An Algorithm for the Generation of Categorical Variables Under Cluster Sampling

Abstract

This work describes the underlying model of an algorithm apt to simulate bivariate clusters of categorical variables. Clusters of $k$ individuals (each individual represented by a vector $(x, y)$, $x = 1, \ldots, c$, $y = 1, \ldots, r$) are modeled as the $k$ first steps of a Markov chain. The model controls the $r \times c$ vector of parent joint probabilities, $\pi$, the number of categories of $x$ and $y$, and the transition matrix, $M$, that rules the intra-cluster correlation. The size, $k$, depends on them. This paper offers exact formulas for an unbiased estimate of $\pi$, and the covariance matrix of its elements. The logic under the model extends immediately to clusters of higher dimensions.

Key Words: Clustered Data, Complex Survey, Markov Chain, Simulation Study

1 INTRODUCTION

Standard statistical methods has been developed under the assumption that simple random sampling is the scheme used to collect the data. Although assuming independence of observations eases the obtention of theoretical results, in both social and health sciences as well
as in economic studies the independence assumption is infrequently met. Human populations are grouped into clusters, some of them are selected to conform the sample, which violates the independence assumption. The lack of independence is not a problem when complex samples are used with descriptive purposes but some issues arise when they are used with inferential purposes and the sample design is not taken into account in the analysis. The procedures used to make inferences from complex surveys are mostly based on presumptions with some theoretical and empirical basis. In the absence of suitable theorems these presumptions could be evaluated through empirical studies.

Most of the empirical studies consist in the simulation of complex samples analyzed by the procedures under investigation, and the results are used to understand the behavior of the procedures. A key point to consider in these studies is the method used to generate the data, since conclusions might depend on it. Several models for generating cluster samples have been developed in the last two decades for categorical variables such as those presented by Cohen (1976), Brier (1980), and Thomas and Rao (1987). Other alternatives has been proposed as an extension to the Brier’s model by Scott and Rao (1981) and one based on a modified logistic normal distribution proposed by Thomas, Singh and Roberts (1995). However these models
except the last one, do not depend on meaningful parameters needed for the interpretation of contingency tables built from cluster samples.

The purpose of this paper is to present a model suitable for investigating the behavior of inferential procedures for analyzing bivariate contingency tables when the population is grouped into clusters. The underlying ideas can also be applied to tables with more than two dimensions. The model presented in this paper can generate data sets where relevant parameters vary in ranges that are similar to the ones of complex data found in real problems.

2 DEFINITIONS AND NOTATION

The relevant parameters for the analysis of bivariate contingency tables derived from complex survey are:

i. the number of rows and columns of the table.

ii. the size of the cluster \(k\) and the number of cluster intersample \((n)\).

iii. the population probabilities of the table.

iv. the generalized design effects of rows and columns of the table.

v. the generalized joint design effects.

vi. the coefficient of variation of the generalized joint design effects.
The sampling scheme to be considered is a one stage cluster sampling with equal sized clusters \((k)\) and all the elements in the cluster being selected in the sample. Row and column variables of the table are identified with \(X\) and \(Y\) respectively, taking values in the sets \(\mathbb{R} = \{1,2,\ldots,r\}\) and \(\mathbb{C} = \{1,2,\ldots,c\}\).

Each observation can be identified with a vector \((x, y)'\) taking values in \(\{(i, j), \ i \in \mathbb{R}; \ j \in \mathbb{C}\}\). The values taken by the \(k\) observations of a cluster can be arranged in a \(2 \times k\) matrix, with each column associated with an observation, the first row with variable \(X\) and the second row with variable \(Y\).

Let \(\pi_{ij}\) the probability of the cell \((i, j)\) in a \((r \times c)\) contingency table, with \(i \in \mathbb{R}\) and \(j \in \mathbb{C}\). Marginal probabilities of row and columns are represented as \(\pi_{i+}\) and \(\pi_{+j}\). In matrix notation, joint and marginal distributions are expressed as \(\pi = (\pi_{11}, \ldots, \pi_{rc})'\), \(\pi_R = (\pi_{1+}, \ldots, \pi_{r+})'\), and \(\pi_C = (\pi_{+1}, \ldots, \pi_{+c})'\).

Let \(\hat{\pi}\), \(\hat{\pi}_R\), and \(\hat{\pi}_C\), be consistent estimators of \(\pi, \pi_R\), and \(\pi_C\) obtained from a sample that takes into account the complex design. Let \(V\) be the covariance matrix of \(\hat{\pi}\) taking into account the design and \(\hat{V}\) a consistent estimator. Let \(V_R\), \(V_C\), \(\hat{V}_R\), and \(\hat{V}_C\) be defined in analogous way with respect to \(\hat{\pi}_R\) and \(\hat{\pi}_C\). Furthermore, let \(P = (1/n)(\text{diag}(\pi) - \pi \ \pi')\) be the \(rc \times rc\) multinomial covariance matrix and \(\hat{P} = (1/n)(\text{diag}(\hat{\pi}) - \hat{\pi} \ \hat{\pi}')\) an estimator based on \(\hat{\pi}\). De-
fine $P_R, P_C, \hat{P}_R,$ and $\hat{P}_C$ with respect to $\hat{\pi}_R$ and $\hat{\pi}_C$ in similar way.

Let $\lambda_{R(i)}$, i=1,..,(r-1), $\lambda_{C(j)}$, j=1,..,(c-1) and $\lambda_m$, m=1,..,(rc-1) be the eigenvalues of the design effects matrices $D_R=\hat{P}_R^{-1} \hat{V}_R$, $D_C=\hat{P}_C^{-1} \hat{V}_C$, and $D=\hat{P}^{-1} \hat{V}$ respectively. They are named generalized design effects. The dot (.) denotes a trimmed matrix, obtained by deleting the last row and column of the matrix. Let $\bar{\lambda}_R$, $\bar{\lambda}_C$, and $\bar{\lambda}$ the means of $\{ \lambda_{R(i)} \}$, $\{ \lambda_{C(j)} \}$, and $\{ \lambda_m \}$ respectively. Let $a(\lambda)$ be the coefficient of variation of $\{\lambda_1, \ldots, \lambda_{(rc-1)}\}$.

3 THE MODEL

The model depends, explicitly on the following parameters:

i. $r$ and $c$, the number of rows and columns of the table.

ii. $\pi_{ij}$, $i \in \mathbb{R}$ $j \in \mathbb{C}$, the true probabilities of the contingency table. $\pi_{ij}$ can model situations of independence as well as departures from it.

iii. $k$, the number of units in each cluster.

iv. the conditional probability that a unit have of belonging to the $(i, j)$ cell of the contingency table, given that the previous unit of the same cluster is in the cell $(i', j')$. There are $(rc \times rc)$ of their kind.

