Review of three latent class cluster analysis packages: Latent gold, poLCA, and MCLUST

Article *in* The American Statistician · February 2009

DOI: 10.1198/tast.2009.0016 · Source: RePEc

Dominique Haughton
Bentley University
125 PUBLICATIONS **1,828** CITATIONS
SEE PROFILE

Pascal Legrand
ESC - Clermont
12 PUBLICATIONS **173** CITATIONS
SEE PROFILE

Review of Three Latent Class Cluster Analysis Packages: Latent GOLD, poLCA, and MCLUST

Dominique HAUGHTON, Pascal LEGRAND, and Sam WOOLFORD

This article reviews three software packages that can be used to perform latent class cluster analysis, namely, Latent GOLD®, MCLUST, and poLCA. Latent GOLD® is a product of Statistical Innovations whereas MCLUST and poLCA are packages written in R and are available through the web site *http://www.r-project.org*. We use a single dataset and apply each software package to develop a latent class cluster analysis for the data. This allows us to compare the features and the resulting clusters from each software package. Each software package has its strengths and weaknesses and we compare the software from the perspectives of usability, cost, data characteristics, and performance. Whereas each software package utilizes the same methodology, we show that each results in a different cluster solution and suggest some rationales for deciding which package to use.

KEY WORDS: Latent class models; Latent GOLD; MCLUST, Mixture models; poLCA.

1. INTRODUCTION

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 dataset 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 increased the feasibility of using LCA to perform cluster analysis.

The basic latent class cluster model is given by

$$P(y_n|\theta) = \sum\nolimits_1^S \pi_j P_j\left(y_n|\theta_j\right),$$

where y_n is the nth observation of the manifest variables, S is the number of clusters, and π_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 θ_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 and McCutcheon (2002). Because LCA is based upon a statistical model, maximum likelihood estimates can be used to classify cases based upon what is referred to as their posterior probability of class membership. In addition, various diagnostics are available to assist in the determination of the optimal 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, Keiningham, Askoy, and Hsu (2007), Malhotra, Person, and Bardi Kleiser (1999), Bodapati (2008), and Pancras and Sudhir (2007). Applications of LCA to cluster analysis have been explored in Hagenaars and McCutcheon (2002).

As opposed to traditional approaches to cluster analysis such as hierarchical or k-means cluster analysis or data mining approaches such as Kohonen maps, latent class cluster analysis is a model based approach that offers a variety of model selection tools and probability based classification through a posterior probability of membership. The increasing availability of latent class cluster analysis statistical software has increased the interest in the latent class methodology. To date, there are no papers in the literature that compare some of the latent class cluster analysis software packages to see how they differ and whether they yield similar results.

The current article is intended to compare three packages that perform latent class cluster analysis. Latent GOLD® is a commercially available software package available from Statistical Innovations Inc. *(www.statisticalinnovations.com)* that performs a variety of latent class analyses including cluster analysis. MCLUST and poLCA are R software packages that are freely distributed programs *(www.r-project.org)*. In particular, for each of the software packages we will review the operation of the package and the results obtained.

The remainder of the article presents the dataset that will be used to compare the software packages and review the use and results of Latent GOLD,® poLCA, and MCLUST respectively. Finally, we compare our results and discuss our findings.

2. DATASET USED IN THE REVIEW

Our dataset includes variables described in Table 1 and arranged into five groups. This dataset was used in its entirety in Deichmann et al. (2006), Deichmann, Eshghi, Haughton, Woolford, and Sayek (2007), and Eshghi, Haughton, Legrand, Skaletsky, and Woolford (2008); further discussion of the dataset can be found in these references. For the purposes of

Dominique Haughton, Pascal Legrand, and Sam Woolford are on the Data Analytics Research Team (DART), Bentley University, 175 Forest Street, Waltham, MA 02452-4705 (E-mail: Dhaughton@bentley.edu). Pascal Legrand is also affiliated with DART/Groupe ESC—Clermont/CRCGM (France).

DOI: 10.1198/tast.2009.0016

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

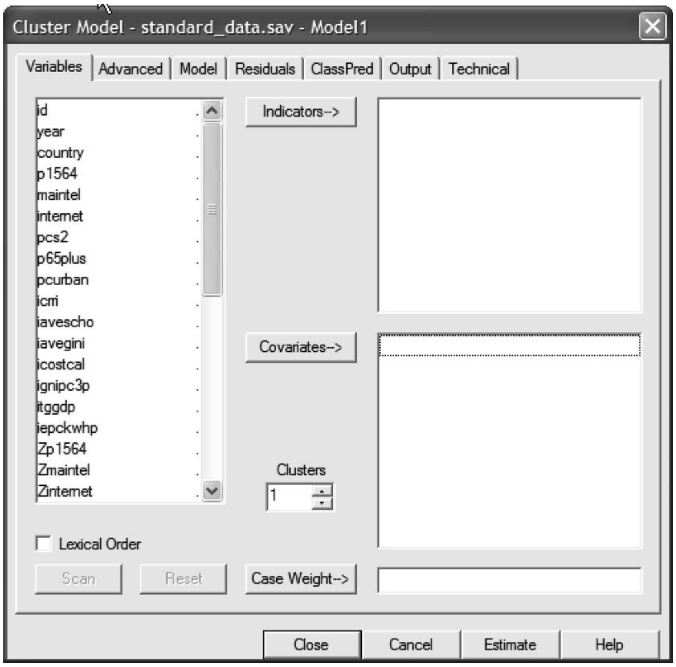
NOTE: ITU, International Telecommunications Union; PRS, Political Risk Services.

this article we are only considering the most recent data from 2003.

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 50 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 evolution.

Our economic variable 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”), the cost of a 3-minute phone call (“costcall”), 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 country’s population that dwells in an urban setting (“urban”).

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 Group in their International Country Risk Guide publications. Political Risk Services’ International Country Risk Guide (ICRG) includes political risk, economic risk, and financial risk measures. The ICRG also reports a measure of composite risk that is a simple function of the three base indices. The guide can be purchased from *http://www.prsgroup.com/ICRG.aspx*. For a critique, please see *http://www.duke.edu/~charvey/Country_risk/pol/pol.htm.* 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 SubSaharan African states to Scandinavian scores in the mid80s. The risk variable is included in our analysis as

