Latent Class Cluster for Clustering Villages Based on Socio-economic Indicators in 2018

A Riyanto¹,², H Kuswanto¹ and D D Prastyo¹

¹ Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Jl. Arif Rahman Hakim Surabaya 60111, Indonesia
² BPS – Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta 10710, Indonesia

E-mail: andreas_riyanto@bps.go.id

Abstract. Latent class analysis (LCA) is a statistical method used to classify units into unobserved (latent) variable classes, so-called clusters. In this article, we introduce LCA to demonstrate its use for village groupings. We applied LCA on data processed from the census conducted by BPS-Statistics Indonesia, namely the village potential (Podes) data 2018. The Podes data is used to extract the socio-economic characteristics of a village. The empirical results of the LCA application show that the villages in the Banyumas Regency are grouped into three classes/clusters.

1. Introduction
Statistics Indonesia (Badan Pusat Statistik abbreviated as BPS) always strives to produce complete, accurate, and up-to-date statistical data. Routinely and in a certain period, BPS conducts censuses and surveys. For some reason, a census cannot be done every time. Instead of providing a solution to describe the population, a survey is needed. The national socio-economic survey (Susenas) is one of the surveys conducted by BPS to support the availability of socio-economic data. The Susenas is the primary support for helping the government to drive and measure the achievement of national development and international development goals known as Sustainable Development Goals (SDGs). A development that takes into account intergenerational sustainability and maintains the balance of nature is a global agenda of the SDGs. In achieving the SDGs target, credible and up to date data support is needed. As a form of support in providing such data, BPS conducts Susenas in providing basic statistical data for measuring the achievement of SDGs.

In conducting surveys, such as Susenas, there are two types of errors. The first type of error is sampling error, i.e., the errors related to the sampling design used. The second type is a non-sampling error, which is not associated with the sampling design [1]. An example of this second type of error is a non-response case. Cases of non-response in the implementation of Susenas mostly occur in eastern provinces of Indonesia, such as Papua Province. Non-response cases in the area usually occur in villages with quite difficult geographical conditions (backward villages) such that a sample replacement is usually needed. This condition gave rise to the idea of how to reduce the backward villages affected by the survey sample by grouping villages so that the backward villages that were later affected by the survey sample could be replaced with other villages from the same group.

In statistics, several grouping techniques can be used, including cluster analysis and model-based clusters. Each of these techniques is closely related to the type of data used. In this study, the data used
are categorical. To grouping objects according to categorical data, an appropriate statistical tool is needed, one of them, namely latent class analysis (LCA) [2]. LCA is a model-based method that identifies groups of individuals using categorical latent variables [3][4]. LCA is also a powerful statistical technique that can be used to classify individuals who have specific characteristics [5]. This analysis is also called model-based clustering [6][7], probabilistic clustering [8], and latent class cluster (LCC) [9]. Parameter estimation in LCA is required to find estimators from population parameters whose magnitude is unknown. The parameter estimation method that is often used is the maximum likelihood method and the maximum posterior method, given that the corresponding distribution is known because it typically produces an unbiased estimator and minimum variance [10]. The existence of latent variables causes the maximum likelihood estimation method cannot be used directly. Hence, it is necessary to modify or augmented data so that the maximum likelihood method can be used more simply. The technique commonly used to solve the maximum likelihood estimation in the latent class model is the Expectation-Maximization (EM) algorithm. The EM algorithm is slower to converge than Newton Raphson's algorithm. However, it is simpler because it does not require a second derivative matrix of the likelihood function [11].

2. Latent Class Cluster

Over the past years, there has been a renewed interest in cluster analysis using LCA. Lazarsfeld firstly proposed the LCA, where it is a special case of model-based grouping, for discrete multivariate data [12]. The LCA can be categorized into three types, namely the latent class cluster (LCC), latent class factor (LCF), and latent class regression (LCR) [13]. This study uses an LCC to identifies clusters by grouping objects that have the same characteristics. The purpose of the LCC is to determine the number of classes required to explain the associations between the observed variables and allocate objects to latent classes [14].

The LCC has many similarities with grouping methods for multivariate data, such as cluster analysis. The main difference between cluster analysis and the latent class cluster is that cluster analysis uses the distance approach to declare an object close to its center, but the LCC uses the probability of posterior grouping that is estimated through the maximum likelihood method to express the proximity of an object to its center [13].