The intraclass correlation of the members of the same cluster is a
function of the parameters of the model allowing to simulate different degrees of clustering. A closed expression for an unbiased estimator of the true probability vector \( \pi = (\pi_{11}, \ldots, \pi_{rc})' \) from a random sample of clusters generated by the model can be found, as well as its variance. This variance is a function of the parameters described in i-iv.

### 3.1 Description of the model

Let \( \nu(u) \) be random variable denoting the bivariate response of the \( u^{th} \) observation in the cluster, \( u = 1, \ldots, k \). For any \( u \), \( \nu(u) \) takes values in the set \( \{(i, j); i \in \mathbb{R}; j \in \mathbb{C}\} \). To avoid double subscripts the vectors in this set will be noted with a single subscript \( h = c(i - 1) + j, h \in \mathbb{RC}, \mathbb{RC} = \{1, 2, \ldots, rc\} \). The distribution of \( \nu(u) \) depends on \( u \). For each cluster of size \( k \), the first observation comes from the probability distribution:

\[
a_{\nu(1)}, \nu(1) \in \mathbb{RC}, \sum_{\nu(1)} a_{\nu(1)} = 1 \tag{1}
\]

The observations that follow are outcomes of a Markov chain with initial probabilities \( \{a_{\nu(1)}\} \) and transition matrix

\[
M = ((p_{hh'})) \quad h, h' \in \mathbb{RC} \tag{2}
\]

where \( p_{hh'} \) is the conditional probability

\[
pr[\nu(u) = h' | \nu(u - 1) = h] \tag{3}
\]
with \( u = 2, \ldots, k \). So, each cluster is a different realization of the \( k \) paths of the same Markov process which generates, equally distributed independent clusters.

The probability of a generic cluster \( \{ \nu(1), \ldots, \nu(k) \} \) is,

\[
P_{k,\nu(1),\nu(2),\ldots,\nu(k)} = a_{\nu(1)} P_{\nu(1)\nu(2)} P_{\nu(2)\nu(2)} \cdots P_{\nu(k-1)\nu(k)}
\]

with \( \nu(,) \in \mathbb{R}C \). Also the probability that the first observation in the cluster has the value \( h \) and the \( u^{th} \) observation the value \( h' \) is,

\[
P_{k,\nu(1)=h,\nu(u)=h'} = \sum_{\nu(1), \ldots, \nu(k) \neq \nu(1), \nu(u)} P_{k,\nu(1)=h,\nu(2), \ldots, \nu(u)=h', \nu(u+1), \ldots, \nu(k)} h, h' \in \mathbb{R}C
\]

Expression (5) is the probability of passing from the initial state \( h \) to the state \( h' \) in exactly \((u-1)\) steps. That is,

\[
P_{k,\nu(1)=h,\nu(u)=h'} = a_h p_{hh'}^{(u-1)} h, h' \in \mathbb{R}C, \ u = 1, \ldots, k
\]

or, in matrix notation,

\[
((P_{k,\nu(1)=h,\nu(u)=h'}))_{h,h'\in\mathbb{R}C} = D_{a}(a) M^{(u-1)}
\]

where \( D_{a}(a) \) is a diagonal matrix with non-zero elements equal to the initial probabilities. Furthermore,

\[
P_{k, \nu(u)=h} = \sum_{\nu(1), \ldots, \nu(u) \neq \nu(u)} P_{k, \nu(1), \ldots, \nu(u)=k, \ldots, \nu(k)} = \sum_{h' \in \mathbb{R}C} P_{k,\nu(1)=h', \nu(u)=h} = \sum_{h' \in \mathbb{R}C} a_{h'} p_{hh'}^{(u-1)}
\]
Since the sample of clusters is a simple random one, the probabilities \( P_{k, \nu(1), \ldots, \nu(k)} \) can easily be estimated. The probabilities corresponding to each cluster are not the main goal, instead the interest lies on the probabilities \( \{\pi_h, h \in \mathbb{R}_C\} \) of a random observation from the whole population having the characteristic \( h \). However, it is possible to obtain the later probabilities from the former. The probability of obtaining an observation with characteristic \( h \) in the population composed by all the clusters of size \( k \) is,

\[
\pi_h = \sum_{\nu(1), \ldots, \nu(k)} P_{k, \nu(1), \ldots, \nu(k)} \Pr(h|\nu(1), \ldots, \nu(k)), \ h \in \mathbb{R}_C \tag{9}
\]

where \( \Pr(h|\nu(1), \ldots, \nu(k)) \) is the probability of selecting an observation in the category \( h \) when \( (\nu(1), \ldots, \nu(k)) \) are fixed values. This probability is the ratio between \( m(h|\nu(1), \ldots, \nu(k)) \) the number of observations in the category \( h \) in the cluster and \( k \). That is,

\[
k \pi_h = \sum_{\nu(1), \ldots, \nu(k)} P_{k, \nu(1), \ldots, \nu(k)} m(h|\nu(1), \ldots, \nu(k)) \quad h \in \mathbb{R}_C \tag{10}
\]

or alternatively,

\[
k \pi_h = \sum_{u=1}^{k} P_{k, \nu(u)=h} = \sum_{u=1}^{k} \sum_{h \in \mathbb{R}_C} a_h P_{h|h}^{(u-1)} \tag{11}
\]

The second equality is a consequence of (3.1.8) and can be written emphasizing \( \{a_h; h \in \mathbb{R}_C\} \) as

\[
k \pi_h = a_h + a_1 \sum_{i=1}^{k-1} p_{1h}^i + a_2 \sum_{i=1}^{k-1} p_{2h}^i + \ldots + a_{rc} \sum_{i=1}^{k-1} p_{(rc)h}^i \tag{12}
\]
This system is crucial in the simulations, as it will be shown in the next paragraph. A general algorithm can be obtained to generate bivariate clusters according the preceding model. It can be described in three steps:

i. A transition matrix \( M = ((p_{hh'})) \) \( h, h' \in \mathbb{R} \mathbb{C} \) and the vector of population probabilities \( \{\pi_h; h \in \mathbb{R} \mathbb{C}\} \) are set to convenient values. Also, the dimension \( (k) \) of the clusters to be generated is also set.

ii. The chosen values are used to conform the system (12) and this is solved for the \( a' s \). The system can be incompatible, or even if it is compatible can produce invalid solutions. The structure of the system ensures that \( \sum_h a_h = 1 \), but it does not grant that all the roots will be positive. If some are not, the problem is said to be not viable. Only viable problems are used in step (iii).

iii. Clusters are simulated in the following way: the first element is the outcome of a discrete random variable with distribution \( \{a_h; h = 1, \ldots, rc\} \), the following element comes from the conditional distribution, given the obtained value of the first, and so on.
3.2 Estimation of $\pi_h$. Expected value and covariance matrix

Suppose that the model is used for generating a random simple sample of $n$ clusters. If the clusters have size $k$, an unbiased estimator of $\pi_h$ is,

$$\hat{\pi}_h^{(k)} = \sum_{u=1}^{k} \hat{P}_{k,\nu(u)=h}/k$$

(13)

$\hat{P}_{k,\nu(u)=h}$ is a linear combination of the probabilities $P_{k,\nu(1),\nu(2),\ldots,\nu(k)}$.

