Bayesian Analysis (2008) **3**, Number 4, pp. 717–736

Model-Based Inferences from Adaptive Cluster Sampling

V.E. Rapley* and A. H. Welsh†

Abstract. Adaptive cluster sampling is useful for exploring populations of rare plant and animal species which cluster together because it allows sampling effort to be concentrated in areas where observed values are high. This allows more useful data to be collected with less effort than simpler sampling methods which ignore the population structure. In this paper, we take a model based approach in a Bayesian framework to make inference about the number of individuals in a sparse, clustered population. This approach allows us to use knowledge of the population to inform both the sampling design and inference, thereby making coherent use of the data in the analysis and resulting in improved population estimates. The methodology is compared to the design-based modified Horvitz-Thompson estimator through analysis of the examples presented in the defining paper of Thompson (1990).

Keywords: Informative sampling, MCMC, spatial sampling, zero-inflated count data

1 Introduction

Thompson (1990, 1992, 1991a,b, 2002) and Thompson and Seber (1996) introduced adaptive cluster sampling as a refined method for estimating the size of sparse, clustered populations of plants or animals. Thompson (1990) formalised the appealing strategy of increasing survey effort around where a plant or animal of interest is found, developed a design-based analysis of the resulting sampling scheme and showed that adaptive cluster sampling is more efficient than the simpler, traditional grid cell sampling. In this paper, we model the data obtained by adaptive cluster sampling and then develop a model-based Bayesian analysis for adaptive cluster sampling. The use of the Bayesian framework is a natural extension of the key idea behind adaptive cluster sampling which incorporates the prior knowledge that a population is clustered into the inference as well as into the sampling design.

Suppose that the plants or animals of interest are spatially distributed in a region R. Suppose further that in the same way as Thompson (1990), we superimpose a grid on R. Following the construction of Thompson, we take the grid to be regular so that the M grid cells are square and of equal size. Let Q denote the set of all M grid cells and let N_i denote the grid cell count which is the number of plants or animals of interest in the ith grid cell, $i = 1, \ldots, M$. The objective is to estimate $N = \sum_{i=1}^{M} N_i$, the total

*S3RI,University Of Southampton, Southampton, UK ,mailto:ronni.rapley@gmail.com
†Centre for Mathematics and its Applications, Australian National University, Canberra, Australia, http://wwwmaths.anu.edu.au/~welastat/

 DOI:10.1214/08-BA327

number of plants or animals of interest in R.

Grid cell sampling methods involve the selection of a subset q of $m < M$ grid cells and the observation of N_i in the selected grid cells. For sparse, clustered populations, most of the samples with small m/M will consist of mainly empty grid cells (for which $N_i = 0$), resulting in many samples yielding poor estimates of N. Adaptive cluster sampling overcomes this difficulty by allowing an increase in survey effort in the neighbourhood of any nonempty grid cells (for which $N_i \geq 1$) in the sample. Specifically, when we find a nonempty grid cell, we also survey the neighbours of that grid cell and continue to survey neighbours of nonempty neighbours until we obtain a set of contiguous nonempty grid cells surrounded by empty grid cells; selected empty grid cells attract no additional survey effort. Intuitively, this type of sampling scheme maximises survey effort where it is most valuable and minimises survey effort elsewhere. It performs better for sparse, clustered populations than standard grid cell sampling methods. However, to be effective, adaptive cluster sampling does require some prior knowledge about the structure of the underlying population, including the number, size and spatial extent of clusters. This information (which may be obtained from a preliminary survey) is implicitly balanced against practical issues such as difficulties in marking, accessing and surveying grid cells and is then incorporated into the sampling design through the choice of grid cells.

Following Thompson (1990), sets of contiguous nonempty grid cells and their neighbouring empty grid cells are together called clusters. The set of contiguous nonempty grid cells within a cluster is called a network. We also define empty cells as networks of size one so the networks (unlike clusters which can overlap) form an exhaustive partition of R. The distinction between clusters and networks is fundamental to adaptive cluster sampling. The insight in Thompson (1990) was to base the analysis on networks rather than clusters and to treat the empty edge units as unobserved unless they were selected in their own right rather than simply as neighbours. For initial grid cells selected by simple random sampling without replacement, Thompson (1990) derived inclusion and joint inclusion probabilities for the networks observed in the sample. He then used these probabilities to construct design-unbiased estimators of N and to construct estimates of their design-variances.