For example \((x_1, x_2, \ldots, x_p)\) shows a vector of \(p\) observed (or manifest) dichotomous or polytomous variables. The value of \(x_{ih}\) becomes the value of the \(h\)th sample element/object for the \(i\)th variable \((h=1,2,\ldots,n)\). The row vector \(x_i' = (x_{1h}, x_{2h}, \ldots, x_{ph})\) is called the response pattern of \(h\)th object. In the latent class models, the factor space is assumed to consist of \(K\) classes. For each class \(k\) has a corresponding probability, \(\eta_k\). This specification leads to prior probabilities because it gives prior probabilities to observe data \(x\). The joint distribution of the observed variables is as follows [14]:

\[
f(x_i) = \sum_{k=1}^{K} \eta_k g(x_i | k)
\]

where \(g(x_i | k)\) is a function of the distribution of manifest variables in the form of dichotomous or polytomous variables. The conditional distribution is an element that belongs to class \(k\) given \(x\). This distribution is referred to as posterior probability and can be written:

\[
h(k | x_i) = \frac{\eta_k g(x_i | k)}{f(x_i)}
\]

In LCA, the variables are usually assumed local independence. Violations of this assumption often cause the incompatibility of latent class models. The variables are then modeled for each variable within
each group with a multinomial density. The multinomial density of a single variable $x$ (with categories $1, 2, ..., c$) for group $k$ is then [12]:

$$g(x|k) = \prod_{s=1}^{c}(\pi_{s(k)})^{x_{s}}$$

where $x_{s} = \begin{cases} 1; & \text{if the response falls in category } s, \text{ for } s=1,\ldots,c \\ 0; & \text{otherwise,} \end{cases}$

with $\pi_{s(k)}$ is the probability of the variable taking value $s$ in group $k$, and $c$ is the number of possible values or categories the variable can take.

Since the observed variables have been taken as conditionally independent given each class $k$, their joint group density can be written as a product of their individual group density. We can write the joint group density as:

$$g(x_{k}|k) = \prod_{i=1}^{p} g(x_{ik}|k)$$

$$g(x_{k}|k) = \prod_{i=1}^{p} \prod_{s=1}^{c_{i}} \left( \pi_{s(k,i)} \right)^{x_{s(i)}}$$

where $x_{s(i)} = \begin{cases} 1; & \text{if the response falls in category } s, \text{ for } s=1,\ldots,c_{i} \\ 0; & \text{otherwise,} \end{cases}$

with $\pi_{s(k,i)}$ is the probability of the variable $i$ taking value $s$ in group $k$, and $c_{i}$ is the number of possible values or categories the $i$th variable can take. The overall density is then the weighted sum of these individual product densities, i.e.:

$$f(x_{k}) = \sum_{k=1}^{K} \left( \eta_{k} \prod_{i=1}^{p} \prod_{s=1}^{c_{i}} \left( \pi_{s(k,i)} \right)^{x_{s(i)}} \right)$$

The model parameters $(\pi_{s(k,i)}, \eta_{k})$ where $i = 1,\ldots, p; k = 1,\ldots, K; s = 1,\ldots, c_{i}$) are estimated from the data with maximum likelihood. The log-likelihood for a random sample of size $n$ is written as:

$$L = \sum_{h=1}^{n} \log f(x_{h})$$

$$L = \sum_{h=1}^{n} \log \sum_{k=1}^{K} \left( \eta_{k} \prod_{i=1}^{p} \prod_{s=1}^{c_{i}} \left( \pi_{s(k,i)} \right)^{x_{s(i)}} \right)$$

The log-likelihood can be maximized using the EM algorithm [11] under the constraint that: $\sum_{k=1}^{K} \eta_{k} = 1$ and $\sum_{i=1}^{c_{i}} \pi_{s(k,i)} = 1$.

From the partial derivatives of the log-likelihood given in (5), the solution for maximum likelihood is obtained as follows. The estimated prior class probabilities are written in equation (6).

$$\hat{\eta}_{k} = \frac{\sum_{h=1}^{n} \hat{h}(k|x_{h})}{n}$$

The estimated conditional probability that $x_{i} = s$ given class $k$ for the multinomial variables is
\[
\hat{\pi}_{a(i)} = \sum_{h=1}^{n} x_{a(i)} \hat{m}_n \hat{k} \left( k | x_n \right) / (n \hat{m}_n) .
\]  
(7)

For the posterior probabilities \( h(k | x_n) \) are estimated by:
\[
\hat{h}(k | x_n) = \frac{\hat{\eta}_n \hat{g}(x_n | k)}{\hat{f}(x_n)} .
\]  
(8)