Since the selection of clusters is completely at random, each of the above probabilities is estimated as the relative frequency of response $\nu(1), \ldots, \nu(k)$ in the sample of clusters. Designating with $\hat{P}_{k,\nu(1),\nu(2),\ldots,\nu(k)}$ each of these frequencies, it results,

$$\hat{P}_{k,\nu(u)=h} = \sum \hat{P}_{k,\nu(1),\ldots,\nu(k)} m(h|\nu(1),\ldots,\nu(k))/k,$$

(14)

and $\hat{\pi}_h^{(k)}$ is an unbiased estimator of $\pi_h$. The covariance matrix of $\{\hat{\pi}_h^{(k)}\}$ is given by the following theorem.

**Theorem** Under the model described in the preceding section, if $\hat{\pi}_h^{(k)}$, $\hat{\pi}_h^{(k)}$, and $\hat{\pi}_h^{(k)}$, $\hat{\pi}_h^{(k)}$, are computed from a sample of $n$ clusters of size $k$, as in (3.2.1),

$$cov(\hat{\pi}_h^{(k)}, \hat{\pi}_h^{(k)}) = \frac{1}{nk^2} [k\pi_h(\delta_{hh'} - k\pi_{h'}) + \sum_{u \neq t} P_{k,\nu(u)=h,\nu(t)=h'}]$$

(15)

where $P_{k,\nu(u)=h,\nu(t)=h'}$ is the probability that the $u^{th}$ observation has the value $h$ and the $t^{th}$ has the value $h'$ in clusters of size $k$; $\delta_{hh'} = 1$
if \( h = h' \) and \( \delta_{hh'} = 0 \) otherwise.

**Proof.**

Let \( k = 2 \),

\[
2\pi_h^{(2)} = \hat{P}_{2,\nu(1)=h} + \hat{P}_{2,\nu(2)=h},
\]

where,

\[
\hat{P}_{2,\nu(1)=h} = \hat{P}_{2,\nu(1)=h,\nu(2)=\bar{h}} + \hat{P}_{2,\nu(1)=h,\nu(2)=h},
\]

\[
\hat{P}_{2,\nu(2)=h} = \hat{P}_{2,\nu(1)=\bar{h},\nu(2)=h} + \hat{P}_{2,\nu(1)=h,\nu(2)=h},
\]

and \( \bar{h} \) denotes a response different from \( h \).

\( \hat{P}_{2,\nu(1)=h,\nu(2)=\bar{h}} \) and \( \hat{P}_{2,\nu(1)=\bar{h},\nu(2)=h} \) are the estimated relative frequencies for the events \( \{\nu(1) = h, \nu(2) = \bar{h}\} \), \( \{\nu(1) = \bar{h}, \nu(2) = h\} \) and \( \{\nu(1) = h, \nu(2) = h\} \) which are mutually exclusive. For a given \( h \), the absolute frequencies in a sample of size \( n \) are distributed as a multinomial distribution with parameters \( \{n, \hat{P}_{2,\nu(1)=h,\nu(2)=\bar{h}}, \hat{P}_{2,\nu(1)=\bar{h},\nu(2)=h}, \hat{P}_{2,\nu(1)=h,\nu(2)=h}\} \).

From this distribution it can be obtained

\[
\text{var}(\hat{P}_{2,\nu(t)=h}) = \frac{P_{2,\nu(t)=h}(1 - P_{2,\nu(t)=h})}{n} \quad t = 1, 2,
\]

\[
cov(\hat{P}_{2,\nu(1)=h}, \hat{P}_{2,\nu(2)=h'}) = -\left[ \frac{P_{2,\nu(1)=h}P_{2,\nu(2)=h'}}{n} \right] + \frac{P_{2,\nu(1)=h,\nu(2)=h'}}{n},
\]

Using (16), (19) and (20)

\[
\text{var}(\pi_h^{(2)}) = \frac{1}{4} \left\{ \frac{2\pi_h(1 - 2\pi_h)}{n} \right\} + \frac{P_{2,\nu(1)=h,\nu(2)=h}}{2n}
\]
For the covariance between $2\hat{\pi}_h^{(2)}$ and $2\hat{\pi}_{h'}^{(2)}$ with $h \neq h'$ ($\in RC$), the mutually exclusive events to consider are

\[
\{\nu(1) = h, \nu(2) = h\}, \{\nu(1) = h, \nu(2) = h'\}, \{\nu(1) = h, \nu(2) = hh'\} \tag{22}
\]

\[
\{\nu(1) = h', \nu(2) = h\}, \{\nu(1) = h', \nu(2) = h'\}, \{\nu(1) = h', \nu(2) = hh'\} \tag{23}
\]

\[
\{\nu(1) = hh', \nu(2) = h\}, \{\nu(1) = hh', \nu(2) = h'\}, \{\nu(1) = hh', \nu(2) = hh'\} \tag{24}
\]

where $hh'$ means that the pair $\{\nu(1), \nu(2)\}$ is different from $hh'$. In a sample of $n$ clusters its frequencies are random variables distributed as a multinomial whose parameters are $n$ and the probabilities of the nine events above.

Given that,  

\[
2\hat{\pi}_h^{(2)} = \hat{P}_{2,\nu(1)=h,\nu(2)=h} + \hat{P}_{2,\nu(1)=h,\nu(2)=h'} + \hat{P}_{2,\nu(1)=h,\nu(2)=hh'} + \hat{P}_{2,\nu(1)=h',\nu(2)=h} + \hat{P}_{2,\nu(1)=h',\nu(2)=h'} \tag{25}
\]

and analogous formula holds for $2\hat{\pi}_{h'}^{(2)}$ it results,

\[
cov(\hat{\pi}_h^{(2)}, \hat{\pi}_{h'}^{(2)}) = -\left[\frac{\pi_h\pi_{h'}}{n}\right] + \frac{P_{2,\nu(1)=h,\nu(2)=h'}}{4n} + \frac{P_{k,\nu(1)=h',\nu(2)=h}}{4n} \tag{26}
\]

The complete inductive principle will be used to prove (3.2.3). Because the marginal distributions of a $k$-variate multinomial distribution are
denote the generic parameter probabilities of a \((k - 1)\) variate multinomial distribution. Then, if \(u < k\),

\[
P_{k,\nu(u)=h} = \sum_{\nu(1) \ldots \nu(k-1) \neq \nu(k)} \sum_{\nu(k)} P_{k,\nu(1) \ldots \nu(u) \ldots \nu(k)} = \sum_{\nu(1) \ldots \nu(k-1) \neq \nu(u)} P_{(k-1),\nu(1) \ldots \nu(u) \ldots \nu(k-1)} = P_{(k-1),\nu(u)=h}, \quad u < k
\]

