NY
A6	Bates College	ME
A7	Boston College	MA
A8	Boston University	MA
A9	Bowdoin College	ME
A11	Bryn Mawr College	PA
A12	Bucknell University	PA
A19	Carleton College	MN
A20	Carnegie Mellon University	PA
A23	Colby College	ME
A24	Colgate University	NY
A25	Colorado College	CO
A26	Colorado, University of—Boulder	CO
A27	Connecticut, University of	CT
A29	Delaware, University of	DE
A30	Denver, University of	CO
A31	Emory University	GA
A34	George Mason University	VA
A35	Georgetown University	DC
A38	Grinnell College	IA
A40	Illinois, University of—Urbana-Champaign	IL
A41	Indiana University	IN
A42	Iowa State University	IA
A43	Iowa, University of	IA
A44	James Madison University	VA
A45	Lafayette College	PA
A46	Lehigh University	PA
A47	Lewis and Clark College	OR
A48	Macalester College	MN
A49	Marquette University	WI
A50	Mary Washington College	VA
A51	Maryland, University of—College Park	MD
A53	Massachusetts, University of—Amherst	MA
A55	Michigan State University	MI
A56	Michigan, University of	MI
A57	Middlebury College	VT
A58	Minnesota, University of—Twin Cities	MN

Table 4.2 One hundred schools selected from *The Fiske Guide* for analysis.

A59	Mount Holyoke College	MA
A60	New Hampshire, University of	NH
A61	New Jersey, The College of	NJ
A62	New York University	NY
A63	North Carolina State University	NC
A64	North Carolina, University of—Chapel Hill	NC
A65	Northeastern University	MA
A66	Northwestern University	IL
A67	Notre Dame, University of	IN
A68	Oberlin College	OH
A69	Oregon State University	OR
A70	Oregon, University of	OR
A71	Pennsylvania State University	PA
A73	Pittsburgh, University of	PA
A75	Puget Sound, University of	WA
A76	Purdue University	IN
A77	Reed College	OR
A78	Richmond, University of	VA
A79	Rutgers University	NJ
A80	Smith College	MA
A85	Tufts University	MA
A86	Vanderbilt University	TN
A87	Vassar College	NY
A88	Vermont, University of	VT
A89	Villanova University	PA
A90	Virginia Polytechnic Institute and State University	VA
A91	Virginia, University of	VA
A92	Wake Forest University	NC
A93	Washington University in St. Louis	MO
A94	Washington, University of	WA
A95	Wellsley College	MA
A96	Whitman College	WA
A97	Willamette University	OR
A98	William and Mary, College of	VA
A99	Wisconsin, University of--Madison	WI
B1	Alabama, University of --Tuscaloosa	AL
B4	Auburn University	AL
B21	Charleston, College of	SC
B22	Clemson University	SC
B32	Florida State University	FL

Table 4.2 (Continued).

B33	Florida, University of	FL
B36	Georgia Institute of Technology	GA
B37	Georgia, University of	GA
B54	Miami, University of	FL
B81	South Carolina, University of	SC
B84	Tennessee, University of--Knoxville	TN
C10	Brown University	RI
C28	Cornell University	NY
C39	Harvard University	MA
C52	Massachusetts Institute of Technology	MA
C72	Pennsylvania, University of	PA
C74	Princeton University	NJ
C83	Stanford University	CA
C100	Yale University	CT
D13	California, University of--Berkeley	CA
D14	California, University of--Davis	CA
D15	California, University of--Irvine	CA
D16	California, University of--Los Angeles	CA
D17	California, University of--San Diego	CA
D18	California, University of--Santa Barbara	CA
D82	Southern California, University of	CA

Table 4.2 (Continued).

Each of the four groups is a strongly connected component in the directed graph of 300 schools given in *The Fiske Guide*. If a group is strongly connected, then there exists at least one directed path from any school in the group to any of the other schools in the same group. In other words, any one school is considered to be a competitor to all of the other schools in the group. The distance between each pair of schools measures the magnitude of the competitiveness. The shorter the distance, the more competitive the schools are. For example, if the distance between schools S and T is shorter than the distance between schools S and R, then T is more likely a competitor of S than R.

4.2 Experimental Design

We implement our modified Sammon mapping method in C/C++. Our code reads in an $n \times n$ asymmetric distance matrix and generates random starting coordinates for each data item i ($i = 1,\ldots,n$) to be plotted in the mapped D-dimensional (usually, two dimensional) space. We use the error function E_1 and the associated updating rule given in Section 3.1 to adjust the coordinates iteratively to minimize the value of the error function. If no improvement is found after a certain number of iterations, the modified SM method is considered converged and the obtained configuration is the resulting Sammon map. Currently, we employ the steepest gradient method as the optimization procedure when applying the modified SM method to the asymmetric distance matrix.

We realize that the resulting configuration is a local minimum, multiple experiments with different random starts are necessary to look for an approximate solution. Meanwhile, we tested several different magic factors ranging from 0.1 to 0.8 in our experiments. Tables 4.3 and 4.4 provide the sensitivity analysis of the magic factor. The values listed in the two tables are average values of five experiments with different random starts on each data set. The minimum average error measure(s) and the maximum average order preservation coefficient(s) of each data set can be found in boldface in Tables 4.3 and 4.4 respectively. For example, on the data set of 100 schools, the minimum error measure is 28829.5600 (see Table 4.3), which is associated with a magic factor of 0.6. The error measure associated with a magic factor of 0.6 is the second minimum (28829.5600). Their corresponding average order preservation coefficients are tied at the value of 0.4720, which is the maximum coefficient obtained

Magic Factor	100Schools	30Schools	ASchools	BSchools	CSchools	DSchools
0.1	28830.5200	773.6160	5688.3040	**18.3131**	6.1831	5.5503
0.2	28835.6200	773.3146	**5688.1640**	**18.3131**	**6.0645**	5.5502
0.3	28846.9600	773.9876	5688.6180	**18.3131**	6.0677	5.5490
0.4	28830.3200	773.0448	5688.3020	**18.3131**	**6.0645**	5.5490
0.5	28831.8400	**773.0162**	5706.6340	**18.3131**	6.0677	5.5490
0.6	**28829.5600**	773.9876	5709.3380	18.4980	6.1799	**5.5369**
0.7	28831.6200	773.9776	5815.3500	18.4980	6.0677	5.5490
0.8	28835.4400	773.0448	5760.8320	**18.3131**	**6.0645**	5.5490

Table 4.3 Average error measures obtained from the modified SM method.

Magic Factor	100Schools	30Schools	Aschools	Bschools	Cschools	Dschools
0.1	**0.4720**	**0.6065**	0.5943	0.7562	**0.5683**	**0.5552**
0.2	0.4718	0.6064	**0.5944**	0.7562	0.5677	0.5533
0.3	0.4714	**0.6065**	0.5943	0.7562	0.5661	0.5524
0.4	**0.4720**	**0.6065**	**0.5944**	0.7562	0.5677	0.5524
0.5	**0.4720**	**0.6065**	0.5939	0.7562	0.5661	0.5524
0.6	**0.4720**	**0.6065**	0.5940	0.7539	0.5661	0.5524
0.7	0.4719	**0.6065**	0.5916	0.7539	0.5661	0.5524
0.8	0.4718	**0.6065**	0.5937	**0.7565**	0.5677	0.5524

Table 4.4 Average order preservation coefficients obtained from the modified SM method.

for the 100 school data. When considering the overall performance on all six data sets in terms of both error measures and order preservation coefficients, each tested magic factor turns to give either the best error measure or the best order preservation coefficient. In other words, the experiments are not sensitive to the magic factors. Therefore, as suggested by Sammon (1969), we use 0.4 as the step size to adjust the locations of data items iteratively in the mapped space.

For each data set examined in the following sections, five experiments with different random starts are performed. We choose the best maps in terms of error measures and order preservation measures for comparison. Listed in the tables are average values of five experiments on each data set.

We point out that the Sammon maps generated with different starting configurations are usually similar, that is, the relative relationships among schools are roughly the same in each map. In our experiments, we keep the substitute value of infinity distance as what is used in Condon's work, which is slightly larger than the longest distance in the distance matrix.

We apply our modified Sammon mapping method to six data sets which are: American college selection data set with 100 schools, each of the four strongly connected groups (A, B, C, D), and 30 schools selected from A. We experiment with five different starting configurations for each data set and then compare the average performance of our modified SM method to that of the standard SM method and that of Merino's method. We use the typical resulting Sammon map of each data set to illustrate the similarities and differences of the resulting maps generated by these three methods.

In Figure 4.2, we show the map of all 100 colleges and universities that was generated by the modified SM method. In Figures 4.3 and 4.4, we show the Sammon maps of all 100 schools that were generated by the standard SM method and by Merino's method respectively with the standard error function (1) given in Section 3.1. In Figures 4.5 to 4.19, we show Sammon maps generated by our modified procedure, the standard procedure and Merino's procedure for five data sets (A, 30 schools from A, B, C, and D). We discuss the maps and results in the next section.

4.3 Discussion of the Results

Shown in Figures 4.2, 4.3, and 4.4 are Sammon maps of all 100 schools generated by the modified, the standard, and Merino's methods respectively. In Figure 4.2, the group of B schools and group of D schools are separated from the group of A schools and the group of C schools. This is consistent with the existing structure of the data set that consists of four strongly connected groups of schools. The most interesting phenomenon is that the group of C schools is located in the center of the map and is surrounded by some schools of group A, e.g., Tufts University (A85), New York University (A62), Boston College (A7), Columbia University (A5) and Georgetown University (A35). From our experimental results, it shows that the more chances a university is considered as a competitor by other universities, the more likely this university will be placed in the center or near the center of the map. Boston College (A7) and New York University (A62) have more chances to be considered competitors by many other universities so that they, we think they are popular, are placed in the center of the maps (see Figures 4.2, 4.3,

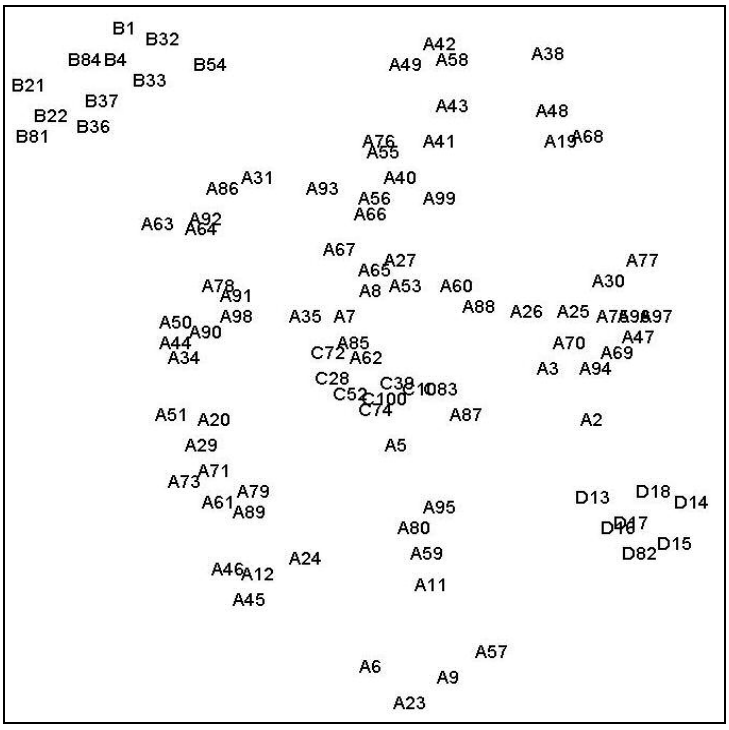

Figure 4.2 Map of 100 schools generated by the modified SM method.

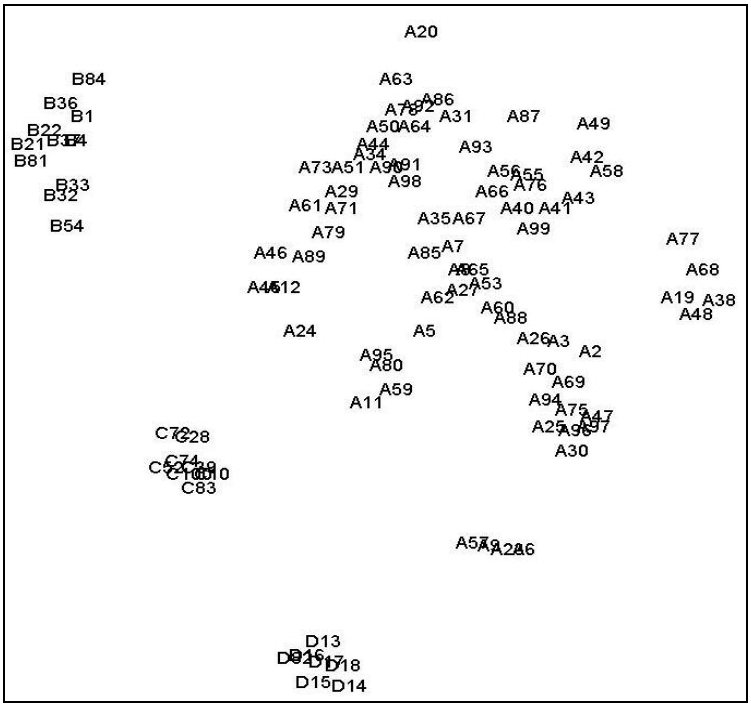

Figure 4.3 Map of 100 schools generated by the standard SM method.

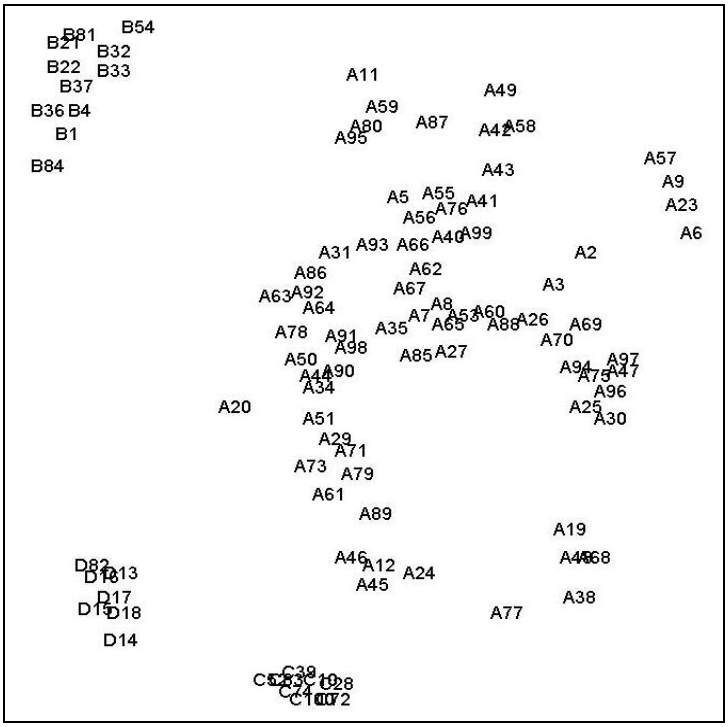

Figure 4.4 Map of 100 schools generated by Merino's method.

and 4.4). In Figure 4.2, the location of group C tells us that group C and some popular A schools have some in common to some extents. For example, group C consists of Ivy League schools which have high education quality and expensive tuition costs, etc., and some A schools that surround group C (i.e., New York University) provide qualified education the same as or no worse than group C schools provide. This relationship between group C and these schools of group A is reasonable in practice. However, this relationship cannot be detected from the maps generated by the standard and Merino's methods, where there is no clue that schools in group C are competitors of schools in group A.

In addition, in Figure 4.2, Group B is closer to group A than to other groups. Group D is also closer to group A than to other groups. In the maps generated by the standard and Merino's methods (Figures 4.3 and 4.4), we see that the four strongly connected groups (As, Bs, Cs, Ds) of schools are separated from each other. It is hard to determine which group(s) is (are) closer to another group; in other words, the relationships between groups are not as clear as they are in the map generated by the modified SM method.

Besides, in terms of inner group relationship, schools in each group are pushed close together in Figures 4.3 and 4.4 so that it becomes difficult to ascertain within-group relationships from the maps. Most of relationships that can be seen in the maps generated by the standard and Merino's methods can also be detected in the map generated by our modified SM method. For example, in Figures 4.3 and 4.4, the University of Maryland (A51) is close to schools A73 (University of Pittsburgh), A29 (University of Delaware), and A34 (George Mason University). These relationships are still preserved in Figure 4.2.

However, the map generated by the modified SM method has its limitations. It seems that the modified SM method might not represent some local structures as precisely as other two SM methods. For example, although these four schools that are considered close competitors by each other can still be thought of forming a cluster (see Figure 4.2), A6, A23, A9 and A57 are not placed together in Figure 4.2 as closely as they are, while in maps generated by the other two methods the relationship between these four schools is represented more clearly.

Regarding the differences between the maps generated by the standard and Merino's methods, there is no significant evidence that one outperforms the other. This observation is also confirmed in our performance measurements that are discussed later in this section. The general structures of the maps generated by these two methods are similar and it is reasonable to see local differences due to different random starting configurations and other factors such as stopping criterion etc.

Figures 4.5, 4.6 and 4.7 show the Sammon maps of group A generated by the modified SM, the standard SM method, and Merino's method, respectively. In Figures 4.8, 4.9 and 4.10, we show the Sammon maps of the subset of 30 schools from group A generated by these three methods. As compared to groups B, C, and D, group A and the 30 schools from group A have a relatively large number of schools.

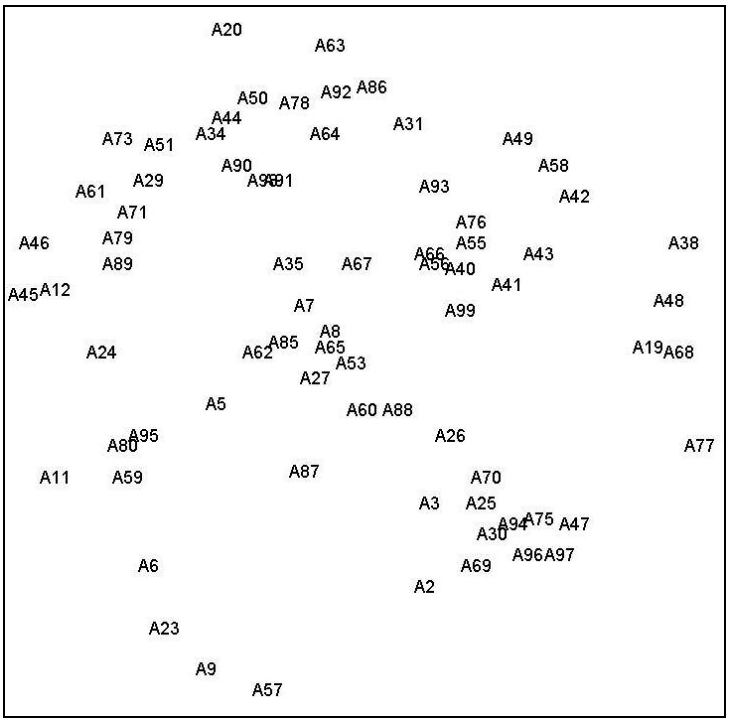

Figure 4.5 Map of group A generated by the modified SM method.

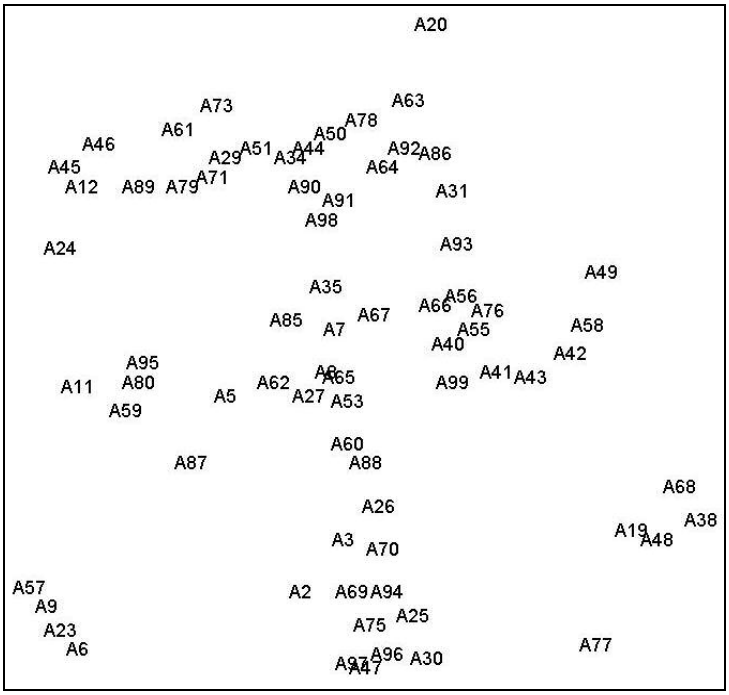

Figure 4.6 Map of group A generated by the standard SM method.

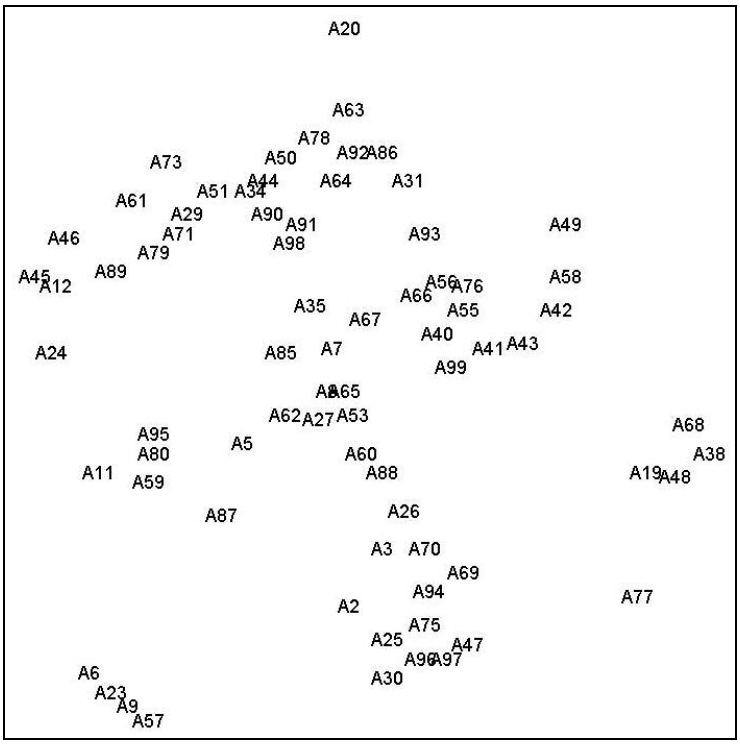

Figure 4.7 Map of group A generated by the Merino's method.

