Bayesian Analysis (2016) **11**, Number 2, pp. 519–544

A Mixture Model for Rare and Clustered Populations Under Adaptive Cluster Sampling

Kelly C. M. Gonçalves* and Fernando A. S. Moura†

Abstract. Rare populations, such as endangered species, drug users and individuals infected by rare diseases, tend to cluster in regions. Adaptive cluster designs are generally applied to obtain information from clustered and sparse populations. The aim of this work is to propose a unit-level mixture model for clustered and sparse populations when the data are obtained from an adaptive cluster sample. Our approach considers heterogeneity among units belonging to different clusters. The proposed model is evaluated using simulated data and a real experiment in which adaptive samples were drawn from an enumeration of a waterfowl species in a 5,000 km^2 area of central Florida. The results show that the model is efficient under many settings, even when the level of heterogeneity is low.

Keywords: informative sampling, Poisson mixture, RJMCMC.

1 Introduction

In many research studies, it is difficult to observe individuals or collect information from them, such as in surveys of rare diseases, elusive individuals or unevenly distributed individuals. According to McDonald (2004), rare populations present a few individuals that are sparsely distributed in clusters across a large region. In those cases, the use of conventional sampling methods is not recommended due to the high costs of locating such individuals and the low precision achieved by employing design-based estimators. For instance, suppose that the individuals of interest are spatially distributed in a region upon which we superimpose a regular grid with N cells. Let Y_i denote the grid cell count, for example, the number of endangered plants or animals of interest in the ith grid cell, where $i = 1, \ldots, N$. The objective is to estimate the population total $T = \sum_{i=1}^{N} Y_i$.

Grid cell sampling methods involve the selection of a subset with $n < N$ grid cells and the observation of the Y_is for the selected grid cells. For rare and clustered populations, most of the samples would consist mainly of empty grid cells, yielding poor estimates of T. To overcome this difficulty, Thompson (1990) introduced adaptive cluster sampling as a refined method for estimating the size of rare and clustered populations. The scheme is useful for exploring such populations because it allows sampling efforts to be focused on the surrounding non-empty grid cells in the sample. As stated in Thompson and Seber (1996), adaptive sampling refers to designs in which the procedure for selecting units to include in the sample may depend on the values of the variable of interest observed during the survey. For instance, in a survey to assess the abundance of a rare

*Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal Fluminense (UFF), RJ, Brazil, kelly@est.uff.br

†Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil, fmoura@im.ufrj.br

 DOI: 10.1214/15-BA961

animal species, neighboring sites may be added to the sample whenever the species is encountered during the survey.

An adaptive sampling design begins with an initial probability sample of units, which is selected using a standard sample design. Then, when it has found a non-empty grid cell, it also surveys the neighbors of that cell and continues to survey neighbors of non-empty cells until it obtains a set of contiguous non-empty grid cells surrounded by empty grid cells. Selected empty grid cells attract no additional survey effort. This procedure allows the collection of more useful data than simpler sampling methods that ignore the population structure. Although the initial sample size is known, the final sample size is a random variable and depends on the variable of interest. Therefore, to be effective at moderate cost, this plan requires some prior knowledge of the structure of the underlying population; see Thompson and Seber (1996) for further details.

The final sample consists of empty grid cells that are selected from the initial sample, non-empty grid cells selected in the first stage and their non-empty neighbors. Thompson (1990) refers to the sets of contiguous non-empty grid cells and their neighboring empty grid cells as clusters. The set of contiguous non-empty grid cells within a cluster is called a network. Empty cells are also defined as networks of size one. Thus, the population may be partitioned not only into grid units but also into networks. It should be noted that the final sample consists of a set of a random number of non-empty networks and empty networks.

For the particular case in which the initial sample is a simple random sample without replacement, Thompson (1990) derived inclusion probabilities for the networks observed in the sample and used these probabilities to construct design-unbiased estimators of T and their variances. Although the initial selection is without replacement, the same network can be selected more than once, a problem that Thompson (1990) resolved by allowing multiple inclusions of networks.

The first insight in Thompson (1990) is to base the analysis on networks and to treat the empty edge units of the clusters as unobserved. Following this approach, the estimator obtained does not use all the data information, and therefore it is not a function of the minimal sufficient statistics. Thompson (1990) further suggested the construction of estimators based on the complete set of observations by using the Rao–Blackwell theorem for taking their conditional expectations given the minimal sufficient statistic, which are the complete set of observations.

Adaptive cluster sampling has been performed on real problems and has been shown to be more efficient than traditional grid cell sampling in various areas. For example, Roesch (1993) and Philippi (2005) showed that this method is a viable alternative for sampling forests with rare plants, Smith et al. (1995) evaluated the methodology for rare species of waterfowl, and Conners and Schwager (2002) applied it to hydroacoustic surveys in fisheries.

The first attempt to model data obtained by adaptive cluster sampling and to develop a model-based Bayesian analysis was provided by Rapley and Welsh (2008). The use of the Bayesian framework is a natural extension of the key idea behind adaptive cluster sampling, which incorporates the prior knowledge of a clustered population into

the inference, as well as into the sampling design. Their approach is based on modeling at the network level. They developed a model for the network counts that considers the informativeness of the adaptive cluster sampling design with respect to the number of counts. However, a crucial aspect of their approach is that they do not model the spatial locations of the networks. This lack of spatial information does not entail any loss of information about the total population because, under the model, the population size does not depend on where the networks are located. They thereby address a potentially difficult problem and are able to proceed relatively simply.

Although the formulation developed by Rapley and Welsh (2008) has certain practical advantages, it does not permit the incorporation of more complex structures, such as spatial dependence between units. Their model assumes homogeneity across all units, even those belonging to different networks, which is equivalent to assuming that the expected total in a network is proportional to its size. However, these assumptions might not be realistic in all real situations.

The aim of this work is to propose a unit-level mixture model for clustered and sparse populations when the data are obtained from an adaptive cluster sample. Our proposed mixture model considers heterogeneity among units belonging to different clusters. As in Rapley and Welsh (2008), our model formulation does not include the additional information provided by the edge units. Although, we are violating the sufficiency principle, in our case, the information provided by the edge units is negligible because the criterion used for adaptively sampling the neighboring units of a unit is that it be non-empty, i.e., $Y_i > c = 0$.

The paper is organized as follows. Section 2 presents the proposed model for estimating the population total of rare and clustered populations from samples selected using an adaptive cluster sampling design. It also discusses prior distributions that may be used in this case. The inference developed specifically for fitting the proposed model is discussed in Section 3, where we also assess the convergence of the Markov Chain Monte Carlo (MCMC) methods by applying informal and formal convergence criteria. Section 4 presents a simulation study for assessing the estimation of model parameters under different scenarios. It also presents a prior sensitivity analysis of the two possible prior distributions of the parameter that controls the degree of homogeneity among units belonging to different clusters. A comparison of our approach with that of Rapley and Welsh (2008) through design-based and model-based perspectives under different scenarios is presented in Section 5. Finally, Section 6 presents some conclusions and suggestions for further research.

2 A Poisson mixture model for unit counts

The basic mixture model for independent scalar or vector observations Y_i, $i = 1, \ldots, n$ is given by

$$Y_i \sim \sum_{j=1}^{k} w_j f(\cdot \mid \phi_j),\ i = 1, \ldots, n, \tag{1}$$

where $f(\cdot \mid \boldsymbol{\phi})$ is a given parametric family of densities indexed by a scalar or a vector $\boldsymbol{\phi}$. In general, the objective of the analysis is to make inferences about the unknowns: the number of groups k, the parameters ϕ_js and the components' weights w_j, $0 < w_j < 1$, $\sum_{j=1}^{k} w_j = 1$. The mixture model in (1) is invariant to permutation of the labels $j = 1, \ldots, k$. Therefore, it is important to adopt unique labeling to ensure identifiability. For example, we can impose an ordering constraint on the ϕ_js, such as $\phi_1 < \phi_2 < \cdots < \phi_k$.

Viallefont et al. (2002) suggested a Poisson mixture model for coping with rare events. The interest in this class of models arises here because it is applicable to heterogeneous populations consisting of groups $j = 1, \ldots, k$ of sizes proportional to w_j, from which a random sample may be drawn. The identity of the group from which each observation is drawn is unknown. As stated in Richardson and Green (1997), due to computational costs, it is natural to regard the group label ϵ_i for the ith observation as a latent variable and rewrite (1) as the following hierarchical model

$$Y_i \mid \phi_j, \epsilon_i = j \sim f(\cdot \mid \phi_j), \text{ with } P(\epsilon_i = j) = w_j,\ i = 1, \ldots, n,\ j = 1, \ldots, k.$$

Let us consider a region $\boldsymbol{\Omega}$ containing a sparse, clustered population of size T. We superimpose a regular grid on $\boldsymbol{\Omega}$ to partition it into N squares. A grid cell is non-empty if it contains at least one observation and empty otherwise. Let X be the number of non-empty grid cells in $\boldsymbol{\Omega}$. Let $R \leq X$ be the number of non-empty networks, and let $\mathbf{C} = (C_1, \ldots, C_R)'$ denote the number of non-empty grid cells within each network, such that $X = \sum_{j=1}^{R} C_j$. As there are $N - X$ empty grid cells, which are defined to be empty networks of size one, there are $N - X + R$ networks in $\boldsymbol{\Omega}$. Thus, it is possible to extend the R-vector $\mathbf{C}$ to the vector $\mathbf{Z} = (\mathbf{C}', \mathbf{1}'_{N-X})'$ of dimension $N - X + R$, where $\mathbf{1}'_{N-X}$ is a vector of ones with dimension $N - X$. Let $\mathbf{Y} = (Y_1, \ldots, Y_X)'$ denote a vector of cell counts, where its elements are the number of observations within each non-empty unit; then, $Y_i \geq 1$. The primary goal is to make inferences about the total population $T = \sum_{i=1}^{X} Y_i$.

The proposed mixture model assumes that the R non-empty network mixture components are heterogeneous, with weights w_j that are proportional in each case to the number of grid cells inside the networks C_j. Let us define the latent allocation variable ϵ_i such that $P(\epsilon_i = j) = w_j = C_j/X$, $i = 1, \ldots, X$ and $j = 1, \ldots, R$.

