Journal of Data Science **9**(2011), 271-291

# Identifying Groups: A Comparison of Methodologies

Abdolreza Eshghi[1], Dominique Haughton[1,2],
Pascal Legrand[3], Maria Skaletsky[1] and Sam Woolford[1]
[1]*Bentley University,* [2]*Université of Toulouse I and* [3]*Groupe ESC Clermont-CRCGM*

*Abstract*: This paper describes and compares three clustering techniques: traditional clustering methods, Kohonen maps and latent class models. The paper also proposes some novel measures of the quality of a clustering. To the best of our knowledge, this is the first contribution in the literature to compare these three techniques in a context where the classes are not known in advance.

*Key words*: Cluster analysis, Kohonen maps, latent class analysis, self-organizing maps.

## 1. Introduction

Identifying groups of observations that appear to be similar across a number of variables is a common multivariate technique that is used in a wide range of applications. The resulting groupings are often used to develop an understanding of the underlying structure of the data set. Identifying market segments based on the characteristics of consumers is one such application. Classical hierarchical cluster analysis and k-means cluster analysis are traditionally the methodologies used for this purpose. More recently, other methodologies such as Kohonen maps and latent class analysis have become feasible as software for implementing them has become available. To date, it does not appear that these methodologies have been compared in a setting in which the underlying structure of the data may not be known a-priori.

The intent of this paper is to extend the results of a recent paper (Deichman *et al.* (2006)) which utilized Kohonen maps to identify the digital divide between groups of countries based on patterns of economic, social and cultural factors. In particular, we will reanalyze a subset of the variables used in the earlier paper and compare the Kohonen map results with the results obtained using traditional cluster analysis methods and latent class methods. This analysis will extend existing results in two directions:

- To the best of our knowledge, no results in the literature have compared all the methodologies presented in this paper;
- Second, in order to compare these results, which are obtained for data with an unknown group structure, we have developed measures of homogeneity and heterogeneity for the results which can be used to provide an objective means by which the methodologies can be compared.

This paper also extends results in Gelbard, Goldman and Spiegler (2007) and Magidson and Vermunt (March 2002) by including a wider range of methodologies and applying them to a data set where classes are not known in advance; in addition we propose measures of clustering quality when classes are not known.

*Summary of findings*

Our results indicate that the traditional cluster analysis, Kohonen maps and latent class analysis yield different groupings both in terms of the number of groups and in terms of group membership. This in turn suggests that each methodology could lead to potentially different interpretations of the underlying structure of the data. Consequently, it may be important for researchers to pay more attention to the underlying assumptions associated with each methodology in order to choose the best methodology for a given application.

We develop our findings by first providing a brief overview of each of the methodologies we apply and then introducing the measures by which we will compare the results. Next we provide a description of the data set we analyze and discuss the results that are obtained using each of the clustering methodologies. Finally we discuss our findings and conclusions.

## 2. Background and Literature Review

### 2.1 Cluster Analysis (Standard Approach)

The standard approach to cluster analysis is widely known so we only very briefly recall the main ideas here. The hierarchical (agglomerative) approach to cluster analysis involves linking cases by looking at all possible pairs of cases and linking those in the pair with the smallest distance, then continuing in this manner until all cases lie in one big cluster. This of course involves defining a distance between cases, typically based on a number of variables available for each case, and a distance between clusters.

The list of successive merges in an agglomerative hierarchical cluster analysis (along with measures of homogeneity of the resulting clusters at each stage) can be examined to help decide when the merging process should be stopped. The

process is stopped when homogeneity measures exhibit a "large" drop in value; this admittedly ad-hoc procedure typically suggests a suitable number of clusters for the data set and is widely used. If the data set is too large to allow a hierarchical clustering, an analysis can be performed on a sample of the data.

Non-hierarchical clustering works by assuming a known number N of clusters, and by choosing N well-separated cases or cluster means from a previous hierarchical clustering as initial seeds. Cases from the data set are then assigned to that seed to which the case is closest. Once all cases have been assigned to a seed, new seeds get computed as current cluster means, and the process of assigning cases to clusters is repeated, until cluster means do not change much anymore, at which point the procedure is said to have converged.

It is worth noting that cases merged in a hierarchical clustering cannot be separated, while such a separation can occur in a non-hierarchical clustering. Many analyses involve a hierarchical clustering to obtain a suitable number of clusters and initial seeds, and are then refined by a non-hierarchical clustering. In any event, the problem of identifying a suitable number of clusters is notoriously difficult in many situations.

### 2.2 Kohonen Maps

A Kohonen Self-Organizing Map (SOM) is an exploratory data analysis technique that projects a multi-dimensional data set onto a space with a small dimension (typically a two-dimensional plane). SOMs thus allow for the convenient visualization of a data set and the effective identification of groups that share similar characteristics. In this sense, a Kohonen map might be compared to a factor-cluster analysis, in which variables are first summarized by the creation of factors, and the factors are then used to cluster observations. As we will see, one advantage of the Kohonen approach is the self-organizing feature of the map, a very powerful property that makes estimated components vary in a monotonic way across the map.

The Kohonen map algorithm will be described later in the paper, but in essence, the methodology is a special case of a competitive neural net. It begins with a typically two-dimensional grid (although one-dimensional and three-dimensional Kohonen maps are also encountered), with the number of positions on the grid decided by the user in advance. To each position on the grid corresponds an initial vector (which is not chosen among the data vectors in the data set, but instead in a way to be described further later) of a dimension equal to the number of variables involved in the clustering. In the initial state of the grid, one can envision each position on the grid with an underlying vector, initially unrelated to the data, but to be modified as the algorithm proceeds. Indeed, at each step of the algorithm, an actual data vector is considered and the underlying

vector which is closest (in Euclidean distance) to that data vector is modified (as well as its neighbors) so as to lie closer to the data vector. In this manner, each data vector in turn influences values of the underlying vectors on small portions of the grid. At some point the modified underlying vectors do not change much, and the algorithm has converged. The components of these underlying vectors at convergence play the role of estimated values at each grid point for each of the variables in the analysis. One powerful property of the SOM algorithm, which is quite striking and yet is often ignored in literature where Kohonen maps are discussed as clustering methods, is that the values of each component of the modified underlying vectors at convergence are arranged in an approximately monotonic way on the grid, hence the appellation *Self Organizing*.