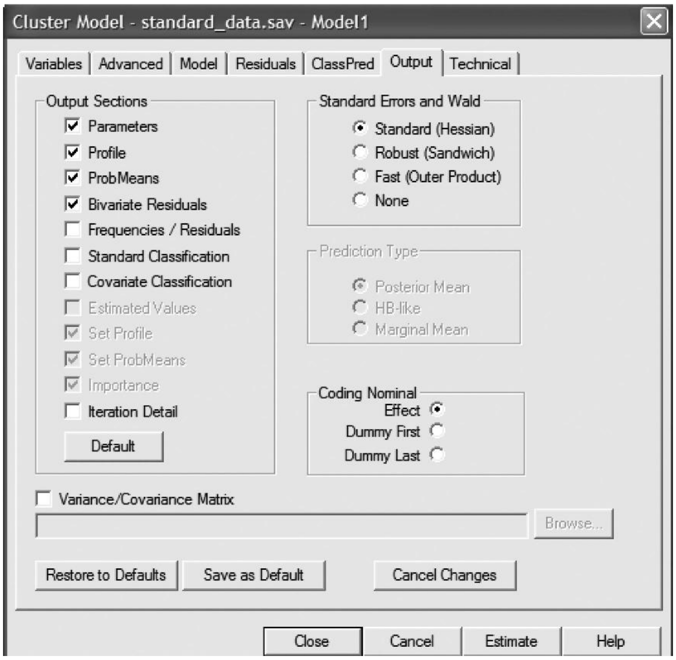


id
year
country
p1564
maintel
internet
pcs2
p65plus
pcurban
icrri
iaveschо
iavegini
icostcal
ignipc3p
itggdp
iepckwhp
Zp1564
Zmaintel
Zinternet

Clusters: 1

☐ Lexical Order

Figure 1. Cluster model window



Output Sections
☑ Parameters
☑ Profile
☑ ProbMeans
☑ Bivariate Residuals
☐ Frequencies / Residuals
☐ Standard Classification
☐ Covariate Classification
☐ Estimated Values
☑ Set Profile
☑ Set ProbMeans
☑ Importance
☐ Iteration Detail

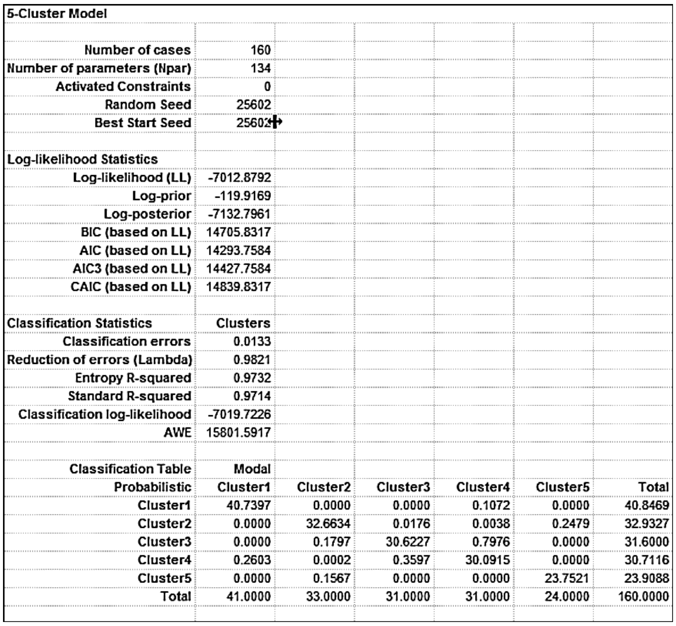
Standard Errors and Wald
● Standard (Hessian)
○ Robust (Sandwich)
○ Fast (Outer Product)
○ None

Prediction Type
● Posterior Mean
○ HB-like
○ Marginal Mean

Coding Nominal
Effect ●
Dummy First ○
Dummy Last ○

☐ Variance/Covariance Matrix

Figure 2. Output window



5-Cluster Model						
Number of cases	160					
Number of parameters (Npar)	134					
Activated Constraints	0					
Random Seed	25602					
Best Start Seed	25602					
Log-likelihood Statistics						
Log-likelihood (LL)	-7012.8792					
Log-prior	-119.9169					
Log-posterior	-7132.7961					
BIC (based on LL)	14705.8317					
AIC (based on LL)	14293.7584					
AIC3 (based on LL)	14427.7584					
CAIC (based on LL)	14839.8317					
Classification Statistics	Clusters					
Classification errors	0.0133					
Reduction of errors (Lambda)	0.9821					
Entropy R-squared	0.9732					
Standard R-squared	0.9714					
Classification log-likelihood	-7019.7226					
AWE	15801.5917					
Classification Table	Modal					
Probabilistic	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Total
Cluster1	40.7397	0.0000	0.0000	0.1072	0.0000	40.8469
Cluster2	0.0000	32.6634	0.0176	0.0038	0.2479	32.9327
Cluster3	0.0000	0.1797	30.6227	0.7976	0.0000	31.6000
Cluster4	0.2603	0.0002	0.3597	30.0915	0.0000	30.7116
Cluster5	0.0000	0.1567	0.0000	0.0000	23.7521	23.9088
Total	41.0000	33.0000	31.0000	31.0000	24.0000	160.0000

Figure 3. Summary model results.

sensible proxy for regularity quality and the rule of law as used in Chinn and Fairlie (2004, 2007), because these variables were not available to us.

Our data were collected from 160 countries. The following variables were fully populated in our dataset: p1564, p65plus, urban, maintel, internet, and computers. For missing cells in other variables we imputed values by regressing predictors on other predictors (but not on ‘‘internet’’ and ‘‘computers’’), as was done in Deichmann et al. (2006, 2007). The percentage of imputed values ranged between 4.6% (for the variable ‘‘trade’’) and 29.4% (for the variable ‘‘electric’’). After imputation, data from 160 countries resulted in a sample size of 480.

3. THE LATENT GOLD® PACKAGE

In this section we discuss a latent class analysis of our data using Latent GOLD® 4.0, a commercially available LCA software package (see Vermunt and Magidson (2005)). Currently LatentGOLD® 4.5 is available from Statistical Innovations; it adds some additional capabilities to LatentGOLD® 4.0 and is available for $995 for a single license. The software provides a number of additional latent class analyses other than just cluster analysis. The User’s Guide and Technical Guide are available at the Statistical Innovations web site, are well written, easy to use, and provide a much more extensive description of the features available in the software for performing cluster analysis than we will provide here. The website also contains numerous tutorials and white papers further highlighting the issues around latent class cluster analysis. For the intent of this article and the results indicated, there are no material differences between Latent GOLD® 4.0 and Latent GOLD® 4.5. The software is Windows based and menu driven. Our analysis was performed on a standard Lenovo T60 laptop computer with an Intel Centrino Duo processor. Installation of the software was straightforward with the disc provided by Statistical Innovations.

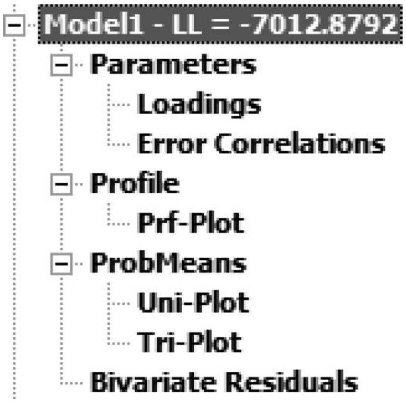


Figure 4. Output options.

