# Clustering of categorical data: a comparison of a model-based and a distance-based approach

Laura Anderlucci [1]

*Department of Statistical Sciences,*
*University of Bologna, Italy*

Christian Hennig [2]

*Department of Statistical Science,*
*University College London, UK*

[1] Electronic address: `laura.anderlucci@unibo.it`; Corresponding author
[2] Electronic address: `ucakche@ucl.ac.uk`

## Abstract

For clustering multivariate categorical data, a latent class model-based approach (LCC) with local independence is compared with a distance-based approach, namely partitioning around medoids (PAM). A comprehensive simulation study was evaluated by both a model-based as well as a distance-based criterion. LCC was better according to the model-based criterion and PAM was sometimes better according to the distance-based criterion. However, LCC had an overall good and sometimes better distance-based performance as PAM, though this was not the case in a real data set on tribal art items. ~~Both methods produced significantly more homogeneous clusters than the truth.~~

# 1 Introduction

This paper is about cluster analysis with multivariate categorical data. It has often been noted that cluster analysis is not a well defined problem. “Clusters” are groups of data points that belong together in some sense, but there are various possible meanings of “belonging together”. In the present paper we consider particularly two different meanings. In model-based (latent class) clustering (e.g., Vermunt and Magidson (2002)), the “true” clusters are defined by parametric probability distributions that can be interpreted to generate homogeneous points, and the whole data set is modelled by a mixture of such distributions. On the other hand, there is a long tradition to regard as clusters data subsets that have small within-cluster distances and large separation from other clusters (e.g., Everitt et al. (2011), Sec. 1.4), which can be called a distance-based cluster concept. It is not obvious, and depends on details such as the specific parametric model chosen, to what extent these cluster concepts coincide.

For multivariate categorical data, a standard parametric model used in latent class clustering is a locally (i.e., within-clusters) independent product of multinomial distributions (Vermunt and Magidson (2002)). We will compare this to partitioning around medoids (PAM, Kaufman and Rouseeuw (1990)), a distance-based clustering method that does not attempt to fit a mixture distribution. Both methods can legitimately be applied to the same data. There is not much literature guiding users about whether to use one or the other, and so it can be presumed that any of the two methods is preferred by some users for the same kind of application. It is however not obvious at all that both methods serve the same aims. Users may be interested in finding well separated clusters with low within-cluster distances and also, at the same time, they would like to be reassured that the “true” clusters (corresponding to homogeneous mixture components) are found if data actually were generated from a latent class model. But if this succeeds, will clusters actually be well separated and have low within-cluster dissimilarities? Or are the two aims conflicting? The local independence assumed in the latent class model allows a nice and well interpretable reduction of the number of parameters to fit an empirical distribution, but it does not necessarily guarantee distance-based homogeneity. These issues have been hardly addressed in the literature with the notable exception of Celeux and Govaert (1991), who related the latent class-likelihood to a similarity-based criterion for binary data.

Comparing the different methods is not straightforward either, because typical objective functions measuring clustering quality adhere to either of the two paradigms. They may measure misclassification compared to the “true” latent class mixture model, or they may measure within-cluster homogeneity and between-cluster separation in a distance based way. In the first case this will favour the model-based approach, in the second case it will favour the distance-based approach. We face this dilemma by investigating how well the methods work according to objective criteria that apparently prefer the respective other approach, and to what extent the approaches coincide.

When presenting this work, we have been confronted on several occasions by statisticians claiming that finding the “true” clustering is really the only legitimate aim and that distance-based quality should be seen as a by-product only. In real applications, this is however not so clear because there is no guarantee that there is an underlying true model defining a “true” clustering at all. Misclassification compared to the “true” clusters is not observable (unless the truth is known and unsupervised classification is no longer of real interest), whereas cluster homogeneity and separation are observable. Even if data are in fact generated from a latent class mixture, in an application one may still be interested in finding homogeneous and separated clusters in the first place. So the misclassification rate is not the unique “correct” quality measure as opposed to a distance-based one, but this depends on the aim of clustering.

In Section 2, we introduce the applied methods formally. In Section 3, we present the simulated data generating processes. In Section 4, we present and discuss the simulation results and in Section 5, the methods are compared on real data on tribal art objects. Section 6 concludes the paper with a discussion.

## 2 Methods

A well known model-based clustering method for categorical data is the Latent Class Clustering (LCC) (Vermunt and Magidson (2002)): it assumes that data are generated by a mixture of underlying probability distributions, where each mixture component represents a single cluster (i.e. latent class). In the case of nominal variables, the underlying model is a mixture of multinomial distributions.

Consider the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}_i \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_p$, $i = 1, \ldots, n$, where $\mathcal{U}_j$, $j = 1, \ldots, p$, are unordered finite sets. Data can be represented in a $p$-way contingency table that cross-classifies a sample of $n$ individuals with respect to $p$ manifest polytomous variables. The LCC states that the observed relationships - if any - among the $p$ variables can be explained by a $K$-class latent structure (i.e. a latent polytomous variable $K$), so that each of the $n$ individuals is in only one of the $K$ classes with respect to this variable, and locally, i.e., within the $k^{th}$ latent class, the manifest variables are mutually independent (Goodman (1974)). How to determine the number $K$ of latent classes in the study population is an unresolved issue. Currently, applied researchers use a combination of criteria to guide the decision; such criteria include agreement with substantive theory and the combination of statistical information criteria, like Akaike Information Criterion (AIC; Akaike (1987)) and Bayesian Information Criterion (BIC; Schwartz (1978)); for further references see Nylund et al. (2007).

The model is described by equation 1:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}, \mathbf{a}_k), \tag{1}$$

with $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$, i.e., the mixing proportions sum to 1. The probability mass function $f(\mathbf{x}, \mathbf{a}_k)$ describes a multinomial distribution with parameters $\mathbf{a}_k = (\mathbf{a}_k^{jl},\ l = 1, \ldots, m_j,\ j = 1, \ldots, p)$:

$$f(\mathbf{x}, \mathbf{a}_k) = \prod_{j=1}^{p} \prod_{l=1}^{m_j} (\mathbf{a}_k^{jl})^{\mathbf{x}^{jl}}, \tag{2}$$

with $\sum_{l=1}^{m_j} \mathbf{a}_k^{jl} = 1$. The generic polytomous variable $j$ $(j = 1, \ldots, p)$ consists of $m_j$ categories, and $m = \sum_{j=1}^{p} m_j$ indicates the total number of levels.

The parameters of the model (1) can be estimated by maximum likelihood, using for example the EM algorithm. Then, for each $\mathbf{x}_i$, its posterior class-membership probabilities are computed from the estimated model parameters and its observed score; units are thus assigned to the class with the highest posterior probability.

A clustering approach based on distance, instead, does not require an underlying model. Nevertheless it cannot be considered as a totally assumption-free option, because the definition incurs implicit assumptions about the nature of the clusters to be found.

The idea is to evaluate distances among objects by a defined dissimilarity measure and, basing on it, to allocate units to the closest group. In other words, the aim is to partition the observations in such a way that objects within the same group are similar to each other, whereas objects in different groups are as dissimilar as possible.

Since different values of a nominal variable should not carry numerical information (unless there are interpretative reasons that can justify it), categorical variables were replaced with binary indicator variables for all their values~~, which means that $q$ above is the number of all categories of all $p$ categorical variables~~.