There have been a number of discussions of the relative merits of the design-based and model-based approaches to the analysis of survey data; see for example Royall (1976); Brus and de Gruijter (1997); R. Valliant and Royall (2000); Brewer (1963); Thompson (2002) etc. The philosophical differences between the approaches are important but are not our present concern. At a pragmatic level, in very simple, general terms, the design-based approach trades off efficiency for wide applicability while the model-based approach which is usually more efficient when the assumed model holds, trades off wide applicability for increased efficiency. A Bayesian approach to inference is then a natural way to work with the model as it allows the incorporation of the prior knowledge used in constructing the sampling design into the analysis as well. Our viewpoint is that there is no model-based approach to the analysis of data collected by adaptive cluster sampling so it is interesting and challenging, both philosophically and practically, to develop such an approach.

While we believe that the specific distributions we use in this paper are widely applicable, it is perhaps useful to emphasize that these can be changed and our main contribution is the clarification and development of a general approach to the model-based analysis of adaptive cluster samples. Our approach involves (i) the use of the insight in Thompson (1990), that the analysis can be based on networks rather than clusters; (ii) the use of a spatial hurdle (Mullahy 1986), two-part (Heilbron 1994) or conditional (A.H. Welsh and Lindenmayer 1996) model in which we first model presence/absence and then conditional on presence, model abundance to accommodate the fact that most network counts are zero; (iii) the setting up of a model for the population of network counts which deals with the fact that, because the size and number of networks is not known in advance, adaptive cluster sampling is informative about the network counts; (iv) the introduction of prior distributions for the unknown parameters in the model to incorporate prior beliefs about the population such as sparseness, clustering etc. and to facilitate the use of the Markov-Chain Monte Carlo (MCMC) methods (see for example Tanner (1996)) to integrate out the unobserved stochastic quantities in the model; and (v) the construction of posterior predictive distributions for making inference about N. Our treatment of informative sampling is informed by the literature on successive sampling discovery models (for example, G. Kaufman and Kruyt (1975); Nair and Wang (1989); West (1996)). Our modelling is informed by the likelihood approach of Breckling et al. (1994) to modelling survey data and the final model can be interpreted as a Bayesian modification to that approach. Finally, as we use the insight in Thompson (1990) explicitly by modelling network rather than grid cell counts, our model depends explicitly on the grid size. This enables us to finesse any need to discuss the underlying point process and to derive the distribution of grid cell/network counts (which would also depend on the grid size) from it. Thus working with networks has roughly the same importance to the model-based analysis as it does to the design-based analysis of Thompson (1990).

We introduce the model and use it to construct predictions of the population size N in Section 2. We discuss how to fit the model in Sections 3 and 4. We then illustrate its use on examples like those suggested by Thompson (1990) in Section 5. We conclude with a brief discussion of our experience with the model-based approach to adaptive cluster sampling in Section 6.

2 A Model for Network Counts

Recall that we consider a region R containing a sparse, clustered population of N points. We superimpose a regular grid on R so that R is partitioned into M square, equal sized grid cells. A grid cell is nonempty if it contains at least one point and empty otherwise. Let $X \leq M$ be the number of nonempty grid cells in R. Let $P \leq X$ be the number of nonempty networks in R and let $\mathbf{Y} = (Y_1, \ldots, Y_P)$ denote the number of nonempty grid cells within each network so that $X = \sum_{i=1}^{P} Y_i$. As there are $M - X$ empty grid cells which are defined to be empty networks of size one, there are $M - X + P$ networks in R. The variables $Y_1, \ldots, Y_P$ describe the number of grid cells in each nonempty network. It is notationally convenient to extend the P-vector $\mathbf{Y}$ to the $(M - X + P)$-vector