The Sammon map of group A generated by the modified SM method has roughly similar structure as has the Sammon maps generated by the standard SM and Merino's methods. For example, the locations of most A schools are roughly same in these three maps (see Figures 4.5, 4.6 and 4.7). Grinnell College (A38), McAlester College (A48), Carleton College (A19), and Oberlin College (A68) are close to each other and are located in right middle/bottom of the maps away from the other A schools, so that they can be considered a cluster. Wellsley College (A95), Smith College (A80), Mount Holyoke College (A59), and Bryn Mawr College (A11) can be considered another cluster for the same reason. Bates College (A6), Colby College (A23), Bowdoin College (A9)

and Middlebury College (A57) are a third cluster that is placed in left bottom of the maps.

The within-group relationships of schools are illustrated similarly in the maps generated by these three methods. For example, the order relationships between Middlebury College (A57), Bowdoin College (A9), Colby College (A23) and Bates College (A6) are obvious, i.e., Middlebury College and Bates College have the longest pairwise distance of all pairwise distances of these four schools.

There are some local differences between the map generated by the modified method and the other two methods. For example, the distance between Arizona State University (A2) and University of Arizona (A3) is smaller than the distance between Arizona State University and Colorado College (A25) because, in the asymmetric distance matrix, $d_{3,2} < d_{25,2}$ and $d_{2,3} < d_{2,25}$. In Figure 4.5, it is clear that Arizona State University is closer to University of Arizona than it is to Colorado College, while in the map generated by Merino's method (Figure 4.7), it appears that Arizona State University is closer to Colorado College.

As another example to show local differences between these three maps. The distance between Minnesota University (A58) and Marquette University (A49) seems equal to the distance between Minnesota University and Iowa State University (A42) in the map generated by the modified method, while in the maps generated by the other two methods Minnesota University is obviously closer to Iowa State University. However, in the original asymmetric distance matrix, $d_{58,49} < d_{58,42}$, i.e. $1 < 4$, and $d_{49,58} > d_{42,58}$, i.e. $5 > 1$, so that the relationships are not as obvious as they are in maps.

For the 30 schools from group A, the Sammon maps generated by the modified, the standard SM and Merino's methods are visually different (see Figures 4.8, 4.9 and 4.10). However, we can see that A41, A76, A56, A93, A31, etc. are placed in the same sequence in these three maps but in different direction, i.e., clockwise in Figures 4.8 and 4.10, and counter-clockwise in Figure 4.9. Therefore, the general structures of the maps generated by these three SM methods are actually similar. As another example, Reed College (A77), Carleton College (A19), University of Washington (A94), Lafayette College (A45), and Carnegie Mellon University (A20), which are located as outliers in the map generated by the modified method, are still outliers and apart from other schools of group A in the other two maps.

The maps generated by the modified SM method for groups B, C, and D have few visual differences from the maps generated by the standard SM and Merino's methods. In these three types of maps, schools are scattered about and the relative relationships among schools are roughly the same. For example, in Figures 4.11, 4.12 and 4.13, Auburn University (B4) is about the same distance from University of Alabama (B1) and University of Georgia (B37).

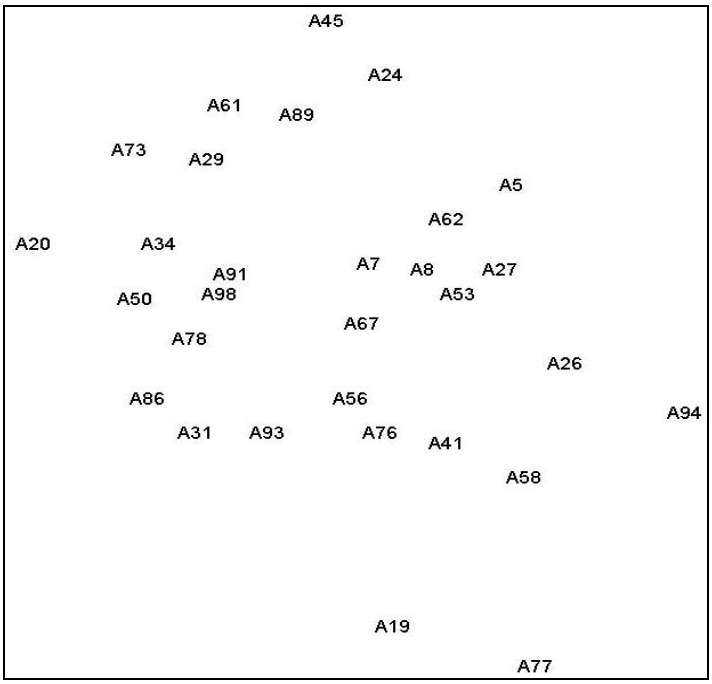

Figure 4.8 Map of 30 schools from group A generated by the modified SM method.

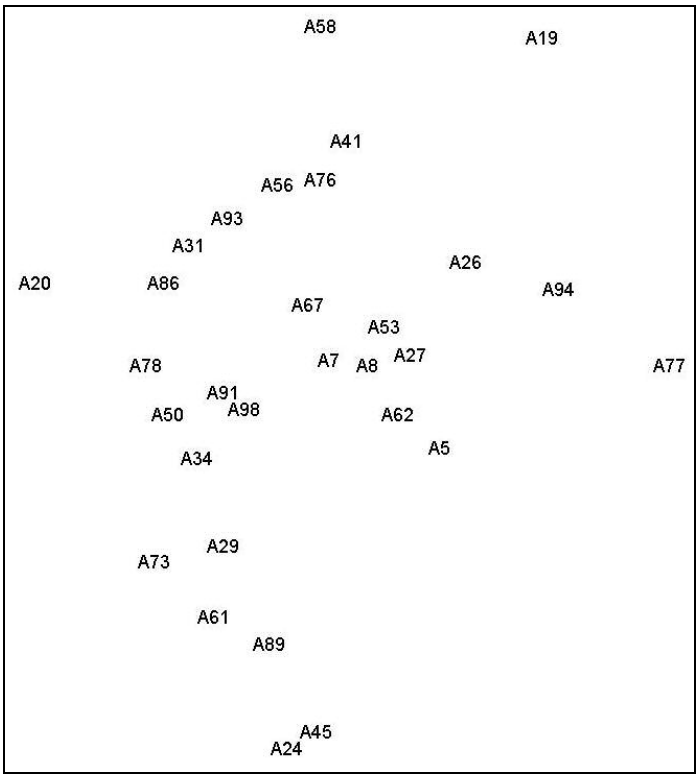

Figure 4.9 Map of 30 schools from group A generated by the standard SM method.

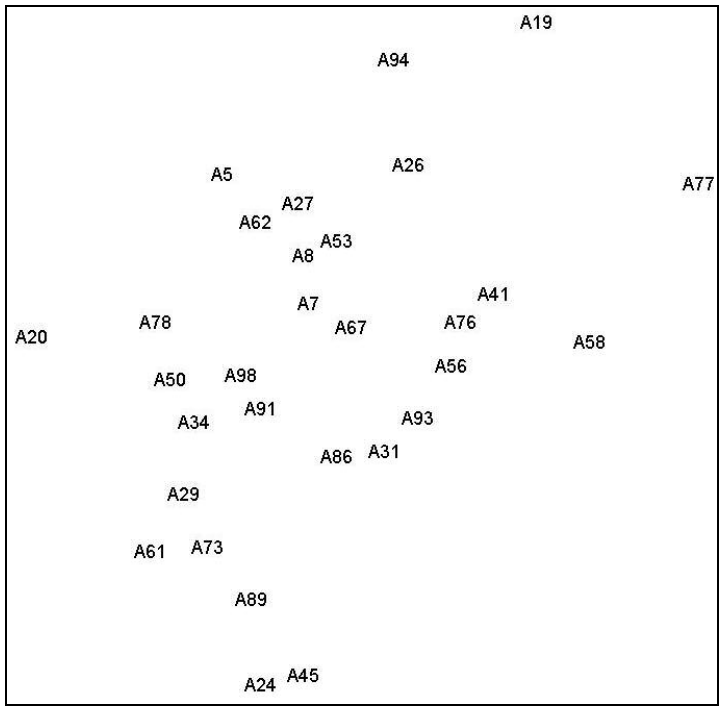

Figure 4.10 Map of 30 schools from group A generated by Merino's method.

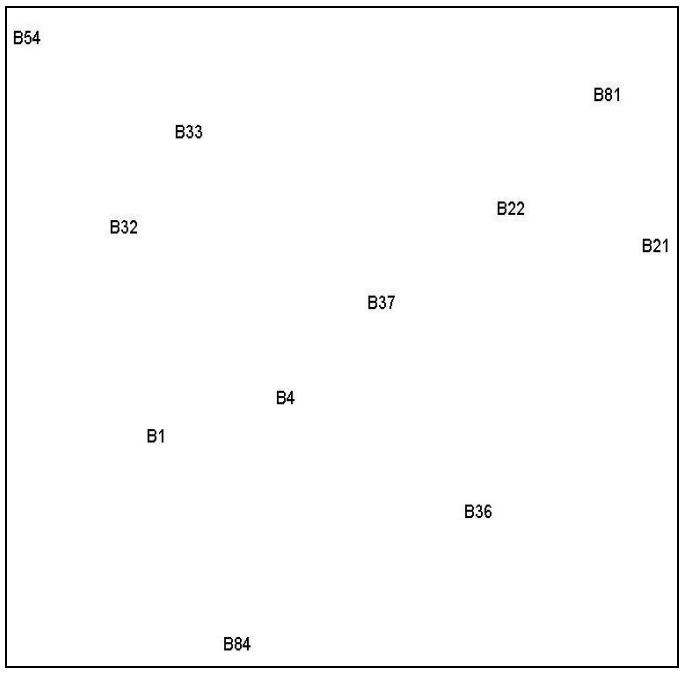

Figure 4.11 Map of group B generated by the modified SM method.

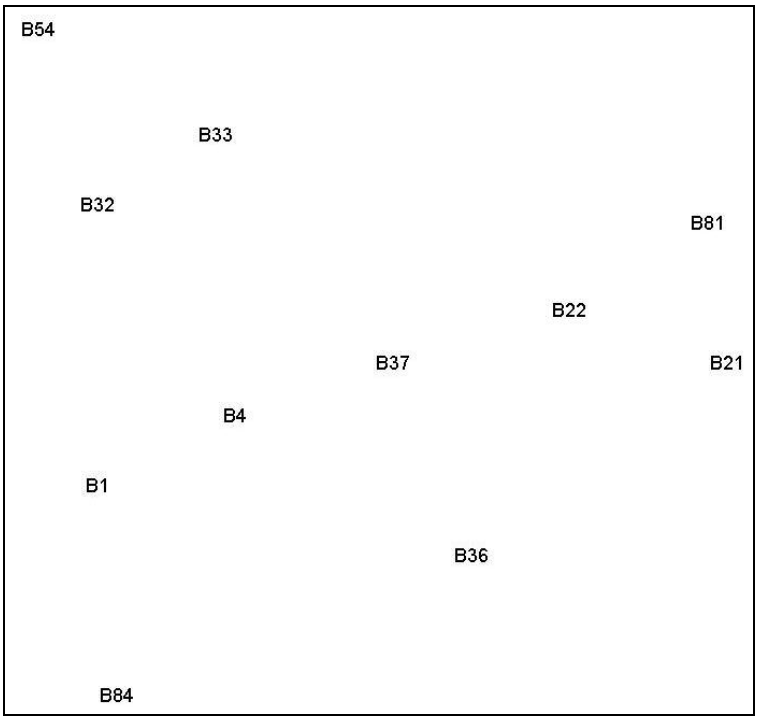

Figure 4.12 Map of group B generated by the standard SM method.

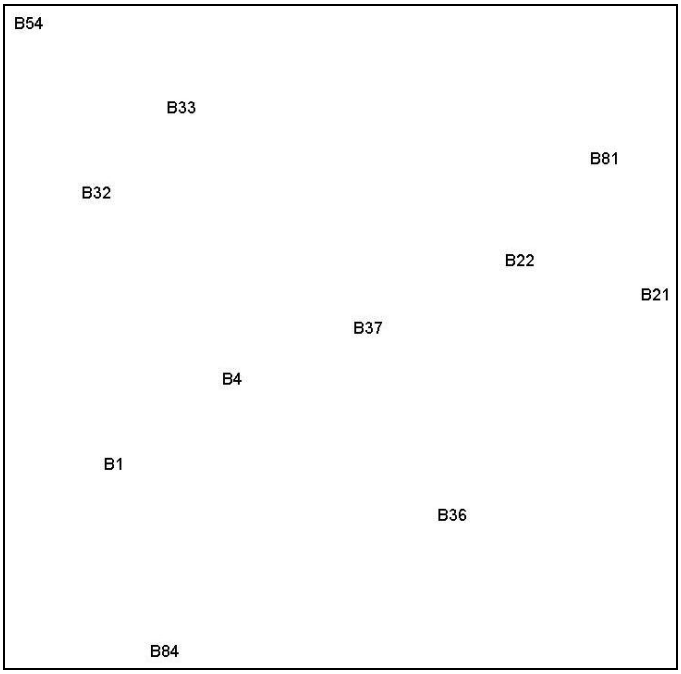

Figure 4.13 Map of group B generated by Merino's method.

There are some differences in the relative relationships among schools in the maps generated by the modified SM method and the maps obtained by the standard SM and Merino's methods. For example, in the asymmetric data set of group C, Harvard (C39) and Yale (C100) are considered competitors by all other schools so that it is desired to place these two schools in the center of the maps and let other schools scatter about. Only in the map (Figure 4.14) generated by the modified SM method Harvard and Yale are located near the center and surrounded by other schools. UPenn (C72), MIT (C52), and Cornell (C28) that are less likely considered competitors by other schools are located in the marginal area of the maps in Figures 4.14, 4.15 and 4.16.

The pairwise distances between Berkeley (D13) and Los Angeles (D16) are the same with the pairwise distances between Berkeley and San Diego (D17) in the asymmetric distance matrix and we expect that Berkeley is equally away from school Los Angeles and San Diego. In Figures 4.17, 4.18 and 4.19, the distances from Berkeley to Los Angeles and to San Diego are approximately equal. For data sets with small size, the modified SM method yields results visually similar to those generated by the standard and Merino's methods.

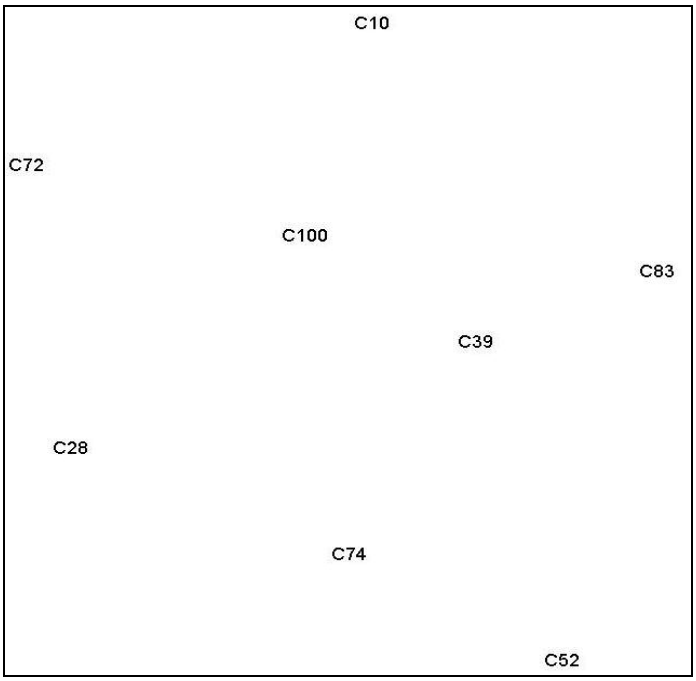

Figure 4.14 Map of group C generated by the modified SM method.

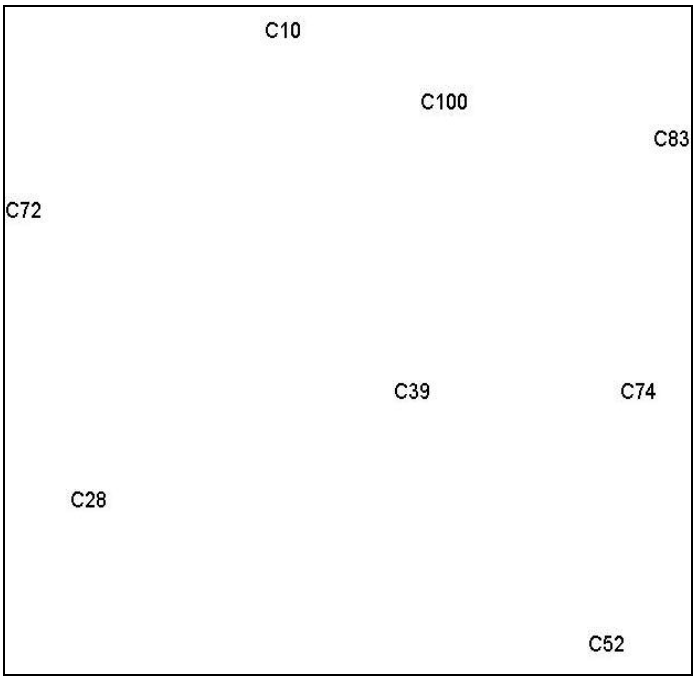

Figure 4.15 Map of group C generated by the standard SM method.

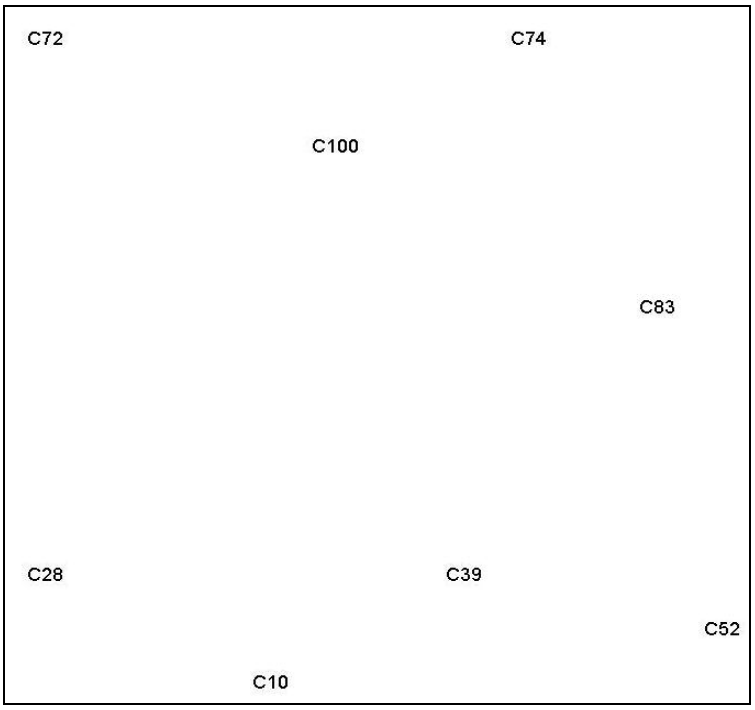

Figure 4.16 Map of group C generated by Merino's method.

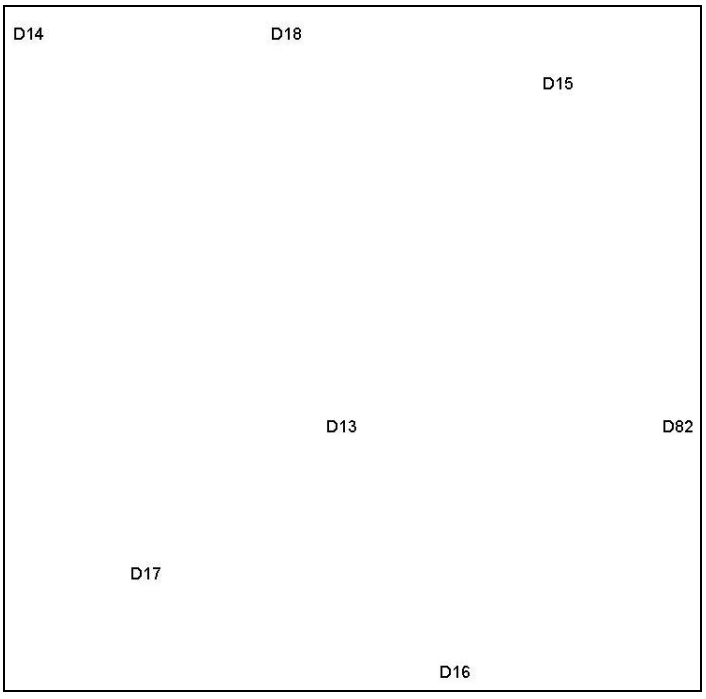

Figure 4.17 Map of group D generated by the modified SM method.

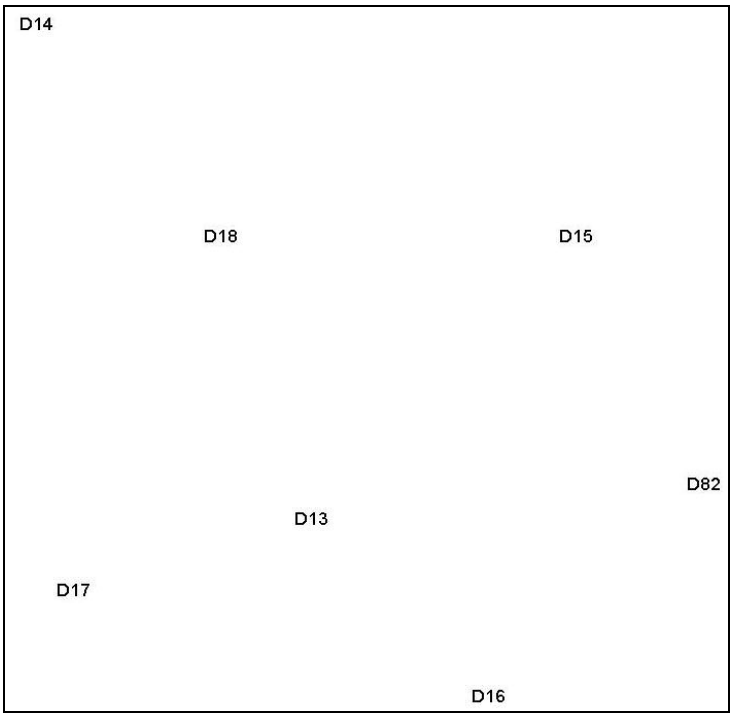

Figure 4.18 Map of group D generated by the standard SM method.