The dissimilarity measure used in this context is the Manhattan (or city block or $L_1$) distance for $q$ variables, defined by:

$$\begin{aligned} d_M(i,j) &= |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{iq} - x_{jq}| \\ &= \sum_{l=1}^{q} |x_{il} - x_{jl}| \end{aligned} \tag{3}$$

In this context, the $q$ above coincides with the sum of all categories over all $p$ categorical variables.

There are various dissimilarity measures that could have been used (see Section 3.2.2 of Everitt et al. (2011); formula (3.1) there actually amounts to the $L_1$-distance divided by $p$); the decision of using the $L_1$ was motivated by the fact that there was no prior knowledge about the variables and by choosing this measure what mattered was the number of disagreements. In this case, the city block distance is equivalent to the Euclidean distance.

One of the distance-based methods that can be viewed as an alternative to the LCC is partition around medoids (PAM; Kaufman and Rouseeuw (1990)). The idea is to find $K$ representative "central" objects for the clusters. Specifically, they are those units for which the average dissimilarity to all the objects of the same cluster is minimal. The objective function, for a given dissimilarity measure d, is

$$g(\mathbf{x}_1^*, \dots, \mathbf{x}_K^*) = \sum_{i=1}^{n} \min_{j \in \{1,\dots,K\}} d(\mathbf{x}_i, \mathbf{x}_j^*), \tag{4}$$

Each of $\mathbf{x}_j^*$ is called the **medoid** of the cluster. After finding the set of medoids, each object of the data set is assigned to the nearest medoid. It is similar to the more popular $k$-means algorithm (e.g., Everitt et al. (2011), Sec. 5.4, see also Section 6), but here the centers are members of the data set and not the cluster means.

In the simulation study, results from the two clustering methods were compared according to the Adjusted Rand Index and the Average Silhouette Width (for further details on the simulation study see Anderlucci (2012)).

The Adjusted Rand Index (ARI) is a measure of similarity between two data clusterings; it takes values in $[-1, 1]$ (Hubert and Arabie (1985)) and is adjusted in order to have an expected value of 0 for unrelated clusters, while the unadjusted version (Rand (1971)) yields a value between 0 and 1.

Given a set of n elements $S = \{O_1, \dots, O_n\}$ and two partitions of S to compare, $U = \{u_1, \dots, u_R\}$ and $V = \{v_1, \dots, v_C\}$, the following is defined:

- a, the number of pairs of elements in S that are in the same set in U and in the same set in V;
- b, the number of pairs of elements in S that are in different sets in U and in different sets in V;
- c, the number of pairs of elements in S that are in the same set in U and in different sets in V;
- d, the number of pairs of elements in S that are in different sets in U and in the same set in V;

The Adjusted Rand Index is calculated as:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{5}$$

In the simulation study, the ARI was used to compare the classifications yielded by a model-based and a distance-based clustering approach with respect to what was recorded as true cluster membership, defined by the underlying model.

The Average Silhouette Width (ASW), described in Kaufman and Rouseeuw (1990), is a measure that emphasises the separation between the clusters and their neighbouring clusters. For a partition of $\mathbf{x}$ into clusters $C_1, \ldots, C_K$ let

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

be the so called ‘silhouette width’, where

$$\begin{aligned} a(i) &= \frac{1}{|C_h| - 1} \sum_{\mathbf{x}_j \in C_h} d(\mathbf{x}_i, \mathbf{x}_j), \\ b(i) &= \min_{\mathbf{x}_i \notin C_l} \frac{1}{|C_l|} \sum_{\mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

for $\mathbf{x}_i \in C_h$. The $a(i)$ is the average dissimilarity of object $\mathbf{x}_i$ to all other objects of the same cluster $C_h$, while $b(i)$ is the neighbour of $\mathbf{x}_i$; this is like its second-best choice.

It is worth to note that the ASW is often used to estimate the number of clusters. Since in this context the number of latent classes was assumed to be fixed and known, it was used to compare the quality of the clustering. The assumption that the number of latent classes is known is very strong and it rarely happen in practise, but here the objective is to compare a model-based and a distance-based clustering approach with respect to the underlying true clustering structure and in terms of clustering quality in specific situations all other features being equal, so that differences between the two methods can be better detected and evaluated.

Furthermore, since LCC is by definition aimed to recover the true classification, it could be seen as favoured by the ARI; whereas, since PAM is a distance-based approach, it was expected to be favoured by the ASW.

## 3 Simulation setup

A simulation study serves as a basis to understand similarities and differences in terms of classification performances of the two approaches and to detect, if any, different roles played by features of the data sets. Simulations consisted of generating several data sets from different parameterizations and structures (i.e. 2000 data sets for each setting) with the LatentGold® software. The model's parameter values were fixed according to a simulation scheme (i.e. full factorial design) that allowed for examining the impact of the following aspects:

- number of latent classes (2/3/5): we generated data from models with 2 and 5 latent classes, and in a few cases from 3 latent classes (namely when the too small number of variables and levels would not have allowed for 5 identified classes);
- number of observed variables (4/12) and number of their categories (2/4/8): data has been generated from models with small and large number of variables; the variables considered each time were respectively only binary, only 4-levels, only 8-levels variables and with a different number of categories, specifically
  - 4 variables, namely one binary variable, two variables with three categories and one variable with four categories, for a total of 11 binary variables ;
  - 12 variables, namely three binary variables, three variables with three levels, four variables with four categories and two variables with eight categories, for a total of 44 binary variables;
- entity of mixing proportions (extremely different/equal): data sets were generated according to models that have allowed for different mixing proportions (e.g. a model with two clusters can have $\pi_1 = 0.85$ and $\pi_2 = 0.15$, while a model with five clusters can have $\pi_1 = 0.10$, $\pi_2 = 0.15$, $\pi_3 = 0.20$, $\pi_4 = 0.25$, and $\pi_5 = 0.30$) and for clusters supposed to have about the same size (e.g. a model with two clusters can have $\pi_1 = \pi_2 = 0.5$, while a model with five clusters can have $\pi_1 = \ldots = \pi_5 = 0.2$);

- expected cluster separation (clear/unclear): parameters values have been chosen with the idea of having, on one hand, a situation where clusters do not have a clear characterization (hence one would expect to have overlapped clusters) and, on the other hand, a situation where clusters have an evident characterization (therefore one would expect to have clearly separated clusters). In other words, in the former case parameterizations of different clusters are very similar (e.g. in the case of binary variable $j = 1$, for cluster $k = 1$ the $a_k^{jl}$ parameters of the probability function can be $a_1^{11} = 0.6$, $a_1^{12} = 0.4$, while for cluster $k = 2$ they can be $a_2^{11} = 0.65$, $a_2^{12} = 0.35$); in the latter case, different clusters have very different parameterizations (e.g. in the case of binary variable $j = 1$, for cluster $k = 1$ the $a_k^{jl}$ parameters of the probability function can be $a_1^{11} = 0.8$, $a_1^{12} = 0.2$, while for cluster $k = 2$ they can be $a_2^{11} = 0.2$, $a_2^{12} = 0.8$, so that cluster 1 is likely to contain mostly observations with level 1 of the considered variable, while cluster 2 is likely to mostly contain observations with level 2);

- number of units for each data set (small samples/big samples): for each of the previous framework we generated data sets with a small number of units, typically 100, but in a few cases 200 or 500, depending on the sample size needed in order to estimate the model), and a big number of units, namely 1000.