A thorough introduction to Kohonen maps can be found in Kaski and Kohonen (1995), and a comprehensive overview of SOM methods and case studies is available in Kohonen (2001). Since the introduction of SOMs by Kohonen (1982), researchers have applied the techniques to a multitude of areas represented by an extensive bibliography with more than 5000 articles available on the SOM web site (http://www.cis.hut.fi/research/som-bibl/).

### 2.3 Latent Class Analysis

Latent class analysis (LCA) is a method for analyzing the relationships among manifest data when some variables are unobserved. The unobserved variables are categorical, allowing the original data set to be segmented into a number of exclusive and exhaustive subsets: the latent classes. Traditional LCA involves the analysis of relationships among polytomous manifest variables. Recent extensions of LCA allow for manifest variables that represent nominal, ordinal, continuous and count data (see Kaplan (2004)). The availability of software packages to perform LCA has increased the feasibility of using LCA to perform cluster analysis.

The basic latent class cluster model is given by

$$P(y_n|\theta) = \sum_{1}^{s} \pi_i P_j(y_n|\theta_j),$$

where $y_n$ is the nth observation of the manifest variables, S is the number of clusters and $\pi_j$ is the prior probability of membership in cluster $j$. $P_j$ is the cluster specific probability of $y_n$ given the cluster specific parameters $\theta_j$. The $P_j$ will be probability mass functions when the manifest variables are discrete and density functions when the manifest variables are continuous. For a more complete definition see Hagenaars, McCutcheon (2002). Since LCA is based upon a statistical model, maximum likelihood estimates can be used to classify cases based upon their posterior probability of class membership. In addition, various diagnostics are available to assist in the determination of the number of clusters.

LCA has been used in a broad range of contexts including sociology, psychology, economics, and marketing. LCA is presented as a segmentation tool for marketing research and tactical brand decision in Finkbeiner and Waters (2008). Other applications in market segmentation are given in Cooil et.al. (2007), Malhotra *et al.* (1999), Bodapati (2008), and Pancras and Sudhir (2007). Applications of LCA to cluster analysis have been explored in Hagenaars and McCutcheon (2002).

### 2.4 Comparison of Cluster Methodologies

Other authors have contrasted some of the clustering methodologies we consider in this paper. Magidson and Vermunt (March 2002), show that LCA offers several advantages over k-means cluster analysis for known deviations from the typical assumptions required for k-means clustering. Gelbard, Goldman and Spiegler (2007) performed an empirical comparison between Kohonen maps and a variety of other clustering algorithms including various hierarchical methods and k-means clustering on four data sets with known group membership. In comparing the results of each clustering methodology to the actual group membership, the authors conclude that non-hierarchical methods typically performed better than Kohonen maps which generally outperformed the hierarchical methods.

This paper adds to this literature by providing a comparison of all three clustering methodologies using the same data albeit with unknown group membership. In order to perform this comparison using our data, we need a measure which can be used to evaluate the performance of each of the methodologies. We suggest that at least two measures be typically used to evaluate the efficacy of a cluster analysis:

- The homogeneity of the observations within each cluster
- The heterogeneity of the clusters

Nguyen and Rayward-Smith (2008) provide an overview of forty-five metrics which can be used to measure cluster quality; their extensive analysis does not identify a single best measure for determining cluster quality. Consequently we approach the problem from a fundamental data analysis perspective and propose measures of homogeneity and heterogeneity which provide a way to assess the results of each clustering methodology on a common basis. We will compare our measures with those proposed in Nguyen and Rayward-Smith (2008) as we go along.

To measure the homogeneity of the resulting clusters, we use the basic idea of the within sum of squares and compute the variation in each cluster across all variables (averaged by the number of variables). We then sum the average

variation across the clusters. In particular, for cluster $i$, $i = 1, \cdots, C$, let

$$s^2(i) = \sum_{j=1}^{J} \sum_{k=1}^{N(i)} (x(ijk) - x(ij.))^2 / (J(N(i) - 1))$$

where $x(ijk)$ = observation $k, k = 1, \cdots, N(i)$, for variable $j$, $j = 1, \cdots, J$ in cluster $i$, $i = 1, \cdots, C$ and

$$x(ij.) = 1/N(i) \sum_{k=1}^{N(i)} x(ijk).$$

We can then define the homogeneity measure (similar to the equation referred to *f13* in Nguyen and Rayward-Smith (2008)) as:

$$S^2 = \sum_{i=1}^{C} s^2(i).$$

To measure the heterogeneity of the resulting clusters we use two measures. The first measure captures the separation of clusters by considering the squared Euclidean distance between the center of the clusters and aggregating the distances between all combinations of cluster centers. In particular we define the distance between two clusters $i_1$ and $i_2$ as the distance between their centers,

$$d^2(i_1, i_2) = \sum_{i=1}^{J} (x(i_1 j.) - x(i_2 j.))^2.$$

The total distance between clusters (equal to the expression referred to as *f14* in Nguyen and Rayward-Smith (2008)) is then defined as

$$D^2 = \sum_{i_1 < i_2} d^2(i_1, i_2).$$

A second measure can be defined by considering the distance from the observations in one cluster and the centers of all the other clusters. This can be defined as

$$s^2(i) = \sum_{j=1}^{J} \sum_{\substack{h=1 \\ h \neq i}}^{C} \sum_{k=1}^{N(i)} (x(ijk) - x(hj.))^2 / ((N(i) - 1) * (C - 1))$$

with the resulting measure of heterogeneity defined by

$$H^2 = \sum_{i=1}^{C} \frac{s^2(i)}{C}.$$

This second measure (which can be considered as a variation on Nguyen and Rayward-Smiths measure referred to as *f39* in that paper) considers how far, on average, the points in one cluster are from the other clusters.

## 3. Data

Our data set consists of 160 countries for which ten variables (described in Table 1) are available; these variables can be arranged into five groups described below. This data set is a subset of the data set used in Deichmann *et al.* (2006). The objective of a cluster analysis in this context is to identify groups of countries which tend to form naturally on the basis of similarity in values of the ten variables. Ultimately one then obtains main profiles in levels of information technology exposure and other dimensions that countries tend to follow. Because our goal in this paper is to compare clustering methodologies rather than expand on the interpretation of these profiles, we refer to Deichman *et al.* (2006) for more such interpretative details.