From (13),

\[
k\bar{\pi}^{(k)}_h = \sum_{u=1}^{k} \hat{P}_{k,\nu(u)=h} = \sum_{u=1}^{k-1} \hat{P}_{(k-1),\nu(u)=h} + \hat{P}_{k,\nu(k)=h} = (k-1)\bar{\pi}^{(k-1)}_h + \hat{P}_{k,\nu(k)=h}
\]

Then,

\[
cov(k\bar{\pi}^{(k)}_h, k\bar{\pi}^{(k)}_{h'}) = cov[(k-1)\bar{\pi}^{(k-1)}_h, (k-1)\bar{\pi}^{(k-1)}_{h'}] +
\]

\[
+ cov[(k-1)\bar{\pi}^{(k-1)}_h, \hat{P}_{k,\nu(k)=h'}] + cov[\hat{P}_{k,\nu(k)=h}, (k-1)\bar{\pi}^{(k-1)}_{h'}] +
\]

\[
+ cov[\hat{P}_{k,\nu(k)=h}, \hat{P}_{k,\nu(k)=h'}]
\]

It can be seen from (29) that,

\[
cov[(k-1)\bar{\pi}^{(k-1)}_h, \hat{P}_{k,\nu(k)=h'}] = \sum_{u=1}^{(k-1)} cov(\hat{P}_{(k-1),\nu(u)=h}, \hat{P}_{(k),\nu(k)=h'}) =
\]
\[= \frac{1}{n} \left\{ \sum_{u=1}^{(k-1)} P_{k,\nu(u)=h,\nu(k)=h'} - \sum_{u=1}^{(k-1)} P_{k,\nu(u)=h} P_{k,\nu(k)=h'} \right\} \]  

And, 

\[\text{cov}(\hat{P}_{k,\nu(k)=h}, \hat{P}_{k,\nu(k)=h'}) = \frac{1}{n} \{ P_{k,\nu(k)=h} (\delta_{hh'} - P_{k,\nu(k)=h'}) \} \]  

Using (15) for \( \text{cov}[(k-1)\hat{\pi}_h^{(k-1)}, (k-1)\hat{\pi}_h'^{(k-1)}] \), (31) and (32) for the proper terms in (30) plus some algebra, the theorem is proved.

### 3.3 Univariate Clusters

In this section, it is convenient to use the notation stated in Section 2 where each observation is identified with double subscript \((i, j)\).

Ignoring one of the variables of the model presented in the previous section, clusters become univariate. Associated to the cluster \(\{\nu(1), \ldots, \nu(k)\}, \nu(u) \in RC, u = 1, \ldots, k\), there are two univariate clusters \(\{x(1), \ldots, x(k)\}, x(u) \in R, u = 1, \ldots, k\), and \(\{y(1), \ldots, y(k)\}, y(u) \in C, u = 1, \ldots, k\), that are the first and the second rows of the \(2k\) matrix that correspond to \(\{\nu(1), \ldots, \nu(k)\}\). Now, \(\nu(u) = (x(u), y(u))'\). It can be seen that \(\{x(1), \ldots, x(k)\}\) can be generated as the first \(k\) steps of a Markov chain of the type described in Section 3.1, but with parameters,

\[\pi_{i+} = \sum_{j=1}^{c} \pi_{ij} \quad i \in R, \quad a_{i+}^{(x)} = \sum_{j=1}^{c} a_{ij} \quad i \in R, \quad M_x = (p_{ii'}^{(x)})_{ii' \in R} \]  

\[(33)\]
with
\[ p_{i'i'}^{(x)} = \sum_j \sum_{j'} a_{ij} p_{ij'i'}^{(x)} a_{i'j'}^{(x)} \quad i, i' \in R. \] (34)

In the same way, the second row Markov chain has the following parameters,
\[ \pi_{+j} = \sum_{i=1}^{r} \pi_{ij} \quad j \in C, \quad a_{+j}^{(y)} = \sum_{i=1}^{r} a_{ij} \quad j \in C, \quad M_y = ((p_{jj'}^{(y)}))_{j,j' \in C} \] (35)

with
\[ p_{jj'}^{(y)} = \sum_i \sum_{i'} a_{ij} p_{ij'i'}^{(y)} a_{i'j'}^{(y)} \quad j, j' \in C. \] (36)

The probability that corresponds to an \( x \)-cluster is
\[ P_{k,x(1),\ldots,x(k)} = \sum_{y(1),\ldots,y(k)} P_{k,\nu(1),\ldots,\nu(k)} \] (37)
and for the \( y \)-cluster is
\[ P_{k,y(1),\ldots,y(k)} = \sum_{x(1),\ldots,x(k)} P_{k,\nu(1),\ldots,\nu(k)} \] (38)

From the univariate clusters, \( \pi_{i+} \quad i \in R \), and \( \pi_{+j} \quad j \in C \) can be estimated as and
\[
\hat{\pi}_{i+} = \sum_{u=1}^{k} \sum_{x(1),\ldots,x(k)} \hat{P}_{k,x(1),\ldots,x(u)=i,...,x(k)} = \sum_{u=1}^{k} P_{k,x(u)=i} \] (39)
\[
\hat{\pi}_{+j} = \sum_{u=1}^{k} \sum_{y(1),\ldots,y(k)} \hat{P}_{k,y(1),\ldots,y(u)=j,...,y(k)} = \sum_{u=1}^{k} P_{k,y(u)=j} \] (40)
The vectors \( \hat{\pi}_{1+}, \ldots, \hat{\pi}_{r+} \) and \( \hat{\pi}_{+1}, \ldots, \hat{\pi}_{+c} \) were named \( \hat{\pi}_R \) and \( \hat{\pi}_C \) in Section 2. Formula (15) specializes for univariate clusters to:

\[
\text{cov}(\hat{\pi}_{i+}, \hat{\pi}_{i'}) = \frac{1}{nk^2} \left[ k\pi_{i+} (\delta_{ii'} - k\pi_{i'}) + \sum_{u \neq t} P_{k,x(u) = i,x(t) = i'} \right] \tag{41}
\]

and

\[
\text{cov}(\hat{\pi}_{+j}, \hat{\pi}_{+j'}) = \frac{1}{nk^2} \left[ k\pi_{+j} (\delta_{jj'} - k\pi_{+j'}) + \sum_{u \neq t} P_{k,y(u) = j,y(t) = j'} \right] \tag{42}
\]

where

\[
P_{k,x(u) = i,x(t) = i'} = \sum_{x(1), \ldots, x(k) \neq x(u), x(t)} P_{k,x(1),\ldots,x(u) = i,\ldots,x(t) = i',\ldots,x(k)} \tag{43}
\]

and

\[
P_{k,y(u) = j,y(t) = j'} = \sum_{y(1), \ldots, y(k) \neq y(u), y(t)} P_{k,y(1),\ldots,y(u) = j,\ldots,y(t) = j',\ldots,y(k)} \tag{44}
\]
4 USE OF THE MODEL FOR GENERATION OF CASES

It would be interesting to keep control on the parameters associated to the univariate clusters because the marginal generalized design effects ($\lambda_{R1}, \ldots, \lambda_{Rr}$) and ($\lambda_{C1}, \ldots, \lambda_{Cc}$) depend on them. It is complicated to fix the first member of (34) and (36) and then, to solve for $\{p_{ij,i'j'}\}$ to get the transition matrix ($M$) of the bivariate clusters.