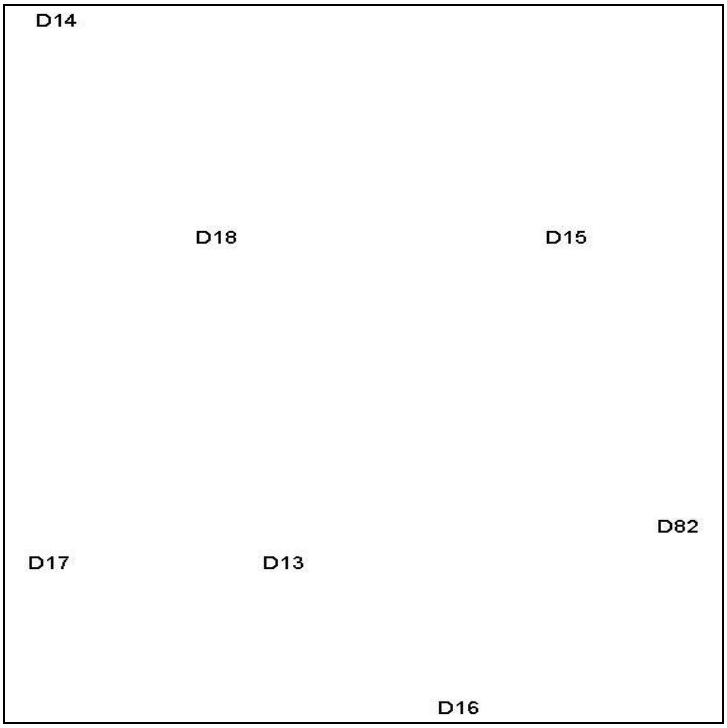

Figure 4.19 Map of group D generated by Merino's method.

The standard and Merino's methods take as inputs symmetrized distance matrices, which are derived from original asymmetric distance matrices by averaging. The modified SM method takes original asymmetric distance matrices as inputs directly. Two performance measurements are employed to assess these three methods' capability of representing original asymmetric distance data. We use order preservation coefficient to show how effectively a SM method preserves order relationships of original data. We use error measurement to indicate how precisely original distances between data are represented by a SM method. The represented order relationships among schools and the resulting error measures are expected to be different in the maps generated by the three SM methods because their input distances and rationales behind the methods are different.

As shown in Tables 4.5 and 4.6, the order preservation coefficients of maps generated by the standard and Merino's SM methods are very close. The error measures of these two SM methods do not change dramatically (see Tables 4.7 and 4.8). These can be explained by their taking the same symmetrized distance matrices. Another explanation may be the asymmetry coefficients introduced in Merino's method (described in section 2.5). In these six data sets most of schools have similar chances to be considered competitors by others and asymmetry coefficients of schools fall in a narrow range of values. Introducing similar asymmetry coefficients will not make the objective function better off and therefore asymmetry coefficients in these cases have little impact on the optimization results.

By incorporating the upper triangular part and the lower triangular part of the original asymmetric distance matrix into the objective function and the associated

Data Set	Standard SM	Modified SM	% Improvement of Standard over Modified
100 Schools	0.482	0.472	2.10
30 Schools	0.619	0.607	1.97
A	0.605	0.594	1.71
B	0.759	0.756	0.34
C	0.580	0.568	2.14
D	0.567	0.552	2.59

Table 4.5 Order preservation coefficients for the six data sets (standard vs. modified).

Data Set	Merino's SM	Modified SM	% Improvement of Merino's over Modified
100 Schools	0.475	0.472	0.70
30 Schools	0.616	0.607	1.51
A	0.603	0.594	1.45
B	0.759	0.756	0.36
C	0.582	0.568	2.52
D	0.567	0.552	2.59

Table 4.6 Order preservation coefficients for the six data sets (Merino's vs. modified).

updating rules, the search procedure of the modified SM method takes into account the entire original distance matrix instead of the symmetrized distance matrix and looks for an optimal configuration for the entire distance matrix. Therefore, the error measures of maps generated by the modified SM method are significantly smaller than the error measures of maps generated by the standard and Merino's SM methods (see Tables 4.7 and 4.8). It seems that the modified SM method reduces the distance error measures

Data Set	Standard SM	Modified SM	% Improvement of Modified over Standard
100 Schools	54906.620	28830.320	47.49
30 Schools	999.067	773.045	22.62
A	6926.616	5688.302	17.88
B	20.192	18.313	9.31
C	6.547	6.065	7.37
D	6.048	5.549	8.25

Table 4.7 Error measures for the six data sets (standard vs. modified).

Data Set	Merino's SM	Modified SM	% Improvement of Modified over Merino's
100 Schools	59774.640	28830.320	51.77
30 Schools	1024.385	773.045	24.54
A	7096.426	5688.302	19.84
B	19.927	18.313	8.10
C	6.615	6.065	8.33
D	6.073	5.549	8.62

Table 4.8 Error measures for the six data sets (Merino's vs. modified).

proportionally to the size of data sets. For example, in 100-school data set, the modified SM method reduced the error measure over the standard SM method by 47.49% (100 × (54906.620 – 28830.320)/ 54906.620). In the remaining data sets, the improvements range from 8.25% to 22.62%. We point out that the values given in the tables of performance measures are averages from the five experiments on each data set.

The modified SM method produces slightly smaller order preservation coefficients than the other two SM methods (large coefficients are better). This may be

due to asymmetric distances in original data sets. When the whole asymmetric distance matrix is taken into account, the search procedure goes through the upper triangular part and the lower triangular part separately to find an improved result. This adds more constraints to the optimization process and introduces more difficulty to achieve a better result because some constraints are conflict due to asymmetry.

Although the modified SM method did not do a better job than the standard SM and Merino's SM methods in preserving the order relationships in our American college selection data, the maps generated by the modified SM method are still considered better than the maps generated by the standard SM method and Merino's method. The modified SM method is capable of preserving order relationships with similar accuracy and reducing the distance errors with significant improvement as compared to the standard and Merino's SM method. The maps generated by the modified SM method show us the intra-group relationships, which were not reflected in the maps generated by other two methods. For example, it is interesting to see that schools of group C mixed with several A schools in the maps generated by the modified SM method. In addition, the generated maps by the modified method are more readable -- schools were not squeezed as tightly in Figure 4.2 as they were in Figure 4.3.

In our study of analyzing American college selection data using different Sammon mapping visualization techniques, currently, the modified SM method seems to be able to generate visualization maps with higher quality as compared to the standard and Merino's methods. The modified method yields results that are with significantly reduced distance errors and reasonably preserved order relationships compared to the

results generated by the other two methods. We hope that our modified SM method can be robust on other applications.

Chapter 5

Visualizing Canadian Ranked College Data

The modified SM method was shown to be good at recovering the structure of American college data at least comparable to the standard and Merino's methods. In American college selection data, the overlaps of each school are not ordered so that there is no indication of ranks among the overlaps. The generated visualization maps are unable to provide students with further details such as which school among the overlaps is the closest competitor. If ranking information is incorporated into the data set, decisions by students can be made more easily and effectively.

Data sets with ranking information can be collected in many fields. In marketing, ranked data can be gathered from customers who give ranks of different brands of products of the same category, i.e., car brands. Consider another example in which a survey is sent to apartment managers in a metropolitan area. The survey asks the managers to identify top 10 rival apartment buildings. The managers respond with a top 10 listing in which the first building is the most competitive rival, the second building is next most competitive rival, and so on. For a person seeking an apartment, a visual map of competitors can help narrow the search effectively and naturally.

Our work in this chapter includes building a visual model of asymmetric data sets that incorporate ranking information. The Canadian ranked college data set is the one that has ranking information and will be analyzed in the following sections. The modified SM method will be applied to this data set to see if ranking information is represented accurately. For comparison, the standard and Merino's SM methods will also be applied to the data set.

5.1 Description of Canadian Ranked College Data

The idea of visualizing universities originates with the work of Yin (2002) and Condon et al. (2002). Yin proposed ViSOM (visualization-induced SOM) to detect clusters of universities in the United Kingdom. Condon et al. created visual maps of 100 American universities that can be used to view patterns and clusters and gain insights.

We collected data from surveys that were sent to the admission directors of undergraduate programs in 52 Canadian universities. The directors were asked to list the five most competitive rivals in terms of overall education quality. We have received responses from 44 universities and we list these in Table 5.1. The competitors of each Canadian university are given in Table 5.2.

Key	School
1	Acadia University
2	Bishop's University
3	Brandon University
4	Brock University
5	Carleton
6	Concordia University
7	Dalhousie University
8	Lakehead University
9	Laval University
10	McGill University
11	McMaster University
12	Memorial University of Newfoundland
13	Mount Saint Vincent University
14	Nipissing University
15	Queens University
16	Simon Fraser University
17	St. Francis Xavier University
18	Univeristy College of Cape Breton
19	Universite de Moncton
20	Universite de Sherbrooke
21	Universite du Quebec a Rimouski
22	Universite du Quebec en Outaouais
23	University of Alberta
24	University of British Columbia
25	University of Calgary
26	University of Guelph
27	University of Lethbridge
28	University of Manitoba
29	University of Montreal
30	University of New Brunswick
31	University of Ontario Institute of Technology
32	University of Ottowa
33	University of Prince Edward Island
34	University of Regina
35	University of Saskatchewan
36	University of Toronto
37	University of Victoria
38	University of Waterloo

Table 5.1 44 Canadian universities collected from surveys.

39	University of Western Ontario
40	University of Windsor
41	Wilfrid Laurier
42	Laurentian University
43	University of St. Anne
44	York University

Table 5.1 (continued).

Key	School	1st	2nd	3rd	4th	5th
1	Acadia University	38	36	10	15	23
2	Bishop's University	1	41	0	0	0
3	Brandon University	28	25	34	35	0
4	Brock University	11	26	41	44	39
5	Carleton	36	32	10	15	38
6	Concordia University	10	29	20	7	0
7	Dalhousie University	10	36	15	24	23
8	Lakehead University	26	4	39	38	0
9	Laval University	29	20	10	6	0
10	McGill University	36	24	15	7	44
11	McMaster University	36	15	38	39	10
12	Memorial University of Newfoundland	7	30	1	17	13
13	Mount Saint Vincent University	1	17	12	0	0
14	Nipissing University	4	41	8	0	0
15	Queens University	36	24	10	26	38
16	Simon Fraser University	24	37	38	26	36
17	St. Francis Xavier University	7	1	0	0	0
18	Univeristy College of Cape Breton	7	17	13	30	0
19	Universite de Moncton	30	9	32	7	43
20	Universite de Sherbrooke	29	9	2	0	0
21	Universite du Quebec a Rimouski	9	20	0	0	0
22	Universite du Quebec en Outaouais	32	20	0	0	0
23	University of Alberta	36	24	10	0	0
24	University of British Columbia	36	10	15	23	39
25	University of Calgary	36	24	15	10	23

Table 5.2 Competitors of 44 Canadian universities.

26	University of Guelph	11	38	15	36	39
27	University of Lethbridge	23	25	34	35	37
28	University of Manitoba	23	35	36	15	0
29	University of Montreal	36	10	32	24	23
30	University of New Brunswick	7	12	1	0	0
31	University of Ontario Institute of Technology	39	36	15	44	38
32	University of Ottowa	11	39	15	29	7
33	University of Prince Edward Island	1	17	30	0	0
34	University of Regina	23	24	28	0	0
35	University of Saskatchewan	34	23	25	15	24
36	University of Toronto	39	24	23	15	10
37	University of Victoria	24	16	25	23	0
38	University of Waterloo	36	39	15	41	11
39	University of Western Ontario	36	10	15	38	11
40	University of Windsor	39	11	41	36	15
41	Wilfrid Laurier	39	38	26	11	36
42	Laurentian University	4	40	8	32	5
43	University of St. Anne	7	17	30	0	0
44	York University	36	15	39	41	0

Table 5.2 (continued).

5.2 Modeling Steps

In the Canadian university data set that we have collected, some universities specified their competitors that were not on the list of the 44 universities. We did not include universities that did not respond and yet were selected as rivals by other universities. For example, York University chose University of Toronto as its top competitor, Ryerson University was the second, and Queens, Western Ontario, and Wilfrid Laurier were third, fourth, and fifth, respectively. Ryerson University did not respond to our survey, so it is not on the list. Therefore, in the rival list of York University, we excluded Ryerson and replaced it with Queens as the second most

competitive school, and then moved Western Ontario and Wilfrid Laurier up one rank, respectively.

In order to create a visual map of the Canadian college data set, we need to generate a distance matrix from the data set and then input the distance matrix into the modified SM method. Using procedures that are described in Chapter 4, we start by creating a 44×44 0-1 adjacency matrix, where entry $s_{ij} = 1$ (row i and column j) if school j is a competitor of school i. Next, we adjust the entries with the value of 1 to reflect the ranking information. In terms of adjacency or distance, the more competitive school j is to school i, the smaller the value of the s_{ij} entry is. The entry of the most competitive school is 1 by default. We set the gap between entries of two consecutive competitors in ranking is 0.5, in other words, the entry of the second competitor is 1.5, and the entry of the third competitor is 2, and so forth.

After adjusting the adjacency matrix, we construct the distance matrix using a shortest path procedure and replace infinity entries with an appropriate value, i.e., 25% larger than the longest distance. After the distance matrix is constructed, we apply the modified SM method that is described in Chapter 3 to Canadian ranked college data trying to detect some interesting relationships of Canadian schools.

5.3 Discussion of the Results

The map shown in Figure 5.1 is generated by the modified SM method. The distance gap, which is used to separate two consecutive competitors, is set to 0.5 and the substitute value of infinity distance is 10.5, which is 25% larger than the longest finite distance i.e. 8.5. It provides us with a general view of the structure of Canadian colleges.

Universities such as Toronto (36), Queens (15), and Western Ontario (39) that are frequently considered competitors by other universities are located in the center of the map. Other universities such as Waterloo (38) and British Columbia (24) with high frequencies are also placed near the center. Therefore, the map tells us that universities placed in or near the center are those that are considered popular. Besides, it shows in the map roughly five clusters of universities – one in the center and the other four surrounding the center.

The map generated by the modified SM method also reveals some ranking information of competitors of a particular school. For example, the top five competitors of Queens University (15) are Toronto (36), British Columbia (24), McGill (10), Guelph (26), and Waterloo (38), where Toronto is the most competitive school to Queens. In Figure 5.1, among these five competitors, Toronto is the closest to Queens. Although Waterloo is not the school farthest from Queens, it is the second farthest school among these five competitors. Guelph is the school farthest from Queens, because the top rival of Guelph is McMaster and Queens is its third top. The asymmetric characteristic of a data set makes it very difficult to generate a map that is a completely accurate representation of the data set. A map that shows roughly similar structures of the data set can be generated by visualization techniques.

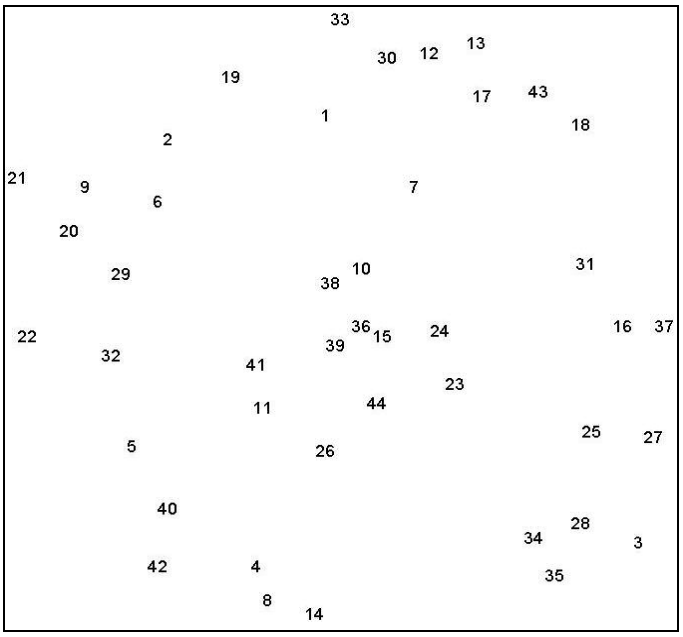

Figure 5.1 Sammon map of the Canadian ranked college generated by the modified method. Gap is set to 0.5.

Figures 5.2 and 5.3 show the maps generated by the standard and Merino methods. Again the maps generated by these two methods are similar to each other due to the same reasons that have been discussed in previous chapter. Universities located in the center of Figure 5.1 are still placed in the center of these two maps. Other universities scatter about surrounding popular universities.

Differences between the map generated by the modified method and the maps generated by the other two methods can be summarized as follows. Universities surrounding the center are separated more evenly in the maps generated by latter two methods. This makes it harder to detect clusters of universities. Universities in the center of the maps generated by the latter two methods squeeze more tightly than they are in the

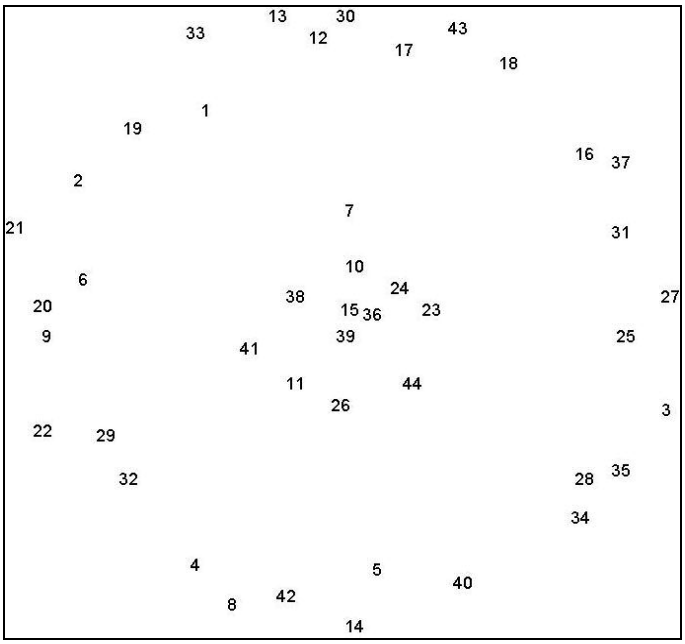

Figure 5.2 Sammon map of the Canadian ranked college generated by the standard method. Gap is set to 0.5.

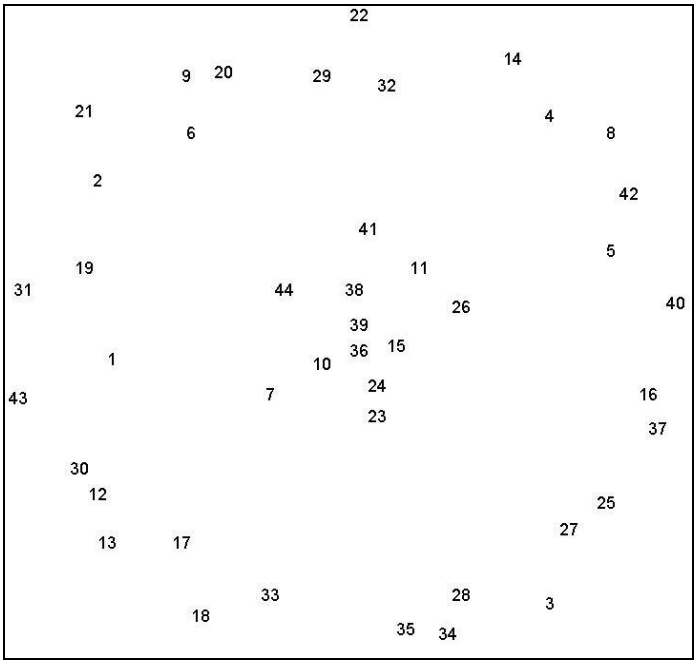

Figure 5.3 Sammon map of the Canadian ranked college generated by Merino's method. Gap is set to 0.5.

map generated by the modified method. Some interesting insights may be hidden in the maps of Figures 5.2 and 5.3. For example, all five competitors of Dalhousie (7) are located in the center of maps and therefore it is reasonable to place Dalhousie near the center. However, most of the universities that consider Dalhousie a competitor are located in the top middle area of the maps. It is expected that Dalhousie would be near these universities while keeping near the center universities. Only in the map generated by the modified SM map the location of Dalhousie reminds us of the relationships while in other two maps it is not clear that we can detect the relationships between Dalhousie and these universities that choose Dalhousie as a competitor.

In terms of insights that possibly will be gained from visualization maps, the modified SM method provides us with a more reasonable map. In terms of performance measures, the modified SM method most of the time in our experiments does slightly better than the standard and Merino's methods in order relationship preservation (see Tables 5.3 and 5.4) and always outperforms the other two methods in error measurement (see Tables 5.5 and 5.6).

In order to see if the gap value affects the performance of these three methods, we experimented with three different gap values: 0.2, 0.5, and 1. 0.5 is the one that is used in Figures 5.1, 5.2 and 5.3. Given five random starts, we used each gap value on the Canadian college data. Tables 5.3 and 5.4 provide average order preservation measures of these three methods, and Tables 5.5 and 5.6 provide average error measures of these three methods. In Figures 5.4 and 5.5 are given generated maps by the modified method with gap values of 0.2 and 1.0 respectively. Comparing these two figures to Figure 5.1, we can see that although there are some local differences between these three maps, the

Gap Value	Standard SM	Modified SM	% Improvement of Modified over Standard
Gap-0.2	0.555	0.560	0.99
Gap-0.5	0.546	0.548	0.46
Gap-1.0	0.537	0.542	0.89

Table 5.3 Order preservation measures of different gap values of Canadian ranked college data (standard vs. modified).