Table 1: Description of Variables

| Variable | Description | Year(s) | Source | Group |
|---|---|---|---|---|
| **computers** | Number of computers per 100 people | 2001-03 | ITU | Digital Dev. |
| **internet** | Number of Internet users per 10,000 | 2001-03 | ITU | Digital Dev. |
| **Income** | GNI per capita in international ppp dollars | 2001-03 | World Bank | Economic |
| **Maintel** | Number of main telephone lines per 100 | 2001-03 | World Bank | Infrastructure |
| **Electric** | Electricity consumption kwh/capita | 2001-03 | World Bank | Infrastructure |
| **p1564** | Percentage of population age 15-64 | 2001-03 | World Bank | Demographic |
| **p65plus** | Percentage of population 65 and older | 2001-03 | World Bank | Demographic |
| **School** | Average years of schooling of adults | 2001-03 | World Bank | Demographic |
| **Urban** | Urban population as percent of total | 2001-03 | World Bank | Demographic |
| **risk** | Composite Risk Rating Index | 2001-03 | PRS Group | Risk |

Note: "ITU"= International Telecommunications Union. "PRS"= Political Risk Services.

The first group, referred to as Digital Development, includes the number of Internet users per 10,000 population (see for example Dimitrova and Beilock, (2005)), and the number of computers per 100 inhabitants (ranging from less than one in the developing world to more than fifty in Europe).

The four remaining groups, Economic, Infrastructure, Demographic and Risk correspond to the commonly agreed upon factors that explain variations across countries in their digital development.

Our economic variable, as prominent in the literature, consists of the income level of a country ("income"), measured by the GNI (Gross National Income) per capita in international ppp (Purchasing Power Parity) dollars.

The level of infrastructure is measured by variables on the number of main telephone lines per 100 population ("maintel"), as well as the level of electricity consumption ("electric").

The demographic structure of a country is measured by variables on the percentage of people between the ages of 15 and 64 ("p1564") and those 65 and over ("p65plus"), the average number of years of schooling of adults ("school"), and the percentage of each countrys population that dwells in an urban setting ("urban"). The relevance of age, gender, education, and other cultural traits is established by the micro-level studies discussed above (Mendoza and Toledo (1997), Kubicek (2004)).

In order to capture the risk related to the political situation in each country, we include the Composite Risk Rating Index ("risk") compiled by the Political Risk Services Group1 in their International Country Risk Guide publications. This index measures not only cyclical economic risks but also the political soundness of each country. Higher values represent lower risks. For example, the data range from scores in the 50s in Sub-Saharan African states to Scandinavian scores in the mid-80s. The risk variable is included in our analysis as a sensible proxy for regularity quality and the rule of law as used in Chinn and Fairlie (2004, 2007), since these variables were not available to us.

Our data were collected from 160 countries. The country codes are listed in Appendix 2. The following variables were fully populated in our dataset: p1564, p65plus, urban, maintel, internet, and computers. For missing cells in other variables we imputed[2] values by regressing predictors on other predictors (but not on "internet" and "computers"), as was done in Deichmann *et al.* (2006, 2007).

## 4. Methodology and Analysis

### 4.1 Cluster Analysis

*Methodology and analysis*

A hierarchical cluster analysis procedure was performed using our 10 variables, which were standardized for the analysis. Both statistical packages SAS and SPSS were used for the analysis.

Various statistics commonly used for evaluation of number of clusters did not lead to a clear cut result. Indeed, the RSQ (RVsquare) and SPRSQ (Semi partial R-square) do not appear to have any noticeable "jumps" as one keeps adding one more cluster. The Pseudo F-statistic graph (omitted here) does not feature any "peaks" at high values. Instead, it displays a gradual decrease in value as the number of clusters increases.

The pseudo t-squared statistic graph (omitted here) indicates that a choice of eight clusters is reasonable for the data, since the value of the pseudo t-squared statistic at eight clusters is slightly lower than at nine clusters, and further variations (beyond nine clusters) of the pseudo t-squared statistic seem to essentially behave as random noise.

The cubic clustering criterion (CCC) graph (omitted here) does not display any clear peaks either. However, the CCC value is slightly higher at eight clusters than it is at nine clusters. We will therefore adopt the solution with eight clusters. The cluster means from this hierarchical clustering were then used as seeds for a k-means (or non-hierarchical) clustering.

Table 2 displays the resulting list of clusters and cluster members.

Table 2: Cluster (k-means) membership

| |
|---|
| **Cluster 1**: Algeria, Egypt, Morocco, Belize, Ecuador, Grenada, Guyana, St. Lucia, China, India, Indonesia, Kyrgyzstan, Mongolia, Philippines, Sri Lanka, Thailand, Vietnam, Albania, Syria, Samoa, Tonga, Maldives, Botswana, Cape Verde, Djibouti, Gabon, Bolivia, El Salvador, Paraguay, Namibia, Swaziland, Honduras |
| **Cluster 2**: Mauritius, Seychelles, Tunisia, Costa Rica, Dominica, Jamaica, St. Vincent, Suriname, Iran, Jordan, Turkey, Fiji, French Polynesia, Lybia, South Africa, Argentina, Brazil, Chile, Colombia, Mexico, Panama, Peru, Uruguay, Venezuela, Lebanon, Oman, Saudi Arabia, Trinidad & Tobago, Malaysia |
| **Cluster 3**: Cuba, Armenia, Georgia, Moldova, Serbia & Montenegro, Ukraine |
| **Cluster 4**: Pakistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Chad, Comoros, Congo, Cote d'Ivoire, Equatorial Guinea, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Kenya, Madagascar, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Senegal, Sudan, Tanzania, Togo, Uganda, Zambia, Zimbabwe, Nicaragua, Bangladesh, Bhutan, Cambodia, Laos, Myanmar, Nepal, Yemen, Papua New Guinea, Solomon Islands, Vanuatu, Guatemala, Central African Republic |
| **Cluster 5**: Romania, Russia, Barbados, St. Kitts and Nevis, Bulgaria, Croatia, Czech Republic, Greece, Hungary, Latvia, Lithuania, Poland, Portugal, Slovak Republic |
| **Cluster 6**: Israel, Cyprus, Estonia, Italy, Malta, Slovenia, Spain, Belgium, Ireland, Japan, Austria, France |
| **Cluster 7**: Canada, United States, Japan, Korea, Denmark, Finland, Germany, Iceland, Luxembourg, Netherlands, Norway, Sweden, Switzerland, United Kingdom, Australia, New Zealand, Hong Kong, Singapore |
| **Cluster 8**: Bahrain, Brunei, Kuwait, Qatar, United Arab Emirates, Macao |