Instead, the search for $M$, whose elements satisfy (34) and (36) can be simplified if $M$ is looked for among the matrices such that

$$M = M_1 \otimes M_2$$

(45)

Where $M_1=\{(p_{ii'})\}_{i,i' \in R}$, and $M_2=\{(p_{jj'})\}_{j,j' \in C}$. In this case, $p_{ij,i'j'} = p_{ii'}p_{jj'}$.

The search for $M$ is reduced to looking for matrices $M_1$ and $M_2$. In the case of 3x3 contingency tables it is proposed to use matrices of the following type

$$M_t = \begin{bmatrix}
\eta_1^{(t)} & \gamma^{(t)}(1-\eta_1^{(t)}) & (1-\gamma^{(t)})(1-\eta_1^{(t)}) \\
\gamma^{(t)}(1-\eta_2^{(t)}) & \eta_2^{(t)} & (1-\gamma^{(t)}+\gamma^{(t)}\eta_1^{(t)}-\eta_2^{(t)}) \\
(1-\gamma^{(t)})(1-\eta_1^{(t)}) & (1-\gamma^{(t)}+\gamma^{(t)}\eta_1^{(t)}-\eta_2^{(t)}) & 2\gamma^{(t)}(1-\eta_2^{(t)}+\eta_1^{(t)}+\eta_2^{(t)})-1
\end{bmatrix}$$

(46)

with $0 < \eta_1^{(t)} < 1$, $0 < \eta_2^{(t)} < 1$, $0 < \gamma^{(t)} < 1$, $t = 1, 2$.

The matrices $M_t$ are symmetric and their columns and rows add
to 1. Taking $\eta_1^{(t)} = \eta_2^{(t)} = \frac{1}{3}$, $\gamma = \frac{1}{2}$, $M_t = \frac{1}{3} J$, for $t = 1, 2$, produces $M = M_1 \otimes M_2 = \frac{1}{9} J$ where $J$ is a $9 \times 9$ matrix with all elements equal to 1. Therefore, the elements in the same cluster would be independent, and the complex sample would become a simple random one with the joint distribution of the frequencies of the states being multinomial and the design effects equal to 1.

In opposite way, if $M_1 = M_2 = I$, the identity matrix, the elements in the same cluster would be all the same and the intraclass correlation would be 1. Playing with the relative sizes of the diagonal elements and the off diagonal ones of the transition matrices, different degrees of clustering can be found.

The method to generate cases (that is, set of parameters) of interest can be described in the following steps:

a. Select $k$, the size of the cluster; the number of rows ($r$) and columns ($c$).

b. Select the parameters of the $r \times c$ contingency table $\{\pi_{ij}; i = 1, \ldots, r; j = 1, \ldots, c\}$. From it, the marginal probabilities $\{\pi_{i+}; i = 1, \ldots, r\}$ and $\{\pi_{+j}; j = 1, \ldots, c\}$ are established. Also compute the matrices $P$, $P_R$ and $P_C$ defined in Section 2.

c. Select the transition matrices $M_x$ and $M_y$.

d. Compute $M = M_x \otimes M_y$; compute the distribution of initial prob-
abilities \( \{a_{ij}; i = 1, \ldots, r; j = 1, \ldots, c\} \) as solution of (12) using 
\( \{\pi_{ij}; i = 1, \ldots, r; j = 1, \ldots, c\} \) and \( M \). If some of the initial probabilities are negative, disregard the case. If not, compute 
\( \{a_{i+}; i = 1, \ldots, r\} \) and \( \{a_{+j}; j = 1, \ldots, c\} \) and continue to step e.

e. From \( M, M_x \) and \( M_y \) and \( a_{ij}, a_{i+} \) and \( a_{+j} \) compute the parameters of the distribution of the bivariate and univariate clusters 
\( \{P_{k,\nu(1),\ldots,\nu(k)}\}, \{P_{k,x(1),\ldots,x(k)}\}, \{P_{k,y(1),\ldots,y(k)}\} \). From them and results of step b compute, according to formulas (15), (41), (42), (43) and (44), the matrices \( V, V_R \) and \( V_C \) defined in Section 2. After that, compute \( D = \dot{V}\dot{P}^{-1}, \quad D_R = \dot{V}_R\dot{P}_R^{-1} \) and \( D_C = \dot{V}_C\dot{P}_C^{-1} \).

f. Obtain the eigenvalues of \( D, D_R \) and \( D_C \), \( \lambda = \{\lambda_h, h = 1, \ldots, (rc-1)\}, \lambda_R = \{\lambda_{Rh}, i = 1, \ldots, (r-1)\}, \lambda_C = \{\lambda_{Cj}, j = 1, \ldots, (c-1)\} \), respectively. They are called the generalized joint design effects, row marginal design effects and column marginal design effects, respectively.

g. Compute \( \bar{\lambda}, \bar{\lambda}_R, \bar{\lambda}_C \) and \( CV(\lambda) \), that are the averages and the coefficient of variation of the design effects computed in f.

5 EVALUATION OF THE MODEL

Two aspects of the model were evaluated: flexibility and capacity to
generate populations with the specified parameters.

To evaluate the flexibility, a database with a very large set of cases was generated, all of them representing viable cases for bivariate clusters of 5 elements whose observations could be exhibited in a 3x3 table. The cases were created according to the values of marginal matrices of the type (4.1).

Each 3x3 case of the database can be identified by the following parameters: $k, \{\pi_{ij} ; i = 1, 2, 3 ; j = 1, 2, 3\}, \{\eta_1^{(x)}, \eta_2^{(x)}, \gamma^{(x)}\}, \{\eta_1^{(y)}, \eta_2^{(y)}, \gamma^{(y)}\}$. The parameters $k, \{\pi_{ij} ; i = 1, 2, 3 ; j = 1, 2, 3\}$ and the $\eta$'s were freely chosen. The restrictions were that they have to lead to positive solutions of equation (12) and positive values for the marginal transition matrices.

In the 3x3 case, $\bar{\lambda}_R$ is the average of $\lambda_{R1}$ and $\lambda_{R2}$, the two marginal generalized design effects. The same is $\bar{\lambda}_C$ with respect to the columns. $\bar{\lambda}$ is the average of, in this case, eight joint design effects: $\lambda_i, i = 1, \ldots, 8$; $CV(\lambda)$ is its coefficient of variation. As an example, to construct the data sets the values of the population probabilities $\pi_{ij}$ were set to
The values $\eta_1$ and $\eta_2$ for each marginal matrix were made to vary between 0.05 and 0.95 in 0.05 increments while $\gamma$ varied between $1/8$ and $7/8$ in $1/8$ increments. A case was considered viable if all the elements of the marginal matrices were positive and resulted in positive values for the initial probabilities (positive solutions to the equation (12)). Among the $6.385.729 \times 7^2$ possible cases, 1.713.905 were retained as viable. For each combination, the mean of the marginal and joint generalized design effects with the coefficient of variation were computed. The flexibility of the model to generate different situations was evaluated through the variety of scenarios that the model was able to generate.