Gap Value	Merino's SM	Modified SM	% Improvement of Modified over Merino's
Gap-0.2	0.559	0.560	0.14
Gap-0.5	0.543	0.548	0.88
Gap-1.0	0.545	0.542	-0.64

Table 5.4 Order preservation measures of different gap values of Canadian ranked college data (Merino's vs. modified).

Gap Value	Standard SM	Modified SM	% Improvement of Modified over Standard
Gap-0.2	2413.83	2100.84	12.97
Gap-0.5	3921.53	3336.99	14.91
Gap-1.0	7001.58	5414.63	22.67

Table 5.5 Error measures of different gap values of Canadian ranked college data (standard vs. modified).

Gap Value	Merino's SM	Modified SM	% Improvement of Modified over Merino's
Gap-0.2	2344.01	2100.84	10.37
Gap-0.5	3972.14	3336.99	15.99
Gap-1.0	6675.48	5414.63	18.89

Table 5.6 Error measures of different gap values of Canadian ranked college data (Merino's vs. modified).

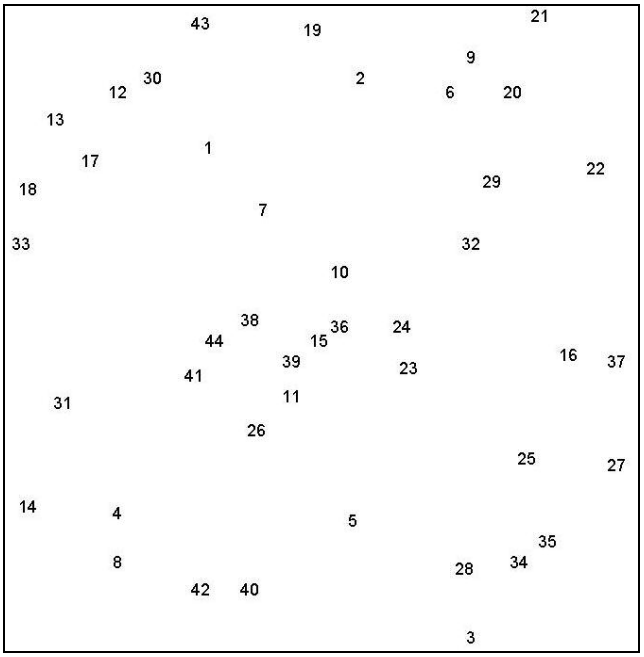

Figure 5.4 Sammon map of the Canadian ranked college generated by the modified method. Gap is set to 0.2.

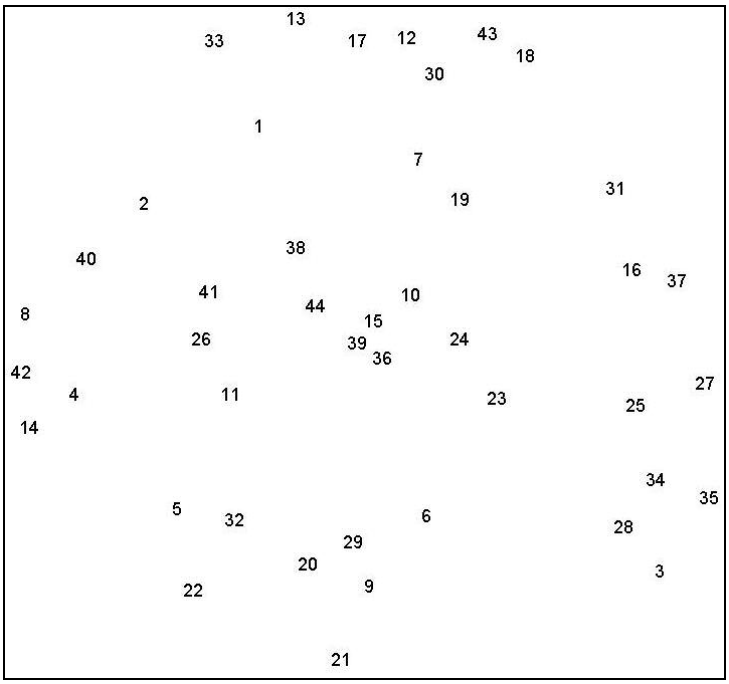

Figure 5.5 Sammon map of the Canadian ranked college generated by the modified method. Gap is set to 1.0.

general structure of the data set is kept similar in each of these maps, i.e., popular universities in the center and surrounded by other clusters of universities. Different gap values have effect on constructing maps especially on local details however the general structure of the data set is not changed dramatically.

To analyze Canadian ranked college data, the modified SM method seems better than the standard and Merino's methods. The modified method reduces distance errors significantly and preserves the order relationships reasonably well compared to other two methods. The modified method also helps gain some interesting insights that can hardly be detected in maps generated by other two methods. We have done some sensitivity analysis of gap values and it seems to us that gap values do not dramatically affect the general structure of the data set.

So far, we have analyzed two visualization applications using the modified Sammon mapping method. As compared to the modified SM method, the self-organizing map, which is a neural-network based method, can also be employed to visualize and analyze data sets. In the following chapters, we will discuss some applications using SOMs.

Chapter 6

Self-Organizing Maps: State Sponsored Murder Data Set

6.1 Introduction

As a neural-network based unsupervised method, self-organizing maps (SOMs) are mainly used for clustering, which is one of activities of data analysis in data visualization applications. In our previous applications, we used the Sammon mapping method to visualize two college data sets and analyze intra-cluster and inter-cluster relationships among clusters. In this chapter, we will use SOMs to analyze clusters in the state sponsored murder data set and to analyze the sport records data in the next chapter.

For discovering clustering information hidden in data sets, there are a few methods that have been proposed in the literature such as hierarchical clustering methods (single linkage, average linkage, and complete linkage), K-means clustering, and Kohonen's self-organizing maps (SOMs) (1995). Among these methods, Kohonen's SOM has received increased attention in the literature in recent years. Some recently proposed clustering methods such as ViSOM (Yin, 2002) are based on Kohonen's SOM. Several software packages (i.e., Viscovery SOMine) with SOM-based clustering procedures have been released. A natural question arises: How well do these software packages perform? We want to make sure that we choose the right clustering procedure

to analyze datasets. Therefore, in the first part of this chapter, we evaluate the performance of four software implementations of SOM-based clustering methods and determine the best SOM-based procedure to be used in our clustering analysis. In the second part of this chapter, we apply the chosen SOM-procedure to the state sponsored murder data set.

6.2 Evaluating SOM-based Methods

Four clustering implementations are compared based on their performances: Ward clustering, modified Ward clustering, single linkage clustering, and classic SOM. The first three clustering methods are implemented in a commercial package, Viscovery SOMine 4.0 from Eudaptics Software (www.eudaptics.com). These three SOM-based clustering methods make use of the representation of the data set given by Kohonen's SOM scheme. SOM-Ward clustering uses Ward's classic minimum distance method (Ward, 1963). In SOM-modified Ward clustering, the classic Ward method is modified to use a different distance measure (Viscovery, 2002). SOM-single linkage uses an adaptation of the classic single-linkage clustering algorithm (Viscovery, 2002). Classic SOM clustering is implemented in a research package, SOM_Pak from the Helsinki University of Technology (SOM_Pak, 1997).

Mangiameli et al. (1996) conducted a comprehensive evaluation of seven hierarchical clustering algorithms and the SOM network generated by the commercial software package NeuralWorks. Our study can be viewed as an extension of their work. We assess the clustering performance of four procedures in two current SOM software

packages. To our knowledge, the performance of these packages has not been reported in the open literature.

The standard approach to studying the performance of a clustering procedure is to apply the procedure to a problem for which the clusters are already known. This approach allows the researcher to measure the method's success in assigning data points to their correct clusters. We adopt this approach and evaluate the performance of the clustering methods on 96 data sets that we construct. The clustering methods that we evaluate are applied to data sets in which the clusters are well separated. Figure 6.1 presents a two-dimensional plot of a four-cluster data set.

6.2.1 Constructing data sets

Four experimental factors are used to characterize each data set: the number of clusters (three, four, five, and six), number of dimensions (three and four), number of data points (50, 100, 150, and 200), and amount of intra-cluster dispersion (low, medium, and high). Using this design, we construct $4 \times 2 \times 4 \times 3 = 96$ data sets, each of which exhibits both external isolation and internal cohesion (see Cormack (1971), Mangiameli et al. (1996), and Milligan (1980)). External isolation means that the members of one cluster are separated from members of another cluster by empty space. Internal cohesion means that members of the same cluster are similar (close) to each other.

We use a procedure that is similar to the method proposed by Milligan (1980, 1985) to construct data sets. The first step is to determine the cluster lengths and the cluster boundaries for the first dimension of the variable space. For each cluster, the

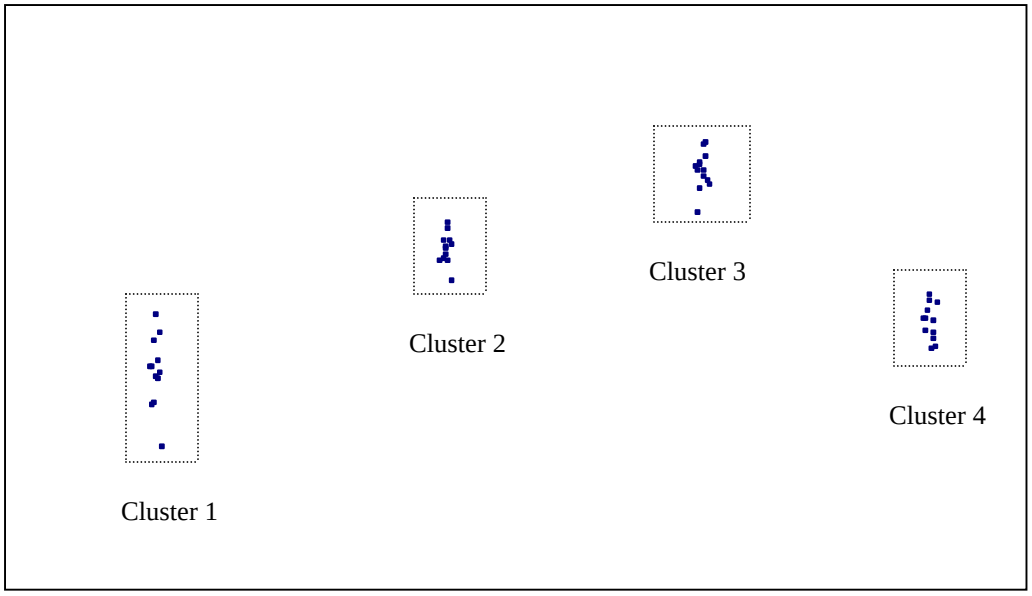

Figure 6.1 Example of a four-cluster data set.

cluster length is selected randomly from the uniform distribution on the interval (10, 40). In order to achieve external isolation, the boundaries of adjacent clusters are separated by an amount selected randomly from the uniform distribution on the interval (0.25, 0.75). The mean of each cluster is taken to be the midpoint of the cluster's boundaries, and the standard deviation of each cluster is set to 0.5.

The second step is to specify the characteristics of the clusters in the remaining dimensions. We select the cluster lengths randomly from the uniform distribution on the interval (10, 40) and then select the cluster boundaries randomly. This makes it possible that cluster boundaries overlap with each other, unlike in the first dimension where external isolation is guaranteed. The mean of each cluster is taken to be the midpoint of the cluster's boundaries. Three levels of intra-cluster dispersion are used here: low, medium, and high; these are the same levels specified in Mangiameli et al. (1996). At the

low, medium, and high levels of dispersion, the standard deviation of the cluster is set equal to 1/12, 1/6, and 1/3 times the cluster length, respectively. The level of the intra-cluster dispersion indicates the density of data points around the cluster boundaries. The higher the intra-cluster dispersion, the higher is the density of data points near the cluster boundaries. Thus, internal cohesion decreases as the intra-cluster dispersion increases.

We generate data points in each cluster from a multivariate normal distribution with mean vector given by the midpoints of the cluster boundaries. The diagonal elements of the variance-covariance matrix are given by the squares of the standard deviations, and all of the off-diagonal elements are equal to zero. We discard data points that fall outside the cluster boundaries.

6.2.2 Measuring performance

In order to evaluate the performance of a clustering procedure, we use two measures: the cluster recovery rate and the Rand statistic. The cluster recovery rate is defined to be the proportion of times a clustering procedure correctly recovers the cluster structure, that is, the percentage of times a procedure correctly determines the cluster membership of each data point. The Rand statistic (Rand, 1971) is a widely used performance metric (Milligan, 1981). The definition of the Rand statistic can be illustrated using the notation given in Table 6.1. Cell *A* is the number of pairs of points in the data set that are from the same cluster and are correctly assigned by a clustering procedure to the same cluster. Cell *B* is the number of pairs of points that are from different clusters and are correctly assigned by a clustering procedure to different

Clustering Procedure Solution	Correct Solution	
	Pair in Same Cluster	Pair Not in Same Cluster
Pair in Same Cluster	A	C
Pair Not in Same Cluster	D	B

Table 6.1 Pairwise classification notation.

clusters. In Cells *C* and *D*, clustering errors are counted. The Rand statistic provides the proportion of correct pairwise classifications for the data set and is given by $(A + B)/(A + B + C + D)$. If the solution generated by a clustering procedure is correct, then the Rand statistic equals one; if the generated solution is incorrect, then the value of the Rand statistic will be less than one. Clearly, the larger the value of the Rand statistic, the better the solution.

6.2.3 Comparison results and conclusions

We applied the three SOM-based clustering procedures in Viscovery (Ward, modified Ward, and single linkage, all with default settings), the classic SOM clustering procedure in SOM_Pak (there are no default settings; for each run, we had to specify values for several parameters found in the package), and the *K*-means algorithm in Clementine from SPSS to each of our 96 data sets. We specified the number of clusters in each data set as input to each procedure.

The cluster recovery rates for the five procedures are given in Table 6.2. We see

SOM-Ward	SOM-Modified Ward	SOM-Single Linkage	SOM-Classic	*K*-Means
92.7	91.7	82.3	14.6	80.2

Table 6.2 Cluster recovery rates (in %).

that two of Viscovery's procedures, SOM-Ward and SOM-modified Ward, recover the true clusters more than 90% of the time, while Viscovery's SOM-single linkage and the *K*-means algorithm recover the clusters about 80% of the time. SOM-classic performs poorly, only recovering the clusters about 15% of the time.

In Table 6.3, we show the effect of intra-cluster dispersion on the cluster recovery rates of the five procedures. As the intra-cluster dispersion increases, thereby reducing the internal cohesion of the clusters, we see that the cluster recovery rates decrease. At all three levels of dispersion, SOM-Ward performs the best, closely followed by SOM-modified Ward.

In addition to the cluster recovery rate, we also examine the performance of the five procedures by calculating the Rand statistic. The average value of the Rand statistic is given in Table 6.4. We apply SOM-Ward to the eight data sets with low dispersion and three clusters, calculate the Rand statistic for each data set, and then average over the eight data sets (we see that the entry is 1). The average Rand statistic for SOM-Ward for all 32 data sets with low dispersion is 1 (this is the row average for SOM-Ward in the first row of Table 6.4).

In Table 6.4, at the low and medium levels of dispersion, we see that all three of Viscovery's procedures perform better than SOM-classic and *K*-means. At the high level

	Level of Dispersion		
Procedure	Low	Medium	High
SOM-Ward	100	94	84
SOM-Modified Ward	100	91	84
SOM-Single Linkage	100	91	56
SOM-Classic	19	16	9
K-Means	88	78	75

Table 6.3 Cluster recovery rates (in %) by level of dispersion.

	Number of Clusters				Row
Clustering Procedure	3	4	5	6	Average
	Low Intra-Cluster Dispersion				
SOM-Ward	1.000	1.000	1.000	1.000	1.000
SOM-Modified Ward	1.000	1.000	1.000	1.000	1.000
SOM-Single Linkage	1.000	1.000	1.000	1.000	1.000
SOM-Classic	0.846	0.899	0.911	0.886	0.886
K-Means	1.000	1.000	0.988	0.965	0.988
	Medium Intra-Cluster Dispersion				
SOM-Ward	0.994	0.989	1.000	1.000	0.996
SOM-Modified Ward	0.995	0.986	1.000	0.997	0.995
SOM-Single linkage	1.000	0.999	1.000	0.999	0.999
SOM-Classic	0.898	0.878	0.893	0.893	0.890
K-Means	0.995	1.000	0.988	0.931	0.979
	High Intra-Cluster Dispersion				
SOM-Ward	0.914	0.954	0.984	1.000	0.963
SOM-Modified Ward	0.951	0.971	0.996	1.000	0.980
SOM-Single linkage	0.946	0.977	0.992	0.991	0.977
SOM-Classic	0.915	0.931	0.876	0.898	0.905
K-Means	1.000	0.960	0.990	0.950	0.975

Table 6.4 Values of the Rand statistics.

of dispersion, the three procedures in Viscovery and *K*-means perform nearly the same and SOM-classic is close behind. It appears that as the level of dispersion increases and when the number of clusters is small (three or four), the performance of each of the Viscovery procedures deteriorates somewhat (the value of the average Rand statistic decreases). We note that SOM-classic and *K*-means seem to be less affected than Viscovery's three procedures by increases in intra-cluster dispersion.

Both SOM-classic and *K*-means require the user to specify the number of clusters (that is why we input the number of clusters into all five procedures). Viscovery, however, does not have this requirement. After inputting the data set, Viscovery can determine the number of clusters and the assignment of points to clusters. This is a desirable feature of the package since in practice a user usually does not know how many clusters to specify in advance. We applied Viscovery to each of the 96 data sets and let it determine the number of clusters. In Table 6.5, we give the cluster recovery rates for the procedures. For each procedure, the recovery rate drops about 10 percentage points from the recovery rate generated by Viscovery when cluster size was specified. These results are still competitive with the recovery rate from *K*-means (80.2%) when *K*-means has the advantage of knowing the true cluster size.

In this study, we assessed the performance of four SOM-based clustering procedures that are implemented in commercial and research software. The three procedures in Viscovery SOMine 4.0 performed generally well in clustering. We found that Viscovery's procedures performed slightly better than the *K*-means algorithm and much better than the procedure in SOM_Pak. In addition, when clusters were well separated (i.e., exhibited external isolation), the clustering procedures in Viscovery were

SOM-Ward	SOM-Modified Ward	SOM-Single Linkage
83.3	82.3	71.9

Table 6.5 Cluster recovery rates (in %) for Viscovery (number of clusters is not specified).

fairly effective at determining the appropriate number of clusters in a data set. This feature may help Viscovery users who are not sure of the number of clusters to determine.

6.3 Self-Organizing Maps: the State Sponsored Murder Data Set

As we discussed in the previous section, Viscovery has useful features such as helping determine the number of clusters and recovering cluster structures of data sets. In this section, we apply Viscovery SOM-Ward procedure to a state sponsored murder (also called genocide and politicide) data set. Genocides and politicides refer to actions committed by governing elites or, in the case of civil war, either of the contending authorities that are intended to destroy a national, ethnical, racial, religious, or political group (Harff, 2003).

The genocide and politicide data set examined in this section includes 28 historical cases of genocide and politicide that began between 1955 and 2002 in independent countries with populations greater than 500,000. The genocide and politicide data were originally studied for identifying independent variables (risk factors or pre-conditions) to distinguish countries that have genocides and politicides from those that do not (Harff, 2003). It consists of 25 countries that have or have no prior genocides

and politicides and six variables that are identified as risk factors contributing to genocide and politicide. The genocide and politicide data set under consideration is given in Table 6.6.

Among the six risk factors (variables) listed, upheaval in political context is defined as an abrupt change in the political community resulted from the formation of a state or regime through violent actions, defeat in international war, or rewriting of state boundaries. Minority elite is designed to reveal the information about interethnic disputes over access to political power. A positive value under the column of 'Minority Elite' (i.e., 'Yes') indicates that elite ethnicity is a recurring issue of political conflicts, which possibly leads to genocide or politicide. Exclusionary ideology refers to a belief system that establishes some cardinal principle that maintains efforts to restrict, persecute, or eliminate certain categories of people. Elite with 'exclusionary ideology' is more apt to eliminate groups. The type of regime is another risk factor that has vital intervening effects to cause genocide and politicide. Elite in autocratical regime is likely to opt for restricting citizens' participation, especially political opponents' participation. The level of trade openness is also an indicator of genocide and politicide. Historical records have shown that armed conflicts and adverse regime changes are more likely to occur in poor countries, especially those countries in Africa and Asia.

Countries in this data set are listed according to their number of positive risk factors. If a country has many positive risk factors (i.e., six or five risk factors), it will be ranked high. For example, Iraq is the only country with six positive risk factors, which indicates that Iraq has the greatest potential to have future genocides and politicides and therefore Iraq is listed in the first place in the data set. By comparison,

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness
Iraq	Yes	High	Yes	Yes	Autocracy	Very Low
Afghanistan (2000)	Yes	Very High	Yes	Yes	Autocracy	Very Low
Afghanistan (2002)	Yes	Very high	No	No	No effective regime	Very Low
Burma	Yes	High	No	Yes	Autocracy	Very Low
Burundi	Yes	Very High	Yes	No	Autocracy	Low
Rwanda	Yes	High	Yes	No	Autocracy	Low
Congo-Kinshasa	Yes	Very High	Yes	No	No effective regime	Medium
Somalia	Yes	Very High	No	No	No effective regime	Very Low
Sierra	No	Very High	Yes	No	No effective regime	Low
Ethiopia	Yes	High	Yes	No	Autocracy	Medium
Uganda	Yes	High	No	No	Autocracy	Low
Algeria	Yes	Very High	No	Yes	Autocracy	Medium
Liberia	No	High	No	No	Autocracy	Low
Pakistan	Yes	Medium	No	No	Autocracy	Low
China	Yes	Medium	No	Yes	Autocracy	Medium
Sri Lanka	Yes	High	No	No	Partial democracy	High
Philippines	Yes	Very High	No	No	Democracy	High
Colombia	No	Very High	No	No	Partial democracy	Low
Turkey	No	High	No	Yes	Partial democracy	Medium
India	No	High	No	No	Democracy	Low
Israel	No	Very High	No	Yes	Democracy	High
Indonesia	Yes	Medium	No	No	Partial democracy	Medium
Russia	Yes	Low	No	No	Partial democracy	Medium
Nigeria	No	Low	No	No	Partial democracy	High
Nepal	No	Medium	No	No	Partial democracy	Medium
Macedonia	No	None	No	No	Partial democracy	High

Table 6.6 Genocide and politicide data set from Harff (2003).