The Latent GOLD® User’s Guide provides a 10 step process for carrying out a latent class cluster analysis, which provides a useful approach to using the software to perform cluster analysis. The program will read a number of different data formats but we opted to use an SPSS (.sav) file, which is read directly by Latent GOLD®. Once a dataset is opened and the user indicates a latent class cluster analysis, the key cluster model window opens (see Fig. 1).

Latent GOLD® allows the user to choose the variables that will be used to create the latent clusters and can model multiple data types (nominal, ordinal, and continuous) in the same model. It also allows for the user to designate some variables as covariates to improve the classification. The user can define a specific number of clusters to fit or a range of clusters and a case weight can also be assigned in the analysis. Additional windows allow the user to refine the model parameter assumptions and manage the output (see Fig. 2).

There are also a number of estimation parameters that control the estimation algorithms (a combination of expectation maximization (EM) algorithm and Newton-Raphson algorithms to obtain a maximum likelihood solution) that can be changed in the Technical window to help ensure convergence. Once the user has specified the variables to be used in the model, the output desired and any estimation parameters, the user clicks on **Estimate**, the model(s) are fit, and the output is generated.

The output includes a number of summary statistics (Fig. 3) for the model fit that include various log-likelihood and related statistics in addition to several classification statistics. The Classification Table compares the assignment of cases to clusters based on their modal posterior probabilities of membership as compared with their posterior probability assignment.

Additional graphical and tabular outputs (Fig. 4) assist the user in interpreting the resulting clusters. This output includes:

Estimation and statistical tests of the model parameters indicating their ability to differentiate the clusters:

The size of the clusters;

The marginal distribution of values of each variable within each cluster (Profile) along with line plots showing how the clusters differ on the values of the variables;

Table 1. Latent GOLD® cluster membership

Cluster 1: Angola, Bangladesh, Benin, Bhutan, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Rep, Chad, Comoros, Congo, Côte 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, Kyrgyzstan, Libya, Maldives, Mongolia, Morocco, Namibia, Paraguay, Peru, 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 Emirates

Aggregated cluster membership probabilities (ProbMeans) and triplots to visualize the information; and

Bivariate residuals to identify potential inadequacies in the model.

The software allows for the interactive evaluation of results and the estimation of more refined models. As the user reviews model output and identifies model refinements, additional models may be fitted and added to the open analysis. The various summary statistics can be used to help choose between competing models. The user can also obtain an output file that includes the posterior probabilities of membership for each case.

In using Latent GOLD® to perform a cluster analysis of our dataset, 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 13-cluster solutions were generated using Latent GOLD®. 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 R^2 values are all above 0.5. Based on the bivariate residuals 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

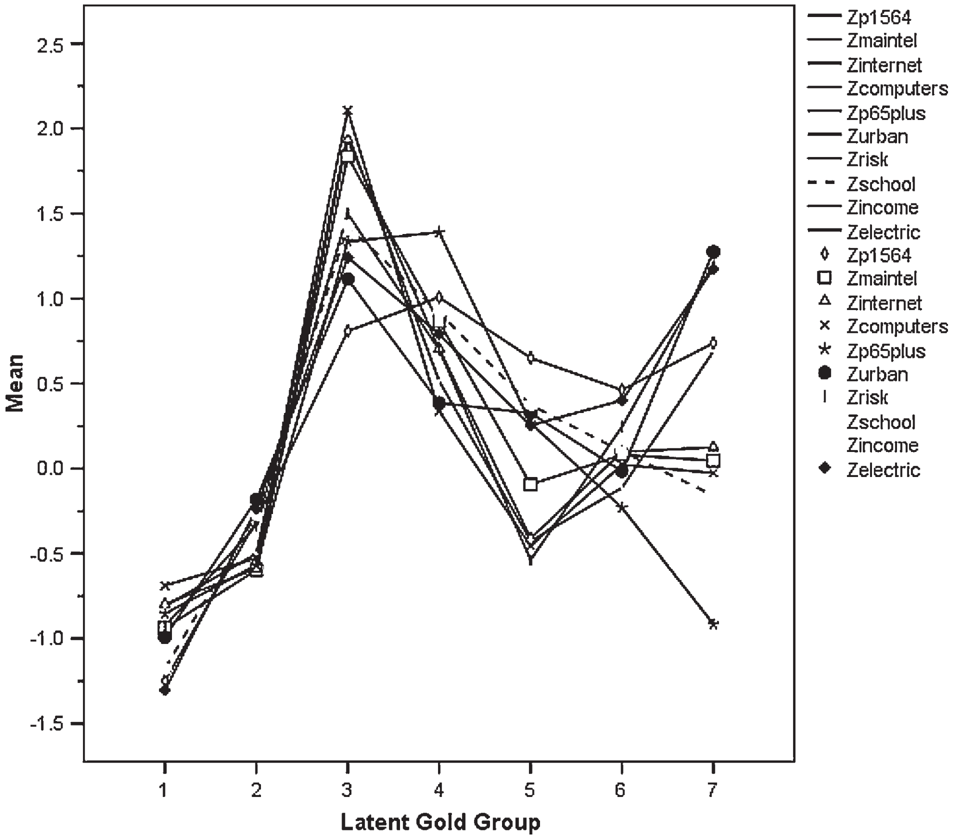


Figure 5. Cluster means of standardized variables (Latent GOLD).

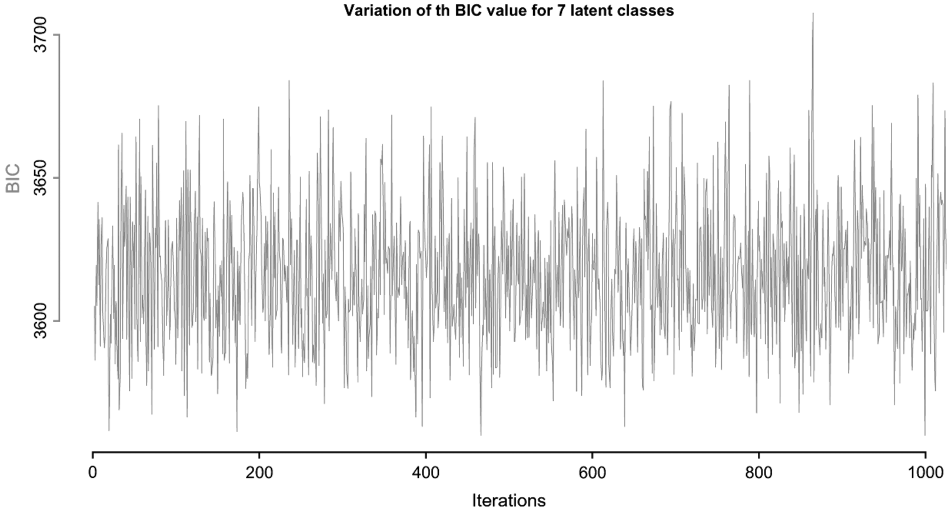


Figure 6. BIC fluctuations.

Bayesian information criterion (BIC) = 2072.2, log-likelihood = −612.3, and Akaike information criterion (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.