\( \hat{g}(x_n | k) \) and \( \hat{f}(x_n) \) use the estimated values of each parameter:
\[
\hat{g}(x_n | k) = \prod_{i=1}^{K} \left( \hat{\pi}_{a(i)} \right)_{k(i)}
\]  
(9)

\[
\hat{f}(x_n) = \sum_{k=1}^{K} \hat{\eta}_n \hat{g}(x_n | k) .
\]  
(10)

The EM algorithm works by selecting the initial value for the posterior probability \( h(k | x_n) \). Using equations (6) and (7), a first approach is obtained for the model parameters. Then using equation (8), a new estimate is obtained for \( h(k | x_n) \). Next, equations (6) and (7) are re-used, so a second approach is obtained for the model parameters. The iteration is done until convergence is attained.

Goodman [15] addresses the problem of information sufficiency to estimate the number of parameters in the model. The conditions needed to be able to identify when there are \( K \) classes and \( p \) variables with the number of categories \( c = (c_1, ..., c_p) \) is
\[
\prod_{i=1}^{p} c_i > \left( \sum_{i=1}^{p} c_i - p + 1 \right) K .
\]

For selecting the number of latent classes, model selection techniques can be applied because statistical models were used in this analysis. Each different value of the number of latent classes defines a different model for the data. In this case, the Bayes factors [16] is used to choose the best model (and the corresponding number of classes). The log-likelihood ratio test can not be used for comparing models with different number of classes because these conditional tests do not follow the \( \chi^2 \) distribution. Therefore, the model selection criteria can be used as an alternative. Some of the model selection criteria can be used in multivariate analysis, such as proposed by Akaike, Schwarz, and Kashap [17]. The model selection criteria usually used in latent class clusters are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) and are defined as follows:
\[
AIC = -2 \left[ \max L \right] + 2m .
\]  
(11)

and
\[
BIC = -2 \left[ \max L \right] + 2m \log(n) .
\]  
(12)

where \( m \) is the number of model parameters, and \( n \) is the number of observations.

Smaller AIC and BIC values present the optimum balance between model compatibility with many parameters, so a better model is a model with minimum AIC and BIC values. The BIC is more appropriate for latent class models because of its simplicity [18][19][20]. The final classification of objects into classes by the model is based on the estimation of posterior probabilities, \( \hat{h}(k | x_n) \), as written in equation (8).
3. Data
The data that are used in this study come from a census that was done by BPS. This data is the village potential data 2018 in the Banyumas Regency. The village potential (Podes) data is one of the territorial (spatial) data held by BPS, which emphasizes the potential of an area. The Podes data collection has objectives, among others: providing data on the existence and development potential of the village, identifying villages that have geographical difficulties, providing data for the calculation of village development indicators, etc. [1]. The village potential data contains many sectors, including population and employment sectors, housing and environment sector, education and health sector, information and communication sector, etc. This research focuses on the clustering villages based on socio-economic indicators so that it only uses village potential data in the socio-economic sector. The LCA computation was conducted using R codes from the poLCA package.

TABLE 1. Summary of variables

| Variable | Description | Value | % |
|----------|-------------|-------|---|
| X₁       | Main source of income for the majority of the population | 1= agriculture, 2= non-agriculture | 75.53, 24.47 |
| X₂       | Cooking fuel for the majority of families in the village | 1= LPG gas, 2= non-LPG gas | 98.49, 1.51 |
| X₃       | Temporary landfill | 1= yes, used, 2= yes, not used, 3= no | 75.98, 2.11, 71.91 |
| X₄       | Toilet facility usage of the majority of families | 1= toilet, 2= non-toilet | 77.09, 0.91 |
| X₅       | Final disposal site for the stool of the majority of families | 1= tank, 2= non-tank | 71.60, 28.40 |
| X₆       | The existence of springs in the village | 1= yes, managed, 2= yes, not managed, 3= no | 49.85, 14.80, 35.35 |
| X₇       | The existence of the high school and equivalent education institutions | 1= yes, 2= no | 28.70, 71.30 |
| X₈       | The existence of health facilities | 1= yes, 2= no | 99.09, 0.91 |
| X₉       | Extraordinary events or disease outbreaks in the past year | 1= yes, 2= no | 1.81, 98.19 |
| X₁₀      | The widest type of road surface | 1= asphalt/concrete, 2= others | 99.70, 0.30 |
| X₁₁      | The existence of public transportation | 1= yes, with fixed routes, 2= yes, without fixed routes, 3= no | 75.53, 6.04, 18.43 |
| X₁₂      | The existence of residents who use handphone | 1= mostly residents, 2= a small number of residents, 3= nothing | 99.40, 0.60, 0.00 |