Macedonia is very unlikely to have future genocides and politicides according to the fact that Macedonia has no positive risk factors.

We apply Viscovery to the genocide and politicide data set to detect clusters of countries that have similar pre-conditions that may result in future genocides and politicides. Viscovery needs numerical values as input, so that the categorical values in

the genocide and politicide data set need to be transformed to numerical values. We transformed the categorical values using a commonly used approach that assigns sorted numerical values to categorical values based on the description of the categorical values (Ritter & Kohonen, 1989). The transformed data are given in Table 6.7. The notation for the transformed genocide and politicide data set is given in Table 6.8. For example, in the second column (prior genocides and politicides), 'No' is assigned by 0 and 'Yes' is assigned by 1.

We input the transformed values into Viscovery and used the software's default settings. We obtained the map shown in Figure 6.2. This map has five clusters of countries. These five clusters are formed based on six variables. For example, Macedonia, Russia, Nigeria, Nepal, Indonesia, and Sri Lanka are grouped into the same cluster. These six countries share several common characteristics. For example, they all are partial democratic countries. Half of them (Macedonia, Nigeria, and Sri Lanka) have a high level of trade openness. Countries in this cluster have relatively infrequent political upheavals except Sri Lanka. Half of this group (i.e., Russia, Indonesia and Sri Lanka) has prior genocides or politicides. Overall, this cluster can be viewed as a group of countries where genocide and politicide is less likely to take place.

India, Colombia, Philippines, Israel and Turkey are clustered together. Countries in this cluster almost have no prior genocides or politicides except that the Philippines has one. Members of this cluster have frequent political upheavals, as their levels of political upheavals are either 'High or Very High'.

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness
Iraq	1	3	1	1	1	0
Afghanistan (2000)	1	4	1	1	1	0
Afghanistan (2002)	1	4	0	0	0	0
Burma	1	3	0	1	1	0
Burundi	1	4	1	0	1	1
Rwanda	1	3	1	0	1	1
Congo-Kinshasa	1	4	1	0	0	2
Somalia	1	4	0	0	0	0
Sierra	0	4	1	0	0	1
Ethiopia	1	3	1	0	1	2
Uganda	1	3	0	0	1	1
Algeria	1	4	0	1	1	2
Liberia	0	3	0	0	1	1
Pakistan	1	2	0	0	1	1
China	1	2	0	1	1	2
Sri Lanka	1	3	0	0	2	3
Philippines	1	4	0	0	3	3
Colombia	0	4	0	0	2	1
Turkey	0	3	0	1	2	2
India	0	3	0	0	3	1
Israel	0	4	0	1	3	3
Indonesia	1	2	0	0	2	2
Russia	1	1	0	0	2	2
Nigeria	0	1	0	0	2	3
Nepal	0	2	0	0	2	2
Macedonia	0	0	0	0	2	3

Table 6.7 Transformed genocide and politicide data.

Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness
0 = No 1 = Yes	0 = None 1 = Low 2 = Medium 3 = High 4 = Very high	0 = No 1 = Yes	0 = No 1 = Yes	0 = No effective regime 1 = Autocracy 2 = Partial democracy 3 = Democracy	0 = Very low 1 = Low 2 = Medium 3 = High

Table 6.8 Notation for the transformed genocide and politicide data set.

Figure 6.2 Resulting SOM map of the genocide and politicide data set.

The level of trade openness of this group of countries is relatively high among the five clusters: more than half of member countries in this group have active trade openness. In addition, countries in this cluster are either democratic or partially democratic. Summaries of the remaining clusters are given in Table 6.9.

The genocide and politicide data set we examine in this section includes the six variables given in Table 6.6. Some researchers suggest including a country's per capita income, which they claim is the best predictor of the ethnic insurgencies and civil wars and which underlies Harff's work. We consider gross domestic product (GDP) per capita and number of prior genocides and politicides. GDP per capita is a purchasing power parity

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness
Cluster 1						
Macedonia	0	0	0	0	2	3
Nigeria	0	1	0	0	2	3
Russia	1	1	0	0	2	2
Nepal	0	2	0	0	2	2
Indonesia	1	2	0	0	2	2
Sri Lanka	1	3	0	0	2	3
Cluster 2						
Pakistan	1	2	0	0	1	1
Uganda	1	3	0	0	1	1
Somalia	1	4	0	0	0	0
Afghanistan (2002)	1	4	0	0	0	0
Liberia	0	3	0	0	1	1
Cluster 3						
Sierra	0	4	1	0	0	1
Congo-Kinshasa	1	4	1	0	0	2
Ethiopia	1	3	1	0	1	2
Burundi	1	4	1	0	1	1
Rwanda	1	3	1	0	1	1
Cluster 4						
India	0	3	0	0	3	1
Colombia	0	4	0	0	2	1
Philippines	1	4	0	0	3	3
Israel	0	4	0	1	3	3
Turkey	0	3	0	1	2	2
Cluster 5						
China	1	2	0	1	1	2
Algeria	1	4	0	1	1	2
Burma	1	3	0	1	1	0
Iraq	1	3	1	1	1	0
Afghanistan (2000)	1	4	1	1	1	0

Table 6.9 Cluster profiles of the genocide and politicide data set.

basis divided by population. In the modified data set M1 given in Table 6.10, we include GDP per capita. In Table 6.11, we give modified data set M2 that includes GDP per capita and number of prior genocides and politicides. The GDP per capita of each country can be found at www.cia.gov. Of these 25 countries, Israel has the highest GDP per capita ($19000) while Somalia has the lowest GDP per capita ($550). Due to large differences in GDP per capita among the countries, it is necessary to scale these values and place them into several categories. Each GDP value is divided by the maximum GDP value ($19000) and then classified into one of five categories according to its scaled GDP value. The categories are given in Table 6.12. For example, the scaled GDP value of Iraq is 0.1263 (2400/19000), which is given 1 in the transformed modified genocide and politicide data sets shown in Tables 6.13 and 6.14. The corresponding visual maps generated by Viscovery are shown in Figures 6.3 and 6.4. The associated significance values (or cluster indicators) are 62, 38, and 55 for data sets O, M1, and M2 respectively. We point out that significance values are recommended by Viscovery to help determine the appropriate number of clusters of a data set. The larger the significance values, the better the choice of a particular number of clusters. The cluster profiles of the data sets M1 and M2 are given in Tables 6.15 and 6.16.

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita (US$)
Iraq	Yes	High	Yes	Yes	Autocracy	Very Low	2400
Afghanistan (2000)	Yes	Very High	Yes	Yes	Autocracy	Very Low	700
Afghanistan (2002)	Yes	Very high	No	No	No effective regime	Very Low	700
Burma	Yes	High	No	Yes	Autocracy	Very Low	1660
Burundi	Yes	Very High	Yes	No	Autocracy	Low	600
Rwanda	Yes	High	Yes	No	Autocracy	Low	1200
Congo-Kinshasa	Yes	Very High	Yes	No	No effective regime	Medium	610
Somalia	Yes	Very High	No	No	No effective regime	Very Low	550
Sierra	No	Very High	Yes	No	No effective regime	Low	580
Ethiopia	Yes	High	Yes	No	Autocracy	Medium	750
Uganda	Yes	High	No	No	Autocracy	Low	1260
Algeria	Yes	Very High	No	Yes	Autocracy	Medium	5300
Liberia	No	High	No	No	Autocracy	Low	1100
Pakistan	Yes	Medium	No	No	Autocracy	Low	2100
China	Yes	Medium	No	Yes	Autocracy	Medium	4400
Sri Lanka	Yes	High	No	No	Partial democracy	High	3700
Philippines	Yes	Very High	No	No	Democracy	High	4200
Colombia	No	Very High	No	No	Partial democracy	Low	6500
Turkey	No	High	No	Yes	Partial democracy	Medium	7000
India	No	High	No	No	Democracy	Low	2540
Israel	No	Very High	No	Yes	Democracy	High	19000
Indonesia	Yes	Medium	No	No	Partial democracy	Medium	3100
Russia	Yes	Low	No	No	Partial democracy	Medium	9300
Nigeria	No	Low	No	No	Partial democracy	High	875
Nepal	No	Medium	No	No	Partial democracy	Medium	1400
Macedonia	No	None	No	No	Partial democracy	High	5000

Table 6.10 Modified genocide and politicide data set 1 (M1).

Country Name	Number of Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita (US$)
Iraq	2	High	Yes	Yes	Autocracy	Very Low	2400
Afghanistan (2000)	1	Very High	Yes	Yes	Autocracy	Very Low	700
Afghanistan (2002)	1	Very high	No	No	No effective regime	Very Low	700
Burma	1	High	No	Yes	Autocracy	Very Low	1660
Burundi	3	Very High	Yes	No	Autocracy	Low	600
Rwanda	2	High	Yes	No	Autocracy	Low	1200
Congo-Kinshasa	2	Very High	Yes	No	No effective regime	Medium	610
Somalia	1	Very High	No	No	No effective regime	Very Low	550
Sierra	0	Very High	Yes	No	No effective regime	Low	580
Ethiopia	1	High	Yes	No	Autocracy	Medium	750
Uganda	2	High	No	No	Autocracy	Low	1260
Algeria	1	Very High	No	Yes	Autocracy	Medium	5300
Liberia	0	High	No	No	Autocracy	Low	1100
Pakistan	2	Medium	No	No	Autocracy	Low	2100
China	3	Medium	No	Yes	Autocracy	Medium	4400
Sri Lanka	1	High	No	No	Partial democracy	High	3700
Philippines	1	Very High	No	No	Democracy	High	4200
Colombia	0	Very High	No	No	Partial democracy	Low	6500
Turkey	0	High	No	Yes	Partial democracy	Medium	7000
India	0	High	No	No	Democracy	Low	2540
Israel	0	Very High	No	Yes	Democracy	High	19000
Indonesia	2	Medium	No	No	Partial democracy	Medium	3100
Russia	2	Low	No	No	Partial democracy	Medium	9300
Nigeria	0	Low	No	No	Partial democracy	High	875
Nepal	0	Medium	No	No	Partial democracy	Medium	1400
Macedonia	0	None	No	No	Partial democracy	High	5000

Table 6.11 Modified genocide and politicide data set 2 (M2).

Prior Genocides or Politicides	0 = No; 1 = Yes
Number of Prior Genocides or Politicides	Actual number of genocides or politicides
Upheaval	0 = None; 1 = Low; 2 = Medium; 3 = High; 4 = Very high
Minority Elite	0 = No; 1 = Yes
Exclusionary Ideology	0 = No; 1 = Yes
Type of Regime	0 = No effective regime; 1 = Autocracy; 2 = Partial democracy; 3 = Democracy
Trade Openness	0 = Very low; 1 = Low; 2 = Medium; 3 = High
GDP per Capita	0 = if the scaled value < 0.1; 1 = if the scaled value >0.1 and < 0.2; 2 = if the scaled value >0.2 and < 0.3; 3 = if the scaled value >0.3 and < 0.4; 4 = if the scaled value >0.4

Table 6.12 Notation of the transformed genocide and politicide data with added variables.

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita (US$)
Iraq	1	3	1	1	1	0	1
Afghanistan (2000)	1	4	1	1	1	0	0
Afghanistan (2002)	1	4	0	0	0	0	0
Burma	1	3	0	1	1	0	0
Burundi	1	4	1	0	1	1	0
Rwanda	1	3	1	0	1	1	0
Congo-Kinshasa	1	4	1	0	0	2	0
Somalia	1	4	0	0	0	0	0
Sierra	0	4	1	0	0	1	0
Ethiopia	1	3	1	0	1	2	0
Uganda	1	3	0	0	1	1	0
Algeria	1	4	0	1	1	2	2
Liberia	0	3	0	0	1	1	0
Pakistan	1	2	0	0	1	1	1
China	1	2	0	1	1	2	2
Sri Lanka	1	3	0	0	2	3	1
Philippines	1	4	0	0	3	3	2
Colombia	0	4	0	0	2	1	3
Turkey	0	3	0	1	2	2	3
India	0	3	0	0	3	1	1
Israel	0	4	0	1	3	3	4
Indonesia	1	2	0	0	2	2	1
Russia	1	1	0	0	2	2	4
Nigeria	0	1	0	0	2	3	0
Nepal	0	2	0	0	2	2	0
Macedonia	0	0	0	0	2	3	2

Table 6.13 Transformed modified genocide and politicide data set 1 (M1).

Country Name	Number of Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita (US$)
Iraq	2	3	1	1	1	0	1
Afghanistan (2000)	1	4	1	1	1	0	0
Afghanistan (2002)	1	4	0	0	0	0	0
Burma	1	3	0	1	1	0	0
Burundi	3	4	1	0	1	1	0
Rwanda	2	3	1	0	1	1	0
Congo-Kinshasa	2	4	1	0	0	2	0
Somalia	1	4	0	0	0	0	0
Sierra	0	4	1	0	0	1	0
Ethiopia	1	3	1	0	1	2	0
Uganda	2	3	0	0	1	1	0
Algeria	1	4	0	1	1	2	2
Liberia	0	3	0	0	1	1	0
Pakistan	2	2	0	0	1	1	1
China	3	2	0	1	1	2	2
Sri Lanka	1	3	0	0	2	3	1
Philippines	1	4	0	0	3	3	2
Colombia	0	4	0	0	2	1	3
Turkey	0	3	0	1	2	2	3
India	0	3	0	0	3	1	1
Israel	0	4	0	1	3	3	4
Indonesia	2	2	0	0	2	2	1
Russia	2	1	0	0	2	2	4
Nigeria	0	1	0	0	2	3	0
Nepal	0	2	0	0	2	2	0
Macedonia	0	0	0	0	2	3	2

Table 6.14 Transformed modified genocide and politicide data set 2 (M2).

Figure 6.3 Resulting SOM map of the first modified data set (M1).

Figure 6.4 Resulting SOM amp of the second modified data set (M2).

Country Name	Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita US$
Cluster 1							
Israel	0	4	0	1	3	3	4
Turkey	0	3	0	1	2	2	3
Colombia	0	4	0	0	2	1	3
Cluster 2							
Indonesia	1	2	0	0	2	2	1
Russia	1	1	0	0	2	2	4
Sri Lanka	1	3	0	0	2	3	1
Philippines	1	4	0	0	3	3	2
Cluster 3							
Nigeria	0	1	0	0	2	3	0
Nepal	0	2	0	0	2	2	0
Macedonia	0	0	0	0	2	3	2
India	0	3	0	0	3	1	1
Liberia	0	3	0	0	1	1	0
Cluster 4							
Pakistan	1	2	0	0	1	1	1
Uganda	1	3	0	0	1	1	0
Afghanistan (2002)	1	4	0	0	0	0	0
Somalia	1	4	0	0	0	0	0
Cluster 5							
Algeria	1	4	0	1	1	2	2
China	1	2	0	1	1	2	2
Burma	1	3	0	1	1	0	0
Iraq	1	3	1	1	1	0	1
Afghanistan (2000)	1	4	1	1	1	0	0
Cluster 6							
Rwanda	1	3	1	0	1	1	0
Burundi	1	4	1	0	1	1	0
Ethiopia	1	3	1	0	1	2	0
Congo-Kinshasa	1	4	1	0	0	2	0
Sierra	0	4	1	0	0	1	0

Table 6.15 Cluster profiles of the first modified data set (M1).

Country Name	Number of Prior Genocides or Politicides	Upheaval	Minority Elite	Exclusionary Ideology	Type of Regime	Trade Openness	GDP per Capita US$
Cluster 1							
Russia	2	1	0	0	2	2	4
Nigeria	0	1	0	0	2	3	0
Nepal	0	2	0	0	2	2	0
Macedonia	0	0	0	0	2	3	2
Cluster 2							
Colombia	0	4	0	0	2	1	3
Israel	0	4	0	1	3	3	4
Turkey	0	3	0	1	2	2	3
India	0	3	0	0	3	1	1
Sri Lanka	1	3	0	0	2	3	1
Philippines	1	4	0	0	3	3	2
Cluster 3							
Algeria	1	4	0	1	1	2	2
Iraq	2	3	1	1	1	0	1
Afghanistan (2000)	1	4	1	1	1	0	0
Burma	1	3	0	1	1	0	0
China	3	2	0	1	1	2	2
Cluster 4							
Burundi	3	4	1	0	1	1	0
Rwanda	2	3	1	0	1	1	0
Congo-Kinshasa	2	4	1	0	0	2	0
Sierra	0	4	1	0	0	1	0
Ethiopia	1	3	1	0	1	2	0
Cluster 5							
Indonesia	2	2	0	0	2	2	1
Pakistan	2	2	0	0	1	1	1
Uganda	2	3	0	0	1	1	0
Afghanistan (2002)	1	4	0	0	0	0	0
Liberia	0	3	0	0	1	1	0
Somalia	1	4	0	0	0	0	0

Table 6.16 Cluster profiles of the second modified data set (M2).

The number of clusters in each map is either five or six, and the corresponding significance values are large. This suggests that five or six clusters may be good enough to represent the actual cluster structure in the data sets. Although visual maps (Figures 6.3 and 6.4) of the two modified data sets (M1, and M2) are different from the map in Figure 6.2 of the original data set (O), they still have some elements in common.

The cluster made up of China, Algeria, Burma, Iraq, and Afghanistan (2002) appears in all three maps, as does the cluster of Sierra, Congo, Ethiopia, Burundi, and Rwanda. Cluster membership of other countries is to some extent similar in these clustering results. It indicates that the two new variables do not significantly affect the clustering structure of observed countries. This is possibly due to the small size of the genocide and politicide data set. The six variables listed in Harff's work may be enough to determine the clustering structure of the data set and may also be sufficient to forecast the future genocides or politicides.

However, in Harff's listing, Iraq was the country most likely to have future genocides, or politicides, and followed by Afghanistan, Burma, Burundi, Rwanda, Congo and Somalia. Algeria and China were not close to Iraq in the list, while in our clustering results these two countries are always in the same cluster with Iraq and countries such as Afghanistan and Burma. If from the perspective of clustering, Algeria and China would be ranked close to Iraq, Burma, and Afghanistan, rather than several ranks down from them in the list.

There might be several factors causing differences between our clustering results and Harff's results. One could be the weight assigned to each variable during

preprocessing, where all variables are equally weighted. Another reason might be lack of further domain knowledge.

Comparing our clustering results with Harff's work, we conclude that countries in the same cluster are equally likely to have future genocides or politicides. This is different from what we see in Tables 6.6 and 6.7, which has the forecasted ordering of countries for possible future genocides or politicides (Harff, 2003).

Our current result might be helpful to future genocide and politicide research. The variables 'GDP per capital' and 'number of prior genocides and politicides' have been shown to be of little influence in determining the clustering structures of the data sets, though possibly due to the small size of the data sets. This is not in favor of other scholars' suggestion that per capita income is the best indicator of genocides and politicides, and therefore somehow reinforces Harff's conclusion regarding identification of risks factors. Moreover, the listing differences may raise research questions such as if clustering offers a reasonable alternative view of the genocide and politicide data set, and to what extent will clustering contribute to forecasting countries' genocides and politicides in the coming years.

Chapter 7

Self-Organizing Maps: Best Values in Colleges

The prospect of graduating with a lot of debt may be one of the many worries that concern new college students. It becomes necessary for those college applicants and their parents to critically screen colleges to see which colleges give students the best value for their money. *Kiplinger's Personal Finance* gave a best value list of 100 public colleges (Kiplinger, 2003) and a best value list of 100 private colleges (Kiplinger, 2004) that combine great academics and reasonable costs. In each list, the colleges judged to be the best value is ranked number one with lower rank values preferred.

As we discussed in previous chapters, SOMs are capable of discovering the hidden structures to help a decision maker understand a data set. SOMs have been widely used to cluster data and gain insight into data sets using visual maps. In this chapter we will apply Viscovery SOMine (2002) to analyze public and private colleges to determine which colleges are the best values and compare our rankings to those given by Kiplinger.

7.1 Data Description and Preprocessing

We used two data sets from *Kiplinger*'s *Personal Finance* (2003 and 2004): a public college data set and a private college data set. The public college data set contains the top 100 public colleges considered to be the best values. Similarly, the private college data set contains the top 100 private colleges considered to be the best values. A total of 11 variables are included in each data set and seven of them are common to both data sets. The seven common variables are listed in Table 7.1. The four additional variables are listed in Table 7.2.

Except for the variable Enrollment, other variables in the two data sets indicate either the academic quality or the financial cost of a school. Typically, college applicants look for suitable colleges in terms of both academic quality and financial cost. The variable Enrollment tells us the number of students who were enrolled in the current academic year. College applicants can get a rough idea of how many students may be admitted by each college. However, the variable Enrollment does not provide us the percentage of applicants who may be admitted – which is known as Admission Rate. The variable Admission Rate reflects a school's academic quality and it affects applicants' decisions. Typically, a low admission rate implies a high academic quality. Compared to the variable Enrollment, the variable Admission Rate speaks about schools' academic quality more directly. Therefore, we included the variable Admission Rate in our datasets and didn't include the variable Enrollment. In addition, we did not include the variable Aid from Grants and Cost after Non-need-based Aid in the private college data. Values in the Aid from Grants variable tell us that at least 70% students in 83 out of 100 private colleges were awarded aid from grants, and at least 50% students were awarded in

Variable	Description
Enrollment	Number of full-time and part-time undergraduates enrolled at the college during 2002-2003 academic year
Admission Rate	Percentage of applicants who were offered admission
SAT or ACT	Percentage of the 2002-03 freshman class that scored above 600 on the verbal and math parts, separated by slash, or the percentage that scored above 24 on the ACT
Student/Faculty Ratio	Average number of students for each faculty member
4-yr. Graduate Rate	Percentage of 1996-97 freshmen who earned a bachelor's degree in four years or fewer
6-yr. Graduate Rate	Percentage of 1996-97 freshmen who earned a bachelor's degree within six years
Average Debt at Graduation	Average debt a student accumulates before graduation

Table 7.1 Seven common variables in Kiplinger's public and private data sets.