From the combination of all these specific features 128 different settings were obtained, called 'patterns' (for a full description of the simulation settings see the Appendix of Anderlucci (2012)). We designed all simulated setups so that they are identifiable, see Allman et al. (2009). Latent GOLD® was used for performing LCC. In order to find maximum likelihood estimates for the model parameters, it uses both EM and Newton-Raphson algorithms; the estimation process starts with 250 EM iterations. When close enough to the final solution, the program switches to Newton-Raphson, carrying on for other 50 iterations (Vermunt and Magidson (2005)). To avoid local maxima, each process was started from 20 different sets. PAM was also applied, using the `pam` function from the R-package `cluster` with default settings (the dissimilarity matrix and the number of clusters were given as input; since the medoids were not specified, the algorithm first looked for a good initial set of medoids, then it found a local minimum for the objective function, that is, a solution such that there is no single switch of an observation with a medoid that will decrease the objective).

A summary of the simulation results can be found in Tables 1, 2, 3, 4, 5, 6, 7, 8.

# 4 Simulation results

Table 1 and Table 2 contain the average values of the ARI and the ASW for each simulation pattern which involved binary variables only, with unclear and clear cluster separation. Table 1 shows that values of the ARI are generally higher for LCC, given the other data features. Note that differences between the two approaches in terms of ARI become smaller if clusters are expected to be (according to the parametrization that generated the data) clearly separated. Indeed, from Table 2 it is possible to see that their values are really close to each other; nevertheless almost all of these differences are significant, because standard errors (written in brackets) are fairly small because of the large number of repetitions. On the other hand, Table 1 shows that as long as the number of the considered variables is small (i.e. equal to 4) PAM actually outperformed LCC in terms of ASW, even though differences are generally low. Where clusters are expected to be clearly separated (Table 2), the two approaches generally yielded similar results, even though there are cases where LCC was slightly better.

Similar considerations can be done from Tables 3 and 4, where results refer to data sets with number of categories for each observed variable increased to four. The only difference is that PAM performed a little bit better in terms of ASW when clusters were not expected to overlap.

When the number of categories for each observed variable increased to eight, LCC and PAM are less able to find the true clustering (see Table 5), since values of the ARI are lower than those of Table 1 and of Table 3. As previously observed, PAM shows its better performance in terms of ASW when the number of variables is fairly small. This does not hold where clusters are supposed to be separated (Table 6). In those cases, surprisingly PAM performed at most as well as LCC. Values themselves are not low, they are actually very good, but no longer better than those from LCC clustering. When the number of variables is

12, again PAM performed a little bit worse than LCC.

Finally, the case where the variables do not have the same number of categories was considered. In this framework, again LCC outperformed PAM in finding the true clustering, but the outcome of the latter was not much worse (see Table 7). It has to be said that the average performances of the two approaches are much higher if we consider the situations where clusters are supposed to be clearly separated (see Table 8). Indeed, when the clusters are expected to be clearly separated there is no particular evidence to prefer one of the two methods in terms of ASW either, because values are about the same here, too.

Overall, the simulations tell us that, in terms of recovering the 'true' clustering (according to a 'true' unknown model), LCC generally behaves better, yielding better results in terms of ARI, even when the clusters are supposed to overlap. With strongly separated clusters, PAM does not make the results worse, though.

PAM's performances improve when the mixing proportions of the components of the mixture that generate the data are about the same, i.e. when the clusters have about the same size. This is apparently due to the fact that in general PAM seems to prefer equally-sized clusters (similarly to what information criterion clustering does, Celeux and Govaert (1991)).

What is more surprising is that LCC, by trying to put together observations coming from the same distribution, succeeded in getting similar observations together and in separating objects that are very different in a way that is not much worse than what PAM usually does, and actually sometimes it works even better (in 63 over 128 cases, on average). On one hand this is encouraging, because it means that attempting to find the true clustering in a locally independent model-based sense, we often also get clusters that are internally homogenous. On the other hand, this is not a very good result for PAM. There are still situations in which PAM works better than LCC, though, and therefore its use can still be beneficial.

Obviously, it is possible to evaluate the ASW for the true clustering, too. Actually, both LCC and PAM achieved values of the ASW significantly higher than the true clustering. The reason is that random variation sometimes produces observations that are quite atypical for their true cluster. PAM generally attempts to produce compact clusters and can therefore be expected to achieve higher ASW values, but also the “highest posterior probability”-rule applied for LCC will assign atypical observations to another cluster if they are closer (in the sense of the posterior probability) to it, and does therefore tend to produce more compact clusters than the true data generating process.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

From these considerations it follows that performances (in terms of the quality of clustering) of the two approaches highly depend on the features of the data set, even though the direction of the dependence is not always very clear. To fully understand the ‘mechanism’ an analysis of variance on the differences between the indexes we calculated in the LCC and in the PAM clustering outcomes was performed, with the idea that it might help to individuate these determinants.

Operatively, a new data set was arranged; it contained a summary of the whole simulation study. Each record was a single simulation, thus the database had 256000 rows, since there are 128 patterns times 2000 simulations for each setting. For each row the value of the following dependent variables were recorded:

- the difference in terms of the ARI, between LCC and PAM clustering, evaluated with respect to the true class membership;

- the difference in terms of the ASW, between the true clustering and both LCC and PAM;

The included factors were the data features coded as follows:

- number of variables (4,12);
- number of categories (binary,4-levels, 8-levels, mixed number of levels)[1];
- number of clusters (small, large);
- sample size (small, large);
- mixing proportions (extremely different, equal);
- cluster separation (unclear, clear)

The Anova results in Anderlucci (2012) showed that for the difference in the ARI between LCC and PAM all the factors were highly significant, and all the interaction terms - other than the number of categories×the sample size - were significant too. According to the mean square values, the factor with the highest effect on the dependent variable is the interaction term of number of variables×the cluster separation (both approaches work much better when clusters are expected to be clearly separated and they increase their performance with a larger number of variables, but the improvement for PAM over the number of variables is less remarkable if the clusters overlap), followed by the number of latent classes (the larger the better for PAM, the opposite is true for LCC on average), the sample size (the smaller the better for PAM on average) and the entity of the mixing components (PAM works better when clusters have more or less the same size, whereas LCC gives better results on average when the mixing proportions are extremely different). The number of variables and the cluster separation taken as additive effects do not affect the outcome more than the other data features.

Differences in the ASW values of LCC and PAM are influenced by all the data characteristics and the first-order interaction, that were all highly significant. What was more important

[1] We chose this coding rather than the total number of binary variables so that information about the variable structure is preserved, since nominal variables were coded with binary ones only for PAM clustering

in determining the differences was again the interaction between the number of variables and the cluster separation (both approaches work better with clearly separated clusters and a large number of variables, whereas when clusters overlap a smaller number of variables is preferred, more strongly by PAM); among the additive terms, the element with the highest mean square is the mixing proportion term (PAM works better when clusters have about the same size, the opposite is true for LCC on average), followed by the number of categories and the number of variables (both works better with binary variables and with a larger number of variables on average).

[Figure 1 about here.]

Figure 1 shows the mean values of ARI and ASW for the two approaches according to the most influential factor, namely the interaction between the number of variables and the expected cluster separation. The blue lines represent the LCC, whereas the red lines represent the PAM clustering; dashed lines indicate frameworks where clusters overlap whereas solid lines indicate clearly separated clusters.