A second step was to evaluate how closely samples generated by the model were able to estimate the specified parameters. Thus, a group of scenarios was selected from the database of viable cases and samples of $n = 15, 30, 70$ and 100 clusters of size $k = 5$ were generated and their marginal and joint generalized design effects estimated. This process was repeated 1000 times for each sample size and the results

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.109714</td>
<td>0.118857</td>
<td>0.114286</td>
</tr>
<tr>
<td>2</td>
<td>0.118857</td>
<td>0.128762</td>
<td>0.123810</td>
</tr>
<tr>
<td>3</td>
<td>0.091428</td>
<td>0.099048</td>
<td>0.095238</td>
</tr>
</tbody>
</table>
were averaged across the repetitions and tabulated.

5.1 Flexibility of the model

Table 1 exhibits selected values of $\eta_1^{(x)}$ and $\eta_2^{(x)}$ and the joint generalized design effects of the data base of viable cases. It can be seen that the higher are the values of $\eta_1^{(x)}$ and $\eta_2^{(x)}$ the higher are the average joint generalized design effects. Empty cells represent combinations where there were no viable cases. This table shows that the model is able to generate populations with different degrees of conglomeration, in some cases with design effects less than 1. Similar results were obtained for the marginal associated with the variable $Y$.

Table 2 exhibits the average and range of values of the coefficient of variation of the generalized joint design effects classified according to the values of the joint design effects means. Since the table considered was $3 \times 3$, there are eight generalized joint design effects $\{\lambda_1, \lambda_1, \ldots, \lambda_8\}$ for each case. Columns 3 and 4 presents the average values that marginal design effects can take. For instance, for those viable cases whose mean of the joint generalized design effects were set between 2 and 2.20, the coefficient of variation range, approximately, from 20 to 90%. It can be observed that for the average of joint design effects greater that 3 the ranges of coefficients of variations are reduced.
Table 3 presents the average value and range of variation for the mean of the generalized joint design effects for cases that have selected values of the marginal design effects. As expected, marginal design effects are related to the joint effect.

In Table 4 the relation between marginal design effects and coefficient of variation of the joint generalized design effect is presented. In most of the cases it can be seen that the coefficient of variation can take values between 20 and 80%.

It can be seen that the model allows simulate populations (cases) with very different joint and marginal design effects, and coefficients of variation of the generalized joint design effects. Also, for fixed values of the marginal design effects, cases with very different values of the average of the joint design effects and coefficients of variation of the joint generalized design effects can be obtained, as it is seen through the length of the intervals in the tables 3 and 4.

5.2 "Goodness" of the estimates of the parameters of the Model

To evaluate the properties of the estimates obtained through samples from the model, samples from several populations were simulated.
Five different scenarios were considered. Table 5 exhibits the parameters of each of them. These cases were taken to cover a certain range of marginal design effects and coefficients of variation of the joint design effects.

Table 6 shows the results of the simulations. The estimated means of the joint generalized design effects are very close to the parameters chosen for any sample size. For the marginal design effects the parameters are also close, especially for sample sizes greater than 15. The coefficient of variation of the generalized design effects are biased upward but the bias decreases with the sample size.

6 SOME REMARKS

The logic under the proposed model is also valid to generate data sets of clusters that can be displayed in a three or higher dimensional table ones. With respect to the size of the clusters it have been noticed that as the size increases the non viable cases increases too.

The model and its associated algorithm has been used to study the behavior of several tests for independence of two categorical variables (each with three categories). Also it has been used to study the performance of the quasi-score test of Rao, Scott and Skinner (1998) and different versions of the generalized score test of Rotnitzky et al.(1990).
7 References


Table 1: Average Values and Range of the Joint Design Effect ($\lambda$) for selected values of $\eta_1^{(x)}$ and $\eta_2^{(x)}$

<table>
<thead>
<tr>
<th>$\eta_1^{(x)}$</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.73$^1$</td>
<td>0.98</td>
<td>1.08</td>
<td>1.54</td>
<td></td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>(0.62-0.92)$^2$</td>
<td>(0.57-1.48)</td>
<td>(0.87-1.29)</td>
<td>(1.48-1.61)</td>
<td></td>
<td>(0.51-1.88)</td>
</tr>
<tr>
<td>0.25</td>
<td>0.77</td>
<td>0.98</td>
<td>1.29</td>
<td>1.66</td>
<td></td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>(0.69-0.87)</td>
<td>(0.70-1.27)</td>
<td>(1.01-1.66)</td>
<td>(1.60-1.71)</td>
<td></td>
<td>(0.63-2.23)</td>
</tr>
<tr>
<td>0.50</td>
<td>1.00</td>
<td>1.15</td>
<td>1.50</td>
<td>1.99</td>
<td></td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>(0.94-1.03)</td>
<td>(1.00-1.28)</td>
<td>(1.30-1.94)</td>
<td>(1.90-2.12)</td>
<td></td>
<td>(0.94-2.52)</td>
</tr>
<tr>
<td>0.75</td>
<td>1.48</td>
<td>1.64</td>
<td>1.94</td>
<td>2.66</td>
<td>3.39</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>(1.46-1.52)</td>
<td>(1.58-1.68)</td>
<td>(1.87-2.10)</td>
<td>(2.46-3.06)</td>
<td>(3.39-3.39)</td>
<td>(1.46-3.39)</td>
</tr>
<tr>
<td>0.95</td>
<td>2.28</td>
<td>2.40</td>
<td>2.70</td>
<td>3.34</td>
<td>4.40</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td>(2.28-2.29)</td>
<td>(2.40-2.41)</td>
<td>(2.69-2.72)</td>
<td>(3.30-3.40)</td>
<td>(4.30-4.52)</td>
<td>(2.28-4.52)</td>
</tr>
<tr>
<td>Any</td>
<td>1.40</td>
<td>1.23</td>
<td>1.72</td>
<td>2.40</td>
<td>3.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.62-2.29)</td>
<td>(0.57-2.41)</td>
<td>(0.87-2.72)</td>
<td>(1.48-3.40)</td>
<td>(3.10-4.52)</td>
<td></td>
</tr>
</tbody>
</table>

$^1$ Average Value Across viable cases,  $^2$ Minimum-Maximum Values
Table 2: Average Values of the Coefficient of Variation of the Generalized Joint Design Effect \((a(\lambda))\) and the Generalized Marginal Effects \((\bar{\lambda}_x, \bar{\lambda}_y)\) classified according to intervals of the Joint Design Effect \((\bar{\lambda})\).