Data Set	Variable	Description
Public College	In-State Total Costs	Overall cost for residents
	In-State Costs After Aid	Overall cost for residents after subtracting average need-based award
	Total Out-of-State Costs	Overall cost for out-of-state students
	Out-of-State Costs after Aid	Overall cost for out-of-state students after subtracting average need-based award
Private College	Total Cost	Overall cost for college students
	Cost After Need-based Aid	Total cost in 2003-04 academic year after subtracting the average need-based award
	Aid from Grants	Percentage of the average aid package that came from grants or scholarships
	Cost after Non-need-based Aid	2003-04 cost for a student after subtracting non-need-based award

Table 7.2 Four additional variables in Kiplinger's public and private data sets.

98% private colleges. These statistics indicate that the majority of private schools are able to support many of their students. Therefore, the variable Aid from Grants may not help discover private colleges that are the best values. We did not include the variable Cost after Non-need-based Aid in the private college dataset because the Cost after Non-need-based Aid is not available in more than 25% of 100 schools in the data set. Other variables such as Total Cost and Cost after Need-based Aid provide applicants financial cost information.

After selecting appropriate variables for each data set, we show the public and private college data in Tables 7.3 and 7.4, respectively.

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
1	University of North Carolina at Chapel Hill	65/76	35	14	65	79	11,290	6,673	23,138	16,465	11,156
2	University of Virginia	77/84	39	16	81	91	12,640	8,737	28,610	19,873	13,536
3	College of William and Mary	83/85	35	12	80	89	13,024	6,668	27,724	21,056	19,762
4	University of Georgia	54/59	65	13	46	66	10,534	5,245	21,310	16,065	12,906
5	University of Florida	58/67	58	21	49	77	10,611	6,125	21,639	15,514	14,449
6	New College of Florida	93/74	65	11	47	72	10,947	5,613	24,185	18,572	16,645
7	Georgia Institute of Technology	75/95	59	14	18	69	11,340	6,022	23,266	17,244	17,221
8	University of Illinois at Urbana, Champaign	79	60	13	52	76	14,410	8,045	25,446	17,401	14,791
9	Truman State University	82	79	15	39	62	10,609	7,604	14,409	6,805	14,382
10	Virginia Polytechnic Institute and State University	39/52	65	15	36	72	10,122	5,613	20,006	14,393	16,229
11	North Carolina State University	39/60	59	15	25	60	10,688	5,508	22,536	17,028	15,476
12	University of Delaware	41/53	48	12	54	72	13,416	7,666	22,946	15,280	13,610
13	University of Wisconsin, Madison	86	71	14	41	77	13,391	7,842	27,401	19,559	15,904
14	University of Michigan, Ann Arbor	83	49	15	61	82	16,671	8,661	33,473	25,463	16,825
15	University of California, San Diego	47/73	41	19	43	78	16,000	9,043	24,105	17,148	13,275
16	University of California, Berkeley	65/78	25	17	48	83	17,265	8,982	25,584	17,301	14,990
17	University of Washington	42/54	68	11	40	70	13,835	6,926	24,991	18,082	14,500
18	New Mexico Institute of Mining and Technology	75	63	13	12	40	9,714	2,708	16,151	9,145	9,500
19	University of Wisconsin, La Crosse	60	65	21	23	58	10,425	7,327	20,102	12,775	14,306
20	University of Texas at Austin	51/65	61	19	39	71	14,391	9,111	20,999	11,888	16,400
21	University of Oklahoma	69	89	21	19	51	10,139	6,618	16,652	10,034	16,886
22	University of Kansas	55	67	19	26	55	9,673	6,266	17,149	10,883	17,347
23	University of North Carolina at Asheville	48/40	67	14	31	48	8,929	6,308	17,754	11,446	14,547
24	State University of New York at Binghamton	51/73	42	19	69	80	13,587	9,214	19,537	10,323	13,915
25	Colorado School of Mines	89	67	12	31	61	13,780	8,532	26,970	18,438	17,500
26	Auburn University	51	83	16	40	68	10,276	7,308	18,736	11,428	18,585

Table 7.3 Kiplinger's (2003) public college data.

27	Colorado State University	56	77	17	29	62	10,689	6,700	21,161	14,461	16,042
28	College of New Jersey	61/71	48	12	59	80	16,686	13,915	21,261	7,346	5,490
29	Michigan State University	57	67	18	31	66	12,743	8,670	22,703	14,033	18,663
30	Appalachian State University	26/30	64	19	31	60	7,913	4,785	16,834	12,049	13,000
31	Iowa State University of Science and Technology	58	89	16	24	62	11,588	8,736	20,930	12,194	17,119
32	State University of New York College at Geneseo	65/73	49	19	67	79	12,840	10,840	18,790	7,950	15,000
33	Texas A&M University	36/49	68	21	27	69	11,899	7,057	18,979	11,922	15,670
34	University of Texas at Dallas	45/59	53	20	30	53	12,075	7,787	19,155	11,368	NA
35	University of North Carolina at Wilmington	18/26	55	16	34	60	9,715	6,375	19,290	12,915	13,583
36	University of Maryland, College Park	64/77	43	13	33	63	16,304	12,223	22,881	10,658	15,566
37	University of California, Los Angeles	62/75	24	17	40	81	17,616	9,975	26,006	16,031	12,775
38	Louisiana State University and Agricultural and Mechanical College	53	77	21	23	58	10,126	7,406	15,426	8,020	17,569
39	University of Tennessee	49	58	18	24	56	11,681	6,706	20,763	14,057	21,689
40	University of Iowa	59	84	15	34	64	11,763	9,460	22,055	12,595	15,335
41	Rutgers, The State University of New Jersey, New Brunswick	39/55	55	14	44	72	16,519	10,044	23,033	12,989	15,270
42	Clemson University	44/61	52	16	35	69	14,618	11,121	22,216	11,095	14,347
43	University of Colorado at Boulder	65	80	16	38	64	11,937	7,877	28,253	20,376	16,737
44	University of Kentucky	54	82	17	27	58	9,432	5,594	16,112	10,518	NA
45	University of Arkansas	59	86	17	20	45	10,706	7,144	17,456	13,894	14,029
46	Mary Washington College	62/49	60	17	65	75	12,287	9,033	20,035	11,002	13,100
47	Oklahoma State University	49	92	19	22	56	10,677	7,478	16,623	9,145	15,580
48	Kansas State University	49	58	20	18	45	9,626	7,104	16,990	9,886	17,000
49	University of Northern Iowa	39	80	16	30	64	10,634	7,632	17,592	9,960	15,786
50	University of Mississippi	46	80	19	29	48	11,257	6,607	16,167	9,560	14,459
51	James Madison University	33/40	58	17	59	78	13,145	9,320	21,367	12,047	11,786
52	University of California, Davis	42/62	63	19	28	75	16,521	10,334	23,814	13,480	13,507

Table 7.3 (Continued).

53	Miami University	84	77	17	61	80	15,833	12,467	25,603	13,136	17,579
54	Purdue University	29/46	76	16	28	64	14,691	8,723	26,311	17,588	15,677
55	Mississippi State University	50	74	16	19	48	11,130	8,174	16,036	7,862	15,081
56	University of Nebraska, Lincoln	55	90	19	15	51	10,687	7,277	18,269	10,992	15,682
57	Florida State University	34/39	70	22	38	64	10,994	7,035	22,022	14,987	16,372
58	University of California, Irvine	24/55	57	18	34	72	15,635	8,507	22,450	13,944	12,513
59	University of Missouri, Columbia	68	88	18	32	65	13,208	8,138	22,655	14,517	17,137
60	University of Minnesota, Morris	61	82	14	50	76	13,477	8,427	13,477	5,050	9,208
61	University of Alabama	46	85	18	31	61	11,197	8,136	18,357	10,221	18,978
62	University of South Carolina	29/36	70	17	31	58	11,795	8,794	21,133	12,339	15,260
63	St. Mary's College of Maryland	67/60	59	12	58	67	16,908	12,908	23,228	10,320	17,125
64	Michigan Technological University	68	92	11	22	63	14,135	9,339	25,025	15,686	15,711
65	University of California, Santa Barbara	38/55	51	19	44	73	16,154	10,231	24,246	14,015	NA
66	University of Minnesota, Twin Cities Campus	64	74	15	17	53	13,910	7,902	25,540	17,638	NA
67	Mississippi University for Women	56	65	13	21	43	8,649	8,649	13,316	4,667	13,500
68	California Polytechnic State University, San Luis Obispo	39/64	39	19	17	66	11,781	10,429	15,268	4,839	12,842
69	University of Wyoming	42	95	15	22	54	10,863	6,833	16,713	9,880	18,311
70	George Mason University	26/29	66	16	25	48	11,602	7,597	21,442	13,845	14,143
71	University of Central Florida	30/36	62	24	25	49	10,839	8,353	21,867	13,514	14,927
72	Ohio State University	69	74	14	25	59	15,249	11,315	25,113	13,798	15,011
73	Illinois State University	47	81	19	28	55	12,971	7,199	17,441	10,242	13,921
74	University at Buffalo, The State University of New York	28/41	61	14	32	56	13,422	9,956	19,372	9,416	16,255
75	Salisbury University	26/35	50	17	50	68	12,895	9,506	19,783	10,277	14,773
76	University of Massachusetts Amherst	32/38	58	19	41	61	14,480	9,714	23,333	13,619	15,321
77	University of Vermont	35/38	71	14	48	67	17,116	8,258	30,168	21,910	22,425
78	College of Charleston	47/47	60	14	32	52	14,008	11,214	21,270	10,056	15,135
79	Indiana University Bloomington	27/33	81	20	40	65	13,129	8,704	24,164	15,460	16,930
80	Pennsylvania State University University Park Campus	40/60	57	17	43	80	17,017	12,872	26,639	13,767	17,900

Table 7.3 (Continued).

81	State University of New York at Albany	26/35	56	21	52	66	13,574	9,599	19,524	9,925	15,108
82	University of Arizona	27/33	86	19	29	55	11,163	8,506	19,933	11,427	17,340
83	Towson University	18/26	58	19	30	56	12,776	8,651	20,422	11,771	15,530
84	Rutgers, The State University of New Jersey, Camden	19/24	54	11	21	60	16,189	10,111	22,703	12,592	15,223
85	University of Maryland, Baltimore County	46/63	63	17	28	53	16,516	12,946	23,368	10,422	14,500
86	University of Connecticut	33/42	62	17	23	70	14,413	9,141	25,197	16,056	16,093
87	University of Pittsburgh	48/56	55	17	35	60	17,025	12,723	26,337	13,614	20,154
88	State University of New York College at Fredonia	22/25	53	18	47	66	11,782	8,961	17,732	8,771	12,430
89	Stony Brook University, State University of New York	25/50	54	18	30	51	13,645	9,704	19,595	9,891	15,747
90	State University of New York at New Paltz	30/29	40	17	21	52	11,276	8,276	16,176	7,900	15,000
91	University of Maine	22/29	79	15	29	56	12,780	8,162	21,480	13,318	17,917
92	University of New Hampshire	25/32	77	14	48	71	15,779	13,693	26,139	12,446	20,700
93	University of Missouri, Rolla	83	92	14	10	52	14,326	9,416	23,144	13,728	17,991
94	University of California, Santa Cruz	37/39	80	19	40	64	16,877	9,309	24,250	14,941	13,282
95	Rowan University	22/33	44	14	37	63	15,416	10,632	20,812	10,180	NA
96	University of Illinois at Chicago	43	63	15	9	37	14,299	6,399	23,995	17,596	17,000
97	University of Oregon	30/31	86	18	36	59	12,795	9,586	24,231	14,645	22,783
98	Texas Tech University	23/29	69	20	22	51	12,968	9,878	20,048	10,170	13,805
99	Ohio University	49	75	20	43	70	16,514	13,102	24,737	11,635	15,285
100	UC Riverside	16/34	86	19	39	64	16,751	9,932	24,530	17,711	13,226

Table 7.3 (Continued).

Rank	School Name	Admission Rate (%)	SAT or ACT (%)	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	Total Costs ($)	Cost After Need-Based Aid ($)	Average Debt at Graduation ($)
1	California Institute of Technology	21	99/100	3	71	85	32,682	10,981	10,244
2	Rice University	24	89/92	5	68	89	28,350	14,779	12,705
3	Williams College	23	93/93	8	89	94	36,550	14,737	12,316
4	Swarthmore College	24	94/98	8	86	92	38,676	17,386	12,759
5	Amherst College	18	94/92	9	84	94	38,492	14,453	11,544
6	Webb Institute	42	100/100	7	79	83	8,079	5,579	5,700
7	Yale University	8	96/97	7	88	95	38,432	15,729	19,228
8	Washington and Lee University	31	89/89	11	86	89	30,225	15,452	15,634
9	Harvard University	11	90*/90*	8	86	97	38,831	17,456	10,465
10	Stanford University	13	93/95	7	77	93	38,875	17,746	15,782
11	Princeton University	11	95/97	5	91	97	40,169	18,325	12,000
12	Massachusetts Institute of Technology	16	95/100	6	82	91	39,213	19,609	22,855
13	Pomona College	23	98/97	9	83	88	38,130	17,411	15,600
14	Emory University	42	89/94	7	82	87	37,272	19,657	17,675
15	Columbia University	12	91/93	7	83	93	39,493	17,778	15,331
16	Duke University	25	91/94	11	88	93	40,080	19,996	20,025
17	Davidson College	34	86/89	10	89	91	34,706	21,455	13,697
18	Wellesley College	47	88/89	9	84	88	37,419	17,526	15,697
19	Vassar College	31	93/89	9	81	87	37,870	19,404	17,170
20	Haverford College	32	89/90	8	89	92	38,928	17,826	15,253
21	Northwestern University	33	88/92	7	83	92	38,817	20,376	14,551
22	Bowdoin College	25	87/92	10	83	90	38,663	17,773	15,307
23	University of Pennsylvania	21	91/96	6	83	91	39,040	20,596	20,247
24	Johns Hopkins University	35	85/93	8	81	88	39,188	19,142	13,600
25	Cooper Union	14	81/83	7	57	78	14,652	11,167	9,250
26	Washington University	24	93/98	7	75	86	39,253	20,700	NA
27	Dartmouth College	23	92/96	9	87	95	38,898	19,546	NA
28	Claremont McKenna College	28	89/95	7	82	86	37,730	17,988	16,914
29	University of Notre Dame	34	83/91	12	88	95	35,392	18,011	25,595
30	Colgate University	34	80/86	10	85	89	38,820	18,856	12,984
31	The Colorado College	53	66/70	9	72	79	35,275	16,516	13,500
32	University of Richmond	41	74/83	10	79	84	31,679	17,588	16,115
33	Georgetown University	21	87/89	11	86	91	39,182	24,382	20,000
34	Brown University	17	86/90	8	79	94	40,248	20,838	21,700
35	Carleton College	35	88/89	9	82	86	35,288	21,677	14,543
36	Lafayette College	36	63/78	11	79	84	35,713	15,147	17,380

Table 7.4 Kiplingers' (2004) private college data.

37	Middlebury College	27	93/96	11	81	87	39,532	18,288	21,751
38	Grinnell College	65	87/85	10	78	84	31,460	16,585	13,854
39	Illinois Wesleyan University	48	96	12	76	81	30,780	18,858	17,722
40	Bates College	28	90/91	10	82	87	38,932	18,258	17,045
41	Cornell University	29	85/92	9	82	90	38,974	23,122	15,587
42	Wesleyan University	28	89/92	9	76	81	39,127	21,401	23,753
43	Colby College	33	84/89	11	85	88	38,699	18,168	17,270
44	Bucknell University	39	72/84	12	83	87	36,165	19,165	16,000
45	Kenyon College	52	87/81	9	80	84	36,273	17,905	20,850
46	Centre College	78	89	11	71	73	28,529	15,842	14,300
47	Rhodes College	70	95	11	71	73	30,080	18,899	15,100
48	Macalester College	44	87/88	10	71	77	32,847	16,394	NA
49	Barnard College	34	88/88	10	72	84	37,940	17,826	14,030
50	Brandeis University	42	88/88	8	79	85	39,101	22,257	NA
51	College of the Holy Cross	43	71/76	11	88	90	36,851	23,846	16,063
52	Harvey Mudd College	37	97/100	9	75	83	38,880	22,041	20,219
53	Wake Forest University	41	79/86	10	77	87	36,079	21,196	24,769
54	Bryn Mawr College	50	86/75	9	76	80	37,890	18,609	NA
54	Wheaton College	54	84/83	11	70	84	27,076	17,341	15,864
55	Tufts University	27	81/90	9	81	88	39,173	20,115	15,499
56	Oberlin College	33	87/80	10	63	76	37,688	21,081	13,926
57	Mount Holyoke College	52	80/70	10	75	79	38,668	19,268	14,200
58	Furman University	58	65/70	11	74	81	29,430	16,296	17,741
59	St. Olaf College	73	84	13	71	75	29,879	17,458	18,806
60	Brigham Young University	73	86	18	31	73	9,663	7,621	11,000
61	Lehigh University	44	59/85	11	70	84	35,670	19,123	16,972
62	Smith College	53	72/66	9	76	80	37,937	18,466	19,911
63	Beloit College	70	82	11	60	72	30,264	17,452	14,942
64	Taylor University	78	74	15	71	75	24,723	15,678	15,117
65	Union College	45	56/71	11	75	80	36,455	18,431	15,725
66	Hamilton College	35	77/82	10	79	84	38,463	19,474	16,856
67	DePauw University	61	55/60	11	75	79	32,150	15,531	14,481
68	Hillsdale College	82	74	11	53	71	23,353	13,853	14,500
69	Knox College	72	72	12	67	74	30,894	15,494	16,920
70	University of Southern California	30	79/91	10	51	73	37,968	21,606	20,619
71	Trinity College	36	73/79	9	77	83	38,890	19,667	17,000
72	Trinity University	69	71/80	11	65	75	27,086	16,706	NA
73	Gustavus Adolphus College	77	70	13	72	75	27,820	17,609	17,400
74	Vanderbilt University	46	83/90	9	78	84	38,847	20,971	24,023
75	Whitman College	50	80/81	10	60	71	33,776	21,176	15,000
76	Scripps College	58	85/78	12	63	68	36,500	17,984	12,941
77	Franklin and Marshall College	62	62/71	11	78	83	36,580	20,925	19,656

Table 7.4 (Continued).

78	Saint Louis University	72	72	12	52	67	29,983	16,902	14,989
79	Carnegie Mellon University	38	75/95	11	61	77	38,460	24,689	19,195
80	Lawrence University	68	83	11	58	68	32,875	17,882	18,311
81	Connecticut College	35	83/84	11	75	81	37,057	16,930	17,250
82	Case Western Reserve University	78	74/85	8	49	75	32,802	18,323	21,830
84	Dickinson College	51	65/64	13	74	78	36,600	19,753	17,586
85	Kalamazoo College	73	91	12	60	69	30,917	17,947	20,000
86	Saint John's University	87	66	13	67	74	27,272	19,544	20,680
87	Boston College	34	78/85	13	0	86	37,745	24,470	16,732
88	Reed College	55	95/86	10	45	67	37,900	18,804	16,758
89	Bard College	36	85/67	9	59	71	38,282	20,558	15,400
90	University of Rochester	50	79/87	12	65	76	37,246	20,297	NA
91	New York University	28	87/86	11	65	74	40,105	28,282	21,495
92	Villanova University	47	57/74	13	79	84	36,560	26,463	28,217
93	Skidmore College	46	70/69	11	71	75	38,838	21,023	15,560
94	Rose-Hulman Institute of Technology	65	62/91	13	58	71	32,625	28,677	27,000
95	St. John's College	71	95/75	8	63	71	36,635	21,940	20,753
96	Babson College	48	50/77	13	77	81	38,443	21,316	NA
97	Rhode Island School of Design	32	49/55	11	0	87	34,472	26,447	21,125
98	Rensselaer Polytechnic Institute	70	68/92	17	48	75	39,200	22,360	24,590
99	Sarah Lawrence College	40	79/44	6	51	66	42,121	22,847	14,864
100	The George Washington University	40	68/72	14	62	73	40,240	25,866	NA

Table 7.4 (Continued).

As shown in Tables 7.3 and 7.4, there are some missing values, denoted by NA or "0", in the Average Debt columns. For example, there are five NAs in the Average Debt column in the public college data and nine "0"s in the same column in the private college data. Values in the SAT or ACT fields are inconsistent with values in other fields because of their value format. For example, the SAT or ACT value for George Washington University in the private college data (see Table 7.4) is 68/72, which is not a single value needed for a SOM method.

To meet the numerical input requirements of an SOM method, the data sets have to be preprocessed. In the Average Debt column, we used the average value of the column to replace the missing values. For example, we replaced the missing values with the average value of 15480 in the public college data. The average value of 16957 was used to replace the missing values in the private college data. In the SAT or ACT column, many entries that are not in a single value format (i.e., percentages above 600 on the verbal and math parts of SAT are separated by a slash). We used the average percentage instead of the two individual percentages as input. For example, for George Washington University, the average percentage value of 70 for the two individual percentages, that is, 68 and 72, in the SAT or ACT column was used as input. After preprocessing data in the two data sets, we applied Viscovery on the public and private college data sets to see which groups of colleges are real bargains.

7.2 Discussion of Results

The Viscovery SOM map of the public college data is given in Figure 7.1. There are three clusters separated by solid lines. The summary of the clusters is given in Table 7.5. Cluster membership of each public college is given in Table 7.6.