### 4.2 Kohonen Maps

*Methodology and Analysis*

The map in this paper was generated using the software Matlab 6.0 and the SOM Matlab toolkit (http://www.cis.hut.fi/projects/somtoolbox/) yielding two graphs, the U-matrix (Figure 1) and the component matrix (Figure 2), which will be explained in more detail below. Prior to applying the Self Organizing Map (SOM) algorithm, we standardized the variables used in the analysis.

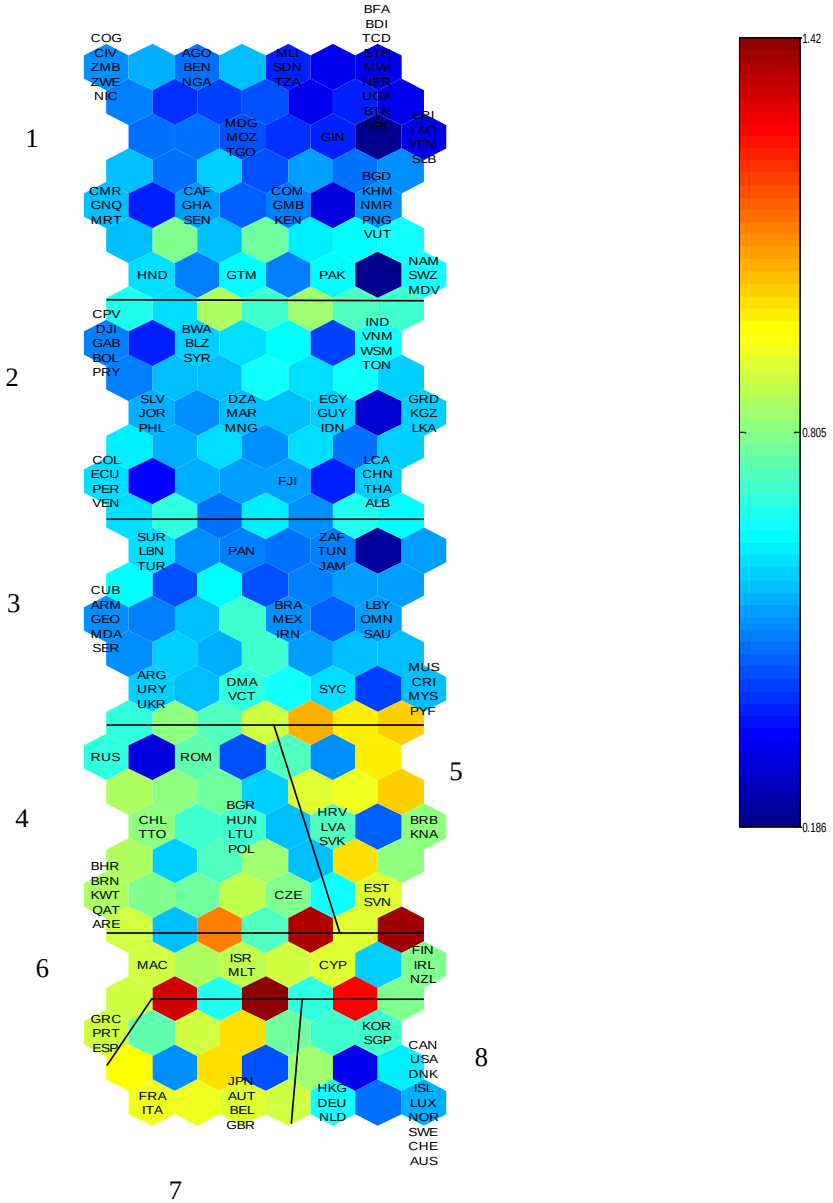


Figure 1: Kohonen Map U-matrix, with groups displayed

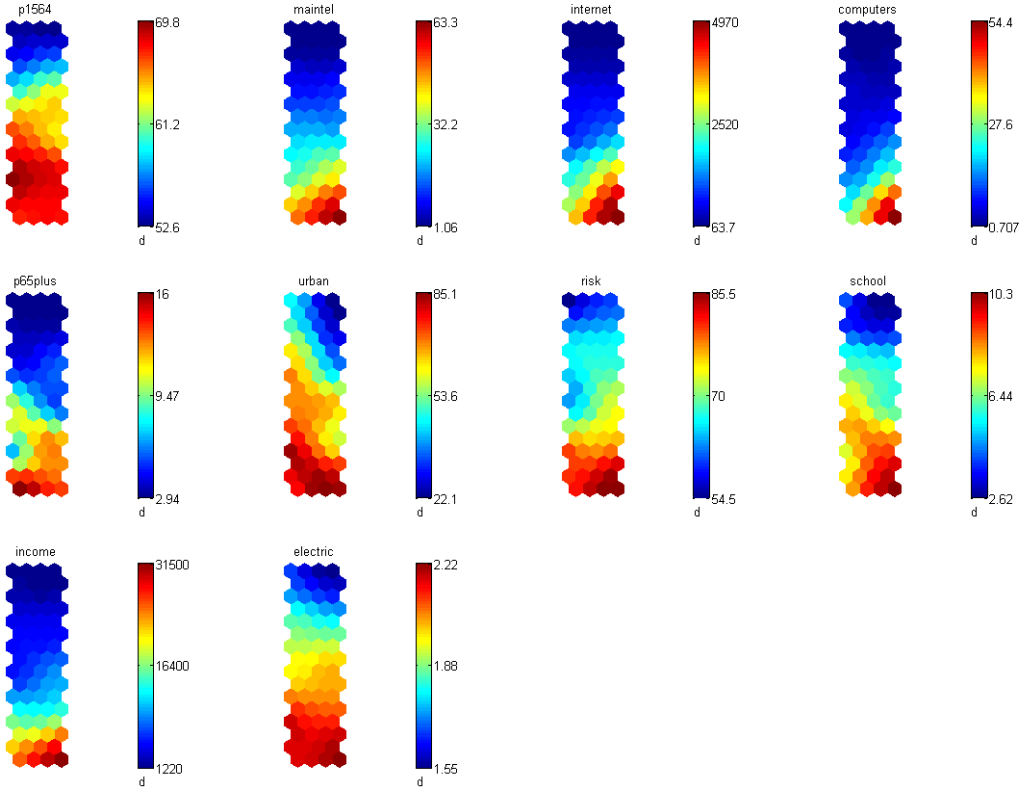


Figure 2: Kohonen components map

We now briefly explain in an intuitive manner how the SOM algorithm functions, and refer the reader to for instance Deichmann *et al.* 2007 for more details. The algorithm first determines a suitable size for the map (that is, a suitable number of rows and columns) on the basis of how much correlation exists among the variables. In the case of this paper, a map with 16 rows and 4 columns was selected, yielding 16 times 4 or 64 positions.

At this point, the algorithm assigns to each of the 64 positions an initial 10-dimensional vector, where the dimension 10 of the vector corresponds to the number of variables used in constructing the map3. In the first iteration, the first country in the data set with its 10-dimensional actual data vector for is considered and the Euclidean distance between the data vector for the country and each of the random vectors is computed. A Best Matching Unit (BMU) is then identified as the map position for which that distance is the smallest. The initial vector for this BMU (and in some variants of the algorithm for units in a neighborhood of the BMU on the grid as well) is then modified as to lie on a segment joining the input data vector and the initial vector.

The data vector for the next country in the data set in then considered, a best matching unit identified and the same process repeated for this next iteration. As iterations proceed, the country data vectors gradually influence the initial vectors in the vicinity of the best matching units. After a number of iterations, the modified vectors do not change much anymore, and the map has converged. These final modified vectors at each grid position are referred to as estimated vectors at each position in what follows.

Once the map has converged, one question that remains is of how the algorithm determines where on the map the countries should be positioned. This is actually done very simply: for each country the Euclidean distance between its (standardized) data vector and each of the 64 estimated vectors obtained at the completion of the algorithm is computed, and a position identified where that distance is smallest. Of course, it is possible that several countries position themselves at the same map location, if they happen to have their (standardized) data vector closest in Euclidean distance to that of the same position on the map. This happens for instance for France and Italy, at the bottom left corner of the U-matrix in Figure 1. It is also clearly possible that some map locations have no country attached to them, even though they do have an estimated vector. This would mean that no country found the estimated vector for that position to be closest to its (standardized) data vector; there were always other positions with closer data vectors.