Table 1 indicates the resulting cluster membership and Figure 5 displays the cluster means for the standardized variables.

4. THE R SOFTWARE ENVIRONMENT

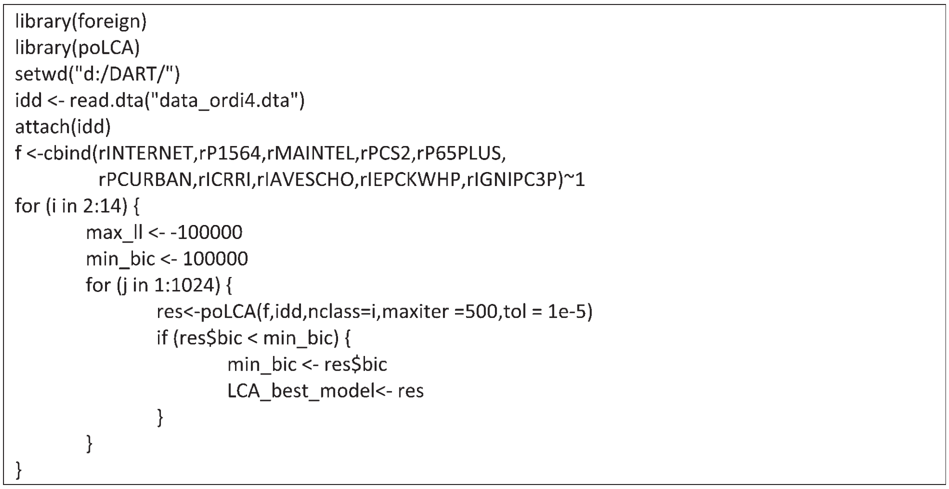
R is a software environment for data manipulation, calculation, and graphical display. It is both an environment and a programming language. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. R runs on a wide variety of platforms (Unix, FreeBSD, Linux, Windows, MacOS). Sources and binaries of R can be downloaded at *http://www.r-project.org*. The installation of R is very simple and a variety of packages can be added directly from the web site. R has a very active development community; many resources can be found including user guides, manuals, script samples, newsgroups, and mailing lists.

R is a command line application. Its integrated object oriented language allows for efficient data manipulation. Whereas the use of R does require programming, scripts can be developed to automate analyses and provide additional functionality. Graphical user interface (GUI)s have been developed for certain applications to avoid user programming (see, for example, Rcommander).

Tests for the R analyses were performed on a HP Compaq 8,510 laptop with an Intel Centrino duo processor. We have used two R packages. The first one, *poLCA*, performs traditional latent class analysis using categorical variables. The second one, *MCLUST*, estimates finite mixture models and allows the use of continuous variables.

4.1 The poLCA Package

The package poLCA is developed by Drew A. Linzer and Jeffrey Lewis of the University of California. This package



```
library(foreign)
library(poLCA)
setwd("d:/DART/")
idd <- read.dta("data_ordi4.dta")
attach(idd)
f <-cbind(rINTERNET,rP1564,rMAINTEL,rPCS2,rP65PLUS,
            rPCURBAN,rICRRI,rIAVESCHO,rIEPCKWHP,rIGNIPC3P)~1
for (i in 2:14) {
          max_ll <- -100000
          min_bic <- 100000
          for (j in 1:1024) {
                    res<-poLCA(f,idd,nclass=i,maxiter =500,tol = 1e-5)
                    if (res$bic < min_bic) {
                              min_bic <- res$bic
                              LCA_best_model<- res
                    }
          }
}
```

Figure 7. R script sample.

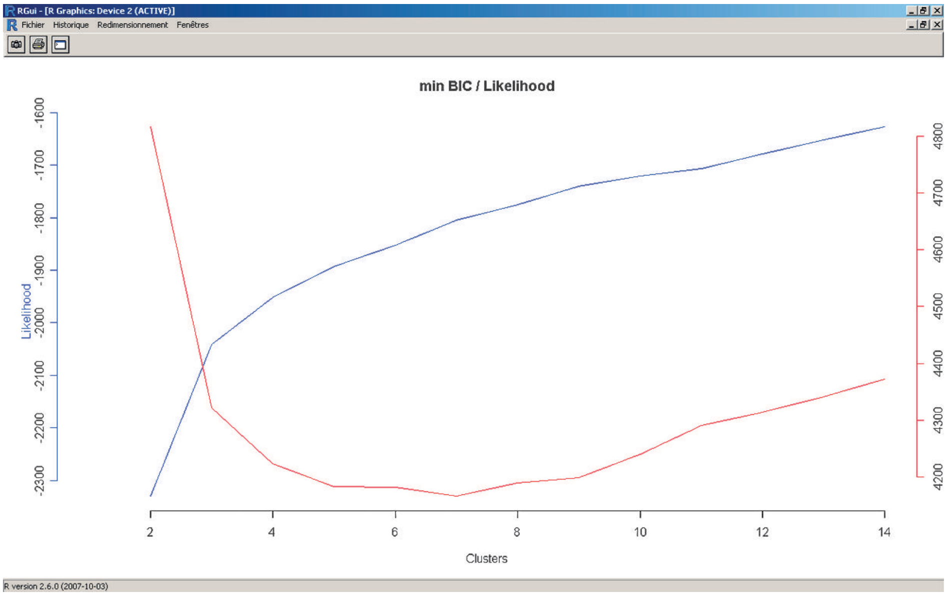


Figure 8. Maximum likelihood and BIC for different numbers of latent clusters.

allows estimation of latent class clusters for polytomous outcome variables. poLCA can also perform latent class regression with categorical data (Linzer and Lewis 2007).

The LCA model is estimated by a call to the *poLCA()* function. It is necessary to specify the model formula (variables selected), the data used, and the number of clusters. The function returns results including the BIC, the AIC, the likelihood, the estimated class-conditional response probabilities, and a matrix containing each observation's posterior class membership probabilities.

The construction of latent class clusters is achieved by maximizing the log-likelihood. For optimization, the Expectation—Maximization (EM) algorithm is used. Results can vary because of the random initialization of this algorithm. Consequently, several iterations are required to counter the risk of identifying a local minimum instead of the global minimum of the BIC. Figure 6 demonstrates the fluctuation in the BIC values across a thousand iterations of the algorithm.

The *poLCA()* function does not search automatically for the best model according to the number of clusters. The R programming language allows the user to automate this search and retain the best model. A sample of the R script used is shown in Figure 7. This R script selects the best LCA model for two to 14-cluster models. The script automates the search for the best model by running the poLCA function 1,024 times for each number of clusters and retaining the model yielding the minimal value of the BIC.

Because poLCA requires polytomous variables, we apply a quartile based conversion to the variables in our dataset. In particular, each original continuous variable is split into four intervals of the same size and a label is assigned to each category. The conversion of continuous variables into categorized ones was performed with Stata (command **xtile).**

To obtain the best classification, we estimated models for one through 14 latent clusters and retained the model giving the lowest BIC criterion among the thousand iterations for each number of clusters. Figure 8 displays the maximum likelihood value and the minimum BIC value for the different numbers of clusters.