Cluster A has 18 schools. The average SAT or ACT percentage of schools in cluster A is 70%, which is the highest among three clusters, indicating that these schools have a good academic atmosphere. Other academic quality variables provide the same insight. For example, schools in cluster A have the lowest average admission rate (53%), the highest average student/faculty ratio (15), the highest average 4-year and 6-year

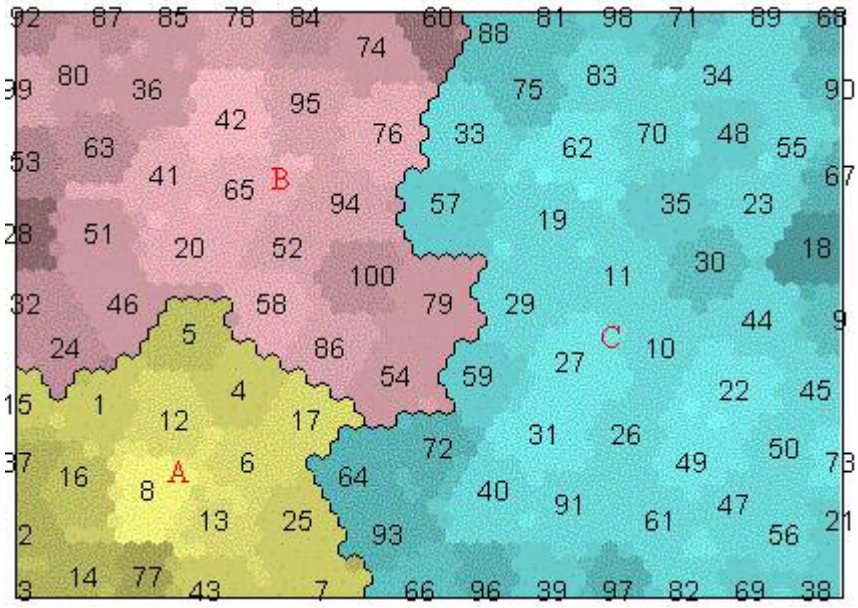

Figure 7.1 Map for the public college data set.

Cluster	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
A	70	53	15	49	75	13,657	7,605	25,845	18,432	15,500
B	48	62	16	43	69	15,238	10,682	22,737	12,158	14,915
C	47	72	17	27	57	11,469	7,737	19,384	11,634	15,798

Table 7.5 Summary of clusters of the public colleges.

Rank	School Name	SAT or ACT (%)	Admis.. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster A											
1	University of North Carolina at Chapel Hill	71	35	14	65	79	11,290	6,673	23,138	16,465	11,156
2	University of Virginia	81	39	16	81	91	12,640	8,737	28,610	19,873	13,536
3	College of William and Mary	84	35	12	80	89	13,024	6,668	27,724	21,056	19,762
4	University of Georgia	57	65	13	46	66	10,534	5,245	21,310	16,065	12,906
5	University of Florida	63	58	21	49	77	10,611	6,125	21,639	15,514	14,449
6	New College of Florida	84	65	11	47	72	10,947	5,613	24,185	18,572	16,645
7	Georgia Institute of Technology	85	59	14	18	69	11,340	6,022	23,266	17,244	17,221
8	University of Illinois at Urbana, Champaign	79	60	13	52	76	14,410	8,045	25,446	17,401	14,791
12	University of Delaware	47	48	12	54	72	13,416	7,666	22,946	15,280	13,610
13	University of Wisconsin, Madison	86	71	14	41	77	13,391	7,842	27,401	19,559	15,904
14	University of Michigan, Ann Arbor	83	49	15	61	82	16,671	8,661	33,473	25,463	16,825
15	University of California, San Diego	60	41	19	43	78	16,000	9,043	24,105	17,148	13,275
16	University of California, Berkeley	72	25	17	48	83	17,265	8,982	25,584	17,301	14,990
17	University of Washington	48	68	11	40	70	13,835	6,926	24,991	18,082	14,500
25	Colorado School of Mines	89	67	12	31	61	13,780	8,532	26,970	18,438	17,500
37	University of California, Los Angeles	69	24	17	40	81	17,616	9,975	26,006	16,031	12,775
43	University of Colorado at Boulder	65	80	16	38	64	11,937	7,877	28,253	20,376	16,737
77	University of Vermont	37	71	14	48	67	17,116	8,258	30,168	21,910	22,425
Average		**70**	**53**	**15**	**49**	**75**	**13,657**	**7,605**	**25,845**	**18,432**	**15,500**

Table 7.6 Cluster profiles of the public colleges.

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster B											
20	University of Texas at Austin	58	61	19	39	71	14,391	9,111	20,999	11,888	16,400
24	State University of New York at Binghamton	62	42	19	69	80	13,587	9,214	19,537	10,323	13,915
28	College of New Jersey	66	48	12	59	80	16,686	13,915	21,261	7,346	5,490
32	State University of New York College at Geneseo	69	49	19	67	79	12,840	10,840	18,790	7,950	15,000
36	University of Maryland, College Park	71	43	13	33	63	16,304	12,223	22,881	10,658	15,566
41	Rutgers, The State University of New Jersey, New Brunswick	47	55	14	44	72	16,519	10,044	23,033	12,989	15,270
42	Clemson University	53	52	16	35	69	14,618	11,121	22,216	11,095	14,347
46	Mary Washington College	56	60	17	65	75	12,287	9,033	20,035	11,002	13,100
51	James Madison University	37	58	17	59	78	13,145	9,320	21,367	12,047	11,786
52	University of California, Davis	52	63	19	28	75	16,521	10,334	23,814	13,480	13,507
53	Miami University	84	77	17	61	80	15,833	12,467	25,603	13,136	17,579
54	Purdue University	38	76	16	28	64	14,691	8,723	26,311	17,588	15,677
58	University of California, Irvine	40	57	18	34	72	15,635	8,507	22,450	13,944	12,513
60	University of Minnesota, Morris	61	82	14	50	76	13,477	8,427	13,477	5,050	9,208
63	St. Mary's College of Maryland	64	59	12	58	67	16,908	12,908	23,228	10,320	17,125
65	University of California, Santa Barbara	47	51	19	44	73	16,154	10,231	24,246	14,015	15,480
74	University at Buffalo, The State University of New York	35	61	14	32	56	13,422	9,956	19,372	9,416	16,255

Table 7.6 (Continued).

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster B	(Continued)										
76	University of Massachusetts Amherst	35	58	19	41	61	14,480	9,714	23,333	13,619	15,321
78	College of Charleston	47	60	14	32	52	14,008	11,214	21,270	10,056	15,135
79	Indiana University Bloomington	30	81	20	40	65	13,129	8,704	24,164	15,460	16,930
80	Pennsylvania State University University Park Campus	50	57	17	43	80	17,017	12,872	26,639	13,767	17,900
84	Rutgers, The State University of New Jersey, Camden	22	54	11	21	60	16,189	10,111	22,703	12,592	15,223
85	University of Maryland, Baltimore County	55	63	17	28	53	16,516	12,946	23,368	10,422	14,500
86	University of Connecticut	38	62	17	23	70	14,413	9,141	25,197	16,056	16,093
87	University of Pittsburgh	52	55	17	35	60	17,025	12,723	26,337	13,614	20,154
92	University of New Hampshire	29	77	14	48	71	15,779	13,693	26,139	12,446	20,700
94	University of California, Santa Cruz	38	80	19	40	64	16,877	9,309	24,250	14,941	13,282
95	Rowan University	28	44	14	37	63	15,416	10,632	20,812	10,180	15,480
99	Ohio University	49	75	20	43	70	16,514	13,102	24,737	11,635	15,285
100	UC Riverside	25	86	19	39	64	16,751	9,932	24,530	17,711	13,226
Average		**48**	**62**	**16**	**43**	**69**	**15,238**	**10,682**	**22,737**	**12,158**	**14,915**

Table 7.6 (Continued).

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster C											
9	Truman State University	82	79	15	39	62	10,609	7,604	14,409	6,805	14,382
10	Virginia Polytechnic Institute and State University	46	65	15	36	72	10,122	5,613	20,006	14,393	16,229
11	North Carolina State University	50	59	15	25	60	10,688	5,508	22,536	17,028	15,476
18	New Mexico Institute of Mining and Technology	75	63	13	12	40	9,714	2,708	16,151	9,145	9,500
19	University of Wisconsin, La Crosse	60	65	21	23	58	10,425	7,327	20,102	12,775	14,306
21	University of Oklahoma	69	89	21	19	51	10,139	6,618	16,652	10,034	16,886
22	University of Kansas	55	67	19	26	55	9,673	6,266	17,149	10,883	17,347
23	University of North Carolina at Asheville	44	67	14	31	48	8,929	6,308	17,754	11,446	14,547
26	Auburn University	51	83	16	40	68	10,276	7,308	18,736	11,428	18,585
27	Colorado State University	56	77	17	29	62	10,689	6,700	21,161	14,461	16,042
29	Michigan State University	57	67	18	31	66	12,743	8,670	22,703	14,033	18,663
30	Appalachian State University	28	64	19	31	60	7,913	4,785	16,834	12,049	13,000
31	Iowa State University of Science and Technology	58	89	16	24	62	11,588	8,736	20,930	12,194	17,119
33	Texas A&M University	43	68	21	27	69	11,899	7,057	18,979	11,922	15,670
34	University of Texas at Dallas	52	53	20	30	53	12,075	7,787	19,155	11,368	15,480
35	University of North Carolina at Wilmington	22	55	16	34	60	9,715	6,375	19,290	12,915	13,583

Table 7.6 (Continued).

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster C	(Continued)										
38	Louisiana State University and Agricultural and Mechanical College	53	77	21	23	58	10,126	7,406	15,426	8,020	17,569
39	University of Tennessee	49	58	18	24	56	11,681	6,706	20,763	14,057	21,689
40	University of Iowa	59	84	15	34	64	11,763	9,460	22,055	12,595	15,335
44	University of Kentucky	54	82	17	27	58	9,432	5,594	16,112	10,518	15,480
45	University of Arkansas	59	86	17	20	45	10,706	7,144	17,456	13,894	14,029
47	Oklahoma State University	49	92	19	22	56	10,677	7,478	16,623	9,145	15,580
48	Kansas State University	49	58	20	18	45	9,626	7,104	16,990	9,886	17,000
49	University of Northern Iowa	39	80	16	30	64	10,634	7,632	17,592	9,960	15,786
50	University of Mississippi	46	80	19	29	48	11,257	6,607	16,167	9,560	14,459
55	Mississippi State University	50	74	16	19	48	11,130	8,174	16,036	7,862	15,081
56	University of Nebraska, Lincoln	55	90	19	15	51	10,687	7,277	18,269	10,992	15,682
57	Florida State University	37	70	22	38	64	10,994	7,035	22,022	14,987	16,372
59	University of Missouri, Columbia	68	88	18	32	65	13,208	8,138	22,655	14,517	17,137
61	University of Alabama	46	85	18	31	61	11,197	8,136	18,357	10,221	18,978
62	University of South Carolina	33	70	17	31	58	11,795	8,794	21,133	12,339	15,260
64	Michigan Technological University	68	92	11	22	63	14,135	9,339	25,025	15,686	15,711
66	University of Minnesota, Twin Cities Campus	64	74	15	17	53	13,910	7,902	25,540	17,638	15,480
67	Mississippi University for Women	56	65	13	21	43	8,649	8,649	13,316	4,667	13,500
68	California Polytechnic State University, San Luis Obispo	52	39	19	17	66	11,781	10,429	15,268	4,839	12,842

Table 7.6 (Continued).

Rank	School Name	SAT or ACT (%)	Admis. Rate	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	In-State Total Costs ($)	In-State Costs After Aid ($)	Total Out-of-State Costs ($)	Out-of-State Costs After Aid ($)	Avg. Debt at Grad. ($)
Cluster C	(Continued)										
69	University of Wyoming	42	95	15	22	54	10,863	6,833	16,713	9,880	18,311
70	George Mason University	28	66	16	25	48	11,602	7,597	21,442	13,845	14,143
71	University of Central Florida	33	62	24	25	49	10,839	8,353	21,867	13,514	14,927
72	Ohio State University	69	74	14	25	59	15,249	11,315	25,113	13,798	15,011
73	Illinois State University	47	81	19	28	55	12,971	7,199	17,441	10,242	13,921
75	Salisbury University	31	50	17	50	68	12,895	9,506	19,783	10,277	14,773
81	State University of New York at Albany	31	56	21	52	66	13,574	9,599	19,524	9,925	15,108
82	University of Arizona	30	86	19	29	55	11,163	8,506	19,933	11,427	17,340
83	Towson University	22	58	19	30	56	12,776	8,651	20,422	11,771	15,530
88	State University of New York College at Fredonia	24	53	18	47	66	11,782	8,961	17,732	8,771	12,430
89	Stony Brook University, State University of New York	38	54	18	30	51	13,645	9,704	19,595	9,891	15,747
90	State University of New York at New Paltz	30	40	17	21	52	11,276	8,276	16,176	7,900	15,000
91	University of Maine	26	79	15	29	56	12,780	8,162	21,480	13,318	17,917
93	University of Missouri, Rolla	83	92	14	10	52	14,326	9,416	23,144	13,728	17,991
96	University of Illinois at Chicago	43	63	15	9	37	14,299	6,399	23,995	17,596	17,000
97	University of Oregon	31	86	18	36	59	12,795	9,586	24,231	14,645	22,783
98	Texas Tech University	26	69	20	22	51	12,968	9,878	20,048	10,170	13,805
Average		**47**	**72**	**17**	**27**	**57**	**11,469**	**7,737**	**19,384**	**11,634**	**15,798**

Table 7.6 (Continued).

graduation rates (49% and 75%) among three clusters. The financial cost is another important concern for students as to choosing schools of the best values. Although the average total costs for out-of-state students of cluster A schools (i.e., Total Out-of-State Cost and Out-of-State Costs after Aid) are the highest, when cluster A schools are compared to cluster B schools and cluster C schools, the average In-State Total Costs ($13657) is the second highest and the average In-State Costs after Aid ($7605) is the lowest. For cluster A, average of the variable Average Debt at Graduation ($15500) is the second highest and it is only about $600 greater than the lowest average debt ($14915) of cluster B schools. Therefore, cluster A schools are the group of schools that not only have excellent education quality but also are financially comparable for students. Perhaps those in-state students who have excellent academic performance are more willing to consider cluster A schools because of the lowest average In-State Costs after Aid.

Cluster C has 52 schools. The overall academic quality of schools in this cluster is worse than schools in the other two clusters. For example, the average SAT or ACT percentage of 47% is the lowest, the 4-year and 6-year graduation rates are the lowest, i.e., 27% and 57% respectively, and the Student/faculty Ratio of 17 is the highest. However, the average total costs for both residents and non-residents are the lowest among the three groups. The average In-State Total Costs is $11469, which is about $2200 less than cluster A schools.

The average Out-of-State Total Costs for cluster B schools is $19384, which is about $6500 less than that of schools in cluster A. The average total costs after aid for in-

state and out-of-state students are also low. It tells that cluster C schools may be good choices for those students who care more about financial costs.

Cluster B has 30 colleges that have good academic quality, i.e., the second highest average values in SAT/ACT percentage, Admission Rate, Student/faculty Ratio, 4-year Graduation Rate, and 6-year Graduation Rate. This cluster of schools has the highest average costs ($15238) for in-state students, the second highest average costs ($22737) for non-resident students, and the lowest Average Debt at Graduation ($14915). Schools in the cluster B may be considered as alternatives to schools in cluster A and schools in cluster C because they are comparable to cluster A schools in terms of good education quality and similar to cluster C schools in terms of less expensive financial costs.

We did not include the Kiplinger's rank variable to generate our SOM maps. The Rank variable is used as school labels in the maps. Fifteen schools in cluster A are from the top 25 public schools in Kiplinger's list, such as University of North Carolina at Chapel Hill, University of Virginia, College of William & Mary, University of Georgia (Table 7.6). Two (i.e., UCLA and University of Colorado-Boulder) of the rest schools come from the middle range, which is between No.26 and No.75, and the last one (University of Vermont) is ranked No.77. Cluster A schools have the highest education quality and require reasonable financial expenditures. Cluster B schools have better academic quality and higher financial expenditures for students than cluster C schools.

When looking at individual schools, we found some schools belonging to the same cluster have different ranks in Kiplinger's list. For example, University of Colorado at Boulder and Colorado School of Mines are in the same cluster A. They have comparable education quality and the financial costs in these two schools are close. In

Kiplinger's ranking, University of Colorado ranks No.43, which is 18 places lower than Colorado school of Mines whose rank is No.25. Additional examples can be found in cluster B schools and cluster C schools. For example, University of California at Riverside and University of Clemson are members of cluster B. However, UC Riverside is ranked No.100 and Clemson is ranked No.42. Although there are some differences between them, most of their educational and financial measures are close. For example, their 4-year and 6-year graduation rates and Average Debt at Graduation are close. Therefore University of Clemson and UC Riverside should have close rankings. In cluster C, University of Maine and Iowa State University have the same situation. Maine is ranked No.31 while Iowa State is ranked No.91 despite their close educational and financial qualities as shown in Table 7.6.

The Viscovery SOM map for the private colleges is given in Figure 7.2. There are five clusters and the summary of the clusters is given in Table 7.7. Cluster memberships of colleges are provided in Table 7.8.

Cluster A has two member colleges: Webb Institute (6) and Copper Union (25). Both colleges have excellent educational qualities: the lowest average Student/faculty Ratio (7), the highest average SAT or ACT percentage (91%) among the five clusters, and the highest average 4-year and 6-year graduation rates (68% and 81% respectively) as well. In addition, the financial costs in both colleges are very low. The average total cost of both schools is $11366, less than half of the second lowest average total cost ($31335) of cluster E schools. The average total cost after need-based aid is the lowest ($8373) among five clusters. The average debt at graduation is the lowest ($7475), which

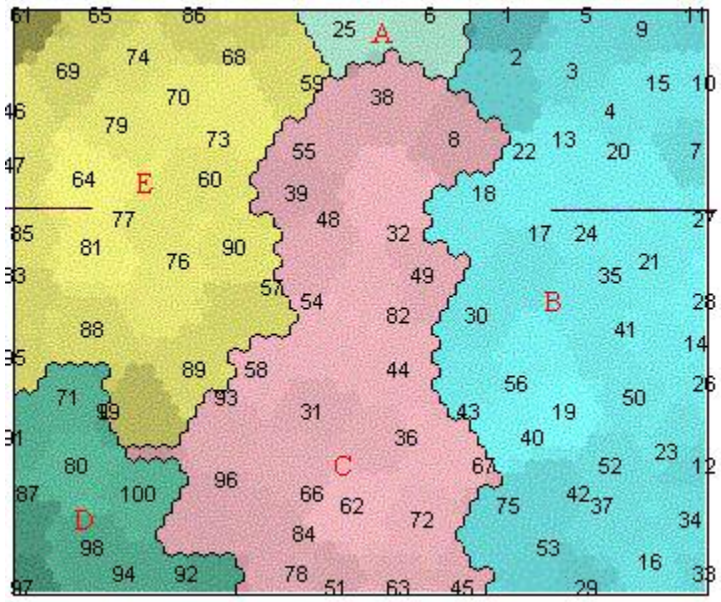

Figure 7.2 Map for the private college data set.

Cluster	Admission Rate (%)	SAT or ACT (%)	Student/ faculty Ratio	4-Year Grad. Rate (%)	6-Year Grad. Rate (%)	Total Costs ($)	Total Costs after Need-based Aid ($)	Avg. Debt at Grad. ($)
A	28	91	7	68	81	11,366	8,373	7,475
B	27	91	8	82	90	38,051	18,949	16,914
C	46	77	11	77	82	35,574	18,567	16,608
D	43	76	13	47	78	37,486	25,429	21,770
E	65	78	11	61	73	31,115	17,749	16,387

Table 7.7 Summary of clusters of the public colleges.

Rank	School Name	Admiss. Rate (%)	SAT or ACT %	Student/ faculty Ratio	4-Year Grad. Rate %	6-Year Grad. Rate %	Total Costs	Cost After Need- Based Aid	Average Debt
Cluster A									
6	Webb Institute	42	100	7	79	83	8,079	5,579	5,700
25	Cooper Union	14	82	7	57	78	14,652	11,167	9,250
Average		**28**	**91**	**7**	**68**	**81**	**11366**	**8373**	**7475**
Cluster B									
1	California Institute of Technology	21	100	3	71	85	32,682	10,981	10,244
2	Rice University	24	91	5	68	89	28,350	14,779	12,705
3	Williams College	23	93	8	89	94	36,550	14,737	12,316
4	Swarthmore College	24	96	8	86	92	38,676	17,386	12,759
5	Amherst College	18	93	9	84	94	38,492	14,453	11,544
7	Yale University	8	97	7	88	95	38,432	15,729	19,228
9	Harvard University	11	90	8	86	97	38,831	17,456	10,465
10	Stanford University	13	94	7	77	93	38,875	17,746	15,782
11	Princeton University	11	96	5	91	97	40,169	18,325	12,000
12	Massachusetts Institute of Technology	16	98	6	82	91	39,213	19,609	22,855
13	Pomona College	23	98	9	83	88	38,130	17,411	15,600
14	Emory University	42	92	7	82	87	37,272	19,657	17,675
15	Columbia University	12	92	7	83	93	39,493	17,778	15,331
16	Duke University	25	93	11	88	93	40,080	19,996	20,025
17	Davidson College	34	88	10	89	91	34,706	21,455	13,697
18	Wellesley College	47	89	9	84	88	37,419	17,526	15,697
19	Vassar College	31	91	9	81	87	37,870	19,404	17,170
20	Haverford College	32	90	8	89	92	38,928	17,826	15,253
21	Northwestern University	33	90	7	83	92	38,817	20,376	14,551
22	Bowdoin College	25	90	10	83	90	38,663	17,773	15,307
23	University of Pennsylvania	21	94	6	83	91	39,040	20,596	20,247
24	Johns Hopkins University	35	89	8	81	88	39,188	19,142	13,600
26	Washington University	24	96	7	75	86	39,253	20,700	16,957
27	Dartmouth College	23	94	9	87	95	38,898	19,546	16,957
28	Claremont McKenna College	28	92	7	82	86	37,730	17,988	16,914
29	University of Notre Dame	34	87	12	88	95	35,392	18,011	25,595
30	Colgate University	34	83	10	85	89	38,820	18,856	12,984
33	Georgetown University	21	88	11	86	91	39,182	24,382	20,000
34	Brown University	17	88	8	79	94	40,248	20,838	21,700
35	Carleton College	35	89	9	82	86	35,288	21,677	14,543

Table 7.8 Cluster profiles of the private college data.