The country locations are presented in Figure 1. The U-matrix represented in that figure contains not only the 64 map positions, but those positions plus an additional hexagon between any two map positions. The color of these intermediate hexagons reflects the Euclidean distance between estimated vectors for the two bordering hexagons. The color scale featured in Figure 1 defines which colors represent which distances. For example, a dark red color corresponds to a large distance, a dark blue color to a small distance and a green color to an intermediate distance.

Turning to the interpretation of the color of intermediate hexagons, we see that for instance in the bottom part of the U-matrix in Figure 1, the estimated vector at the hexagon featuring France and Italy is quite distant (light green) from that at the hexagon featuring Japan, Australia, Belgium and Great Britain. Colors of hexagons that do represent country positions reflect the average distance between the estimated vector at a map hexagon and those of its neighbors. For instance, the green color on the hexagon featuring Finland, Ireland and New Zealand represents an average between the large distance between the estimated vector for this hexagon and that of its neighbor featuring Estonia and Slovenia, the smaller distance between this vector and that of its neighbor featuring Cyprus, and the moderate distance between this vector and that of its neighbor featuring

Korea and Singapore.

The U-matrix and its represented Euclidean distances between map positions is where one can begin to identify clusters in the map as amalgams of hexagons with low distance hexagons separated by "walls" of hexagons with higher distance colors. This of course is the reason why Kohonen maps are also considered as a clustering technique, but with the additional Self Organizing (SO) property.

The "self-organizing" property of the Kohonen algorithm can be visualized on Figure 2, where the estimated values of each of the 10 variables at each of the 64 positions are displayed, each with a color scale, with red representing high estimated values and blue low estimated values of the components. Indeed, one can see the estimated values of the components move monotonically from large values to small values as one moves vertically or diagonally across the map. This property gives the SOM algorithm a very attractive property: one can now identify for instance that moving from top to bottom on the graph seems to imply an increase in wealth, judging by the estimated values of "income". If estimated "income" fluctuated as one looks from top to bottom on the map, it would be much harder to interpret axes on the map.

Examining a component map in more detail, for instance the first one containing the estimated values of p1564 (proportion of population aged 15 to 64), we see that the estimated values increase gradually from blue to yellow to red as one moves vertically from top to bottom on the map. It is useful at this point to remember that the 64 positions are the same on each component map, and on each non-intermediate hexagon of the U-matrix. In other words, for example, the bottom left hexagon represents the same position on the U-matrix of Figure 1 and each of the component maps in Figure 2. The 64 values on the first component map are the first components of each of the 64 final estimated vectors resulting from the algorithm. We see that for all components maps in Figure 2, the estimated values tend to gradually increase from blue to yellow to red as one moves vertically down the map.

To summarize, the clusters are seen on the U-matrix with its featured measures of proximity, while the interpretation of what it means to move up and down or across the map is indicated by the components maps (Figure 2).

The main groups identified by the Kohonen algorithm include the following countries (Table 3).