Table 2. poLCA cluster membership

Cluster 1: Angola, Bangladesh, Benin, Bhutan, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Congo, Côte d'Ivoire, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Ghana, Guinea, Kenya, Lao (People's Democratic Republic), Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Senegal, Solomon Islands, Sudan, Tanzania, Togo, Uganda, Yemen, Zambia

Cluster 2: Algeria, Belize, Bolivia, Botswana, Cape Verde, Colombia, Ecuador, Egypt, El Salvador, Gabon, Gambia, Guatemala, Guyana, Honduras, India, Indonesia, Jordan, Kyrgyzstan, Maldives, Mongolia, Morocco, Namibia, Nicaragua, Oman, Paraguay, Peru, Philippines, Samoa, Swaziland, Syria, Tonga, Vanuatu, Venezuela, Viet Nam, Zimbabwe

Cluster 3: Australia, Austria, Barbados, Belgium, Canada, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Korea (Republic), Luxembourg, Macao, Malta, Netherlands, New Zealand, Norway, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States

Cluster 4: Bulgaria, Croatia, Estonia, Greece, Hungary, Latvia, Lithuania, Poland, Portugal, Romania, Russia, Slovak Republic, St. Kitts and Nevis, Trinidad and Tobago

Cluster 5: Albania, Armenia, China, Cuba, Georgia, Moldova, Panama, Serbia and Montenegro, Sri Lanka, Thailand, Ukraine

Cluster 6: Argentina, Brazil, Chile, Costa Rica, Dominica, Fiji, French Polynesia, Grenada, Iran (Islamic Republic), Jamaica, Lebanon, Libya, Malaysia, Mauritius, Mexico, Seychelles, South Africa, St.Lucia, St.Vincent, Suriname, Tunisia, Turkey, Uruguay

Cluster 7: Bahrain, Brunei Darussalam, Kuwait, Qatar, Saudi Arabia, United Arab Emirates

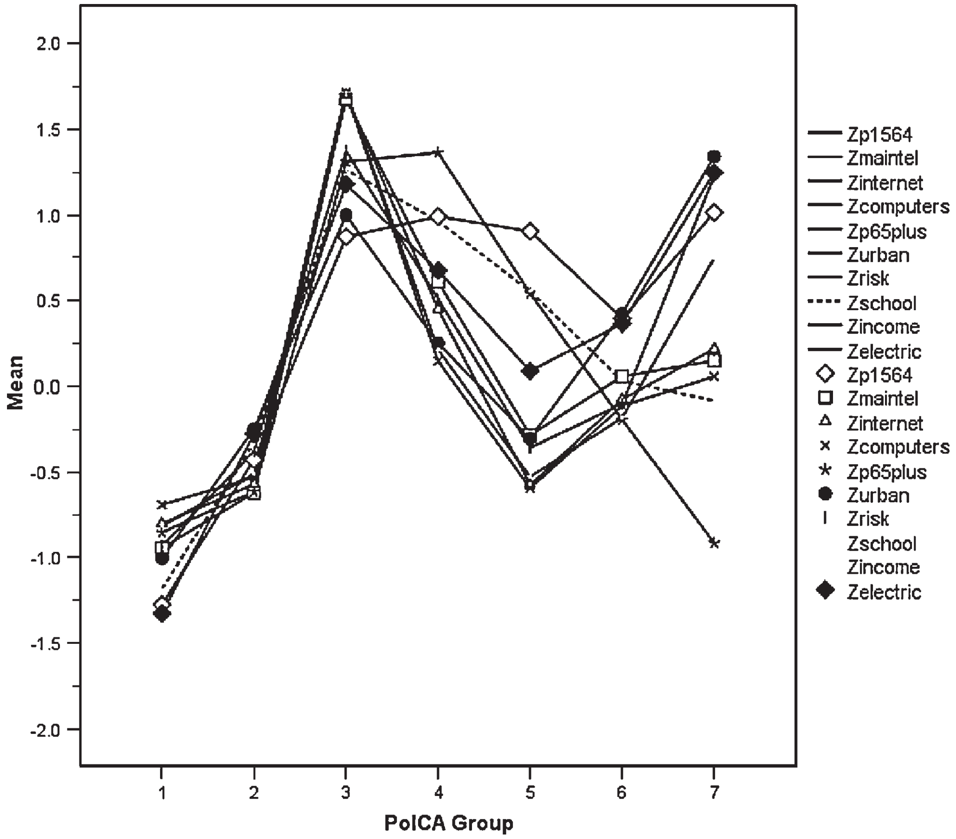


Figure 9. Cluster means of standardized variables (poLCA).

We observe that the optimal value of the BIC is obtained for a model with seven clusters. The minimal value of BIC observed is 3,560.174 (maximum log-likelihood −1,231.968 and AIC: 2,895.937). For this model the classification of countries is presented on Table 2, and the cluster means of standardized variables are displayed on Figure 9.