$\mathbf{Z} = (\mathbf{Y}^T, \mathbf{1}_{M-X}^T)^T$, where $\mathbf{1}_{M-X}$ is the $(M-X)$-vector of ones, by defining $Z_i = Y_i$ for each nonempty network and $Z_i = 1$ for each empty network. Note that $\mathbf{Z}$ and $\mathbf{Y}$ contain the same information about the networks. Let $N_1, \ldots, N_{M-X+P}$ denote the network counts which represent the number of points within each network. Only P of these counts are nonzero so it is convenient to write $(N_1, \ldots, N_{M-X+P}) = (\mathbf{N}^T, \mathbf{0}_{M-X}^T)^T$, where $\mathbf{N}$ is the set of nonzero network counts and $\mathbf{0}_{M-X}$ is the $(M-X)$-vector of zeros. The ultimate goal of the analysis is to make inferences about the population size

$$N = \sum_{i=1}^{M-X+P} N_i = \sum_{i:N_i>0}^{P} N_i$$

which is a random variable in our model-based approach.

We construct a model for the network counts by specifying the joint distribution of X, P, $\mathbf{Y}$ and $\mathbf{N}$ for the entire population and the sampling mechanism which leads to a particular sample $s = \{i_1, \ldots, i_m\}$ of m out of $M - X + P$ networks. A crucial aspect of our approach is that the network structure is determined by X, P and $\mathbf{Y}$ and we do not need to model the spatial locations of the networks. This does not entail any loss of information about N because, under the model, the population size does not depend on where the networks are located. We therefore finesse a potentially difficult problem and are able to proceed relatively simply.

The sampling mechanism describes the selection of networks, so can only be defined conditionally on the network structure described by X, P and $\mathbf{Y}$. That is, we formulate the model using what Breckling et al. (1994) call the target parameterisation. Thompson (1990) computed network inclusion probabilities under an initial simple random sample of grid cells chosen without replacement. This is slightly inconvenient as the same network can be selected more than once, a problem he dealt with by allowing multiple inclusions of networks. In contrast, we treat the sampling mechanism as sampling 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 M grid cells, surveying that grid cell and, if it is nonempty, 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 way until we have selected m networks in the sample. That is, we are effectively sampling networks by probability proportional to size without replacement. Note that the probability of selecting a network in the sample depends on its size Z_i and not on its network count N_i. The sampling is informative because the random variables $\mathbf{Z}$ are only observed for the sampled networks after their selection in the sample.

To motivate the notation for the probability of selecting a given sample, consider a population consisting of ten networks of size $\{5, 5, 1, 1, 1, 3, 3, 1\}$ from which we obtain the (ordered) sample $\{5, 1, 5, 3\}$ by selecting units by probability proportional to size sampling without replacement. The probability of selecting the first unit is the probability of selecting a unit of size 5 (because the two units are indistinguishable) which is 10/20, the probability of selecting a unit of size 1 at the second step given the previous

step is 4/15 and so on so that the probability of selecting the specified sample is

$$\frac{10}{20} \times \frac{4}{15} \times \frac{5}{14} \times \frac{6}{9} = \frac{5 \times 2}{20} \times \frac{1 \times 4}{20-5} \times \frac{5 \times 1}{20-5-1} \times \frac{3 \times 2}{20-5-1-5}.$$

At each step, the numerator is the size of the unit selected at that step times the number of units of that size still unselected in the population at that step and the denominator is the sum of the sizes of the still unselected units in the population at that step. The denominator at a step can also be expressed as the sum of the sizes of all units in the population minus the sum of the sizes of the already selected units in the population up to that step. Thus, in general, the probability of selecting the sample $s = \{i_1, \ldots, i_m\}$ of m networks is

$$[\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}] = \prod_{j=1}^{m} \frac{z_{i_j} \times g_{i_j,j}}{\sum_{i=1}^{M-x+p} z_i - \sum_{k=0}^{j-1} z_{i_k}} \tag{1}$$

where $g_{i,j}$ is the number of networks of size z_i unselected after $j-1$ networks have been selected. The incorporation of $g_{i,j}$ modifies the expressions used in Nair and Wang (1989) and West (1996) for the discreteness of network sizes.