## 5 Real data example

In this section a real data example is presented for comparing LCC and PAM on data that is not so artificially "nice" as simulated data. The data comes from the first existing database of Tribal Art prices, which contains about 20000 records of items sold from 1998 to 2011 by the most important auction houses (see Modugno and Giannerini (2008), Modugno (2012), Modugno et al. (2012)). The database contains 43 variables, from the physical and historical features of the object to the market characteristics. The information is collected from the paper catalogues released from the auction houses before the auctions.

In this illustrative example for each object (here $n = 19165$) only a selection of 5 variables was considered:

- OGG, Type of object, with 12 levels: Furniture, Sticks, Masks, Religious objects, Ornaments, Sculptures, Musical instruments, Tools, Clothing, Textiles, Weapons, Jewels.

- MATP, Material, with 10 levels: Ivory, Vegetable Fibre, Wood, Metal, Gold, Stone, Precious stone, Terracotta/ Ceramic, Textile and hides, Bone/Horn.
- CONT, Continent, with 4 levels: Africa, America, Eurasia, Oceania.
- CPAT, Patina, with 4 levels: Not indicated, Pejorative, Present, Appreciative.
- CLIO Gaps and Repairs, with 3 levels: Missing piece, No significant repair, Generic repair.

Variable selection was performed according to experts' knowledge about features with a high discriminating power for the price among the nominal variables.

The aim is to group the observations into homogenous clusters. LCC and PAM were performed with the number of clusters allowed to vary between 2 and 10. The best LCC model was chosen according to the BIC (K=10). LCC was computed by the function `lcmixed` (R package `fpc`).

Since `pam` is based on the full dissimilarity matrix, the given data set would require too much memory and time. Therefore another function, `clara` (R package `cluster`), was used; CLARA is an approximate version of PAM for Euclidean distances that can be computed for larger data sets (Kaufman and Rouseeuw (1990)). It performs `pam` on several data subsets, assigns the further observations to the closest cluster medoid and selects from these the solution that is best according to the objective function. The function's default values were not changed and, as was done for LCC, the number of clusters was allowed to vary between 2 and 10; the best model was chosen according to the ASW.

Both of the two approaches selected the maximum 10 groups. Since this is just an illustrative example a larger number of clusters was not explored. The two clusterings are quite different: the ARI between them is 0.414. In this case there is no information about any "true partition", so it is not possible to compare the methods with respect to misclassifications. Nevertheless, the quality of the two clusterings can be measured by calculating the ASW, which was 0.243 for LCC (and no better results for $K < 10$) and 0.275 for `clara`. Thus, in this case, CLARA yields to a classification of units that is better than LCC in terms of dissimilarity. A concise description of the clusters follows.

Cluster 1 in LCC contains 339 items, mostly ornaments coming from Oceania. They are generally made of vegetable fiber or precious stone and they do not have a specified patina nor a significant repair. The analogue for CLARA contains 1145 items, they are all masks coming mostly from Africa and Oceania. They are mostly made of wood and they do not have a specified patina nor a significant repair; a few of them have pieces missing.

Cluster 2 in LCC contains 4216 items, mostly masks and sculptures. They are generally made of wood and they come from Africa and Oceania. Most of them do not have a specified patina; the majority does not have a significant repair, but some have piece missing. CLARA's cluster 2 contains 3347 observations, mostly sculptures. The material used are generally wood and metal and they mostly come from Africa and Oceania. All of them do not have a specified patina nor a significant repair.

Cluster 3 in LCC contains 694 objects; mostly ornaments, made of ivory and horn/bones. They generally come from Africa (564) and some of them got an appreciative adjective. They are mostly well preserved, only a few part have pieces missing. CLARA's cluster 3 has 2010 units, mostly sculptures made of wood and ivory, coming from Africa. All of them got an appreciative adjective and the major part is well preserved.

Cluster 4 in LCC contains 972 units, mostly ornaments and jewels, made of gold and precious stone. Almost all of them are from America and do not have a specified patina nor a significant repairs. CLARA's cluster 4 contains 836 objects and has a similar characterization.

Cluster 5 in LCC has 898 units, mostly weapons from Africa, made of metal and wood. Most of them do not have a specified patina nor a significant repair. The analogue in CLARA has 1417 masks, generally made of wood, mostly coming from Africa. The majority has a patina but it does not have a significant repair.

Cluster 6 in LCC contains 4945 observations, mostly sculptures from Africa, made of wood. They are generally well preserved and come with an appreciative adjective. Cluster 6 in CLARA has 3201 units, mostly sculptures made of wood, coming from Africa. They all have a patina and they are generally well preserved.

Cluster 7 in LCC counts 692 objects, mainly textiles and clothes, made of textiles and holes, coming from America. Most of them have no specified patina and no significant repair. The analogue in CLARA has a similar characterization, but it is a bit smaller with 629 units.

Cluster 8 in LCC contains 2020 objects, mostly sculptures made of terracotta/ceramic, coming from America. They generally do not have a specified patina nor a significant repair. CLARA's cluster 8 has a similar characterization, though it is a bit larger (2428 units).

Cluster 9 in LCC counts 2499 units, mainly wooden tools coming from Oceania. They are generally well preserved and got an appreciative adjective. Cluster 9 in CLARA contains 1902 units with a similar characterization.

Finally, cluster 10 in LCC has 1890 units, mostly tools made of terracotta/ceramic, coming from America. Generally, they do not have a specified patina nor a significant repair. Cluster 10 in CLARA is a bit larger (1950 units) and contains only tools, generally made of terracotta/ceramic and coming from America. Also in this case they are generally well preserved and without a specified patina.

[Figure 2 about here.]

Figure 2 is a two-dimensional Multidimensional Scaling (MDS) representation of the example considered, obtained with the function `cmdscale` (library `MASS` of the R statistical software; Mardia et al. (1979)). The size of points is proportional to the number of identical units and different colours refer to the different cluster memberships (yellow=1, orange=2, black=3, grey=4, pink=5, green=6, purple=7, blue=8, brown=9, red=10). The picture on the top refers to the LCC clustering, while the picture on the bottom refers to the CLARA (approximate PAM) clustering. One thing that can clearly be seen is that CLARA, as opposed to LCC, is reluctant to put different large collections of identical units together into the same cluster, resulting in more uniform cluster sizes. Most CLARA clusters are dominated by one such point (used as cluster medoid), whereas LCC has a larger number of smaller clusters without dominating point, and some very large clusters putting several points with many identical units together.

It is difficult to state which of the two is the most useful classification, because it depends on the reasons and objectives that motivate the cluster analysis. Among the main differences of the two classifications, LCC seems to discriminate more in materials and country of the objects, whereas CLARA yields groups mainly distinguished by the kind of objects and their origin. Of course this is just an illustrative example and more insights can rise with a larger number of clusters and with a larger number of variables to be considered.

## 6 Discussion

The main achievement of the present study is to compare methods that are based on different underlying principles and are therefore not comparable in a straightforward manner, but that should be compared because in practice they are applied to the same kind of data with very similar aims.

LCC and PAM refer to two wider classes of clustering methods, respectively model-based and distance-based methods (LCC and PAM). In practice, the choice between the two approaches should be driven by the aims of the researcher, since they are based on very different assumptions.