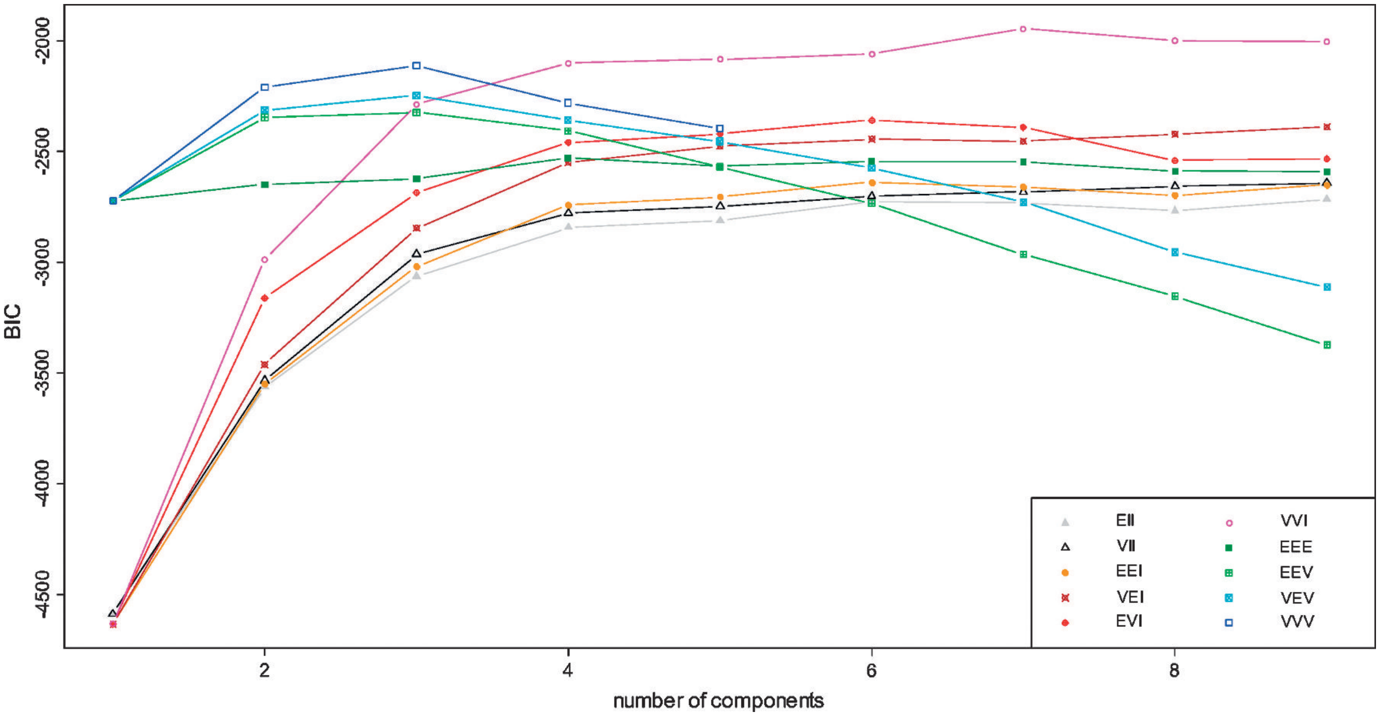


Figure 10. The BIC values according to number of clusters (see x axis) and covariance structures.

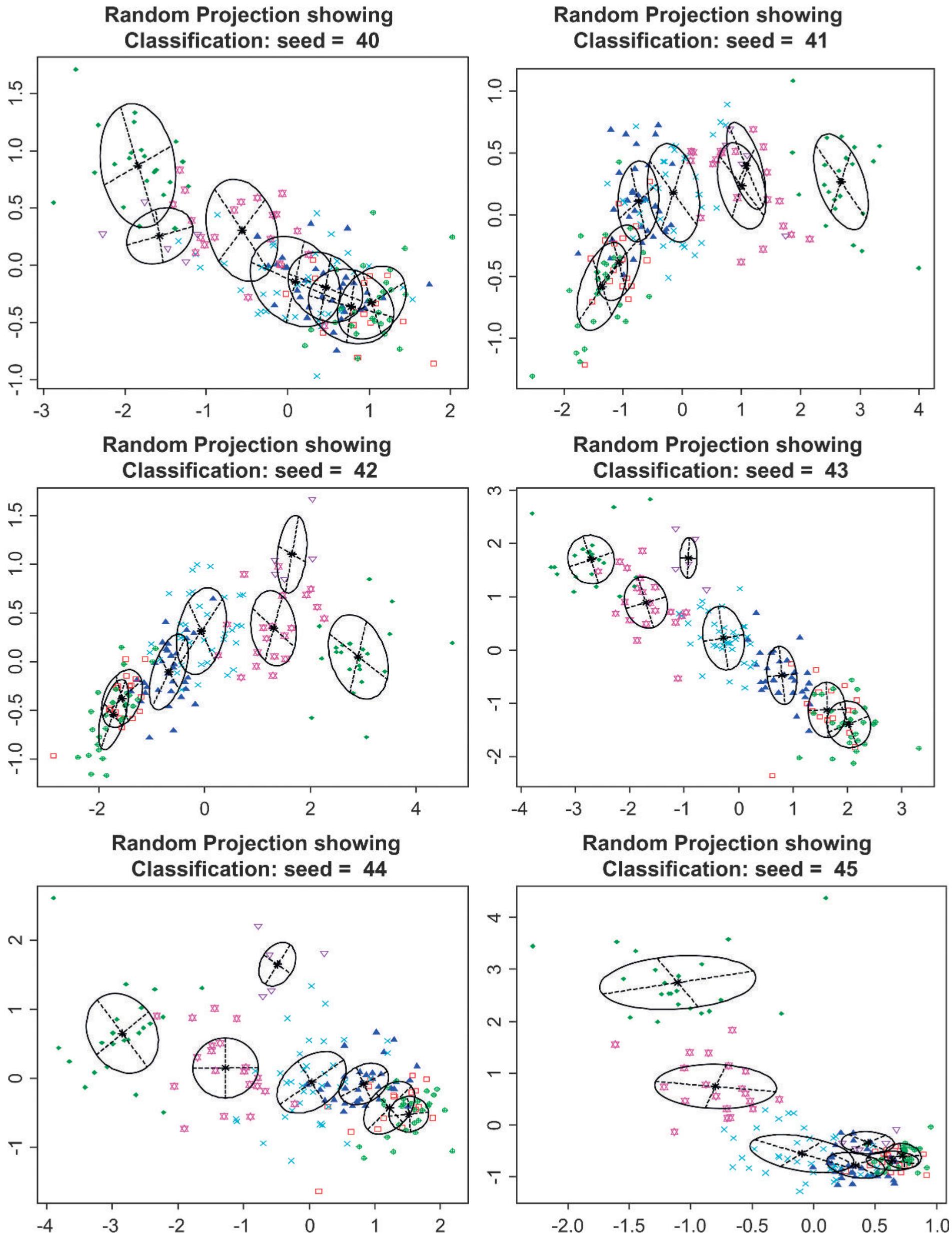
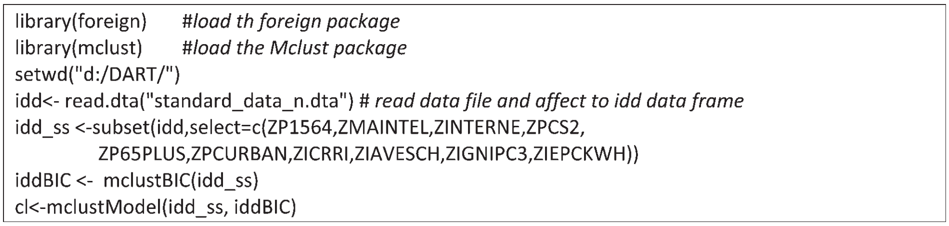


Figure 11. Examples of random projections.

4.2 The MCLUST Package

Because categorization entails a loss of information, we test another R package that performs latent class cluster analysis on continuous data. This package was developed by Chris Fraley and Adrian E. Raftery of the University of Washington and estimates normal mixture models using continuous data (Fraley and Raftery 2002, 2006).

MCLUST is a powerful package that automatically estimates the best mixture model according to different covariance



```
library(foreign)      #load th foreign package
library(mclust)       #load the Mclust package
setwd("d:/DART/")
idd<- read.dta("standard_data_n.dta") # read data file and affect to idd data frame
idd_ss <-subset(idd,select=c(ZP1564,ZMAINTEL,ZINTERNE,ZPCS2,
              ZP65PLUS,ZPCURBAN,ZICRRI,ZIAVESCH,ZIGNIPC3,ZIEPCKWH))
iddBIC <-  mclustBIC(idd_ss)
cl<-mclustModel(idd_ss, iddBIC)
```

Figure 12. R script sample.