<table>
<thead>
<tr>
<th>(\bar{\lambda})</th>
<th>(a(\lambda))</th>
<th>(\bar{\lambda}_x)</th>
<th>(\bar{\lambda}_y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1.00</td>
<td>32.9(4.0-70.4)</td>
<td>0.97(0.51-2.04)</td>
<td>0.87(0.49-1.83)</td>
</tr>
<tr>
<td>1.00-1.20</td>
<td>47.0(2.1-107.9)</td>
<td>1.24(0.51-3.07)</td>
<td>1.25(0.49-2.97)</td>
</tr>
<tr>
<td>1.20-1.40</td>
<td>57.7(13.4-109.7)</td>
<td>1.56(0.51-4.02)</td>
<td>1.60(0.49-3.93)</td>
</tr>
<tr>
<td>1.40-1.60</td>
<td>61.0(17.1-114.9)</td>
<td>1.87(0.51-4.52)</td>
<td>1.88(0.49-4.58)</td>
</tr>
<tr>
<td>1.60-1.80</td>
<td>60.4(19.7-106.7)</td>
<td>2.15(0.61-4.52)</td>
<td>2.13(0.62-4.58)</td>
</tr>
<tr>
<td>1.80-2.00</td>
<td>58.1(20.5-92.6)</td>
<td>2.40(0.87-4.52)</td>
<td>2.37(0.89-4.58)</td>
</tr>
<tr>
<td>2.00-2.20</td>
<td>54.5(19.7-93.1)</td>
<td>2.62(1.15-4.52)</td>
<td>2.60(1.18-4.58)</td>
</tr>
<tr>
<td>2.20-2.40</td>
<td>50.1(18.8-80.0)</td>
<td>2.82(1.43-4.52)</td>
<td>2.82(1.45-4.58)</td>
</tr>
<tr>
<td>2.40-2.60</td>
<td>45.5(18.1-75.6)</td>
<td>3.02(1.72-4.52)</td>
<td>3.02(1.75-4.58)</td>
</tr>
<tr>
<td>2.60-2.80</td>
<td>40.4(17.0-74.4)</td>
<td>3.20(2.00-4.52)</td>
<td>3.21(2.04-4.58)</td>
</tr>
<tr>
<td>2.80-3.00</td>
<td>34.3(15.2-71.9)</td>
<td>3.37(2.28-4.52)</td>
<td>3.39(2.33-4.58)</td>
</tr>
<tr>
<td>3.00-3.50</td>
<td>26.6(12.1-58.1)</td>
<td>3.61(2.57-4.52)</td>
<td>3.64(2.60-4.58)</td>
</tr>
<tr>
<td>&gt;3.50</td>
<td>16.0(5.6-32.7)</td>
<td>4.03(3.29-4.52)</td>
<td>4.06(3.33-4.58)</td>
</tr>
</tbody>
</table>
Table 3: Average Values of the Joint Generalized Design Effect for Selected Combinations of the Marginal Design Effects.

<table>
<thead>
<tr>
<th></th>
<th>( \tilde{\lambda}_y )</th>
<th>( \tilde{\lambda}_x )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;1.0</td>
<td>1.2-1.4</td>
</tr>
<tr>
<td>( \tilde{\lambda}_y )</td>
<td>0.95</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>0.97-1.26</td>
<td>(0.86-1.07)</td>
</tr>
<tr>
<td></td>
<td>1.28</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>1.07-1.39</td>
<td>(0.92-1.17)</td>
</tr>
<tr>
<td></td>
<td>1.40</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>1.31-1.45</td>
<td>(0.97-1.26)</td>
</tr>
<tr>
<td></td>
<td>1.57</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>1.33-1.60</td>
<td>(1.04-1.12)</td>
</tr>
<tr>
<td></td>
<td>1.91</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>1.56-1.89</td>
<td>(1.12-1.33)</td>
</tr>
<tr>
<td></td>
<td>2.15</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>1.75-2.18</td>
<td>(1.17-1.28)</td>
</tr>
<tr>
<td></td>
<td>2.40</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>1.94-2.46</td>
<td>(1.28-1.42)</td>
</tr>
<tr>
<td></td>
<td>2.77</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>2.23-3.16</td>
<td>(1.55-1.71)</td>
</tr>
<tr>
<td>&gt;3.5</td>
<td>2.77</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>2.54-3.16</td>
<td>(1.94-2.23)</td>
</tr>
<tr>
<td></td>
<td>3.49</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>3.07-4.32</td>
<td>(2.77-3.49)</td>
</tr>
</tbody>
</table>
Table 4: Average Coefficient of Variation of the Joint Generalized Design Effect for Selected Combinations of Marginal Design

<table>
<thead>
<tr>
<th>$\bar{\lambda}_y$</th>
<th>&lt;1.0</th>
<th>1.2-1.4</th>
<th>1.6-1.8</th>
<th>2.0-2.2</th>
<th>2.6-2.8</th>
<th>&gt;3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\lambda}_x$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1</td>
<td>29.9</td>
<td>41.2</td>
<td>54.2</td>
<td>66.2</td>
<td>74.5</td>
<td>85.3</td>
</tr>
<tr>
<td></td>
<td>(4.6-58.4)</td>
<td>(10.9-69.0)</td>
<td>(24.5-79.4)</td>
<td>(37.3-94.1)</td>
<td>(51.1-101.8)</td>
<td>(69.2-114.3)</td>
</tr>
<tr>
<td>1.2-1.4</td>
<td>37.9</td>
<td>42.3</td>
<td>50.1</td>
<td>58.3</td>
<td>61.5</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>(11.5-69.0)</td>
<td>(10.8-74.1)</td>
<td>(18.0-79.8)</td>
<td>(27.3-87.3)</td>
<td>(38.0-85.7)</td>
<td>(52.6-81.9)</td>
</tr>
<tr>
<td>1.6-1.8</td>
<td>51.2</td>
<td>49.0</td>
<td>52.0</td>
<td>56.7</td>
<td>56.6</td>
<td>57.8</td>
</tr>
<tr>
<td></td>
<td>(25.9-82.4)</td>
<td>(17.9-80.4)</td>
<td>(17.1-82.1)</td>
<td>(22.1-89.0)</td>
<td>(29.5-86.0)</td>
<td>(41.6-77.1)</td>
</tr>
<tr>
<td>2.0-2.2</td>
<td>61.3</td>
<td>55.3</td>
<td>55.0</td>
<td>56.9</td>
<td>54.2</td>
<td>54.5</td>
</tr>
<tr>
<td></td>
<td>(37.8-94.9)</td>
<td>(26.2-89.0)</td>
<td>(21.4-89.1)</td>
<td>(19.7-91.4)</td>
<td>(23.1-89.1)</td>
<td>(31.9-76.3)</td>
</tr>
<tr>
<td>2.6-2.8</td>
<td>76.9</td>
<td>64.0</td>
<td>59.2</td>
<td>57.3</td>
<td>49.4</td>
<td>40.1</td>
</tr>
<tr>
<td></td>
<td>(51.9-107.9)</td>
<td>(37.3-86.4)</td>
<td>(28.6-85.3)</td>
<td>(22.6-87.9)</td>
<td>(18.8-79.0)</td>
<td>(21.7-59.9)</td>
</tr>
<tr>
<td>&gt;3.5</td>
<td>85.4</td>
<td>67.3</td>
<td>58.3</td>
<td>52.6</td>
<td>39.7</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>(69.7-114.9)</td>
<td>(52.4-81.3)</td>
<td>(40.8-74.1)</td>
<td>(31.3-74.2)</td>
<td>(21.3-59.7)</td>
<td>(5.6-32.6)</td>
</tr>
<tr>
<td>Case #</td>
<td>$\bar{\lambda}_x$</td>
<td>$\bar{\lambda}_y$</td>
<td>$\bar{\lambda}$</td>
<td>$\lambda(a)$</td>
<td>$\eta_1^{(x)}$</td>
<td>$\eta_2^{(x)}$</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>1.301</td>
<td>1.301</td>
<td>1.165</td>
<td>36.37</td>
<td>0.60</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>1.301</td>
<td>1.302</td>
<td>1.176</td>
<td>51.91</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>2.004</td>
<td>2.002</td>
<td>1.667</td>
<td>31.32</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>2.003</td>
<td>1.997</td>
<td>1.636</td>
<td>76.79</td>
<td>0.85</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>1.190</td>
<td>3.009</td>
<td>1.595</td>
<td>74.70</td>
<td>0.20</td>
<td>0.60</td>
</tr>
</tbody>
</table>