The research question that arose was whether both of these approaches lead to similar clusterings and how good the clustering methods that are designed to fulfil one of these aims are in terms of the other one. In order to have a fair 'match', the two clustering outcomes were compared according to two different criteria, one (ARI) based on the 'true' clustering defined by the data generating model, the other one (ASW) based on dissimilarities. In terms of recovering the 'true' clustering, LCC generally behaved better in the simulations, but both methods were similar with clearly separated clusters.

In terms of retrieving homogeneous groups as measured by the ASW, no method always outperformed the other one on average. LCC quite often accomplished to get similar observations together and to separate objects that are very different in a not much worse manner than PAM, and surprisingly sometimes it even outperformed PAM.

Notice that the ASW compares the dissimilarities of observations from other observations of the same cluster with observations of the nearest other cluster, which is not exactly optimized by the PAM criterion, so there is no mathematical necessity that PAM performs better. LCC was certainly helped by the fact that the data were generated by the very same model that is assumed to be true by LCC, though LCC (as well as PAM) even came up with clusters that were significantly more homogeneous than the true ones. Still, the good performance of LCC with respect to the ASW cannot necessarily be expected to be reproduced in real data that will normally deviate to some extent from this model.

Still, the results encourages a user who is interested in finding homogeneous well separated clusters to try LCC. It is very useful about the ASW (and distance-based criteria in general) that it can be evaluated for clusterings of real data without knowledge of the truth, and therefore one can apply various methods and pick the best clustering according to the ASW empirically. In the real data example, PAM did better in this respect than LCC. A similar approach has been taken for mixed-type (categorical/homogeneous/continuous) data in Hennig and Liao (2013).

A number of decisions have been made in this work that could have been made in a different way. One could have used other distance-based criteria such as the Pearson version of Hubert's $\Gamma$ (Hennig and Liao (2013)), other clustering methods such as $k$-means (this is not normally seen as a distance-based method, but our distance measure can be computed as the Euclidean distance on dummy variables) and other distance measures. LCC allows for some relaxation of the local independence assumption, which could be tried out. Future work could be done in these directions.

It may be controversial whether PAM or $k$-means is more appropriate for categorical variables represented by dummies. PAM can be expected to produce clusters that are more homogeneous in terms of the marginal distributions of the variables (although not necessarily in terms of dissimilarities), because $k$-means's centroids are defined by averaging frequencies of the categories within clusters whereas PAM's centroids are members of the data set and therefore represent "pure" categories.

Furthermore results regarding comparing model-based with distance-based methodology in other clustering problems such as mixed type data, functional clustering or time series clustering are still of strong interest, again attempting a fair multicriterion approach to quality measurement, which may come up with some surprises, as in the present study.

# References

Akaike, H. (1987). Factor analysis and aic. *Psychometrika* *52*, 317–332.

Allman, E. S., C. Matias, and J. A. Rhodes (2009, December). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* *37*(6A), 3099–3132.

Anderlucci, L. (2012, February). *Comparing Different Approaches for Clustering Categorical Data*. Ph. D. thesis, Scuola di Dottorato in Scienze Economiche e Statistiche - Alma Mater Studiorum, Università di Bologna, `http://amsdottorato.cib.unibo.it/4302/`.

Celeux, G. and G. Govaert (1991). Clustering Criteria for Discrete Data and Latent Class Models. *Journal of Classification* *8*, 157–176.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011, January). *Cluster Analysis* (5th ed.), Volume 14. John Wiley & Sons, Ltd.

Goodman, L. A. (1974, August). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika* *61*(2), 215.

Hennig, C. and T. F. Liao (2013). How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification. *Appl. Statist., to appear*.

Hubert, L. and P. Arabie (1985). Comparing Partitions. *Journal of Classification* *2*, 193–218.

Kaufman, L. and P. Rouseeuw (1990, January). *Finding Groups in Data*. New York: Wiley.

Mardia, K. V., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press.

Modugno, L. (2012, February). *A Multilevel Model with Time Series Components for the Analysis of Tribal Art Prices*. Ph. D. thesis, Scuola di Dottorato in Scienze Economiche e

Statistiche - Alma Mater Studiorum, Università di Bologna, http://amsdottorato.cib.unibo.it/4301/.

Modugno, L. and S. Giannerini (2008). La prima banca dati dedicata all'arte etnica: prime evidenze empiriche. Sistemaeconomico (2), 45–57.

Modugno, L., S. Giannerini, and S. Cagnone (2012). A multilevel model with time series components for the analysis of tribal art prices. Submitted.

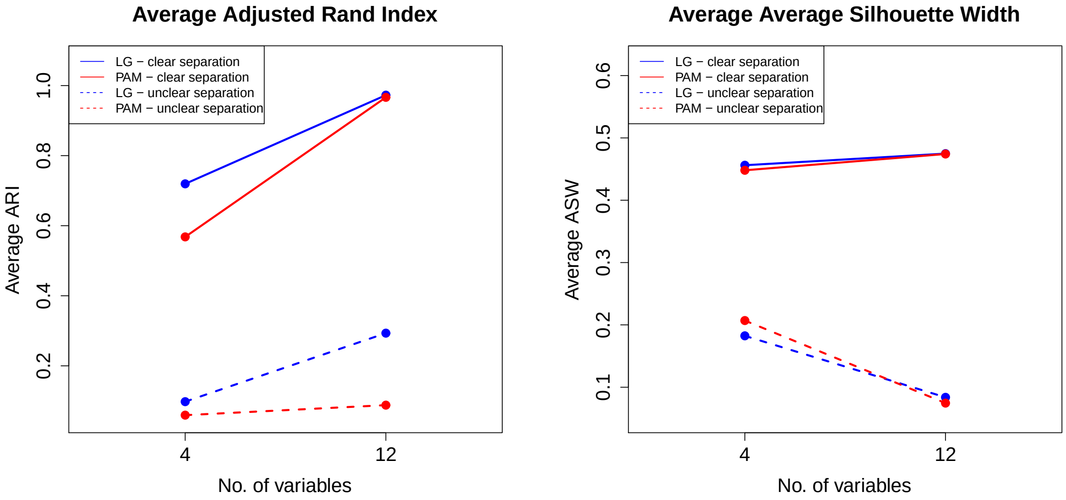
Nylund, K. L., T. Asparouhov, and B. O. Muthén (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. Structural Equation Modeling: An Interdisciplinary Journal 14, 535–569.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66(336), 846– 850.

Schwartz, G. E. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Vermunt, J. K. and J. Magidson (2002). Latent class cluster analysis. In Applied latent class analysis, pp. 89–106. Cambridge: Cambridge University Press.