The mixture model is completed with the hierarchical structure proposed in Rapley and Welsh (2008), where they assign distributions to X, R and $\mathbf{C}$ associated with the non-empty grid cells and then, conditional on the network structure, model the network counts $\mathbf{Y}$ for the non-empty networks.

Our proposed model can be stated as follows

$$\begin{aligned}
Y_i \mid \epsilon_i = j, \lambda_j, X &\sim \text{independent truncated Poisson}(\lambda_j),\ Y_i \geq 1, && \text{(2a)}\\
P(\epsilon_i = j) &= w_j = C_j/X,\ i = 1, \ldots, X \text{ and } j = 1, \ldots, R, && \text{(2b)}\\
\mathbf{C} - \mathbf{1}_R \mid X, R &\sim \text{Multinomial}\left(X - R, \frac{1}{R}\mathbf{1}_R\right),\ \sum_{i=1}^{R} C_i = X, && \text{(2c)}
\end{aligned}$$

$$R \mid X, \beta \sim \text{truncated Binomial } (X, \beta),\ R = 1, \ldots, X, \tag{2d}$$
$$X \mid \alpha \sim \text{truncated Binomial } (N, \alpha),\ X = 1, \ldots, N, \tag{2e}$$

where $\lambda_j/\{1 - \exp(-\lambda_j)\}$ is the mean of the truncated Poisson distribution and $\mathbf{1}_R$ is the R-vector of ones. Note that, to avoid degeneracy, at least one non-empty network is assumed to be in the region. Consequently, all of the distributions are left-truncated at one.

The distributions stated in (2c), (2d) and (2e) are the same as in the model by Rapley and Welsh (2008), but unlike their model, the analysis here is performed at the unit level. In the Rapley and Welsh (2008) model, equations (2a) and (2b) are replaced with independent Poisson distributions truncated at zero: $Y_{\cdot j} \mid \lambda, R, \mathbf{C} \sim$ Poisson (λC_j), where $Y_{\cdot j} = \sum_{i \in U_j} Y_i$ with U_j denoting the set of units that belong to the network $j, j = 1, \ldots, R$. Therefore, our model can accommodate heterogeneity between units that belong to different networks, which is not considered in the approached proposed by Rapley and Welsh (2008).

The adaptive design begins with a simple random sample without replacement of fixed size and then follows with the adaptive design, yielding a particular sample $s = \{i_1, \ldots, i_m\}$ of size m taken from $N - X + R$ networks of the population. Thus, the mechanism depends only on the network structure, described by X, R and $\mathbf{C}$, and needs to be included in the model. However, because the same network can be selected more than once in this procedure, we decided to sample networks directly via a sequential procedure in which the ordered sample of networks is selected without replacement. We implement this sampling procedure by selecting a grid cell in the set of N grid cells, surveying that grid cell and, if it is non-empty, then surveying the entire network containing the selected grid cell. We then remove this network from the population, select one of the remaining grid cells and continue in this fashion until we have selected m networks for the sample. Therefore, using this procedure, networks are sampled with probability proportional to their size without replacement. This method was proposed by Salehi and Seber (1997) and is a modification of the sample design proposed in Thompson (1990) in which networks are sampled only once. Thus, in this paper, we considered the sequential method with fixed sample size and not the precise sampling mechanism proposed by Thompson (1990).

Note that the inclusion probability of a network depends on its size Z_i, and the sampling is informative because the components of the random vector $\mathbf{Z}$ are only observed for the sampled networks after being selected. Thus, the probability of selecting the ordered sample $s = \{i_1, \ldots, i_m\}$ of m networks must be included in the model likelihood. The joint inclusion probability can be deduced as follows.

Let the event $A_{i_j} = \{\text{the network } i_j \text{ be selected in the } j\text{th draw}\}$. Thus, the probability of selecting the ordered sample $s = \{i_1, \ldots, i_m\}$ of m networks can be written as follows

$$\begin{aligned} p(s \mid X, R, \mathbf{C}) = P(\cap_{j=1}^{m} A_{i_j} \mid X, R, \mathbf{C}) &= P(A_{i_1} \mid X, R, \mathbf{C}) \\ &\times \prod_{j=2}^{m} P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C}). \end{aligned} \tag{3}$$

Because the networks are sampled without replacement, the conditional probabilities $P(A_{i_1} \mid X, R, \mathbf{C})$ and $P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C})$ in (3) are, respectively, given by

$$\begin{aligned} P(A_{i_1} \mid X, R, \mathbf{C}) &= \frac{z_{i_1} \times g_{i_1,1}}{\sum_{i=1}^{N-X+R} z_i - z_{i_0}}, \\ P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C}) &= \frac{z_{i_j} \times g_{i_j,j}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}}, \quad j = 2, \ldots, m, \end{aligned} \tag{4}$$

where $g_{i_j,j}$ is the number of unselected networks of size z_{i_j} after $j-1$ networks have been selected and $z_{i_0} = 0$.

Substituting the equations in (4) into (3), we finally have

$$p(s \mid X, R, \mathbf{C}) = \prod_{j=1}^{m} \frac{z_{i_j} \times g_{i_j,j}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}}. \tag{5}$$

The sampling procedure entails observing Y_i for the networks in the sample s. The input variables are split into an observed and an unobserved component, using the subscripts s and $\bar{s}$, respectively. Thus, we have $X = X_s + X_{\bar{s}}$, $R = R_s + R_{\bar{s}}$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_s, \boldsymbol{\epsilon}'_{\bar{s}})'$, $\mathbf{C} = (\mathbf{C}'_s, \mathbf{C}'_{\bar{s}})'$ and $\mathbf{Y} = (\mathbf{Y}'_s, \mathbf{Y}'_{\bar{s}})'$.

Because the sampling procedure is informative, it is useful to divide the joint probability model into two parts: the model for the underlying complete data, including both observed and unobserved components, and the model for the inclusion probability vector, as stated in (5); see Pfeffermann et al. (2006) for further explanation. The complete-data likelihood is defined as the product of these two factors, as stated by Gelman et al. (1995). Thus, we can write the complete-data likelihood as

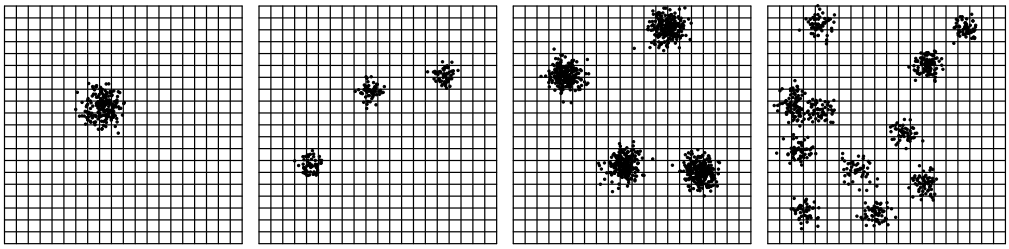
$$\begin{aligned} p(\{i_1, \ldots, i_m\}, X, R, \boldsymbol{\epsilon}, \mathbf{C}, \mathbf{Y} \mid \boldsymbol{\lambda}, \alpha, \beta) &= p(\{i_1, \ldots, i_m\} \mid X, R, \mathbf{C}) p(\mathbf{Y} \mid \boldsymbol{\epsilon}, \boldsymbol{\lambda}, X) \\ &\quad \times p(\boldsymbol{\epsilon} \mid \mathbf{C}, R, X) p(\mathbf{C} \mid R, X) p(R \mid X, \beta) p(X \mid \alpha) \\ &= \prod_{l=1}^{m} \frac{z_{i_l} \times g_{i_l,l}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}} \times \prod_{j=1}^{R_s+R_{\bar{s}}} \prod_{\{i:\epsilon_i=j\}} \frac{\lambda_j^{y_i} \exp(-\lambda_j)}{y_i![1-\exp(-\lambda_j)]} \\ &\quad \times \frac{1}{(X_s + X_{\bar{s}})^{X_s+X_{\bar{s}}}} \prod_{j=1}^{R_s+R_{\bar{s}}} C_j^{C_j} \times \prod_{j=1}^{R_s+R_{\bar{s}}} \frac{1}{(C_j-1)!} \left(\frac{1}{R_s+R_{\bar{s}}}\right)^{C_j-1} \\ &\quad \frac{1}{(R_s+R_{\bar{s}})!} \frac{\beta^{R_s+R_{\bar{s}}}(1-\beta)^{X_s+X_{\bar{s}}-R_s-R_{\bar{s}}}}{1-(1-\beta)^{X_s+X_{\bar{s}}}} \times N! \frac{\alpha^{X_s+X_{\bar{s}}}(1-\alpha)^{N-X_s-X_{\bar{s}}}}{1-(1-\alpha)^N}. \end{aligned} \tag{6}$$

It should be noted that expression (6) is useful for specifying a probability model but does not present the actual likelihood of the data unless the variables are completely observed. The appropriate likelihood of Bayesian inference for the actual information available is obtained by summing over the unknown quantities, which are not otherwise observed in the selected sample. The observed-data likelihood, conditional on $\boldsymbol{\lambda}$, α and β, is given by

$$p(\{i_1, \ldots, i_m\}, X_s, R_s, \boldsymbol{\epsilon}_s, \mathbf{C}_s, \mathbf{Y}_s) = \sum_{\mathbf{Y}_{\bar{s}}} \sum_{\mathbf{C}_{\bar{s}}} \sum_{\boldsymbol{\epsilon}_{\bar{s}}} \sum_{R_{\bar{s}}} \sum_{X_{\bar{s}}} p(\{i_1, \ldots, i_m\}, X, R, \boldsymbol{\epsilon}, \mathbf{C}, \mathbf{Y}).$$

2.1 Prior distributions

In a Bayesian framework, the three unknowns α, β and $\boldsymbol{\lambda}$ are regarded as having been drawn from appropriate prior distributions. Assume that these parameters are independent; then, the joint prior distribution of $(\alpha, \beta, \boldsymbol{\lambda})$ is the product of their marginal prior distributions, described here. Parameter α controls the expected number of non-empty grid cells, and β controls the conditional expected number of non-empty networks. Figure 1 presents an illustration with certain artificial populations generated by model (2) and certain values fixed for α and β. We also arbitrarily fixed $\lambda_j = 10$ for all $j = 1, \ldots, R$; thus, approximately 10 observations are expected in each unit. Because model (2) does not provide information on the locations of the networks or on their shapes, we used model (2) to generate X, R and $\mathbf{Y}$, and conditional on these values, we then sampled from the Poisson cluster process (cf. Diggle (2014), chap. 6, p. 101) to generate the locations. Specifically, the R "parent locations" were generated from a uniform distribution and the numbers of point-objects $Y_{\cdot j}$ were dispersed in relation to the parent locations with a symmetric Gaussian distribution with a fixed standard deviation of 0.6.