Table 3. Mclust clusters

Cluster 1: Argentina, Armenia, Brazil, Chile, China, Costa Rica, Cuba, Dominica, French Polynesia, Georgia, Grenada, Iran (I.R.), Jamaica, Lebanon, Libya, Malaysia, Mauritius, Mexico, Moldova, Oman, Panama, Romania, Russia, Saudi Arabia, Serbia and Montenegro, Seychelles, South Africa, St. Lucia, St. Vincent, Suriname, Thailand, Trinidad & Tobago, Turkey, Ukraine, Uruguay, Venezuela

Cluster 2: Albania, Algeria, Belize, Bolivia, Botswana, Cape Verde, Colombia, Ecuador, Egypt, El Salvador, Fiji, Gabon, Guatemala, Guyana, Honduras, India, Indonesia, Jordan, Kyrgyzstan, Maldives, Mongolia, Morocco, Namibia, Paraguay, Peru, Philippines, Samoa, Sri Lanka, Swaziland, Syria, Tonga, Tunisia, Viet Nam

Cluster 3: Angola, Bangladesh, Benin, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Rep., Chad, Comoros, Congo, Eritrea, Ethiopia, Ghana, Guinea, Lao P.D.R., Madagascar, Malawi, Mali, Mozambique, Myanmar, Nepal, Niger, Nigeria, Tanzania, Uganda, Zambia

Cluster 4: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Hong Kong, Iceland, Ireland, Japan, Korea (Rep.), Luxembourg, Netherlands, New Zealand, Norway, Singapore, Sweden, Switzerland, United Kingdom, United States

Cluster 5: Barbados, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece, Hungary, Israel, Italy, Latvia, Lithuania, Macao, Malta, Poland, Portugal, Slovak Republic, Slovenia, Spain, St. Kitts and Nevis

Cluster 6: Bhutan, Côte d'Ivoire, Djibouti, Equatorial Guinea, Gambia, Kenya, Mauritania, Nicaragua, Pakistan, Papua New Guinea, Senegal, Solomon Islands, Sudan, Togo, Vanuatu, Yemen, Zimbabwe

Cluster 7: Bahrain, Brunei Darussalam, Kuwait, Qatar, United Arab Emirates

structures and different numbers of clusters. The *mclustBIC()* function implements this computation and outputs the BIC value for the different covariance structures and for different numbers of clusters. The best model corresponds to the maximum BIC. A call to the *plot()* function displays these results. Results of the different models are summarized in Figure 10. We have used the normal mixture modeling feature to find the best classification.

The best model is reached with a seven-cluster solution yielding a BIC value of −1,946.037.

A call to the *mclustModel()* function gives additional results about the selected model:
modelName: a character string denoting the model corresponding to the optimal BIC.
n: the number of observations in the data.
d: the dimension of the data (number of variables).
G: the number of mixture components in the model corresponding to the optimal BIC.
bic: the optimal BIC value.
loglik: the log-likelihood corresponding to the optimal BIC.
z: a matrix whose [i,k]th entry is the probability that observation i in the test data belongs to the kth class.

With two variables, it would be possible to represent on a graph the observations and the classification into clusters. When more than two variables are involved, graphical visualization is difficult or impossible. A way to visualize results is to make projections of the observations on a two-dimensional

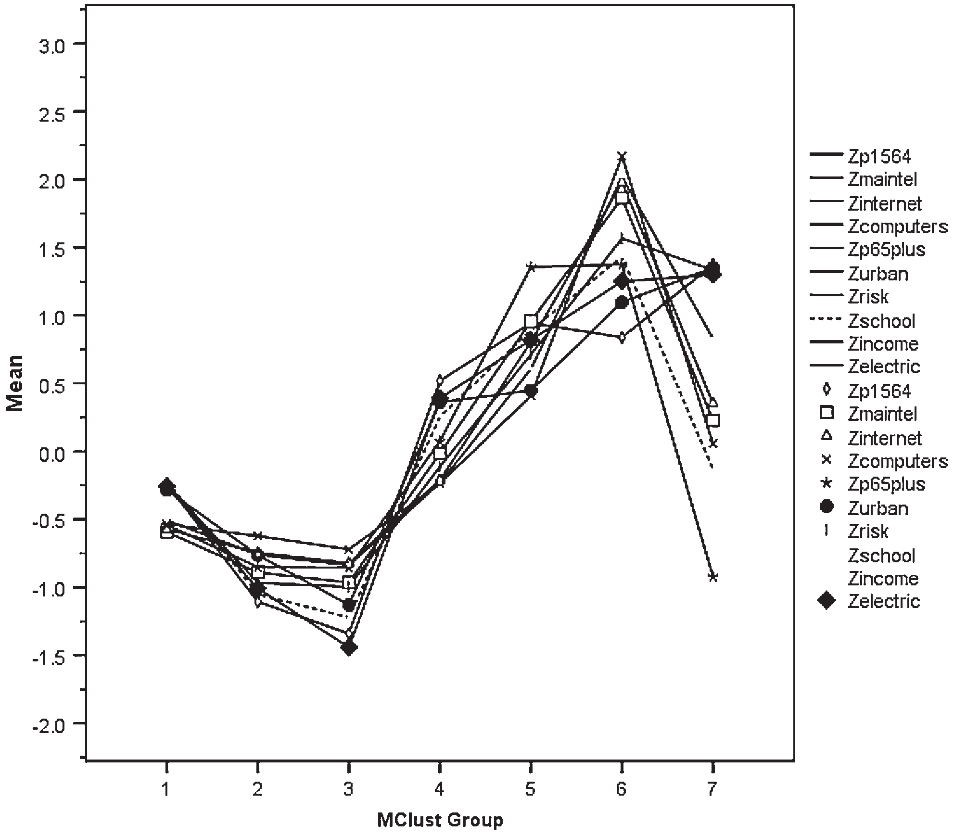


Figure 13. Cluster means of standardized variables (Mclust).

Table 4. Cluster performance results

	Homogeneity	Heterogeneity h1	Heterogeneity h2
poLCA results	1.496690	320.3653	18.38455
MCLUST results	1.334718	438.031	24.25434
Latent GOLD results	1.490479	339.9556	19.23558

plane; but there is an infinite number of possibilities for the choice of a suitable two-dimensional plane. The *randProj()* function computes a projection on a random two-dimensional plane; a seed is required to initialize the random process. Several runs with different values of the seed can be necessary to obtain a useful visualization. This function can plot classification and uncertainty or errors. Examples of classifications corresponding to different seeds are presented on Figure 11.

The seed must be in the range [0;1,000] ; it is of course impossible to plot all the graphs to find the best graphical representation. The R script to use MCLUST is given in Figure 12 and the cluster solution determined by MCLUST is presented in Table 3 and cluster means are displayed on Figure 13.