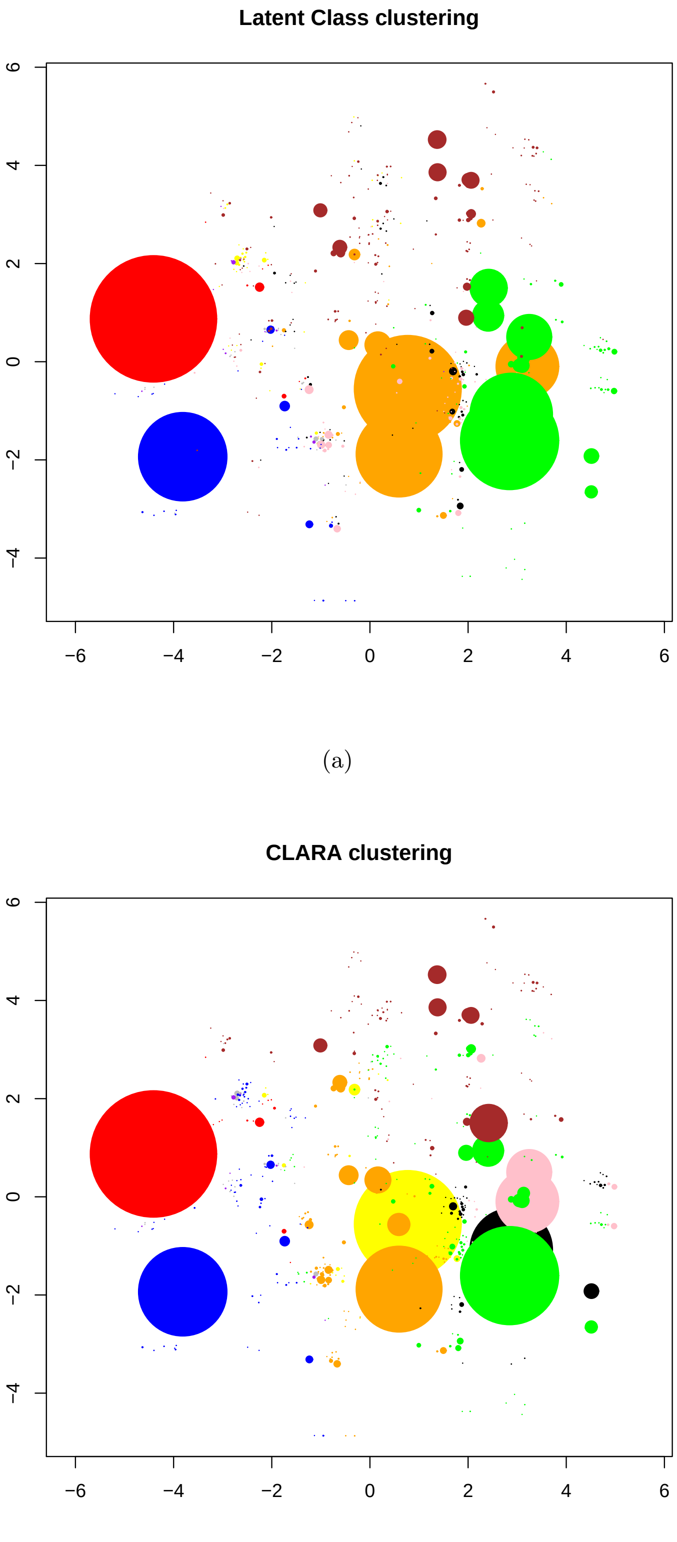
Vermunt, J. K. and J. Magidson (2005). Technical Guide for Latent GOLD 4.0: Basic and Advanced 1.



(a)

(b)

Figure 1: Average values of ARI and ASW according to the no. of variables and the expected cluster separation



(a)

(b)

Figure 2: Multidimensional Scaling of Tribal Art data

Table 1: Average values (and their standard errors) of ARI and ASW: **binary** variables only and **unclear** cluster separation

| No.<br>Var | K | Mixing<br>Prop. | No.<br>obs. | ARI<br>lcc | ARI<br>pam | ASW<br>true | ASW<br>lcc | ASW<br>pam |
|---|---|---|---|---|---|---|---|---|
| 4 | 2 | Dif | Small | .590 (.003) | .440 (.004) | .374 (.001) | .464 (.001) | .490 (.001) |
| | | | Big | .620 (.001) | .505 (.002) | .375 (.000) | .443 (.001) | .496 (.001) |
| | | Eq | Small | .112 (.002) | .089 (.002) | .160 (.001) | .341 (.002) | .336 (.001) |
| | | | Big | .161 (.002) | .009 (.001) | .161 (.000) | .327 (.002) | .305 (.000) |
| | 3 | Dif | Small | .150 (.002) | .094 (.002) | .107 (.001) | .370 (.002) | .395 (.001) |
| | | | Big | .156 (.003) | .062 (.001) | .128 (.000) | .328 (.002) | .364 (.001) |
| | | Eq | Small | .109 (.001) | .114 (.001) | .042 (.001) | .324 (.002) | .373 (.001) |
| | | | Big | .134 (.001) | .120 (.001) | .048 (.000) | .292 (.002) | .369 (.000) |
| 12 | 2 | Dif | Small | .224 (.005) | .005 (.001) | .051 (.000) | .071 (.000) | .091 (.000) |
| | | | Big | .637 (.001) | .007 (.001) | .075 (.000) | .088 (.000) | .049 (.000) |
| | | Eq | Small | .060 (.002) | .026 (.001) | .032 (.000) | .072 (.000) | .060 (.000) |
| | | | Big | .264 (.001) | .025 (.001) | .032 (.000) | .056 (.000) | .050 (.000) |
| | 5 | Dif | Small | .160 (.001) | .149 (.001) | .018 (.000) | .153 (.001) | .147 (.000) |
| | | | Big | .253 (.001) | .177 (.001) | .034 (.000) | .099 (.001) | .128 (.000) |
| | | Eq | Small | .140 (.001) | .137 (.001) | .021 (.000) | .151 (.001) | .147 (.000) |
| | | | Big | .212 (.001) | .168 (.001) | .035 (.000) | .105 (.001) | .129 (.000) |

Table 2: Average values (and their standard errors) of ARI and ASW: **binary** variables only and **clear** cluster separation

| No.<br>Var | K | Mixing<br>Prop. | No.<br>obs. | ARI<br>lcc | ARI<br>pam | ASW<br>true | ASW<br>lcc | ASW<br>pam |
|---|---|---|---|---|---|---|---|---|
| 4 | 2 | Dif | Small | .914 (.002) | .819 (.002) | .728 (.001) | .739 (.001) | .731 (.001) |
| | | | Big | .935 (.000) | .819 (.001) | .729 (.000) | .739 (.000) | .732 (.000) |
| | | Eq | Small | .898 (.001) | .896 (.001) | .754 (.001) | .763 (.001) | .761 (.001) |
| | | | Big | .898 (.000) | .897 (.000) | .754 (.000) | .762 (.000) | .761 (.000) |
| | 3 | Dif | Small | .544 (.002) | .469 (.003) | .446 (.001) | .556 (.001) | .560 (.001) |
| | | | Big | .580 (.001) | .479 (.002) | .450 (.000) | .540 (.001) | .540 (.001) |
| | | Eq | Small | .554 (.002) | .546 (.002) | .469 (.001) | .578 (.001) | .540 (.001) |
| | | | Big | .556 (.001) | .546 (.001) | .470 (.000) | .570 (.001) | .535 (.001) |
| 12 | 2 | Dif | Small | .980 (.001) | .912 (.002) | .398 (.001) | .398 (.001) | .395 (.001) |
| | | | Big | .988 (.001) | .934 (.001) | .399 (.000) | .400 (.000) | .399 (.000) |
| | | Eq | Small | .983 (.001) | .952 (.001) | .411 (.001) | .411 (.001) | .413 (.001) |
| | | | Big | .988 (.000) | .961 (.000) | .412 (.000) | .412 (.000) | .414 (.000) |
| | 5 | Dif | Small | .844 (.001) | .846 (.001) | .461 (.001) | .476 (.001) | .470 (.001) |
| | | | Big | .869 (.000) | .844 (.001) | .467 (.000) | .479 (.000) | .476 (.000) |
| | | Eq | Small | .912 (.001) | .922 (.001) | .536 (.001) | .543 (.001) | .542 (.001) |
| | | | Big | .922 (.001) | .921 (.000) | .540 (.000) | .547 (.000) | .546 (.000) |