(a) $(\alpha, \beta) = (0.05, 0.05)$ (b) $(\alpha, \beta) = (0.05, 0.20)$ (c) $(\alpha, \beta) = (0.20, 0.05)$ (d) $(\alpha, \beta) = (0.20, 0.20)$

Figure 1: Artificial populations generated by the proposed model with some fixed values for α and β, and $\lambda_j = 10$, for all $j = 1, \ldots, R$, in a regular grid with $N = 400$ units.

Because the aim of our approach is to survey sparse populations, when analyzing Figure 1, it is reasonable to assume that both α and β parameters should typically be small. To be uninformative with respect to these parameters, we should choose flat prior distributions. However, we can assign prior distributions that incorporate our knowledge of a rare and clustered population. In particular, we can assume that $\alpha \sim$ Beta(a_α, b_α) and $\beta \sim$ Beta(a_β, b_β) and choose values for the beta distribution's parameters such that α and β are within an interval centered on a small value with high probability. Here, $W \sim \text{Beta}(a, b)$ generically denotes a beta-distributed random variable parameterized with mean $a/(a+b)$ and variance $ab(a+b+1)^{-1}(a+b)^{-2}$.

To ensure identifiability, it is necessary to adopt a unique labeling. For the proposed model in (2), unique labeling can be achieved by imposing a restriction on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_R)'$. However, it should be noted that $\boldsymbol{\lambda}$, although entirely unknown, has components associated with the sample whereby better estimates are expected. Thus, let us define $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_s, \boldsymbol{\lambda}_{\bar{s}})'$ such that $\boldsymbol{\lambda}_s$ refers to the networks observed in the sample and $\boldsymbol{\lambda}_{\bar{s}}$ refers to the unobserved networks. Note that it is necessary to impose a

restriction on $\boldsymbol{\lambda}$ to ensure the identifiability of the model. Nevertheless, this restriction is only necessary for the elements of $\boldsymbol{\lambda}$ associated with the unknown networks, i.e., $\boldsymbol{\lambda}_{\bar{s}}$.

Let us assume the following for $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda} \mid \boldsymbol{\theta} \sim p(\cdot \mid \boldsymbol{\theta}, R) \text{ such that } \lambda_j < \lambda_{j+1}, \quad \forall j \in [R_s + 1, R_s + R_{\bar{s}}),$$

where $p(\cdot \mid \boldsymbol{\theta}, R)$ represents the prior distribution of $\boldsymbol{\lambda}$, which depends on the number of networks in the population R and on the vector of the hyperparameters $\boldsymbol{\theta}$.

We use two different prior distributions for $\boldsymbol{\lambda}$. First, we assume that the joint prior density for $\boldsymbol{\lambda}$ is given by the following

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\theta}, R) = R_{\bar{s}}! \prod_{j=1}^{R} p(\lambda_j \mid \boldsymbol{\theta}) \text{ such that } \lambda_j < \lambda_{j+1}, \quad \forall j \in [R_s + 1, R_s + R_{\bar{s}}). \tag{7}$$

In particular, we consider $p(\lambda_j|\boldsymbol{\theta})$, $j = 1, \ldots, R$ all equal to the density of Gamma(d, ν), with $\boldsymbol{\theta} = (d, \nu)$, and we introduce an additional hierarchical level by allowing ν to follow Gamma(e, f) distribution. Here, Gamma(a, b) generically denotes a gamma distribution with mean a/b and variance a/b^2.

A standard approach for setting a gamma as a weakly informative prior is to choose small values for its two parameters. However, such a distribution has a peak in the neighborhood of zero, which might encourage the inclusion of components with very small Poisson parameters, which would be difficult to estimate in general. Therefore, we used a weakly informative prior based on Viallefont et al. (2002), i.e., Gamma(d, ν) with d greater than one, to avoid an exponential shape without overly reducing the distribution of the Coefficient of Variation (CV). Ideally, the parameter ν should be set based on prior information. Nevertheless, Viallefont et al. (2002) suggested that the prior mean d/ν be equal to the midrange of the observed data. However, in our case, we also consider ν to be unknown, and hence we choose e and f in the prior of ν such that the approximation to the mean of λ_j, $d/(e/f)$, is equal to the midrange of the observed data and that the variance e/f^2 is relatively small. It should be noted that in the case of samples with only empty units, we needed to add a small constant to the midrange of the observed data to ensure that it be positive.

The other prior considered for $\boldsymbol{\lambda}$ is that introduced by Roeder and Wasserman (1997) for normal mixtures as an explicit way to place an informative prior on the distance between two consecutive λ_js. Here, the hyperparameter $\boldsymbol{\theta}$ is τ, a positive constant, and the prior model is given by the following

$$p(\boldsymbol{\lambda} \mid \tau, R) = p(\lambda_R \mid \lambda_{R-1}, \tau) p(\lambda_{R-1} \mid \lambda_{R-2}, \tau) \cdots p(\lambda_1), \tag{8}$$

where $p(\lambda_j \mid \lambda_{j-1}, \tau)$ is $\mathrm{N}_{(\lambda_{j-1}, \infty)}(\lambda_{j-1}, \tau)$, i.e., a normal centered at λ_{j-1} with variance τ^2, truncated to be greater than λ_{j-1} and $p(\lambda_1) \propto 1$. This ordering ensures the identifiability of the model.

Viallefont et al. (2002) illustrated the difficulty of eliciting τ and its clear influence on the posterior distribution of the mixture parameters, as well as on the posterior

distribution of the number of components. For example, if τ is very small compared with the anticipated distance between two consecutive λ_js, there will be a tendency to fit intermediate components between the true ones and hence to find a posterior distribution favoring higher values of R. This strategy yields a low prior probability that any two neighboring components are more than τ standard deviations apart. Based on a simulation study, Roeder and Wasserman (1997) recommend choosing $\tau = 5$ because this choice leads to reasonable density estimates.

3 Inference

The posterior distribution of the parametric vector $\boldsymbol{\Theta} = (X_{\bar{s}}, R_{\bar{s}}, \boldsymbol{\epsilon}_{\bar{s}}, \mathbf{C}_{\bar{s}}, \mathbf{Y}_{\bar{s}}, \alpha, \beta, \boldsymbol{\lambda}, \nu)$ of model (2) cannot be obtained in closed form. Therefore, it is necessary to use some numerical approximation methods. One alternative, which is often used and is feasible to implement, is to generate samples from the marginal distributions of the parameters based on the MCMC algorithm. Nevertheless, this method, as originally formulated, requires the posterior distribution to have a density with respect to some fixed measure. Thus, it cannot be used alone in this case, where the size of the parametric space is also a parameter. We use an approach based on reversible jump MCMC (RJMCMC), which was first proposed in Green (1995) and applied in mixture models with unknown numbers of components by Richardson and Green (1997). The method essentially consists of jumps between the parameter subspaces corresponding to different numbers of components in the mixture.

For the proposed model (2), we used the steps specified below:

(i) update the parameters α, β, $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$;

(ii) update the unobserved variables $X_{\bar{s}}$, $\mathbf{C}_{\bar{s}}$ and $\mathbf{Y}_{\bar{s}}$;

(iii) update the allocation $\boldsymbol{\epsilon}_{\bar{s}}$ such that $\mathbf{C}_{\bar{s}}$ is also updated; and

(iv) combine two networks into one, or split one network into two.

Steps (i)–(iii) are performed using the Gibbs sampler or a Metropolis–Hastings sampler, and they do not change the dimensions of $\boldsymbol{\Theta}$. Note that, because the proposed model (2) is defined only for the non-empty units, it is not possible to update the allocation, thus resulting in networks without any observations. Consequently, this step needs to be restricted such that each network must have at least one observation. The full conditional distributions of the parameters are in Appendix A.

Step (iv) involves changing $R_{\bar{s}}$ by 1 and making the necessary corresponding changes to $(\boldsymbol{\lambda}, \mathbf{C}, \boldsymbol{\epsilon})$. We made a random choice between splitting and combining, with probabilities depending on $R_{\bar{s}}$. Let $\lambda'_j = \lambda_j / \{1 - \exp(-\lambda_j)\}$ be the mean of the truncated Poisson distribution. The combination proposal begins by choosing a pair of components (j_1, j_2) at random, such that $\lambda'_{j_1} < \lambda'_{j_2}$. These two components are merged, forming a new component j^*. Now, we have to reallocate all of the observations with $\epsilon_i = j_1$ or $\epsilon_i = j_2$ and create values for $(w_{j^*}, \lambda'_{j^*})$. They are chosen such that

$$w_{j^*} = w_{j_1} + w_{j_2},$$
$$w_{j^*}\lambda'_{j^*} = w_{j_1}\lambda'_{j_1} + w_{j_2}\lambda'_{j_2},$$

and we must impose $\lambda'_{j-1} < \lambda'_{j_1} < \lambda'_{j_2} < \lambda'_{j+1}$. A component j^* is chosen at random and split into j_1 and j_2. However, there are two degrees of freedom for achieving this step, and hence we need to generate a two-dimensional random vector $\mathbf{u} = (u_1, u_2)$ to specify the new parameters. Viallefont et al. (2002) presented some ways of proposing a split that enforces the positivity constraint on Poisson parameters. In this work, we used the one referenced as "SM2" in their paper. In particular, the proposed model (2) is applicable to non-empty networks; thus, the split proposal also requires that both networks have at least one observation. Therefore, networks with only one observation cannot be chosen to be split. The acceptance probability for the split and combination steps can also be found in Appendix A.