4. Results
The number of clusters is selected based on the lowest AIC value. Based on the result shown in Table 2, we chose to retain the model with three latent classes. The posterior probability of class membership, \( \hat{h}(k|x) \), becomes the basis in determining which villages to enter into clusters. This determination uses the modal allocation method, which marks each object into a class with the highest posterior probability.
TABLE 2. Summary of model results used for model selection

| Model  | LL       | BIC       | AIC       | # par |
|--------|----------|-----------|-----------|-------|
| 1-class| -1275.100| 2635.807  | 2580.201  | 15    |
| 2-classes| -1244.740| 2666.400  | 2551.480  | 31    |
| 3-classes| -1222.692| 2713.618  | 2539.384  | 47    |
| 4-classes| -1217.254| 2794.055  | 2560.507  | 63    |
| 5-classes| -1209.110| 2869.081  | 2576.219  | 79    |

The results of the clustering of villages using the LCC model are presented in table 3. From these clustering, villages are clustered into cluster 1, cluster 2, and cluster 3. Table 3 contains the response patterns of 12 variables in the 2018 Podes data collection in Banyumas Regency, the magnitude of the chances of each response pattern in each group, and the cluster prediction of each response pattern. The table only presents a standard classification for the first ten rows and the last row.

TABLE 3. Summary of villages clustering with three clusters

| No. | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | Prob1 | Prob2 | Prob3 | Pred_Clust |
|-----|----|----|----|----|----|----|----|----|----|-----|-----|-----|-------|-------|-------|-------------|
| 1   | 1  | 1  | 3  | 1  | 1  | 1  | 2  | 1  | 2  | 1   | 3   | 1   | 0.0000 | 0.9930 | 0.0070 | 2           |
| 2   | 1  | 1  | 3  | 1  | 1  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 0.6450 | 0.3431 | 0.0119 | 1           |
| 3   | 1  | 1  | 3  | 1  | 1  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 0.0000 | 0.8632 | 0.1368 | 2           |
| 4   | 1  | 1  | 1  | 1  | 1  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 0.9156 | 0.0514 | 0.0330 | 1           |
| 5   | 1  | 1  | 3  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 1   | 0.7354 | 0.2584 | 0.0062 | 1           |
| 6   | 1  | 1  | 3  | 1  | 2  | 1  | 2  | 1  | 1   | 1   | 1   | 0.9020 | 0.0980 | 0.0000 | 1           |
| 7   | 1  | 1  | 3  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 1   | 0.7354 | 0.2584 | 0.0062 | 1           |
| 8   | 1  | 1  | 3  | 1  | 2  | 1  | 2  | 1   | 1   | 1   | 1   | 0.6450 | 0.3431 | 0.0119 | 1           |
| 9   | 1  | 1  | 2  | 1  | 1  | 1  | 1  | 2   | 1  | 1   | 1   | 0.0000 | 0.0000 | 0.0000 | 1           |
| 10  | 1  | 1  | 3  | 1  | 1  | 1  | 2  | 1  | 2   | 1   | 3   | 1   | 0.0000 | 0.9930 | 0.0070 | 2           |

Suppose, for example, a village in the first row (village-1) is characterized by most of the population's income in the agricultural sector, the fuel for cooking most of the family is LPG gas, does not have a temporary landfill, toilet facility usage of the majority of families are toilet, final disposal site for the stool of majority of families is a tank, springs in the village are available and managed, there is no high school or equivalent, health facilities are available, there are no extraordinary events or disease outbreaks, the widest type of road surface is asphalt, there is no public transportation, and most residents using cell phones. This village-1 is classified in cluster 2 with a chance of 0.9930. It is classified in cluster 2 because the probability of being in this cluster is the highest compared to being in other clusters.

In the second row, the village-2, it can be seen that the characteristics of most of the population's income is in the agricultural sector, the fuel for cooking most of its family is LPG gas, does not have a temporary landfill, toilet facility usage of the majority of families are toilet, final disposal site for the stool of majority of families is a tank, springs in the village are available and managed, there is no high school or equivalent, health facilities are available, there are no extraordinary events or disease outbreaks, the widest type of road surface is asphalt, there is public transportation with fixed routes, and most residents using cell phones. This village-2 is classified in cluster 1 with an opportunity of 0.6450. Likewise, for the following lines.