$\bar{\lambda}_x$: mean of the marginal generalized design effects of variable $X$,

$\bar{\lambda}_y$: mean of the marginal generalized design effects of variable $Y$,

$\bar{\lambda}$: mean of the Joint generalized design effects,

$\lambda(a)$: Coefficient of variation of the joint generalized design effects
Table 6: Average values (on 1000 samples) of the estimated parameters(*) corresponding to the theoretical values (TV) in Table 5.

<table>
<thead>
<tr>
<th>Case #1 (TV)</th>
<th>$\bar{\lambda}_x$</th>
<th>$\bar{\lambda}_y$</th>
<th>$\bar{\lambda}$</th>
<th>$\alpha(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.34(0.31)</td>
<td>1.31(0.29)</td>
<td>1.17(0.14)</td>
<td>75.47(8.35)</td>
</tr>
<tr>
<td>30</td>
<td>1.31(0.21)</td>
<td>1.33(0.20)</td>
<td>1.17(0.11)</td>
<td>58.64(6.82)</td>
</tr>
<tr>
<td>70</td>
<td>1.30(0.13)</td>
<td>1.31(0.13)</td>
<td>1.16(0.07)</td>
<td>47.11(4.90)</td>
</tr>
<tr>
<td>100</td>
<td>1.30(0.11)</td>
<td>1.30(0.11)</td>
<td>1.16(0.06)</td>
<td>44.14(4.07)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case #2 (TV)</th>
<th>$\bar{\lambda}_x$</th>
<th>$\bar{\lambda}_y$</th>
<th>$\bar{\lambda}$</th>
<th>$\alpha(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.35(0.32)</td>
<td>1.33(0.29)</td>
<td>1.19(0.15)</td>
<td>85.18(8.74)</td>
</tr>
<tr>
<td>30</td>
<td>1.31(0.21)</td>
<td>1.32(0.20)</td>
<td>1.17(0.11)</td>
<td>71.33(6.32)</td>
</tr>
<tr>
<td>70</td>
<td>1.30(0.14)</td>
<td>1.31(0.13)</td>
<td>1.18(0.07)</td>
<td>62.60(4.47)</td>
</tr>
<tr>
<td>100</td>
<td>1.30(0.11)</td>
<td>1.31(0.11)</td>
<td>1.17(0.06)</td>
<td>60.73(3.74)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case #3 (TV)</th>
<th>$\bar{\lambda}_x$</th>
<th>$\bar{\lambda}_y$</th>
<th>$\bar{\lambda}$</th>
<th>$\alpha(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2.03(0.41)</td>
<td>2.06(0.41)</td>
<td>1.61(0.21)</td>
<td>70.57(8.09)</td>
</tr>
<tr>
<td>30</td>
<td>2.01(0.27)</td>
<td>2.02(0.28)</td>
<td>1.64(0.16)</td>
<td>53.00(6.26)</td>
</tr>
<tr>
<td>70</td>
<td>2.01(0.18)</td>
<td>2.01(0.17)</td>
<td>1.66(0.10)</td>
<td>41.36(4.53)</td>
</tr>
<tr>
<td>100</td>
<td>2.01(0.14)</td>
<td>2.02(0.15)</td>
<td>1.66(0.09)</td>
<td>38.70(3.88)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case #4 (TV)</th>
<th>$\bar{\lambda}_x$</th>
<th>$\bar{\lambda}_y$</th>
<th>$\bar{\lambda}$</th>
<th>$\alpha(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2.09(0.33)</td>
<td>2.05(0.36)</td>
<td>1.61(0.21)</td>
<td>95.31(10.15)</td>
</tr>
<tr>
<td>30</td>
<td>2.03(0.22)</td>
<td>2.05(0.22)</td>
<td>1.63(0.13)</td>
<td>85.07(6.38)</td>
</tr>
<tr>
<td>70</td>
<td>2.02(0.14)</td>
<td>2.02(0.13)</td>
<td>1.64(0.09)</td>
<td>80.18(4.22)</td>
</tr>
<tr>
<td>100</td>
<td>2.01(0.12)</td>
<td>2.00(0.11)</td>
<td>1.63(0.07)</td>
<td>79.08(3.56)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case #5 (TV)</th>
<th>$\bar{\lambda}_x$</th>
<th>$\bar{\lambda}_y$</th>
<th>$\bar{\lambda}$</th>
<th>$\alpha(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.23(0.30)</td>
<td>3.10(0.43)</td>
<td>1.57(0.20)</td>
<td>95.99(10.86)</td>
</tr>
<tr>
<td>30</td>
<td>1.20(0.19)</td>
<td>3.07(0.24)</td>
<td>1.59(0.13)</td>
<td>84.81(7.64)</td>
</tr>
<tr>
<td>70</td>
<td>1.19(0.12)</td>
<td>3.03(0.16)</td>
<td>1.59(0.08)</td>
<td>79.01(5.09)</td>
</tr>
<tr>
<td>100</td>
<td>1.19(0.10)</td>
<td>3.03(0.13)</td>
<td>1.59(0.07)</td>
<td>77.66(3.93)</td>
</tr>
</tbody>
</table>

(*) bracketed numbers are the sample standard deviations of the estimates.