Although the expression above can be written in terms of λ'_j, the likelihood is expressed in terms of λ_j. Therefore, after step (iv), we need to obtain λ_j from λ'_j by solving the equation $\lambda'_j = \lambda_j/\{1 - \exp(-\lambda_j)\}$. Furthermore, although the target function is invertible, it involves a polynomial with an exponential function for which, in general, it is impossible to obtain an exact analytical solution. When the value of λ_j is sufficiently large, we can approximate λ_j by λ'_j (see Figure 2). However, for cases in which this approximation is not good, we need to use a numerical approximation, such as the Taylor approximation.

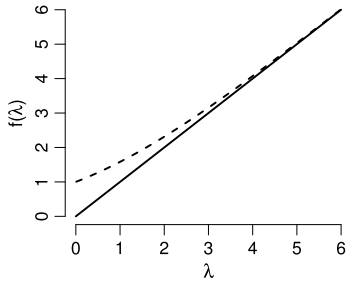


Figure 2: Comparison of the first-order moments of the Poisson distribution (—) and of the Poisson distribution truncated at zero (- - -).

4 Simulation study

To assess the convergence of the RJMCMC estimation, we generated only one population in an area with $N = 400$ units. See Appendix B for details concerning the convergence diagnostic study. The results obtained from this study indicate that the MCMC chains have converged.

To examine the performance of the Bayesian estimator and the influence of the different prior models on the Poisson parameters, we sampled several simulated clustered populations and obtained samples from the posterior distributions of the model param-

eters and population parameters. The population estimates were then compared with the true values to evaluate the model's performance.

4.1 Simulation scenarios

We generated 500 populations for each scenario that we considered. Twelve scenarios were created by varying the values of N, R and X as well as by varying the $\boldsymbol{\lambda}$ components. The values of parameters (α, β) were fixed such that their combinations expressed different degrees of rare and clustered populations. For the first simulation study, we considered only the gamma prior for $\boldsymbol{\lambda}$; thus, we generated the values of the components of $\boldsymbol{\lambda}$ as a gamma distribution with $d = 1.1$ and $\nu = 0.13$. These values of d and ν ensure that the generated populations provide heterogeneous networks. Finally, an adaptive cluster sample of size $m = 0.05 \times N$ was selected from each population without replacement.

Table 1 presents summary statistics with some frequentist measures of the posterior distributions of the model parameters after reaching convergence for each of the twelve evaluated scenarios. Table 1 reports the relative mean square error (RMSE), the relative absolute error (RAE), the empirical nominal coverage of the 95% highest posterior density (HPD) intervals measured in percentages (Cov.) and the respective widths averaged over the 500 simulations (Wid.). Specifically, the summary statistics of the components of the vector $\boldsymbol{\lambda}$ are calculated separately for the observed and non-observed networks. To facilitate future comparisons, the widths presented for the total T and for the components of the vectors $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_{\bar{s}}$ are expressed in ratio form relative to their true values.

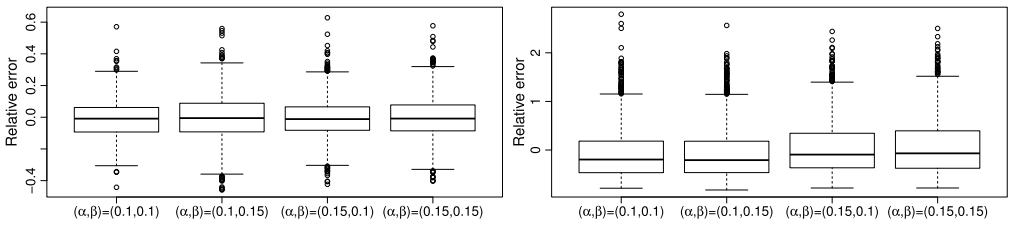
Given the difficulty in making inference on a rare and clustered population, generally, the parameters are not very badly estimated, except for the cases when $N = 200$, $(\alpha, \beta) = (0.10, 0.10)$ and $N = 200$, $(\alpha, \beta) = (0.10, 0.15)$. This is not an unexpected result because these two scenarios typically characterize an extremely small, rare and clustered population and thus reliable estimation might be not achieved. This problem may be overcome by increasing the initial sample size. The coverage of the 95% HPD intervals is close to the nominal level. The less rare and clustered the population is, the narrower the 95% HPD interval is. As expected, the results for λ_j obtained with the samples containing the network j show smaller errors and are more precise than the results that consider the samples in which the network j was not observed. As the value of N increases, the RMSEs and RAEs of most of the parameters decrease. This phenomenon may occur because the number of non-empty networks increases with N, improving the estimates of α and β and consequently of the other parameters. However, for the same reason, for a fixed value of N, the errors decrease as the values of α and β increase.

It is not possible to present the frequentist properties for each λ_j because the value of R was not fixed over the simulations. Figure 3 presents the relative errors of $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_{\bar{s}}$ for all of the networks and all of the simulations for different values of α and β and for $N = 400$. Note that, in all cases, the relative error is approximately zero and is smaller for $\boldsymbol{\lambda}_s$, as expected. Additionally, $\boldsymbol{\lambda}_{\bar{s}}$ is slightly underestimated.

	$N = 200$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$
RMSE	0.21	0.38	0.53	0.56	0.03	0.29	0.22	0.29	0.29	0.39	0.03	0.28
RAE	0.35	0.17	0.25	0.60	0.12	0.46	0.36	0.16	0.35	0.47	0.13	0.45
Cov.	95.0	91.1	96.7	89.5	91.7	87.8	93.8	93.7	98.1	89.7	90.3	87.7
Wid.	1.60	0.20	0.31	0.28	0.58	1.23	1.60	0.19	0.31	0.28	0.57	1.26
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.09	0.20	0.50	0.22	0.02	0.31	0.06	0.10	0.19	0.32	0.02	0.27
RAE	0.24	0.31	0.45	0.40	0.11	0.46	0.21	0.27	0.21	0.47	0.10	0.41
Cov.	94.6	90.9	97.1	90.2	93.6	89.1	97.3	97.0	98.5	90.5	94.1	89.8
Wid.	1.22	0.19	0.21	0.22	0.50	1.33	1.24	0.20	0.23	0.21	0.56	1.51
	$N = 400$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$
RMSE	0.06	0.15	0.42	0.14	0.02	0.29	0.05	0.08	0.15	0.10	0.02	0.31
RAE	0.21	0.32	0.35	0.28	0.10	0.43	0.20	0.23	0.29	0.21	0.12	0.43
Cov.	96.7	91.1	96.0	90.8	94.2	91.0	96.8	95.1	98.1	90.5	94.3	91.8
Wid.	1.04	0.09	0.20	0.19	0.47	1.38	1.05	0.10	0.21	0.18	0.55	1.64
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.04	0.06	0.35	0.04	0.02	0.30	0.05	0.03	0.15	0.03	0.02	0.36
RAE	0.18	0.18	0.39	0.18	0.09	0.42	0.20	0.15	0.21	0.15	0.10	0.43
Cov.	93.4	91.2	96.9	96.7	94.2	93.9	92.4	97.0	98.7	96.5	93.5	95.6
Wid.	0.79	0.11	0.15	0.14	0.45	1.43	0.77	0.11	0.16	0.13	0.51	1.77
	$N = 600$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$
RMSE	0.04	0.05	0.25	0.10	0.02	0.32	0.05	0.03	0.11	0.09	0.02	0.35
RAE	0.17	0.17	0.28	0.12	0.09	0.42	0.20	0.14	0.26	0.11	0.11	0.42
Cov.	96.3	91.8	98.1	98.0	93.5	93.1	92.8	97.5	98.3	97.0	93.8	96.1
Wid.	0.79	0.08	0.22	0.20	0.46	1.40	0.78	0.08	0.23	0.19	0.52	1.70
	$(\alpha, \beta) = (0.15, 0.10)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.05	0.04	0.21	0.06	0.01	0.37	0.09	0.08	0.06	0.05	0.02	0.35
RAE	0.19	0.17	0.30	0.09	0.09	0.44	0.29	0.24	0.18	0.09	0.10	0.43
Cov.	90.4	91.1	98.7	98.9	95.3	96.0	90.0	90.5	98.8	98.4	95.5	96.8
Wid.	0.78	0.08	0.17	0.18	0.43	1.49	0.53	0.08	0.20	0.17	0.53	1.79

Table 1: Summary measurements of the point and 95% HPD interval estimates of the model and population parameters over 500 simulations for different values of α, β and N. The reported results for $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_{\bar{s}}$ are obtained by averaging over the sampled and not sampled λ_js, respectively.

The 500 populations were previously generated by fixing the parameters of λ_j's gamma distribution at $d = 1.1$ and $\nu = 0.13$, yielding a mean of 8.5 and a CV of 95%. The aim was to evaluate the performance of the proposed model with respect to the



(a) Relative error - $\boldsymbol{\lambda}_s$ (b) Relative error - $\boldsymbol{\lambda}_{\bar{s}}$

Figure 3: Relative errors for $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_{\bar{s}}$ over 500 simulations, for $N = 400$ and for different values of α and β.

level of homogeneity. We considered two further CV values, 25% and 50%, with the means fixed at 8.5 for both. Then, we calculated the two sets of values of d and ν. When the CV was fixed at 50%, we obtained $d = 4$ and $\nu = 0.47$; when the CV equaled 25%, the result was $d = 16$ and $\nu = 1.89$.

Figure 4 displays the densities of λ_j for each fixed value of the CV. Note that, as the CV decreases, the prior distribution for λ_j becomes more concentrated and symmetrical around the mean of the distribution; consequently, the networks will become more homogeneous with respect to the total in their units.

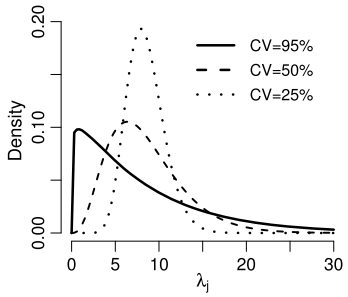


Figure 4: Prior distributions to λ_j used in the simulations obtained by varying the value of the CV of the distribution.