Table 3: Average values (and their standard errors) of ARI and ASW: **4-levels** variables only and **unclear** cluster separation

| No. | K | Mixing | No. | ARI | ARI | ASW | ASW | ASW |
|---|---|---|---|---|---|---|---|---|
| Var | | Prop. | obs. | lcc | pam | true | lcc | pam |
| 4 | 2 | Dif | Small | .080 (.003) | .009 (.001) | .071 (.001) | .165 (.001) | .160 (.000) |
| | | | Big | .232 (.002) | -.003 (.000) | .073 (.000) | .135 (.000) | .142 (.000) |
| | | Eq | Small | .023 (.001) | .012 (.001) | .029 (.000) | .158 (.001) | .152 (.000) |
| | | | Big | .033 (.001) | .009 (.000) | .029 (.000) | .104 (.001) | .133 (.000) |
| | 5 | Dif | Small | .025 (.001) | .023 (.000) | -.053 (.000) | .161 (.001) | .163 (.000) |
| | | | Big | .025 (.000) | .021 (.000) | -.026 (.001) | .076 (.000) | .126 (.000) |
| | | Eq | Small | .053 (.001) | .051 (.001) | -.035 (.000) | .166 (.001) | .168 (.000) |
| | | | Big | .063 (.000) | .057 (.000) | -.014 (.000) | .090 (.001) | .136 (.000) |
| 12 | 2 | Dif | Small | .224 (.005) | .005 (.001) | .051 (.000) | .071 (.000) | .091 (.000) |
| | | | Big | .637 (.001) | .007 (.001) | .075 (.000) | .088 (.000) | .049 (.000) |
| | | Eq | Small | .060 (.002) | .026 (.001) | .032 (.000) | .072 (.000) | .060 (.000) |
| | | | Big | .264 (.001) | .025 (.001) | .032 (.000) | .056 (.000) | .050 (.000) |
| | 5 | Dif | Small | .073 (.001) | .040 (.000) | .001 (.000) | .054 (.000) | .045 (.000) |
| | | | Big | .159 (.001) | .043 (.000) | .010 (.000) | .032 (.000) | .039 (.000) |
| | | Eq | Small | .063 (.001) | .039 (.000) | .000 (.000) | .054 (.000) | .046 (.000) |
| | | | Big | .115 (.001) | .043 (.000) | .006 (.000) | .031 (.000) | .039 (.000) |

Table 4: Average values (and their standard errors) of ARI and ASW: **4-levels** variables only and **clear** cluster separation

| No. | K | Mixing | No. | ARI | ARI | ASW | ASW | ASW |
|---|---|---|---|---|---|---|---|---|
| Var | | Prop. | obs. | lcc | pam | true | lcc | pam |
| 4 | 2 | Dif | Small | .628 (.004) | .169 (.006) | .257 (.001) | .278 (.001) | .197 (.001) |
| | | | Big | .740 (.001) | -.024 (.000) | .258 (.000) | .281 (.000) | .147 (.000) |
| | | Eq | Small | .644 (.002) | .098 (.001) | .200 (.001) | .218 (.002) | .522 (.000) |
| | | | Big | .711 (.001) | .093 (.000) | .200 (.000) | .212 (.000) | .507 (.000) |
| | 5 | Dif | Small | .463 (.002) | .539 (.002) | .245 (.001) | .316 (.001) | .322 (.001) |
| | | | Big | .564 (.001) | .554 (.001) | .254 (.000) | .328 (.000) | .330 (.000) |
| | | Eq | Small | .505 (.002) | .604 (.002) | .269 (.001) | .331 (.001) | .342 (.001) |
| | | | Big | .603 (.001) | .609 (.001) | .278 (.000) | .350 (.000) | .350 (.000) |
| 12 | 2 | Dif | Small | .980 (.001) | .912 (.002) | .398 (.001) | .398 (.001) | .395 (.001) |
| | | | Big | .988 (.001) | .934 (.000) | .399 (.000) | .400 (.000) | .399 (.000) |
| | | Eq | Small | .983 (.001) | .952 (.001) | .411 (.001) | .412 (.001) | .413 (.001) |
| | | | Big | .988 (.000) | .962 (.000) | .412 (.000) | .412 (.000) | .414 (.000) |
| | 5 | Dif | Small | .938 (.001) | .941 (.001) | .332 (.000) | .334 (.000) | .333 (.000) |
| | | | Big | .958 (.000) | .957 (.000) | .335 (.000) | .338 (.000) | .337 (.000) |
| | | Eq | Small | .929 (.001) | .935 (.001) | .328 (.000) | .330 (.000) | .329 (.000) |
| | | | Big | .952 (.001) | .952 (.000) | .330 (.000) | .333 (.000) | .333 (.000) |

Table 5: Average values (and their standard errors) of ARI and ASW: **8-levels** variables only and **unclear** cluster separation

| No. | K | Mixing | No. | ARI | ARI | ASW | ASW | ASW |
|---|---|---|---|---|---|---|---|---|
| Var | | Prop. | obs. | lcc | pam | true | lcc | pam |
| 4 | 2 | Dif | Small | .016 (.001) | -.001 (.001) | .026 (.000) | .099 (.000) | .106 (.000) |
| | | | Big | .032 (.001) | -.003 (.000) | .027 (.000) | .066 (.000) | .095 (.000) |
| | | Eq | Small | .013 (.001) | .006 (.001) | .013 (.000) | .086 (.000) | .076 (.000) |
| | | | Big | .024 (.001) | .005 (.000) | .014 (.000) | .052 (.000) | .063 (.000) |
| | 5 | Dif | Small | .030 (.000) | .024 (.000) | -.016 (.000) | .075 (.000) | .076 (.000) |
| | | | Big | .044 (.000) | .007 (.000) | -.005 (.000) | .045 (.000) | .059 (.000) |
| | | Eq | Small | .031 (.000) | .026 (.000) | -.013 (.000) | .075 (.000) | .075 (.000) |
| | | | Big | .042 (.000) | .009 (.000) | -.004 (.000) | .046 (.000) | .056 (.000) |
| 12 | 2 | Dif | Small | .135 (.004) | .004 (.001) | .037 (.000) | .038 (.000) | .029 (.000) |
| | | | Big | .586 (.001) | .002 (.000) | .038 (.000) | .044 (.000) | .026 (.000) |
| | | Eq | Small | .579 (.002) | .105 (.002) | .046 (.000) | .051 (.000) | .030 (.000) |
| | | | Big | .710 (.001) | .128 (.002) | .046 (.000) | .051 (.000) | .029 (.000) |
| | 5 | Dif | Small | .201 (.001) | .050 (.000) | .015 (.000) | .026 (.000) | .019 (.000) |
| | | | Big | .350 (.001) | .054 (.000) | .016 (.000) | .026 (.000) | .019 (.000) |
| | | Eq | Small | .137 (.001) | .034 (.000) | .010 (.000) | .024 (.000) | .019 (.000) |
| | | | Big | .263 (.001) | .036 (.000) | .012 (.000) | .022 (.000) | .018 (.000) |