### 4.3 Latent Class Analysis

*Methodology and Analysis*

We have opted to conduct the latent class analysis using Latent Gold 4.0®, a commercially available LCA software package (see Vermunt and Magidson

Table 3: Cluster (Kohonen) membership

| |
|---|
| **Kohonen Group 1**: Congo, Cote d'Ivoire, Zambia, Zimbabwe, Nicaragua, Angola, Benin, Nigeria, Mali, Sudan, Tanzania, Burkina Faso, Burundi, Chad, Ethiopia, Malawi, Niger, Uganda, Bhutan, Nepal, Madagascar, Mozambique, Togo, Guinea, Eritrea, Laos, Yemen, Solomon I., Cameroon, Eq. Guinea, Mauritania, Central Af. Rep., Ghana, Senegal, Comoros, The Gambia, Kenya, Bangladesh, Cambodia, Myanmar, P. New Guinea, Vanuatu, Honduras, Guatemala, Pakistan, Namibia, Swaziland, Maldives |
| **Kohonen Group 2**: Cape Verde, Djibouti, Gabon, Bolivia, Paraguay, Botswana, Belize, Syria, India, Vietnam, Samoa, Tonga, El Salvador, Jordan, Philippines, Algeria,Morocco, Mongolia, Egypt, Guyana, Indonesia, Grenada, Kyrgyz rep., Sri Lanka, Colombia, Ecuador, Peru, Venezuela, Fiji, St. Lucia, China, Thailand, Albania |
| **Kohonen Group 3**: Suriname, Lebanon, Turkey, Cuba, Armenia, Georgia, Moldova, Serbia, Argentina, Uruguay, Ukraine, Panama, South Africa, Tunisia, Jamaica, Brazil, Mexico, Iran, Libya, Oman, Saudi Arabia, SYC, Mauritius, Costa Rica, Malaysia, French Pol., Seychelles, Dominica, St. Vincent |
| **Kohonen Group 4**: Russia, Romania, Chile, Trin. & Tob., Bulgaria, Hungary, Lithuania, Poland, Bahrain, Brunei, Kuwait, Qatar, Un.Arab Em., Czech Rep. |
| **Kohonen Group 5**: Croatia, Latvia, Slovakia, Barbados, St. Kitts/Nevis, Estonia, Slovenia |
| **Kohonen Group 6**: Macao, Israel, Malta, Cyprus, Finland, Ireland, New Zealand, Greece, Portugal, Spain |
| **Kohonen Group 7**: France, Italy, Japan, Austria, Belgium, United Kingdom |
| **Kohonen Group 8**: Korea, Singapore, Hong Kong, Germany, Netherlands, Canada, USA, Denmark, Iceland, Luxemburg, Norway, Sweden, Switzerland, Australia |

(2005)). The data for the latent class analysis was standardized to be consistent with the k-means cluster analysis and Kohonen map analysis although this was not technically required.

An initial exploratory latent class analysis was conducted in an attempt to narrow down the number of clusters that would be explored more fully. One to thirteen cluster solutions were generated using Latent Gold 4.0$^{\text{TM}}$. The BIC results indicated that a seven cluster model was optimal. The detailed analysis indicated that all the parameters were highly significant for differentiating the clusters and that the associated values are all above .5. Based on the diagnostics, covariances between urban and electric, computers and internet, and risk and income were also estimated for each cluster. These changes led to a seven cluster model (with associated BIC=2072.2, log likelihood=-612.3 and AIC=1558.6) for which the associated diagnostics indicated that the resulting model provided an adequate fit to the data. The resulting parameters were all highly significant and the $R^2$ values are all above 0.5.

The resulting cluster membership is given in Table 4.

Table 4: Cluster (latent class) membership

| |
|---|
| **Cluster 1**: Angola, Bangladesh, Benin, Bhutan, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Rep, Chad, Comoros, Congo, C?te d'Ivoire, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Kenya, Lao P.D.R., Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nepal, Nicaragua, Niger, Nigeria, Pakistan, Papua New Guinea, Senegal, Solomon Islands, Sudan, Tanzania, Togo, Uganda, Vanuatu, Yemen, Zambia |
| **Cluster 2**: Algeria, Belize, Bolivia, Botswana, Cape Verde, Colombia, Ecuador, Egypt, El Salvador, Gabon, Guatemala, Honduras, India, Indonesia, Jordan, Peru, Kyrgyzstan, Libya, Maldives, Mongolia, Morocco, Namibia, Paraguay, Philippines, Samoa, Sri Lanka, Swaziland, Syria, Tonga, Tunisia, Venezuela, Viet Nam, Zimbabwe |
| **Cluster 3**: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Hong Kong, Iceland, Ireland, Israel, Japan, Korea (Rep.), Luxembourg, Netherlands, New Zealand, Norway, Singapore, Sweden, Switzerland, United Kingdom, United States |
| **Cluster 4**: Barbados, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece, Hungary, Italy, Latvia, Lithuania, Macao, Malta, Poland, Portugal, Romania, Russia, Slovak Republic, Slovenia, Spain, St. Kitts and Nevis |
| **Cluster 5**: Albania, Argentina, Armenia, Brazil, China, Cuba, Dominica, Fiji, Georgia, Guyana, Iran, Lebanon, Moldova, Panama, Serbia and Montenegro, Suriname, Turkey, Ukraine, Uruguay |
| **Cluster 6**: Chile, Costa Rica, French Polynesia, Grenada, Jamaica, Malaysia, Mauritius, Mexico, Seychelles, South Africa, St. Lucia, St. Vincent, Thailand, Trinidad and Tobago |
| **Cluster 7**: Bahrain, Brunei, Darussalam, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emiratesmark, Iceland, Luxemburg, Norway, Sweden, Switzerland, Australia |

## 5. Discussion

In considering the results from the three methodologies applied to the data in Table 1, there are a number of conclusions that can be seen in our results. First and most obviously, the methodologies do not result in the same number of clusters when applying the generally accepted procedures for determination of the number of groups. Cluster analysis and Kohonen mapping suggest eight clusters while latent class analysis suggests seven clusters.

This obviously results in differences in cluster membership between the methods. For situations where the researcher does not know how many clusters may be present in the data this lack of consistency in the results may be problematic in interpreting the results. In general, the cluster centroids are also somewhat variable across the methodologies and are difficult to compare as the number of variables used in the analysis increases.

Using the measures of cluster homogeneity and heterogeneity described earlier in this paper can provide an objective means of evaluating the conceptual performance of the different methodologies. Table 5 provides the measures of homogeneity and heterogeneity based on the results from each methodology using the standardized data .