We generated two other sets of populations by fixing the CVs of the λ_j distributions at 50% and 25%, respectively. The population size was set at $N = 400$, and a $m = 5\%N$ adaptive sample was taken from it.

Table 2 presents summary measurements of the estimators over the 500 populations generated for the two values considered for the CV. Note that, even for the more homogeneous cases, the proposed model (2) has reasonable performance, resulting in parameter estimates with relatively small errors and 95% HPD intervals with coverage probabilities near the fixed nominal levels. The less rare and clustered the population is, the more precisely estimated the model parameters are.

	$CV = 50\%$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$	T	α	β	ν	$\boldsymbol{\lambda}_s$	$\boldsymbol{\lambda}_{\bar{s}}$
RMSE	0.13	0.15	0.52	0.16	0.02	0.04	0.06	0.09	0.18	0.10	0.02	0.03
RAE	0.26	0.32	0.27	0.30	0.10	0.15	0.18	0.24	0.36	0.23	0.11	0.15
Cov.	95.3	87.2	97.0	95.3	94.7	97.0	96.7	95.0	98.2	95.0	94.5	97.6
Wid.	1.38	0.11	0.26	0.91	0.51	1.27	1.24	0.11	0.27	0.82	0.55	1.31
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.03	0.04	0.40	0.08	0.02	0.03	0.03	0.03	0.10	0.06	0.02	0.03
RAE	0.15	0.15	0.50	0.21	0.10	0.12	0.16	0.14	0.26	0.18	0.10	0.13
Cov.	96.5	94.7	97.3	97.8	95.6	98.0	95.8	97.3	98.0	97.5	95.8	97.9
Wid.	0.95	0.11	0.23	0.75	0.48	1.28	0.92	0.11	0.24	0.70	0.53	1.36
	$CV = 25\%$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
RMSE	0.09	0.30	0.50	0.36	0.03	0.08	0.05	0.18	0.12	0.34	0.03	0.08
RAE	0.23	0.48	0.37	0.47	0.13	0.24	0.19	0.37	0.29	0.44	0.14	0.26
Cov.	89.7	86.8	98.0	75.0	85.7	82.2	94.7	90.1	98.2	74.9	85.7	81.0
Wid.	0.96	0.12	0.25	3.01	0.47	0.70	0.91	0.12	0.27	2.83	0.51	0.75
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.03	0.08	0.41	0.25	0.02	0.03	0.04	0.05	0.07	0.19	0.02	0.04
RAE	0.14	0.22	0.49	0.34	0.10	0.15	0.17	0.15	0.21	0.24	0.11	0.17
Cov.	96.6	91.7	97.5	80.8	94.6	94.4	91.9	92.5	98.3	83.2	93.3	94.8
Wid.	0.70	0.12	0.22	2.48	0.46	0.74	0.70	0.12	0.23	2.25	0.50	0.79

Table 2: Summary measurements of the point and 95% HPD interval estimates of the model parameters over 500 simulations obtained by varying the level of homogeneity in $\boldsymbol{\lambda}$ and expressed as the CV of its distribution, for $N = 400$. The reported results for $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_{\bar{s}}$ are obtained by averaging over the sampled and not sampled λ_js, respectively.

The relative errors of T do not vary significantly with the values of the CV, except when $(\alpha, \beta) = (0.10, 0.10)$, for which, on average, smaller numbers of non-empty networks in the generated populations are found. In addition, the relative errors for $\boldsymbol{\lambda}_{\bar{s}}$ are smaller than the errors obtained when the CV is fixed at 95%, although the errors for ν become larger. Furthermore, as the CV decreases, the empirical coverage of the nominal 95% HPD intervals is underestimated, primarily with respect to ν and $\boldsymbol{\lambda}$.

4.2 Prior sensitivity analysis

In this section, we compare the performance of the two prior distributions considered for $\boldsymbol{\lambda}$. To obtain simulation results for each component λ_j of $\boldsymbol{\lambda}$ using a different method from the previous section, the values of R were fixed. The population size was set at $N = 400$ with $(\alpha, \beta) = (0.15, 0.10)$. These settings were chosen to provide populations that are as rare and clustered as possible. Then, we conducted a large number of simulations until we reached 500 populations with $R = 5$; another 500 populations were generated with $R = 6$, followed by another 500 populations with $R = 7$. We considered only these

values of R because the others had much lower probabilities of being generated in this simulation scenario with $(\alpha, \beta) = (0.15, 0.10)$. Furthermore, because we were specifying two different priors for $\boldsymbol{\lambda}$, we fixed the $\boldsymbol{\lambda}$'s components at $(4.5, 4.8, 8.0, 11.3, 13.8)$ for $R = 5$, at $(3.9, 6.4, 6.9, 7.1, 10.5, 14.8)$ for $R = 6$ and at $(4.8, 7.4, 9.5, 10.1, 11.4, 11.7, 14.5)$ for $R = 7$. These values were generated from a uniform distribution defined on the interval $(3, 15)$.

All of the results shown hereafter correspond to 100,000 RJMCMC sweeps, after a burn-in of 10,000 iterations; the chain was then thinned by taking every 10th sample value. We used the same prior distribution for α and β described in the previous section. For $\boldsymbol{\lambda}$, we considered the gamma prior distribution used in the previous simulation study and the prior $\lambda_j \mid \lambda_{j-1} \sim \mathrm{N}_{(\lambda_{j-1},\infty)}(\lambda_{j-1}, \tau)$ with $\tau \in \{1, 5, 10, 20\}$.

Figure 5 depicts the 95% HPD interval obtained for R for each $\boldsymbol{\lambda}$ prior assumed when we fit the model for one of the 500 populations generated. The parameter R was much more sensitive to the value of τ assigned for the normal prior given in (8). In addition, the R posterior distribution was relatively vague when $\tau = 1$. However, as τ increased, this behavior was attenuated.

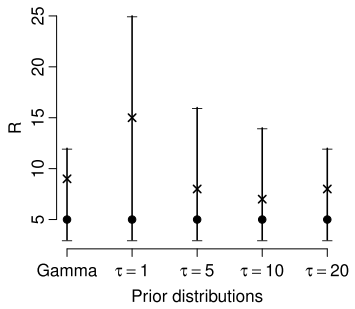
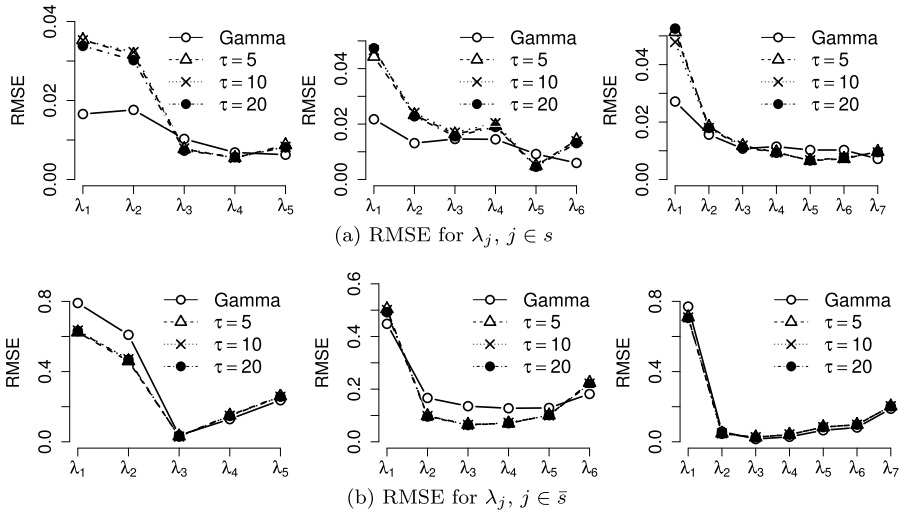


Figure 5: The 95% HPD interval of R for different prior distributions of $\boldsymbol{\lambda}$. The crosses represent the median of the distribution, the circle represents the true value of R, and the line represents the 95% HPD interval.

The gamma prior and the normal one with $\tau = 20$ yielded approximately the same 95% HPD interval for R. This behavior was observed for nearly all of the 500 simulation samples. Thus, henceforth, we do not consider the normal $\boldsymbol{\lambda}$ prior with $\tau = 1$.

Figure 6 presents the RMSE for each λ_j displayed for samples when the network j is observed (a) and when it is not (b) for the four $\boldsymbol{\lambda}$ priors employed. Figure 6 shows that the gamma prior provides a smaller RMSE than the normal prior in most cases, notably for the smaller λ_js. These results do not depend heavily on the values of τ. As expected, the RMSE values of the λ_js for non-sampled networks ($j \notin s$) are greater than the RMSE values of the λ_js for sample networks ($j \in s$).

Because total population prediction is the main aim in this context, we also evaluated the impact of those prior distributions on the posterior distribution of T. Figure 7



(a) RMSE for λ_j, $j \in s$

(b) RMSE for λ_j, $j \in \bar{s}$

Figure 6: RMSE of each λ_j assuming different priors for $\boldsymbol{\lambda}$. The results with the gamma prior distribution and the normal prior distribution with $\tau = 5$, $\tau = 10$ and $\tau = 20$ are represented by the empty circles and the line, the triangles and the dashed line, the cross and the dotted line, and the full circle and the dot-dashed line, respectively.

displays the RMSE of T, the nominal coverage of the 95% HPD interval and its respective width for each considered value of R. We can observe from Figure 7 that the RMSEs obtained using the gamma $\boldsymbol{\lambda}$ prior are always smaller than the RMSEs obtained using the normal $\boldsymbol{\lambda}$ prior. However, the 95% HPD intervals based on the normal $\boldsymbol{\lambda}$ prior have higher coverage than the nominal level and higher width than when using the gamma $\boldsymbol{\lambda}$ prior. Note that for a fixed value of R, the results provided by the normal prior are very similar for all values of τ.