5. COMPARISON OF RESULTS

All three analyses resulted in the same number of clusters in the final solution. To perform a comparison of our results, we use measures that evaluate the performance of each analysis on the basis of quantities typically employed to evaluate the efficacy of a cluster analysis:

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

The measures that we use were originally proposed in (Eshghi, Haughton, Legrand, Skaletsky, and Woolford 2008). In essence, **Homogeneity** is a measure of the variance within the clusters whereas **Heterogeneities h1** and **h2** are different measures of the distance between clusters. Typically, one of the objectives of performing cluster analysis is to obtain resulting clusters that have low variability within clusters while providing a high degree of separation between the clusters. Consequently, we are looking for low values of **Homogeneity** and high values of **Heterogeneity**. These results are presented in Table 4.

Latent Gold® clusters

poLCA clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
1	0	2	0	0	9	12	0	*23*
2	0	1	0	0	9	1	0	*11*
3	40	0	0	0	0	0	0	*40*
4	0	0	0	13	0	1	0	*14*
5	3	30	0	0	1	0	1	*35*
6	0	0	23	8	0	0	0	*31*
7	0	0	0	0	0	0	6	*6*
	43	*33*	*23*	*21*	*19*	*14*	*7*	

Figure 15. Comparison of Latent GOLD® and poLCA clusters.

Table 4 reveals that MCLUST outperformed Latent GOLD®, which outperformed poLCA. It should be noted that each package produces different cluster solutions as indicated by Figures 14, 15, and 16. Figure 14 suggests that cluster 1 in the Latent GOLD® solution is split into two MCLUST clusters (2 and 3) and that clusters 5 and 6 in the Latent GOLD® solution correspond to MCLUST cluster 4. Similarly, Figure 15 and 16 show that each analysis results in a cluster solution that exhibits some fundamental differences with the other solutions.

6. CONCLUSIONS

All three packages provide the user with capabilities to perform latent class cluster analysis and each has strengths and weaknesses. From the perspective of usability, Latent GOLD® is the easiest to use with well written and usable documentation and a GUI interface that eliminates the need for user 'programming'. In addition, Latent GOLD® automatically provides a variety of output, graphics and diagnostics to help the user interpret the resulting clusters and to refine their analysis. The R packages poLCA and MCLUST both require the user to learn how to use the R programming environment and language. Both R packages have the capability to provide posterior probabilities of membership but other diagnostics and analysis capabilities may require the use of additional R packages. The documentation for R is extensive but not as complete as that provided by Latent GOLD.® On the positive

Latent Gold® clusters

Mclust clusers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
1	0	30	0	0	3	0	0	*33*
2	16	1	0	0	0	0	0	*17*
3	27	0	0	0	0	0	0	*27*
4	0	2	0	2	16	14	2	*36*
5	0	0	1	19	0	0	0	*20*
6	0	0	22	0	0	0	0	*22*
7	0	0	0	0	0	0	5	*5*
	43	*33*	*23*	*21*	*19*	*14*	*7*	

Figure 14. Comparison of Latent GOLD and MCLUST clusters.

Mclust clusters

poLCA clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
1	2	0	0	21	0	0	0	*23*
2	2	0	0	9	0	0	0	*11*
3	0	13	27	0	0	0	0	*40*
4	0	0	0	3	11	0	0	*14*
5	29	4	0	2	0	0	0	*35*
6	0	0	0	0	9	22	0	*31*
7	0	0	0	1	0	0	5	*6*
	33	*17*	*27*	*36*	*20*	*22*	*5*	

Figure 16. Comparison of poLCA and MCLUST clusters.

side, however, both R packages are free whereas there is a substantial cost associated with Latent GOLD®, which may make poLCA and MCLUST more easily accessible.

The nature of the data (continuous, discrete or mixed) included in the analysis may also play a role in determining which software is most appropriate. From a performance perspective, it would appear that there is some loss of performance, in terms of heterogeneity and homogeneity of the resulting clusters, when using poLCA with continuous data. In general we might expect this performance loss due to the loss of information resulting from transforming continuous data into categorical data. This performance loss may be able to be reduced in some cases by using the empirical data distribution to help categorize the data based on natural break points in the data. We have not tested to see if this loss of performance would still result if the original data are categorical to start with. It would suggest, however, that the user should consider the variable types when deciding which R package to use. Because Latent GOLD® handles both continuous and discrete variables in the same analysis, this should not be an issue.

For the dataset used here, it appears that the results obtained using Mclust outperform those obtained by Latent GOLD®on the cluster measures that we applied. We draw no conclusions as to whether this would be the result in all cases. Our results would suggest that all three approaches, while resulting in the same number of clusters for this dataset, produce different clusters. Even for Latent GOLD®and MCLUST, which might be expected to give the same results for the same number of clusters, the resulting clusters are structurally different and so could lead to different interpretations.

REFERENCES

Bodapati, A. V. (2008), “Recommendation Systems with Purchase Data,” *JMR, 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 (1)**,** 16–48.

—— (2004), “*The Determinants of the Global Digital Divide: A Cross-Country Analysis of Computer and Internet Penetration*,” IZA Discussion Paper No. 1305, Bonn: Forschungsinstitut zur Zukunft der Arbeit.

Cooil, B., Keiningham, T. L., Askoy, L., and Hsu, M. (2007), “A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating 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 Analytical Approach,” *Journal of Global Information Technology Management,* 9 (4)**,** 47–71.

Deichmann, J. I., Eshghi, A., Haughton, D., Woolford, S., and Sayek, S. (2007), “Measuring the International Digital Divide: An Application of Kohonen Self-Organizing Maps,” *International Journal of Knowledge and Learning,* 3 (6)**,** 552–575.

Dimitrova, D., and Beilock, R. (2005), “Where Freedom Matters: Internet Adoption Among the Former Socialist Countries,” *The International Journal for Communication Studies,* 67 (2)**,** 173–187.

Eshghi, A., Haughton, D., Legrand, P., Skaletsky, M., and Woolford, S. (2008), *Identifying Groups: A Comparison of Methodologies* (preprint).

Finkbeiner, C., and Waters, K. (2008), “Call Every Shot,” in *Marketing Management, American Marketing Association, January/February 2008*, pp. 38–43.

Fraley, C., and Raftery, A. E. (2002), “Model-Based Clustering, Discriminant Analysis and Density Estimation,” *Journal of the American Statistical Association,* 97, 611–631.

Fraley, C., and Raftery, A. E. (2006), *“{MCLUST} Version 3 for {R}: {N}ormal Mixture Modeling and Model-Based Clustering,”* Technical Report Number 504, Department of Statistics, University of Washington, September.

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

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

Linzer, D. A., and Lewis, J. (2007) ‘poLCA: Polytomous Variable Latent ClassAnalysis’, R package version 1.1, http://userwww.service.emory.edu/∼dlinzer/poLCA.

Malhotra, N. K., Person, M., 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 (2)**,** 160–183.

Pancras, J., and Sudhir, K. (2007), “Optimal Marketing Strategies for a Customer Data Intermediary,” *JMR, Journal of Marketing Research,* 44, 560–578.

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