Table 5: Measures of homogeneity and heterogeneity

| | Homogeneity | Heterogeneity $d^2$ | Heterogeneity $h^2$ |
|---|---|---|---|
| **LatentGold results** | 1.490479 | 339.9556 | 19.23558 |
| **Cluster Analysis results** | 1.394168 | 516.7845 | 21.80146 |
| **Kohonen map results** | 1.765216 | 494.027 | 21.25504 |

These results would imply that the traditional cluster analysis provides the most homogeneous clusters while most effectively differentiating between clusters (note that the order of performance remains the same if the homogeneity measure and $d^2$ are normalized for the number of clusters involved). This could be a result of the fact that our performance measures are most consistent with the algorithm used for traditional cluster analysis.

The Kohonen map results represent a more data driven approach yielding the worst results with respect to the homogeneity of the resulting clusters and heterogeneity results that are only slightly worse than traditional cluster analysis but better than those obtained through latent class analysis.

## 6. Conclusion

The results of this analysis would appear to suggest that for data sets that do not have predefined clusters, different clustering methodologies may result in different results which may make any interpretation of the results dependent upon the methodology used. This would suggest that a deeper understanding of the theoretical underpinnings for each methodology may be required to ensure that the assumptions underlying the various algorithms are indeed appropriate for the specific analysis at hand. Since our analysis is limited to a single data set, we cannot conclude that these methodologies would perform in a consistent fashion across all data sets which may be a topic of further research. However these results would appear to suggest that researchers who are attempting to classify a data set into segments need to evaluate carefully the methodology they apply.

## Appendix 1: Brief description of the Kohonen SOM algorithm

The Kohonen algorithm can be briefly described as follows (see for example Kaski and Kohonen 1996). We begin with a grid in the two-dimensional plane where each position i is assigned an arbitrary (random) vector $m_i(0)$ with as many components as there are input variables. At each iteration $t$ the vector of variables $x(t)$ corresponding to one of the observations (in our case a country) updates the current vectors $m_i(t)$ according to the formula $m_i(t+1) = m_i(t) + h_{ci}(t)(x(t) - m_i(t))$, where $c = \arg\min_i(\|x - m_i\|)$ and $h_{ij}(t)$ is a function of $t$ and of the geometric distance on the lattice between position $i$ and position $j$. Typically $h_{ij} \to 0$ as the distance between $i$ and $j$ increases and as more iterations are performed. So the vector $x(t)$ is allowed to update the vector $m_c(t)$ it is closest to as well as some neighbouring vectors $m_i(t)$. The algorithm converges when little or no change occurs in the vectors $m_i(t)$. It is a key feature of Kohonen maps that once the algorithm has converged, the vectors $m_i$ tend to be ordered along the lattice in a "monotonic" way, hence the "self-organizing" appellation; that means that the components of the vectors in each position of the map when the algorithm has converged tend to decrease (or increase) as one moves across the grid. This contributes to an easier interpretation of the dimensions on the map and is an important reason why the technique has met with considerable popularity.

## Appendix 2: Country codes (abbreviation and name)

| | | | | | |
|---|---|---|---|---|---|
| AFG | Afghanistan | DNK | Denmark | KGZ | Kyrgyz Rep. |
| ALB | Albania | DJI | Djibouti | LAO | Lao PDR |
| ARE | United Arab Emirates | DMA | Dominica | LVA | Latvia |
| DZA | Algeria | DOM | Dominican Rep | LBN | Lebanon |
| ASM | Am Samoa | ECU | Ecuador | LSO | Lesotho |
| AGO | Angola | EGY | Egypt | LBR | Liberia |
| ATG | Antigua | SLV | El Salvador | LBY | Libya |
| ARG | Argentina | GNQ | Eq. Guinea | LTU | Lithuania |
| ARM | Armenia | ERI | Eritrea | LUX | Luxembourg |
| AUS | Australia | EST | Estonia | MAC | Macao |
| AUT | Austria | ETH | Ethiopia | MKD | Macedonia |
| AZE | Azerbaijan | FJI | Fiji | MDG | Madagascar |
| BHS | Bahamas, The | FIN | Finland | MWI | Malawi |
| BHR | Bahrain | FRA | France | MYS | Malaysia |
| BGD | Bangladesh | PYF | French Poly. | MDV | Maldives |
| BRB | Barbados | GAB | Gabon | MLI | Mali |
| BLR | Belarus | GMB | Gambia, The | MLT | Malta |
| BEL | Belgium | GEO | Georgia | MHL | Marshall Is. |
| BLZ | Belize | DEU | Germany | MRT | Mauritania |
| BEN | Benin | GHA | Ghana | MUS | Mauritius |
| BMU | Bermuda | GRC | Greece | MEX | Mexico |
| BTN | Bhutan | GRD | Grenada | MDA | Moldova |
| BOL | Bolivia | GUM | Guam | MNG | Mongolia |
| BWA | Botswana | GTM | Guatemala | MAR | Morocco |
| BRA | Brazil | GIN | Guinea | MOZ | Mozambique |
| BRN | Brunei | GNB | Guinea-Bissau | MMR | Myanmar |