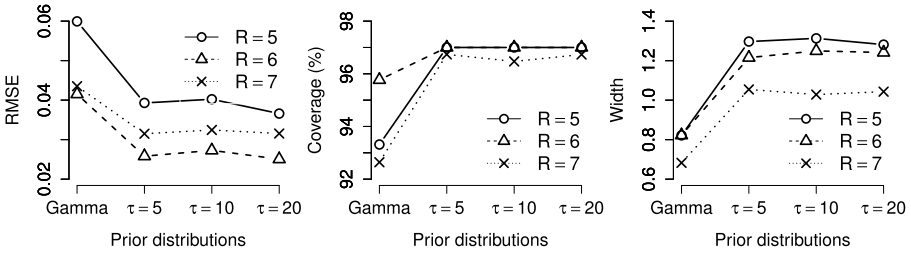


Figure 7: RMSEs, coverages and widths of the 95% HPD intervals for the population total T for each prior distribution assumed for $\boldsymbol{\lambda}$ and for each R fixed. The results for $R = 5$, $R = 6$ and $R = 7$ are represented by the empty circles and the line, the triangles and the dashed line, and the cross and the dotted line, respectively.

5 Comparison with the network model

The mixture model (2) has been presented as an alternative to the model of Rapley and Welsh (2008). The mixture model (2) is principally useful when we cannot assume homogeneity between networks with respect to the number of observations inside them and when the expected number of observations inside any network is not proportional to its respective area size. The key idea of this paper is to improve on the population estimates obtained by Rapley and Welsh (2008) through the use of a model that accounts for heterogeneity between networks. This is accomplished by modeling at the unit level rather than at the network level.

To assess the effectiveness of our methodology, we compared the results of our approach to those obtained in Rapley and Welsh (2008). The first comparison consists of a design-based experiment with a real population, and the second study is a model-based experiment. To fit both models, we assigned the same prior distributions used in Subsection 4.1. To conduct the MCMC and RJMCMC simulations, we generated two chains of 100,000 iterations each, discarded the first 10,000 and then thinned the chain by taking every 90th sample value to obtain 1,000 independent samples.

5.1 A design-based experiment

We evaluated the proposed model (2) by performing a design-based experiment in which adaptive samples were drawn from a real, fixed population. Design-based studies are used in the context of survey sampling inference to evaluate the performance of model-based estimators under repeated samples taken from a real, fixed population where a characteristic of interest is known for all its units. This real population can be a census or a large sample that is assumed, for evaluation purposes, to be the population. The main aim of this design-based experiment is to analyze the frequentist properties of the total estimators using both approaches.

The population used here for design-based evaluation is the same as that described in Smith et al. (1995) and consists of counts of a waterfowl species, called the blue-winged teal, in a 5,000 km^2 area of central Florida in 1992. Figure 8 shows the counts of blue-winged teals in a grid with $N = 200$ units. Note that these counts are sparse and clustered, justifying the use of adaptive sampling.

The study consists of 600 replications of a 10% adaptive sample of the population. Because the blue-winged teal population is extremely sparse and clustered, there are many samples that consist either of only empty networks or networks with only one unit each. These samples are expected to be of limited use in accurately estimating the total population. Thus, the results must ultimately be affected. Henceforth, we will refer to the model of Rapley and Welsh (2008) as the 'network model'. Note that the assumptions of their model are not wholly suitable for the blue-winged teal data. In contrast to their model, our proposed model assumes heterogeneity among units, which seems more reasonable when we analyze Figure 8. Furthermore, note that there are two units with a number of blue-winged teal that differs substantially from the others, and hence if the samples selected do not contain this network, it will be very difficult

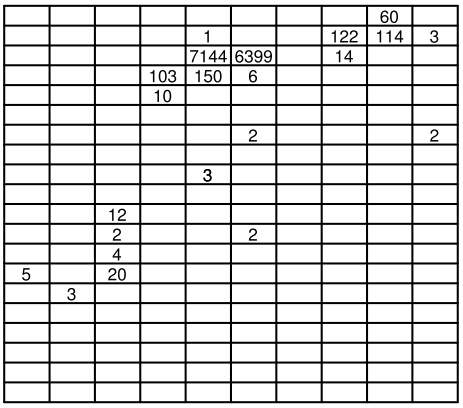


Figure 8: Counts of blue-winged teals in central Florida in 1992 in a grid with $N = 200$ units.

to accurately estimate the total population. Thus, we evaluated the performance of the estimators using the number of elements that are not in this network. After omitting that network from the population, the population total target parameter becomes $T = 365$.

Figure 12 in Appendix C shows the trace plot with the posterior distribution of parameters α and β and the population total when fitting both models for one of the samples selected. The gray line represents the true value of the population total. Both models tend to overestimate the total, but in the network model, this issue is more perceptible. The convergence was also assessed for this selected sample. Table 6 in Appendix C presents the values of the Geweke and Raftery–Lewis criteria. Analyzing Figure 12 and Table 6 leads us to conclude that convergence appears to have been reached. The same conclusion was achieved for all 600 samples selected from this population.

A summary comparison of the population total estimators using the relative bias (RB), RAE, CV and the empirical coverage of the nominal 95% HPD intervals and their widths (expressed as their respective ratios to the true values and averaged over the 600 samples) are presented in Table 3. We also compared the results from our technique and the approach developed by Rapley and Welsh (2008) to the results obtained by applying an unbiased Raj's estimator. Salehi and Seber (1997) offered details on how to apply Raj's estimator in adaptive cluster sampling without replacement. This estimator of the population total is based only on the information contained in the selected networks, i.e., ignoring the information in the edge units. In this case, we used a normal approximation to set the 95% confidence interval to the population total.

The results in Table 3 are obtained considering all 600 replications generated and excluding from the analysis the replications having either only empty networks or networks with only one unit each (in parenthesis).

Table 3 shows that the Bayes estimator produced by the network model's fit seems to be nearly unbiased. Although it is well-known that Raj's estimator is unbiased, both

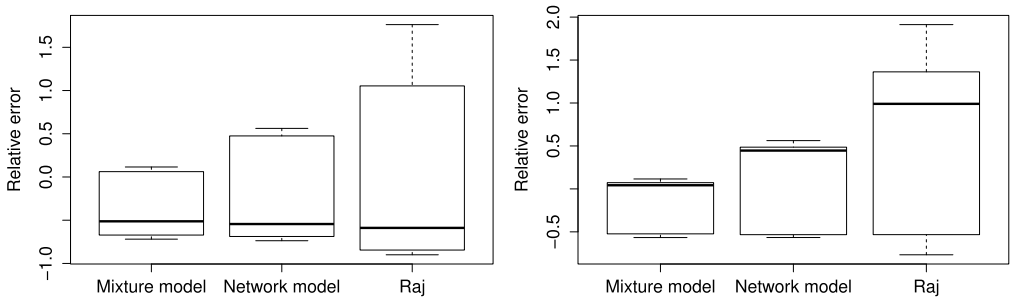
estimators have larger RAEs than our proposed estimator (2). Moreover, our proposed estimator has a smaller variance than the network and Raj's estimators. Raj's estimator has a much larger variance than its counterparts.

The network model produces 95% HPD intervals that have lower nominal coverages than the others, even when excluding the samples with either only empty networks or networks with only one unit each. Furthermore, our proposed model appears to be more efficient when applied to these data.

	RB	RAE	Coverage	Width	$CV(\hat{T})$
Mixture model	–0.10 (–0.04)	0.22 (0.12)	69.8 (89.1)	0.41 (0.47)	0.34 (0.27)
Network model	0.04 (0.15)	0.32 (0.24)	60.1 (77.2)	0.85 (0.93)	0.53 (0.45)
Raj's estimator	–0.02 (0.52)	0.95 (0.98)	69.8 (89.1)	2.79 (4.28)	0.96 (0.89)

Table 3: Summary measurements of the point and interval estimates of the total population, obtained by fitting the proposed and network models and Raj's estimator.

Figure 9 shows the boxplots of the relative errors of the Bayes estimators obtained when fitting each model and Raj's estimator, based on all 600 replications (a) and excluding the replications having either only empty networks or networks with only one unit each (b). Here again, we see that the relative errors obtained for our proposed model are lower than its counterparts.



(a) All samples (b) Restricted samples

Figure 9: Boxplots with the relative errors for T, obtained from the fits of the mixture and network models and from Raj's estimator.

5.2 A model-based experiment

The purpose of this simulation study was to compare the performance of the network and mixture models when the populations were generated according to the mixture model. We considered two scenarios. For the first scenario, we used the same populations of 500 generated in the simulation study presented in Section 4.1 and fitted the network model to evaluate its performance. In particular, we considered the case in which $(\alpha, \beta) = (0.15, 0.10)$. For the second scenario, we generated the components of $\boldsymbol{\lambda}$ according to a gamma distribution with CV=25%. Thus, we expected that the performance of the

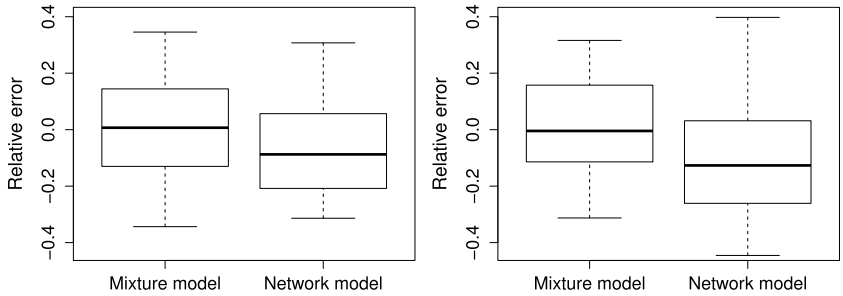
network model would improve because the degree of homogeneity of $\boldsymbol{\lambda}$'s components was higher than in the first scenario (CV=50%).

Table 4 displays some of the frequentist properties of the estimators obtained by fitting the network and mixture models. To facilitate the comparison, the results when fitting the mixture model with the same populations are presented in parentheses in Table 4. Regarding the estimation of T, both models have equivalent performance when CV=25%. However, as the degree of homogeneity decreases, the mixture model performs considerably better than the network model. However, the network model exhibits better performance than the mixture model with respect to the parameter β in both scenarios.