Table 4 shows item response probabilities. The bottom row of the table shows the probability of a latent class, where there are three latent classes formed. Class 1 has a probability of 0.4844, which means that the probability of a village being a member of the group is 48.44%. It can also be said that group 1 consists of 48.44%, group 2 consists of 35.13%, and group 3 consists of 16.43% of the entire village.
Overall, out of 301 villages, 172 of them are members of the first cluster, 90 villages from the second cluster and only 39 villages are grouped in the third cluster. The socio-economic aspects of a village in 2018 in Banyumas Regency (301 villages), in terms of 12 variables, showed the following rankings:

- Rank 1 with independent village qualifications: 39 villages are members of Cluster 3;
- Rank 2 with advanced village qualifications: consisting of 90 villages that are members of Cluster 2;
- Rank 3 with developing village qualifications: 172 villages joined in Cluster 1.

Judging from the proportion of each cluster (number of villages), in 2018, the quality of development achievements in the socio-economic aspects of 301 villages in the Banyumas Regency area was 57 percent or dominant, which were still developing villages (172 out of 301). Only 13 percent belong to independent villages. These empirical findings indicate that there is a relatively large development gap in the Banyumas Regency area.

**TABLE 4. Item response probabilities**

| Variable | Value                  | Cluster 1 | Cluster 2 | Cluster 3 |
|----------|------------------------|-----------|-----------|-----------|
| X1       | 1= agriculture         | 0.7690    | 0.9721    | 0.6691    |
|          | 2= non-agriculture     | 0.2310    | 0.0279    | 0.3309    |
| X2       | 1= LPG gas             | 0.9757    | 0.9862    | 1.0000    |
|          | 2= non-LPG gas         | 0.0243    | 0.0138    | 0.0000    |
| X3       | 1= yes, used           | 0.2430    | 0.0349    | 0.4018    |
|          | 2= yes, not used       | 0.0480    | 0.0000    | 0.0000    |
|          | 3= no                  | 0.7090    | 0.9651    | 0.5982    |
| X4       | 1= toilet              | 0.9794    | 1.0000    | 1.0000    |
|          | 2= non-toilet          | 0.0206    | 0.0000    | 0.0000    |
| X5       | 1= tank                | 0.6372    | 0.7267    | 0.7938    |
|          | 2= non-tank            | 0.3628    | 0.2733    | 0.2062    |
| X6       | 1= yes, managed        | 0.7210    | 0.4213    | 0.1482    |
|          | 2= yes, not managed    | 0.2790    | 0.0504    | 0.0000    |
|          | 3= no                  | 0.0000    | 0.5284    | 0.8518    |
| X7       | 1= yes                 | 0.2424    | 0.0000    | 0.7411    |
|          | 2= no                  | 0.7576    | 1.0000    | 0.2589    |
| X8       | 1= yes                 | 1.0000    | 0.9716    | 1.0000    |
|          | 2= no                  | 0.0000    | 0.0284    | 0.0000    |
| X9       | 1= yes                 | 0.0062    | 0.0101    | 0.0206    |
|          | 2= no                  | 0.9938    | 0.9899    | 0.9794    |
| X10      | 1= asphalt/concrete    | 1.0000    | 0.9905    | 1.0000    |
|          | 2= others              | 0.0000    | 0.0095    | 0.0000    |
| X11      | 1= yes, with fixed routes | 0.9485 | 0.4615 | 0.7867 |
|          | 2= yes, without fixed routes | 0.0515 | 0.0745 | 0.0528 |
|          | 3= no                  | 0.0000    | 0.4640    | 0.1605    |
| X12      | 1= mostly residents    | 0.9931    | 1.0000    | 0.9798    |
|          | 2= a small number of residents | 0.0069 | 0.0000 | 0.0202 |
|          | 3= nothing             | 0.0000    | 0.0000    | 0.0000    |
| Overall  |                       | 0.4844    | 0.3513    | 0.1643    |
TABLE 5. The number of villages by sub district and clustering results