| | | | | | |
|---|---|---|---|---|---|
| BGR | Bulgaria | GUY | Guyana | NAM | Namibia |
| BFA | Burkina Faso | HTI | Haiti | NPL | Nepal |
| BDI | Burundi | HND | Honduras | NLD | Netherlands |
| KHM | Cambodia | HKG | Hong Kong | NCL | New Caledonia |
| CMR | Cameroon | HUN | Hungary | NZL | New Zealand |
| CAN | Canada | ISL | Iceland | NIC | Nicaragua |
| CPV | Cape Verde | IND | India | NER | Niger |
| CAF | Central Afr Rep | IDN | Indonesia | NGA | Nigeria |
| TCD | Chad | IRN | Iran | NOR | Norway |
| CHL | Chile | IRQ | Iraq | OMN | Oman |
| CHN | China | IRL | Ireland | PAK | Pakistan |
| COL | Colombia | ISR | Israel | PAN | Panama |
| COM | Comoros | ITA | Italy | PNG | P. New Guinea |
| ZAR | Congo, DR | JAM | Jamaica | PRY | Paraguay |
| COG | Congo, Rep. | JPN | Japan | PER | Peru |
| CRI | Costa Rica | JOR | Jordan | PHL | Philippines |
| CIV | Cote d'Ivoire | KAZ | Kazakhstan | POL | Poland |
| HRV | Croatia | KEN | Kenya | PRT | Portugal |
| CUB | Cuba | KIR | Kiribati | PRI | Puerto Rico |
| CYP | Cyprus | KOR | Korea, Rep. | QAT | Qatar |
| CZE | Czech Republic | KWT | Kuwait | ROM | Romania |
| RUS | Russia | LKA | Sri Lanka | TTO | Trin & Tobago |
| RWA | Rwanda | KNA | St. Kitts/Nevis | TUN | Tunisia |
| WSM | Samoa | LCA | St. Lucia | TUR | Turkey |
| STP | Sao Tome | VCT | St. Vincent | UGA | Uganda |
| SAU | Saudi Arabia | SDN | Sudan | UKR | Ukraine |
| SEN | Senegal | SUR | Suriname | GBR | United Kindom |
| SYC | Seychelles | SWZ | Swaziland | USA | United States |
| SLE | Sierra Leone | SWE | Sweden | URY | Uruguay |
| SGP | Singapore | CHE | Switzerland | UZB | Uzbekistan |
| SVK | Slovakia | SYR | Syria | VUT | Vanuatu |
| SVN | Slovenia | TJK | Tajikistan | VEN | Venezuela, RB |
| SLB | Solomon Is | TZA | Tanzania | VNM | Vietnam |
| SOM | Somalia | THA | Thailand | VIR | Virgin Islands |
| ZAF | South Africa | TGO | Togo | ZMB | Zambia |
| ESP | Spain | TON | Tonga | ZWE | Zimbabwe |

## References

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research* **45**, 77-93.

Chinn, M.D. and Fairlie, R. (2007). The determinants of the global digital divide: A cross-country analysis of computer and internet penetration. *Oxford Economic Papers* **59**, 16-48.

Chinn, M.D. and Fairlie, R. (2004). The determinants of the global digital divide: A cross-country analysis of computer and internet penetration, bonn: forschungsinstitut zur Zukunft der Arbeit. *IZA Discussion Paper* No.1305.

Cooil, B., Keiningham, T.L., Askoy, L., and Hsu, M. (2007). A longitudinal analysis of customer satisfaction and share of wallet: investigating the mode

rating effect of customer characteristics. *Journal of Marketing* **71**, 67-83.

Deichmann, J., Eshghi, A., Haughton, D. Masnaghetti, M. Sayek. S. and Topi, H. (2006). Exploring break-points and interaction effects among predictors of the international digital divide: an analytic Approach. *Journal of Global Information Technology Management* **9**, 47-71.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., Teebagy, N. and Topi, H. (2006). Understanding eurasian convergence 1992-2000: application of Kohonen self-organizing Maps. *Journal of Modern Applied Statistical Methods* **5**, 72-93.

Dimitrova, D. and Beilock, R. (2005). Where freedom matters: internet adoption among the former socialist countries. *The International Journal for Communication Studies* **67** 173-187.

Finkbeiner, C. and Waters, K. (2008). Call every shot. *Marketing Management*, 38-43.

Gelbard, R., Goldman, O., and Spiegler, I. (2007). Investigating diversity of clustering methods: an empirical comparison. *Data Knowledge and Engineering* **63**, 155-166.

Hagenaars, J.A. and McCutcheon, A.L. (Eds.) (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge

Kaplan, D. (Ed.) (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Sage Publications, Thousand Oaks.

Kaski, S. and Kohonen, T. (1995). Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. *Proceedings of the Third International Conference on Neural Net works in the Capital Markets*. London, England, 11-13.

Kohonen, T. (1982). Analysis of a simple self-organizing process. *Biological Cybernet* **44**, 135-140.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd ed. Springer Verlag, Berlin.

Kubicek, H. (2004). Fighting a moving target: hard lessons from Germany's digital divide program. *IT and Society* **1**, 1-19.

Magidson, J. and Vermunt, J.K. (2002). Latent class modelling as a probabilistic extension of K-means clustering. *Quirks Marketing Research Review*, 77-80.

Magidson, J and Vermunt, J.K. (2002). Latent class models for clustering: a comparison with K-means. *Canadian Journal of Marketing Research* **20**, 37-44.

Malhotra, Naresh K., Person, Mark and Bardi Kleiser, S. (1999). Marketing research: a state of the art review and directions for the twenty first century. *Journal of the Academy of Marketing Science* **27**, 160-183.

Mendoza, M. and Alvarez de Toledo, J. (1997). Demographics and behavior of the chilean internet population. *Journal of Computer-Mediated Communication* **3**, electronic edition.

Nguyen, Q.H and Rayward-Smith, V.J. (2008). Internal quality measures for clustering in metric spaces. *International Journal of Business Intelligence and Data Mining* **3** 4-29.

Pancras, J. and Sudhir, K. (2007). Optimal marketing strategies for a customer data intermediary. *Journal of Marketing Research* **44**, 560V578.

Vermunt, J. K. and Magidson, J. (2005) *Technical Guide for latent Gold Choice 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont Massachusetts.

Receied October 18, 2009; accepted January 18, 2010.

Abdolreza Eshghi
Marketing in the Marketing Department
Bentley University
175 Forest Street, Waltham, MA 02452-4705, USA
aeshghi@bentley.edu

Dominique Haughton
Mathematical Sciences
Bentley University
175 Forest Street, Waltham, MA 02452-4705, USA
dhaughton@bentley.edu

Pascal Legrand
Information Systems
Ecole Supérieure de Commerce de Commerce Clermont
Groupe ESC Clermont V CRCGM, 4 Boulevard Trudaine, 63037 Clermont Ferrand Cedex 1 France
pascal.legrand@esc-clermont.fr

Maria Skaletsky
Business Analytics
Bentley University
175 Forest Street, Waltham, MA 02452-4705, USA
mskaletsky@bentley.edu

Sam Woolford
Mathematical Sciences
Bentley University
175 Forest Street, Waltham, MA 02452-4705, USA
swoolford@bentley.edu