	CV=25%			CV=50%		
	T	α	β	T	α	β
RMSE	0.03 (0.03)	0.05 (0.08)	0.18 (0.41)	0.05 (0.03)	0.04 (0.04)	0.10 (0.40)
RAE	0.17 (0.14)	0.16 (0.22)	0.32 (0.49)	0.21 (0.15)	0.19 (0.15)	0.37 (0.50)
Cov.	96.8 (96.6)	97.1 (91.7)	95.6 (97.5)	95.6 (96.5)	98.1 (94.7)	97.4 (97.3)
Wid.	0.86 (0.70)	0.16 (0.12)	0.19 (0.22)	0.85 (0.95)	0.16 (0.11)	0.18 (0.23)

Table 4: Summary measurements of the point and 95% HPD interval estimates of the network and mixture (in parentheses) models' parameters over 500 simulations where $\boldsymbol{\lambda}$s were generated from a gamma distribution with CV=25% or 50%, for $N = 400$ and $(\alpha, \beta) = (0.15, 0.10)$.

Finally, we present the boxplots of the relative error of T for both models in Figure 10. The conclusion is analogous to the other measurements. In particular, the estimator provided by the network model seems to underestimate T in both scenarios.



(a) CV = 25% (b) CV = 50%

Figure 10: Boxplots with the relative errors for T for the 500 samples, obtained by the fits of the proposed and the network models, for a gamma distribution for $\boldsymbol{\lambda}$ with CV=25% and CV=50%.

Therefore, from those results, we concluded that the performance of the evaluated models becomes similar as the level of homogeneity between networks increases. The main difference is the number of parameters to estimate and the computational effort, which is more significant when fitting the mixture model.

6 Conclusions and suggestions for future work

We have considered the problem of estimating the total number of individuals in a rare and clustered population. Our approach is to model the observed counts in grid cells, selected by adaptive cluster sampling, and then to use model-based analysis to estimate the total population. The proposed model is an alternative to the model of Rapley and Welsh (2008) because it models the grid cells instead of the networks and assumes heterogeneity across units that belong to different networks. Nevertheless, it requires considerable computational effort and should therefore be used only if the data support it. However, simulation studies show that it might become sensible to employ the mixture model as an alternative to the network model as the homogeneity between networks decreases.

More general assumptions can also be considered and modeled within this framework. For example, in the same network, units near the centroid should have higher frequencies than units that are far from the centroid. It is possible to consider this assumption in the proposed model.

Note that the parameters of the response variable associated with the unobserved components present some estimation difficulties. Therefore, the prior distribution should be carefully elicited.

The main findings of this work encourage an extension of the model-based analysis to other adaptive sampling plans, which uncover more information about the population. One example is adaptive cluster double sampling, which was proposed by Felix-Medina and Thompson (2004) and allows the sampler to control the number of measurements of the variable of interest and to use auxiliary information.

Appendix A: Full conditional distributions and acceptance probability for the split or combination moves

Combining the joint likelihood (6) with the prior distribution, we can easily obtain the full conditional distributions of $X_{\bar{s}}, R_{\bar{s}}, \boldsymbol{\epsilon}_{\bar{s}}, \mathbf{C}_{\bar{s}}, \mathbf{Y}_{\bar{s}}, \alpha, \beta, \boldsymbol{\lambda}$ and ν

$$[\alpha \mid \cdot] \propto \frac{\alpha^{X_s+X_{\bar{s}}}(1-\alpha)^{N-X_s-X_{\bar{s}}}}{1-(1-\alpha)^N}\alpha^{a_\alpha-1}(1-\alpha)^{b_\alpha-1},$$

$$[\beta \mid \cdot] \propto \frac{\beta^{R_s+R_{\bar{s}}}(1-\beta)^{X_s+X_{\bar{s}}-R_s-R_{\bar{s}}}}{1-(1-\beta)^{X_s+X_{\bar{s}}}}\beta^{a_\beta-1}(1-\beta)^{b_\beta-1},$$

$$[\lambda_j \mid \cdot] \propto \frac{\lambda_j^{\sum_{\{i:\epsilon_i=j\}} Y_i+d-1}\exp\{-\lambda_j(\nu+C_j)\}}{1-\exp(-\lambda_j)},$$

$$[X_{\bar{s}}, \mathbf{C}_{\bar{s}} \mid \cdot] \propto \prod_{l=1}^{m}\frac{Z_{il}\times g_{il,l}}{\sum_{i=1}^{N-X+R}Z_i-\sum_{k=0}^{l-1}Z_{i_k}}\frac{\alpha^{X_{\bar{s}}}(1-\alpha)^{-X_{\bar{s}}}}{(N-X_s-X_{\bar{s}})!}\frac{(1-\beta)^{X_{\bar{s}}}}{(1-(1-\beta)^{X_s+X_{\bar{s}}})}$$

$$\times (X_s + X_{\bar{s}})^{-(X_s+X_{\bar{s}})} \prod_{\{j:j\in\bar{s}\}} C_j^{C_j} \prod_{\{j:j\in\bar{s}\}} \frac{1}{(C_j-1)!} R^{-(X_{\bar{s}}-R_{\bar{s}})}$$

$$\times \prod_{\{j:j\in\bar{s}\}} \frac{\exp\{-\lambda_j C_j\}}{[1-\exp(-\lambda_j)]^{C_j}},$$

$$Y_i \sim \text{Truncated Poisson}(\lambda_j), \{i : \epsilon_i = j, j \in \bar{s}\},$$

$$[\epsilon_i = j \mid \cdot] \propto \frac{C_j}{X_s + X_{\bar{s}}} \frac{\lambda_j^{Y_i} \exp(-\lambda_j)}{Y_i![1-\exp(-\lambda_j)]}, j \in \bar{s},$$

$$\nu \sim \text{Gamma}\left((R_s + R_{\bar{s}})d + e, f + \sum_{j=1}^{R_s+R_{\bar{s}}} \lambda_j\right).$$

For the split step, to obtain the acceptance probability it is necessary to simulate (u_1, u_2) from distributions with densities g_1 and g_2, respectively. The probability of acceptance, assuming the gamma prior distribution for $\boldsymbol{\lambda}$, is $\min(1, A)$ with

$$A = \frac{\exp\{-(c_{j_1}\lambda_{j_1} + c_{j_2}\lambda_{j_2})\}\lambda_{j_1}^{\sum_{\{i:\epsilon_i=j_1\}} y_i} \lambda_{j_2}^{\sum_{\{i:\epsilon_i=j_2\}} y_i} (1-\exp(-\lambda_{j_1}))^{-c_{j_1}} (1-\exp(-\lambda_{j_2}))^{-c_{j_2}}}{\exp\{-c_{j^*}\lambda_{j^*}\}\lambda_{j^*}^{\sum_{\{i:\epsilon_i=j^*\}} y_i} (1-\exp(-\lambda_{j^*}))^{-c_{j^*}}}$$

$$\frac{p(\{i_{j_1}, i_{j_2}\})}{p(\{i_{j^*}\})} \times \frac{p(R_{\bar{s}}+1)}{p(R_{\bar{s}})} \times \frac{(c_{j^*}-1)!}{(c_{j_1}-1)!(c_{j_2}-1)!} (R_s + R_{\bar{s}})^{-(c_{j_1}+c_{j_2}-c_{j^*})} \times \frac{c_{j_1}^{c_{j_1}} c_{j_2}^{c_{j_2}}}{c_{j^*}^{c_{j^*}}} \times (R_{\bar{s}}+1)$$

$$\times \frac{\nu^d}{\Gamma(d)} \left(\frac{\lambda_{j_1}\lambda_{j_2}}{\lambda_{j^*}}\right)^{d-1} \exp\{-\nu(\lambda_{j_1} + \lambda_{j_2} - \lambda_{j^*})\}$$

$$\times \frac{p_{R_{\bar{s}}|R_{\bar{s}}+1}}{p_{R_{\bar{s}}+1|R_{\bar{s}}} P_{alloc} q(u_1) q(u_2)} \times |J|,$$

where $p_{R_{\bar{s}+1}|R_{\bar{s}}}$ is the probability of choosing the split step, P_{alloc} is the probability that this particular allocation is made, and $|J|$ is the Jacobian of the transformation $(w_{j^*}, \lambda'_{j^*})$ to $(w_{j_1}, w_{j_2}, \lambda'_{j_1}, \lambda'_{j_2})$. For the corresponding combination step, the acceptance probability is $\min(1, A^{-1})$, and simple adaptations must be made because the proposal reduces the number of non-sampled networks by 1.

Appendix B: Assessment of MCMC and RJMCMC with one simulated data set

To check the convergence of the RJMCMC estimation, we generated only one population in an area with $N = 400$ units, fixing $\alpha = 0.15$ and $\beta = 0.10$. The values of the components of $\boldsymbol{\lambda}$ were generated from a gamma distribution centered on 8.5 with a CV equal to 95%, resulting in a gamma distribution with parameters $d = 1.1$ and $\nu = 0.13$. Then, we selected a sample of $m = 20$ networks using the adaptive design. We considered the prior distributions described in Section 2.1. For α and β, we chose $a_\alpha = 3$, $b_\alpha = 15$, $a_\beta = 1$ and $b_\beta = 9$, which parallel the prior distributions considered by Rapley and Welsh (2008). These values are suitable when the only knowledge that can be obtained