Table 6: Average values (and their standard errors) of ARI and ASW: **8-levels** variables only and **clear** cluster separation

| No. | K | Mixing | No. | ARI | ARI | ASW | ASW | ASW |
|---|---|---|---|---|---|---|---|---|
| Var | | Prop. | obs. | lcc | pam | true | lcc | pam |
| 4 | 2 | Dif | Small | .680 (.004) | .322 (.008) | .294 (.001) | .284 (.001) | .213 (.002) |
| | | | Big | .858 (.001) | .061 (.006) | .295 (.000) | .305 (.000) | .145 (.001) |
| | | Eq | Small | .787 (.002) | .802 (.002) | .335 (.001) | .341 (.001) | .343 (.001) |
| | | | Big | .848 (.001) | .801 (.001) | .336 (.001) | .348 (.001) | .343 (.001) |
| | 5 | Dif | Small | .709 (.001) | .768 (.001) | .320 (.001) | .333 (.001) | .344 (.001) |
| | | | Big | .777 (.001) | .262 (.001) | .324 (.000) | .346 (.000) | .189 (.000) |
| | | Eq | Small | .704 (.001) | .766 (.001) | .321 (.001) | .334 (.001) | .344 (.001) |
| | | | Big | .764 (.000) | .172 (.000) | .324 (.000) | .348 (.000) | .149 (.000) |
| 12 | 2 | Dif | Small | .992 (.000) | .984 (.001) | .340 (.000) | .340 (.000) | .339 (.000) |
| | | | Big | .996 (.000) | .987 (.000) | .340 (.000) | .340 (.000) | .339 (.000) |
| | | Eq | Small | .995 (.000) | .995 (.000) | .389 (.000) | .389 (.000) | .389 (.000) |
| | | | Big | .994 (.001) | .996 (.000) | .389 (.000) | .389 (.000) | .389 (.000) |
| | 5 | Dif | Small | .990 (.000) | .990 (.000) | .358 (.000) | .358 (.000) | .358 (.000) |
| | | | Big | .992 (.000) | .991 (.000) | .358 (.000) | .358 (.000) | .358 (.000) |
| | | Eq | Small | .989 (.000) | .989 (.000) | .357 (.000) | .358 (.000) | .357 (.000) |
| | | | Big | .992 (.000) | .991 (.000) | .358 (.000) | .358 (.000) | .358 (.000) |

Table 7: Average values (and their standard errors) of ARI and ASW: **mixed no.-levels** variables only and **unclear** cluster separation

| No. | K | Mixing | No. | ARI | ARI | ASW | ASW | ASW |
|---|---|---|---|---|---|---|---|---|
| Var | | Prop. | obs. | lcc | pam | true | lcc | pam |
| 4 | 2 | Dif | Small | .052 (.002) | .028 (.001) | .067 (.001) | .214 (.001) | .222 (.001) |
| | | | Big | .105 (.002) | .014 (.000) | .160 (.001) | .192 (.000) | .185 (.001) |
| | | Eq | Small | .016 (.001) | .011 (.001) | .028 (.000) | .202 (.001) | .215 (.001) |
| | | | Big | .013 (.000) | .003 (.000) | .029 (.000) | .121 (.002) | .190 (.000) |
| | 5 | Dif | Small | .035 (.001) | .035 (.001) | -.068 (.000) | .217 (.001) | .237 (.001) |
| | | | Big | .031 (.000) | .033 (.000) | -.036 (.000) | .109 (.002) | .209 (.000) |
| | | Eq | Small | .040 (.001) | .042 (.001) | -.059 (.000) | .220 (.001) | .236 (.001) |
| | | | Big | .035 (.000) | .042 (.000) | -.037 (.000) | .109 (.001) | .206 (.000) |
| 12 | 2 | Dif | Small | .121 (.004) | .027 (.001) | .070 (.000) | .084 (.000) | .074 (.000) |
| | | | Big | .521 (.001) | .012 (.001) | .071 (.000) | .089 (.000) | .057 (.000) |
| | | Eq | Small | .281 (.003) | .118 (.002) | .071 (.000) | .093 (.000) | .085 (.000) |
| | | | Big | .517 (.001) | .176 (.002) | .072 (.000) | .091 (.000) | .082 (.000) |
| | 5 | Dif | Small | .132 (.001) | .104 (.001) | .023 (.000) | .058 (.000) | .060 (.000) |
| | | | Big | .331 (.001) | .128 (.000) | .028 (.000) | .055 (.000) | .054 (.000) |
| | | Eq | Small | .145 (.001) | .111 (.000) | .023 (.000) | .061 (.000) | .061 (.000) |
| | | | Big | .319 (.001) | .141 (.001) | .027 (.000) | .055 (.000) | .054 (.000) |

Table 8: Average values (and their standard errors) of ARI and ASW: **mixed no.-levels** variables only and **clear** cluster separation

| No.<br>Var | K | Mixing<br>Prop. | No.<br>obs. | ARI<br>lcc | ARI<br>pam | ASW<br>true | ASW<br>lcc | ASW<br>pam |
|---|---|---|---|---|---|---|---|---|
| 4 | 2 | Dif | Small | .851 (.002) | .802 (.002) | .570 (.001) | .584 (.001) | .577 (.001) |
| | | | Big | .891 (.000) | .806 (.001) | .572 (.000) | .587 (.000) | .578 (.000) |
| | | Eq | Small | .844 (.002) | .856 (.002) | .573 (.001) | .589 (.001) | .588 (.001) |
| | | | Big | .859 (.000) | .855 (.001) | .573 (.000) | .590 (.000) | .588 (.000) |
| | 5 | Dif | Small | .662 (.002) | .689 (.002) | .412 (.001) | .526 (.001) | .525 (.001) |
| | | | Big | .681 (.000) | .694 (.001) | .418 (.000) | .512 (.000) | .527 (.000) |
| | | Eq | Small | .666 (.002) | .703 (.002) | .435 (.001) | .499 (.001) | .501 (.001) |
| | | | Big | .710 (.001) | .703 (.001) | .440 (.000) | .509 (.000) | .505 (.000) |
| 12 | 2 | Dif | Small | .998 (.000) | .998 (.000) | .596 (.000) | .596 (.000) | .596 (.000) |
| | | | Big | .999 (.000) | .997 (.000) | .596 (.000) | .596 (.000) | .596 (.000) |
| | | Eq | Small | .998 (.000) | .998 (.000) | .597 (.000) | .597 (.000) | .597 (.000) |
| | | | Big | .998 (.001) | .998 (.000) | .596 (.000) | .596 (.000) | .596 (.000) |
| | 5 | Dif | Small | .982 (.000) | .985 (.000) | .475 (.000) | .476 (.000) | .476 (.000) |
| | | | Big | .988 (.000) | .986 (.000) | .477 (.000) | .478 (.000) | .477 (.000) |
| | | Eq | Small | .981 (.000) | .985 (.000) | .476 (.000) | .477 (.000) | .477 (.000) |
| | | | Big | .987 (.001) | .986 (.000) | .478 (.000) | .478 (.000) | .478 (.000) |