Rank	School Name	Admiss. Rate (%)	SAT or ACT %	Student/ faculty Ratio	4-Year Grad. Rate %	6-Year Grad. Rate %	Total Costs	Cost After Need- Based Aid	Average Debt
Cluster B	(Continued)								
37	Middlebury College	27	95	11	81	87	39,532	18,288	21,751
40	Bates College	28	91	10	82	87	38,932	18,258	17,045
41	Cornell University	29	89	9	82	90	38,974	23,122	15,587
42	Wesleyan University	28	91	9	76	81	39,127	21,401	23,753
43	Colby College	33	87	11	85	88	38,699	18,168	17,270
50	Brandeis University	42	88	8	79	85	39,101	22,257	16,957
52	Harvey Mudd College	37	99	9	75	83	38,880	22,041	20,219
53	Wake Forest University	41	83	10	77	87	36,079	21,196	24,769
56	Tufts University	27	86	9	81	88	39,173	20,115	15,499
75	Vanderbilt University	46	87	9	78	84	38,847	20,971	24,023
Average		**27**	**91**	**8**	**82**	**90**	**38,051**	**18,949**	**16,914**
Cluster C									
8	Washington and Lee University	31	89	11	86	89	30,225	15,452	15,634
31	The Colorado College	53	68	9	72	79	35,275	16,516	13,500
32	University of Richmond	41	79	10	79	84	31,679	17,588	16,115
36	Lafayette College	36	71	11	79	84	35,713	15,147	17,380
38	Grinnell College	65	86	10	78	84	31,460	16,585	13,854
39	Illinois Wesleyan University	48	96	12	76	81	30,780	18,858	17,722
44	Bucknell University	39	78	12	83	87	36,165	19,165	16,000
45	Kenyon College	52	84	9	80	84	36,273	17,905	20,850
48	Macalester College	44	88	10	71	77	32,847	16,394	16,957
49	Barnard College	34	88	10	72	84	37,940	17,826	14,030
51	College of the Holy Cross	43	74	11	88	90	36,851	23,846	16,063
54	Bryn Mawr College	50	81	9	76	80	37,890	18,609	16,957
55	Wheaton College	54	84	11	70	84	27,076	17,341	15,864
58	Mount Holyoke College	52	75	10	75	79	38,668	19,268	14,200
62	Lehigh University	44	72	11	70	84	35,670	19,123	16,972
63	Smith College	53	69	9	76	80	37,937	18,466	19,911
66	Union College	45	64	11	75	80	36,455	18,431	15,725
67	Hamilton College	35	80	10	79	84	38,463	19,474	16,856
72	Trinity College	36	76	9	77	83	38,890	19,667	17,000
78	Franklin and Marshall College	62	67	11	78	83	36,580	20,925	19,656
82	Connecticut College	35	84	11	75	81	37,057	16,930	17,250
84	Dickinson College	51	65	13	74	78	36,600	19,753	17,586
93	Skidmore College	46	70	11	71	75	38,838	21,023	15,560
96	Babson College	48	64	13	77	81	38,443	21,316	16,957
Average		**46**	**77**	**11**	**77**	**82**	**35,574**	**18,567**	**16,608**

Table 7.8 (Continued).

Rank	School Name	Admiss. Rate (%)	SAT or ACT %	Student/ faculty Ratio	4-Year Grad. Rate %	6-Year Grad. Rate %	Total Costs	Cost After Need- Based Aid	Average Debt
Cluster D									
71	University of Southern California	30	85	10	51	73	37,968	21,606	20,619
80	Carnegie Mellon University	38	85	11	61	77	38,460	24,689	19,195
87	Boston College	34	82	13	0	86	37,745	24,470	16,732
91	New York University	28	87	11	65	74	40,105	28,282	21,495
92	Villanova University	47	66	13	79	84	36,560	26,463	28,217
94	Rose-Hulman Institute of Technology	65	77	13	58	71	32,625	28,677	27,000
97	Rhode Island School of Design	32	52	11	0	87	34,472	26,447	21,125
98	Rensselaer Polytechnic Institute	70	80	17	48	75	39,200	22,360	24,590
100	The George Washington University	40	70	14	62	73	40,240	25,866	16,957
Average		**43**	**76**	**13**	**47**	**78**	**37,486**	**25,429**	**21,770**

Table 7.8 (Continued).

Rank	School Name	Admiss. Rate (%)	SAT or ACT %	Student/ faculty Ratio	4-Year Grad. Rate %	6-Year Grad. Rate %	Total Costs	Cost After Need- Based Aid	Average Debt
Cluster E									
46	Centre College	78	89	11	71	73	28,529	15,842	14,300
47	Rhodes College	70	95	11	71	73	30,080	18,899	15,100
57	Oberlin College	33	84	10	63	76	37,688	21,081	13,926
59	Furman University	58	68	11	74	81	29,430	16,296	17,741
60	St. Olaf College	73	84	13	71	75	29,879	17,458	18,806
61	Brigham Young University	73	86	18	31	73	9,663	7,621	11,000
64	Beloit College	70	82	11	60	72	30,264	17,452	14,942
65	Taylor University	78	74	15	71	75	24,723	15,678	15,117
68	DePauw University	61	58	11	75	79	32,150	15,531	14,481
69	Hillsdale College	82	74	11	53	71	23,353	13,853	14,500
70	Knox College	72	72	12	67	74	30,894	15,494	16,920
73	Trinity University	69	76	11	65	75	27,086	16,706	16,957
74	Gustavus Adolphus College	77	70	13	72	75	27,820	17,609	17,400
76	Whitman College	50	81	10	60	71	33,776	21,176	15,000
77	Scripps College	58	82	12	63	68	36,500	17,984	12,941
79	Saint Louis University	72	72	12	52	67	29,983	16,902	14,989
81	Lawrence University	68	83	11	58	68	32,875	17,882	18,311
83	Case Western Reserve University	78	80	8	49	75	32,802	18,323	21,830
85	Kalamazoo College	73	91	12	60	69	30,917	17,947	20,000
86	Saint John's University	87	66	13	67	74	27,272	19,544	20,680
88	Reed College	55	91	10	45	67	37,900	18,804	16,758
89	Bard College	36	76	9	59	71	38,282	20,558	15,400
90	University of Rochester	50	83	12	65	76	37,246	20,297	16,957
95	St. John's College	71	85	8	63	71	36,635	21,940	20,753
99	Sarah Lawrence College	40	62	6	51	66	42,121	22,847	14,864
Average		**65**	**78**	**11**	**61**	**73**	**31,115**	**17,749**	**16,387**

Table 7.8 (Continued).

is nearly $9000 less than the second lowest average debt at graduation ($16387) of the cluster E. Therefore, cluster A schools can be considered as schools of great value.

Cluster B has 40 schools. From the average Admission Rate, the average SAT or ACT, and other academic-related variables (i.e., SAT/ACT, Student/Faculty Ratio, 4-year and 6-year Graduation Rates), we see that most of schools in cluster B have excellent education quality. All of Ivy League schools are included. Most of them are financially expensive, which is shown by the highest average Total Costs ($38051) (see Table 7.7) among the five clusters. However, if the need-based aid is taken into account, the average total cost is reduced to $18949 and the average Average Debt at Graduation is $16914, which makes cluster B schools comparable to schools in clusters C, D, and E. If students care more about educational quality, then schools in cluster B are worthwhile.

There are 24 colleges in cluster C. Compared to cluster B schools, schools in cluster C are financially less expensive. For example, the average Total Cost ($35574) is about $2400 less than that of cluster B schools. However, the education quality of schools in this cluster is not as exceptional as schools in cluster B. For example, in cluster C, the average SAT or ACT percentage is less than the average SAT or ACT percentage of cluster B by 14 points. In addition, the Student/faculty Ratio and, the 4-year and 6-year graduation rates, are all lower than those of cluster B. Therefore, cluster C schools provide relatively good academic quality and ask for reasonable financial sacrifice.

There are nine schools in cluster D. The overall academic quality of these schools is not as good as the overall academic quality of schools in Cluster C, while the financial costs of cluster D schools are much higher than the financial costs of cluster C. As

shown in Table 7.7, the average academic-related measures are slightly lower than those of cluster C schools. For example, the average 6-year graduation rate of cluster D (78%) is a little worse than the average 6-year graduation of cluster C (82%). The average financial costs in cluster D are the highest among the five clusters. For example, the average Cost after Need-based Aid of this cluster is $25429 which is about $6500 higher than the second highest of cluster B ($18949). In terms of educational quality and financial consideration, schools in cluster D might be considered less competitive than schools in clusters A, B, and C.

Cluster E has 25 colleges. Compared to the other four clusters, the academic quality of schools in this cluster is not as good as schools in other four clusters. The average 4-year and 6-year graduation rates are the lowest among five clusters. However, the average financial cost of schools in cluster E is the second lowest. Since the academic quality of cluster E is close to or slightly worse than that of cluster D and the financial cost of cluster E is much lower than that of cluster D, schools in cluster E might be considered alternatives of schools in cluster D for students who have financial concerns.

Although our analysis of cluster structures agrees with Kiplinger's list in most cases, there are some discrepancies, especially when examining individual schools within each cluster. For example, in cluster B, Vanderbilt University has similar academic and financial measures as Wake Forest (53) and Wesleyan (42). Vanderbilt is ranked No. 75 in Kiplinger's list, where Vanderbilt's Kiplinger rank is 23 places and 32 places lower than Wake Forest and Wesleyan, respectively. Barnard College (49) and Connecticut College (82) in cluster C is another example. These two schools have comparable

academic measures and close financial measures (see Table 7.8). However, the difference between their Kiplinger's ranks is 47. This kind of information can not be discovered on Kiplinger's list. With the help of the SOM visual maps, alternatives can easily be found by examining a school's neighbors.

Our results of the public and private college data sets have shown that Viscovery's SOM map helps identify alternatives of a particular school, which may not be detected from Kiplingers' rankings. For example, Figure 7.1 gives us a visual map of the hidden cluster structure of the public college data. On this map, three clusters are clearly visible, where alternatives of a school can be easily recognized. Colorado State (27) has six neighbors: Virginia Polytechnic Institute and State University (10), North Carolina State University (11), Auburn University (26), Michigan State University (29), Iowa State University (31), and University of Missouri at Columbia (59). If we look for Colorado State's alternatives solely on the Kiplinger's list, we are very likely to think about schools with close ranks to Colorado State as its alternatives. However, not every alternative has a close rank to Colorado State, such as North Caroline State. Therefore, Viscovery can help college students identify alternatives, gain more insights from the data, and facilitate them to make a better decision.

The differences between our results and Kiplinger's list do not mean that the Kiplinger's list is of no value, although our analysis shows that the Kiplinger's list can be misleading. If we could combine the results obtained from our maps and Kiplinger's rankings, we might have better way to analyze and explain the data set to help students.

Chapter 8

Summary and Future Work

In order to visualize a data set with an asymmetric distance matrix, the standard SM method takes as inputs the symmetrized distance matrix by averaging entries in the asymmetric distance matrix. This approach is the simplest way to represent asymmetric data onto a 2-dimentional map. However, some interesting information hidden in asymmetric data may be ignored due to averaging. Merino et al.'s method introduces into the standard SM method an asymmetry coefficient that is expected to reflect asymmetric information. However, when the data set under consideration is not very asymmetric, the asymmetry coefficient defined by Merino et al. has little influence on the resulting map. The map obtained from Merino's method is very similar to the one obtained from the standard SM method. Our modified SM method takes into account the upper triangular part and the lower triangular part of an asymmetric distance matrix simultaneously. It is reasonable to expect that the modified SM method may outperform the standard and Merino's method to some extents.

We applied the modified method to two asymmetric data sets: American college selection data and Canadian ranked college data. From the results obtained on these two data sets, we found that the modified SM method always did a fairly good job at reducing

distance errors and performed reasonably well at preserving order relationships at least comparable to the standard and Merino's methods. Since asymmetric proximity data arise in other research areas such as marketing, psychology, sociology, etc, one research problem could be how to use our modified SM method to visualize these data sets. If such maps could be generated with reasonable interpretability, they might be used to discover relationships between data items that may hardly be detected by other methods, and therefore assist the analysis of the asymmetric data sets in different research and business disciplines.

In terms of helping detect hidden structures, clustering has been widely used in the applications of data visualization. We have found that the clustering procedures in Viscovery outperformed the *K*-means clustering method and the classic SOM method. We have applied Viscovery to the state-sponsored murder data set and got some interesting results. Meanwhile, through analyzing 200 public and private colleges with Viscovery, we have generated several clusters for the public college data set and the private college data set, respectively. These college clustering results are not quite similar to Kiplinger's results. In practice, there are lots of ranking lists trying to give readers the idea of which is the best/worst or which is most likely to happen. However, the ranking lists such as Kiplinger's list do not include alternative information that readers want to look for. Maybe in addition to their original ranking lists, SOM maps showing the clustering information should be included as well to better deliver useful information to readers to help make their decisions.

References

Agrafiotis, D.K., "A new method for analyzing protein sequence relationships based on Sammon maps," *Protein Science*, 6(2), 287-293 (1997).

Apostol, I. and Szpankowski, W, "Indexing and mapping of proteins using a modified nonlinear Sammon's projection," *Journal of Computational Chemistry*, 20, 1049-1059 (1999)

Baker, F.B., and Hubert, L.J., "Applications of combinatorial programming to data analysis: Seriation using asymmetric proximity measures," *British Journal of Mathematical and Statistical Psychology*, 30, 154-164 (1977).

Becker, S. and Le Cun, Y., "Improving the convergence of back-propagation learning with second order methods," *Proceedings of the 1988 Connectionists Models Summer School*, Carnegie-Mellon University: Morgan Kaufmann (1989).

Borg, I. and Groenen, P., *Modern Multidimensional Scaling*, New York: Springer-Verlag (1997).

Buja, A., Swayne, D.F., Littman, M., and Dean, N., "XGvis: Interactive data visualization with multidimensional scaling," under review at *Journal of Computational and Graphical Statistics* (1998).

Chien, Y., *Interactive Pattern Recognition*, New York: Marcel Dekker, Inc. (1978).

Condon, E., Golden, B., Lele, S., Raghavan, S., and Wasil, E., "A visualization model based on adjacency data," *Decision Support Systems*, 33 (4), 349-362 (2002).

Cormack, R.M., "A review of classification," *Journal of the Royal Statistical Society* (Series A), 134, Part 3, 321-367 (1971).

Cox, T.F. and Cox, M.A.A., *Multidimensional Scaling*, London: Chapman & Hall (1994).

Cunningham, J.P., "Free trees and bi-directional trees as representations of psychological distance," *Journal of Mathematical Psychology*, 17, 165-188 (1978).

de Ridder, D. and Duin, R.P.W., "Sammon's mapping using neural networks: A comparison," *Pattern Recognition Letters*, 18(11-13), 1307-1316 (1997).

Der, R., Steinmetz, U., Balzuweit, G., Schüürmann, G., "Nonlinear principal component analysis," Technical Report at the Institute für Informatik, University of Leipzig (1998).

Dykes, J. A., "Dynamic maps for spatial science, a unified approach to cartographic visualization," in *Innovations in GIS 3*, Parker, D. (ed.), 177-187, London: Taylor & Francis (1996).

Fiske, E., *The Fiske Guide to Colleges 2000*, New York: Times Books (1999).

Flexer, A., "On the use of self-organizing maps for clustering and visualization," *Principles of Data Mining and Knowledge Discovery*, 80-88 (1999).

Friedman, J.H., "Exploratory projection pursuit," *Journal of the American Statistical Association*, 82, 249-266 (1987).

Friendly, M. and Denis, D., available at www.math.yorku.ca/SCS/Gallery/milestone (2003).

Fyfe, C. and Baddeley, R., "Non-linear data structure extraction using simple Hebbian networks," *Biological Cybernetics*, 72, 533-541 (1995).

Garrido, L., Gaitan, V., Serra-Ricart, M., and Calbert, X., "Use of multilayer feedforward neural nets as a display method for multidimensional distributions," *International Journal of Neural Systems*, 6, 273-282 (1995).

GRG Solver, Frontline Systems, available at www.solver.com, (2004).

Harff, B., "No lessons learned from the holocaust? Assessing risks of genocide and political mass murder since 1955," *American Political Science Review*, 97 (1), 57-73 (2003).

Hastie, T. and Stuetzle, W., "Principal curves," *Journal of the American Statistical Association*, 84, 502-516 (1989).

Hotelling, H., "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, 24, 417-441 (1933).

Hutchinson, J.W., "NETSCAL: A network scaling algorithm for nonsymmetric proximity data," *Psychometrika*, 54, 25-52 (1989).

Johnson, R.A., and Wichern, D.W., *Applied Multivariate Statistical Analysis*, (4^{th} ed) New Jersey: Prentice-Hall (1998).

Kaski, S., Nikkilä, J., Törönen, P., Castren, E., Wong, G., "Analysis and visualization of gene expression data using self-organizing maps," *IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP-01)*, Baltimore, Maryland (2001).

Kiplinger's best values in private colleges, *Kiplinger's Personal Finance,* Jan 1 2004, 64-71

Kiplinger's best values in public colleges, *Kiplinger's Personal Finance,* Nov 1 2003, 72-79

Klauer, K.C., "Ordinal network representation: Representing proximities by graphs," *Psychometrica*, 54, 737-750 (1989).

Klock, H. and Buhmann, J.M., "Data visualization by multidimensional scaling: A deterministic annealing approach," *Pattern Recognition*, 33(4), 651-669 (1999).

Kohonen, T., *Self-Organization and Associative Memory*, (3rd ed) Berlin: Springer (1989).

Kohonen, T., *Self-Organizing Maps*, Berlin: Springer (1995).

Krumhansl, C.L., "Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density," *Pshchological Review*, 85, 445-463 (1978).

Kruskal, J.B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, 29(1), 1-27 (1964a).

Kruskal, J.B., "Non-metric multidimensional scaling: A numerical method," *Psychometrika*, 29(1), 15-127 (1964b).

Latham, R., *The Dictionary of Computer Graphics and Virtual Reality*, (2nd ed) New York: Springer (1995).

LeBlanc, M. and Tibshirani, R., "Adaptive principal surfaces," *Journal of the American Statistical Association*, 89, 53-64 (1994).

Lee, R.C.T., Slagle, J.R., and Blum, H., "A triangulation methods for the sequential mapping of points from N-space to two-space," *IEEE Transactions on Computers*, 26, 288-292 (1977).

Lerner, B., Guterman, H., Aladjem, M., Dinstein, I., and Romem, Y., "On pattern classification with Sammon's nonlinear mapping – an experimental study," *Pattern Recognition*, 31, 371-381 (1998).

Levin, J., and Brown, M., "Scaling a conditional proximity matrix to symmetry," *Psychometrika*, 44, 239-244 (1979).

Mangiameli, P., Chen, S., West, D., "A comparison of SOM neural network and hierarchical clustering methods," *European Journal of Operational Research*, 93, 402-417 (1996).

Mao, J. and Jain, A.K., "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Transactions on Neural Networks*, 6, 296-317 (1995).

Merino, M. and Munoz, A., "Self organizing map and Sammon mapping for asymmetric proximities," *ICANN 2001*, 429—435 (2001). (Note: Int. Conference on Artificial Neural Networks.)

Milligan, G.W., "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, 43(5), 325-342 (1980).

Milligan, G.W., "A review of Monte Carlo tests of cluster analysis," *Multivariate Behavioral Research*, 16, 379-407 (1981).

Milligan, G.W., "An algorithm for generating artificial test clusters," *Psychometrika*, 50/1, 123-127 (1985).

Oja, M., Kaski, S., and Kohonen, T., "Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum," *Neural Computing Surveys*, 3, 1-156 (2003). (Available at http://www.soe.ucsc.edu/NCS/vol3.html)

Rand, W.M., "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846-850 (1971).

Ritter, H. and Kohonen, T., "Self-organizing semantic maps," *Biological Cybernetics*, 61, 241-254 (1989).

Rodgers, J.L. and Thompson, T.D., "Seriation and multidimensional scaling: A data analysis approach to scaling asymmetric proximity matrices," *Applied Psychological Measurement*, 16, 105-117 (1992).

Saito, T., "Multidiensional scaling to explore complex aspects in dissimilarity judgment," *Behaviormetrika*, 20, 35-62 (1986).

Sammon, Jr., J.W., "A non-linear mapping for data structure analysis," *IEEE Transactions on Computers*, 18, 401-409 (1969).

Shepard, R.N., "The analysis of proximities: Multidimensional scaling with an unknown distance function," *Psychometrika*, 27, 125-140 (1962).

SOM_Pak, available at http://www.hut.fi/Units/CSE/research.html, (1997).

Swayne, D.F., Cook, D., and Buja, A., "XGobi: Interactive dynamic data visualization in the X window system," *Journal of Computational and Graphical Statistics*, 7(1), 113-130 (1998).

Swayne, D.F., Lang, D.F., Buja, A., and Cook, D., "GGobi: Evolving from XGobi into an extensil framework for interactive data visualization," *Journal of Computational Statistics and Data Analysis* (to appear) (2002).

Tufte, E.R., *The Visual Display of Quantitative Information*, Chesire, Connecticut: Graphics Press (1983).

Tversky, A., "Features of similarity," *Psychological Review*, 84, 327-352 (1977).

Tversky, A., and Hutchinson, J., "Nearest neighbor analysis of psychological spaces," *Psychological Review*, 93, 3-22 (1986).

Vesanto, J., "SOM-based data visualization methods," *Intelligent Data Analysis*, 3, 111-126 (1999).

Viscovery SOMine 4.0, Eudaptics Software, available at www.eudaptics.com, (2002).

Ward, J.H., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, 58, 236-244 (1963).

Weeks, D.G., and Bentler, P.M., "Restricted multidimensional scaling models for asymmetric proximites," *Psychometrika*, 47, 201-208 (1982).

White, I., "Comment on 'A nonlinear mapping for data structure analysis'," *IEEE Transactions on Computers*, 21, 220-221 (1972).

Yin, H., "Data visualization and manifold mapping using the ViSOM," *Neural Networks*, 15, 1005-1016 (2002).

Young, F. W., *ViSta: The Visual Statistics System*, Technical Report RM 94-1, L.L. Thurstone Psychometric Laboratory, University of North Carolina (1994).

Zielman, B. and Heiser, W.J., "Models for asymmetric proximities," *British Journal of Mathematical and Statistical Psychology*, 49, 127-146 (1996).