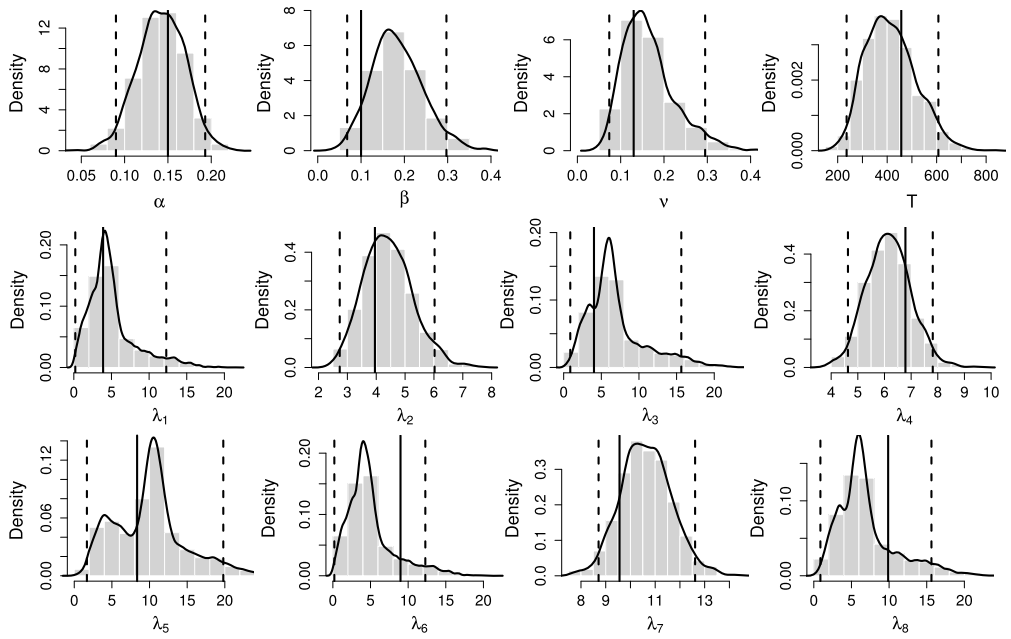


Figure 11: Posterior densities for certain model parameters and the population total T for an artificial population. The vertical solid line is the true value fixed in the simulation, and the dashed line is the 95% HPD interval.

about the underlying population is that it is sparse and clustered. For $\boldsymbol{\lambda}$, we considered only the gamma prior given in (7) used in the generation of the artificial data. The population generated yielded $R = 8$ networks, the value of the population total T was 457, and the networks observed were $s = \{2, 4, 7\}$, labeled such that the components of $\boldsymbol{\lambda}$ were in increasing order.

For the RJMCMC simulations, we generated 100,000 samples from the posterior distribution, discarded the first 10,000, and then thinned the chain by taking every 90th sample value. Figure 11 displays the histogram with the posterior densities of α, β, ν, $\boldsymbol{\lambda}$ and T for the generated population. The posterior densities of $\boldsymbol{\lambda}$'s components are conditional on the posterior samples, for which the estimated value of R is equal to eight. The solid and the dashed lines represent the true value and the 95% HPD interval, respectively. Note that most of the parameters are well estimated, with their true values lying within the 95% HPD interval.

Note that some λ_js associated with unobserved networks have bimodal posterior distributions and lower precision. This behavior is expected in the posterior densities of mixture model parameters obtained by RJMCMC and is generally associated with the labeling at each sweep; see Richardson and Green (1997) for details. For instance, let us consider the case of two normal distributions that are unambiguously labeled. The posterior distribution of the two means could overlap, but the extent of the overlap depends on its separation and the sample size. When the means are well separated, the

labels of the realizations from the posterior obtained by ordering their means generally coincide with the labels of the population. As the separation reduces, "label switching" may occur. This problem can be minimized by choosing to order other parameters of the mixture components, for example, the variance. In our case, this bimodality does not appear in all of the simulations, only in simulations generated by the λ_js that are not well separated. Nevertheless, the bimodality influences neither the convergence of the other parameters nor the most important quantity: the total T.

The λ_js associated with the sampled networks present better estimates than the λ_js associated with the non-sampled networks. This result is expected because we have specific information for the sampled networks.

Two other diagnostics were used to show that convergence was achieved: the Geweke and the Raftery–Lewis diagnostics. The first was proposed by Geweke (1992) and is based on a test of the equality of the means of the first and last parts of the Markov chain. If the samples are drawn from the stationary distribution of the chain, the two means are equal, and Geweke's statistic has an asymptotically standard normal distribution. The second was proposed in Raftery and Lewis (1992) and calculates the number of iterations required to estimate a quantile with a desired accuracy and a certain probability. The minimum length is the required sample size for a chain with no correlation between consecutive samples. An estimate dependence factor of the extent to which autocorrelation inflates the required sample size is also provided. Values for the factor that are larger than 5 indicate strong autocorrelation, which may be due to a poor choice of starting value, high posterior correlations or stickiness of the MCMC algorithm. Table 5 presents the value of Geweke's statistic and the value of the dependence factor. The results for both criteria indicate that the MCMC chains have converged.

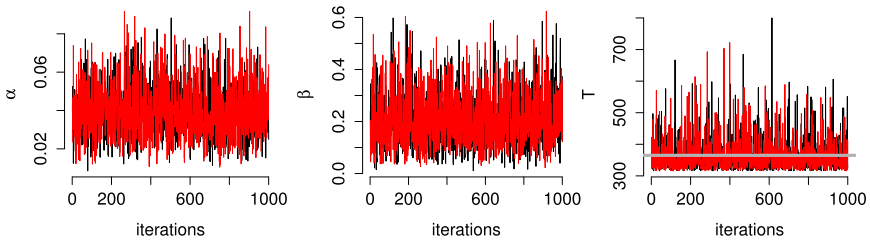
	α	β	ν	T	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
Geweke	0.7	–0.4	–1.6	0.4	1.4	–1.3	1.4	–0.4	1.5	1.5	1.2	1.5
R–L	1.3	1.1	1.1	1.8	0.9	1.0	1.0	1.0	0.9	1.0	1.1	1.1

Table 5: Geweke and Raftery–Lewis (R–L) convergence diagnostics for all of the parameters estimated for the artificial population.

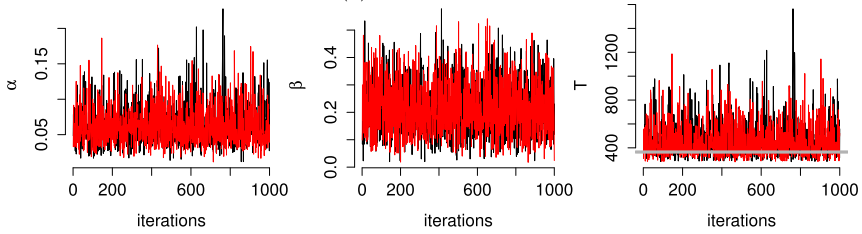
Appendix C: Assessment of MCMC and RJMCMC with real data

	Geweke		Raftery–Lewis	
Param	Mixture	Network	Mixture	Network
α	–0.13	–0.10	1.02	1.21
β	0.72	–0.67	1.15	2.56
T	–1.38	–0.30	3.22	1.33

Table 6: Geweke and Raftery–Lewis convergence diagnostics for some of the parameters estimated for the real population for both models.



(a) Mixture model



(b) Network model

Figure 12: Trace plot with the posterior densities of α, β and T obtained from the fits of the proposed and the network models. The gray line represents the true value of T.

References

Conners, M. and Schwager, S. (2002). "The use of adaptive cluster sampling for hydroacoustic surveys." *ICES Journal of Marine Science: Journal du Conseil*, 59(6): 1314–1325. 520

Diggle, P. J. (2014). *Statistical analysis of spatial and spatio-temporal point patterns.* CRC Press, Chapman & Hall. MR3113855. 525

Felix-Medina, M. H. and Thompson, S. K. (2004). "Adaptive cluster double sampling." *Biometrika*, 91(4): 877. MR2126039. doi: http://dx.doi.org/10.1093/biomet/91.4.877. 539

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman & Hall. MR1385925. 524

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments." In: J. Bernardo, A. Dawid, J. Berger, and A. Smith (eds.), *Bayesian Statistics*. Oxford University Press, New York. MR1380276. 542

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: http://dx.doi.org/10.1093/biomet/82.4.711. 527

McDonald, L. L. (2004). "Sampling rare populations." In: W. Thompson (ed.), *Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters*, chapter 4, 11–42. Island Press Washington, DC, USA. 519

Pfeffermann, D., Moura, F. A. S., and Silva, P. L. N. (2006). "Multi-level modelling under informative sampling." *Biometrika*, 93(4): 943–959. MR2285081. doi: http://dx.doi.org/10.1093/biomet/93.4.943. 524

Philippi, T. (2005). "Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants." *Ecology*, 86(5): 1091–1100. 520

Raftery, A. E. and Lewis, S. M. (1992). "One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo." *Statistical Science*, 7(4): 493–497. 542

Rapley, V. and Welsh, A. (2008). "Model-Based Inferences from Adaptive Cluster Sampling." *Bayesian Analysis*, 3(4): 717–736. MR2469797. doi: http://dx.doi.org/10.1214/08-BA327. 520, 521, 522, 523, 535, 536, 539, 540

Richardson, S. and Green, P. (1997). "On Bayesian analysis of mixtures with an unknown number of components." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 59(4): 731–792. MR1483213. doi: http://dx.doi.org/10.1111/1467-9868.00095. 522, 527, 541

Roeder, K. and Wasserman, L. (1997). "Practical Bayesian density estimation using mixtures of normals." *Journal of the American Statistical Association*, 92(439): 894–902. MR1482121. doi: http://dx.doi.org/10.2307/2965553. 526, 527

Roesch, F. (1993). "Adaptive cluster sampling for forest inventories." *Forest Science*, 39(4): 655–669. 520

Salehi, M. and Seber, G. A. (1997). "Adaptive cluster sampling with networks selected without replacement." *Biometrika*, 84(1): 209–219. MR1450201. doi: http://dx.doi.org/10.1093/biomet/84.1.209. 523, 536

Smith, D., Conroy, M., and Brakhage, D. (1995). "Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl." *Biometrics*, 51(2): 777–788. 520, 535

Thompson, S. K. (1990). "Adaptive cluster sampling." *Journal of the American Statistical Association*, 85(412): 1050–1059. MR1134501. 519, 520, 523

Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive sampling*. Wiley New York. MR1390995. 519, 520

Viallefont, V., Richardson, S., and Green, P. J. (2002). "Bayesian analysis of Poisson mixtures." *Journal of Nonparametric Statistics*, 14(1–2): 181–202. MR1905593. doi: http://dx.doi.org/10.1080/10485250211383. 522, 526, 528

Acknowledgments

This work is part of the Ph.D. thesis of Kelly C. M. Gonçalves under the supervision of Fernando Moura, in the Graduate Program of UFRJ. Kelly has a scholarship from Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES). Fernando Moura received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil, BPPesq). The authors would like to thank the editor, an associate editor and the referees for their very thoughtful and constructive comments.