| Sub District | Cluster 1 | Cluster 2 | Cluster 3 |
|--------------|-----------|-----------|-----------|
| Lumbir       | 7         | 3         | 0         |
| Wangon       | 6         | 2         | 4         |
| Jatilawang   | 5         | 4         | 2         |
| Rawalo       | 2         | 4         | 3         |
| Kebasen      | 11        | 1         | 0         |
| Kemranjen    | 2         | 9         | 4         |
| Sumpiuh      | 2         | 9         | 0         |
| Tambak       | 2         | 6         | 4         |
| Somagede     | 4         | 5         | 0         |
| Kalijapak    | 2         | 8         | 2         |
| Banyumas     | 6         | 4         | 2         |
| Patikraja    | 4         | 8         | 1         |
| Purwojati    | 2         | 7         | 1         |
| Ajibarang    | 14        | 0         | 1         |
| Gumelar      | 6         | 3         | 1         |
| Pekuncen     | 13        | 2         | 1         |
| Cilongok     | 16        | 3         | 1         |
| Karanglewas  | 11        | 1         | 1         |
| Kedung Banteng | 13        | 1     | 0         |
| Baturraden   | 11        | 1         | 0         |
| Sumbang      | 17        | 2         | 0         |
| Kembaran     | 10        | 1         | 5         |
| Sokaraja     | 6         | 6         | 0         |
| Total        | 172       | 90        | 39        |

5. Conclusion
In this article, we have introduced the LCA method for use in village clustering. The proposed method was applied to Podes data in 2018 for Banyumas Regency. The LCA assign to group the villages into mutually exclusive categories that represent different levels. Based on the item response probability of the twelve variables used and on the latent class membership probability value, the villages were grouped into three clusters.

Acknowledgments
This research was conducted based on data supported by BPS. We thank BPS for their support. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of BPS. We are grateful to all those who helped in this research.

References
[1] BPS 2018 Pedoman Pencacahan Desa/Kelurahan Podes 2018 – Buku 2 (BPS, Jakarta)
[2] Butera N M, Lanza S T, and Coffman D L 2013 A Framework for Causal Effect in Latent Class Analysis: Is There a Causal Link Between Early Sex and Subsequent Profiles of Delinquency? Prevention Science, 15 (3) pp. 397-407
[3] Muthen B, Muthen L K 2000 Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling with Latent Trajectory Classes Alcoholism: Clinical and Experimental Research, 24 pp. 882-891
[4] Porcu M, Giambona F 2017 Introduction to Latent Class Analysis with Applications Journal of Early Adolescence, 37 (1) pp. 129-158
[5] Boldero J M, Bell R C, and Et More S M 2010 Do Gambling Activity Patterns Predict Gambling Problem? A Latent Class Analysis of Gambling Forms Among Australian Youth *International Gambling Studies, 10* (2) pp. 151-163
[6] Banfield J D, Raftery A E 1993 Model-based Gaussian and non-Gaussian Clustering *Biometrics, 49* pp. 803-821
[7] Bensmail H, Celeux G, Raftery A E, and Robert C P 1997 Inference in Model Based Clustering *Statistics and Computing, 7* pp. 1-10
[8] Bacher J 2000 A Probabilistic Clustering Model for Variables of Mixed Type *Quality and Quantity, 34* pp. 223-235
[9] Vermunt J K, Magidson J 2000 *Latent GOLD’s User’s Guide* (Statistical Innovations, Inc., Boston)
[10] Dean N, Raftery A E 2010 Latent Class Analysis Variable Selection *Annals of The Institute of Statistical Mathematics, Vol. 62* pp. 11-35
[11] Magidson J, Vermunt J K 2002 A Nontechnical Introduction to Latent Class Models; Statistical Innovations White Paper, in www.statisticalinnovations.com/wp-content/uploads/Magidson2002.pdf
[12] Moustaki I, Papageorgiou I 2004 Latent Class Models for Mixed Variables with applications in Archaeometry *Computational Statistics & Data Analysis, Vol. 48* (3) pp. 659 – 675
[13] Goodman L A 1974 Exploratory Latent Structure Analysis using both Identifiable and Unidentifiable Models *Biometrika, Vol. 61* pp. 215-231
[14] Kass R E, Raftery A E 1995 Bayes Factors *Journal of The American Statistical Association, Vol. 90* pp. 773-785
[15] Sclove S 1985 Application of Model Selection Criteria to Some Problems of Multivariate Analysis *Psychometrika, Vol. 52* pp. 333-343
[16] Lin T H, Dayton C M 1997 Model Selection Information Criteria for Non-Nested Latent Class Models *Journal of Educational and Behavioral Statistics, Vol. 22*(3) pp. 249-264
[17] Forster M R Key Concepts in Model Selection: Performance and Generalizability *Journal of Mathematical Psychology, Vol. 44* pp. 205-231
[18] Posada D, Buckley T R 2004 Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Test *Offord Journal: Society of Systematic Biologists, Vol. 53*(5) pp. 793-808