Next, we specify the joint distribution of X, P, $\mathbf{Y}$ and $\mathbf{N}$ for the entire population. Conceptually, we first model the nonempty/empty network structure (determined by X, P and $\mathbf{Y}$) and then, conditionally on the network structure, model the network counts $\mathbf{N}$ in the non-empty networks. The specific distributions we adopt are chosen to give a simple structure to the model and alternative choices could be made. To avoid degeneracy, we assume that there is at least one network in R and, as nonempty networks consist, by definition, of at least one grid cell, we left-truncate distributions at one. In particular it should be noted that the method of overlaying a distribution of population counts on top of the network framework allows considerable flexibility. If specific knowledge is held about the population of interest this distribution can be modified without changing the underlying structure or altering the inference methodology. Specifically, we first assume that

$$\begin{aligned}
X|\alpha &\sim \text{Truncated binomial}(M, \alpha), \quad X = 1, \ldots, M \\
P|X, \beta &\sim \text{Truncated binomial}(x, \beta) \quad P = 1, \ldots, x \text{ and} \\
\mathbf{Y}|X, P &\sim \mathbf{1}_p + \text{Multinomial}(x-p, \frac{1}{p}\mathbf{1}_p), \quad Y_i = 1, \ldots, x-p, \ \sum_{i=1}^{p} Y_i = x,
\end{aligned}$$

where $\mathbf{1}_p$ is the p-vector of ones. (The offset like formulation is used because it is simpler to use than the truncated multinomial distribution. A similar approach could be used instead of truncation for the binomial distributions.) Using the convention that the

density of a random variable P is denoted $[P]$, we have the population level model

$$
\begin{aligned}
[X, P, \mathbf{Y}|\alpha, \beta] &= [X|\alpha][P|X, \beta][\mathbf{Y}|X, P] \\
&= \binom{M}{x} \frac{\alpha^x (1-\alpha)^{M-x}}{1-(1-\alpha)^M} \binom{x}{p} \frac{\beta^p (1-\beta)^{x-p}}{1-(1-\beta)^x} (x-p)! \prod_{i=1}^{p} \frac{1}{(y_i-1)!} \left(\frac{1}{p}\right)^{y_i-1} \\
&= \frac{M!}{(M-x)!p!} \frac{\alpha^x (1-\alpha)^{M-x}}{1-(1-\alpha)^M} \frac{\beta^p (1-\beta)^{x-p}}{1-(1-\beta)^x} \prod_{i=1}^{p} \frac{1}{(y_i-1)!} \left(\frac{1}{p}\right)^{y_i-1}. \qquad (2)
\end{aligned}
$$

The parameter α controls the expected number of nonempty grid cells ($M\alpha/\{1-(1-\alpha)^M\}$) and β controls the conditional expected number of nonempty networks $X\beta/\{1-(1-\beta)^M\}$. As we are surveying sparse populations, both these parameters will typically be small. The multinomial probabilities are taken as known (because we will not have much sample information to estimate unknown probabilities) and equal (because there is no convincing reason for, or way to describe heterogeneity before doing any sampling) for simplicity. Under our model, the conditional expected network size is $1+(X-P)/P = X/P$.

Finally, we adopt a conditional log-linear model for the counts $\mathbf{N}$ in which

$$
N_1, \ldots, N_P | P, \mathbf{Y}, \boldsymbol{\gamma} \sim \text{ independent truncated Poisson}(\exp\{\gamma_0 + \gamma_1 h(y_i)\}),
$$

where h is a known function and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are unknown parameters. The function h will typically be specified as the identity or the log function; in the latter case, the expected conditional network counts $\exp\{\gamma_0 + \gamma_1 h(y_i)\}$ are a multiplicative function of Y_i. The truncation is needed to take into account the fact that each grid-cell in a network must contain at least one plant or animal of interest so that $N_i \geq Y_i$, $i = 1, ..., P$. We obtain

$$
\begin{aligned}
&[\mathbf{N}|P, \mathbf{Y}, \boldsymbol{\gamma}] \\
&= \prod_{i=1}^{P} \frac{\exp\{-\exp(\gamma_0 + \gamma_1 h(y_i)) + n_i(\gamma_0 + \gamma_1 h(y_i))\}}{n_i![1 - \sum_{j=1}^{y_i} \exp\{-\exp(\gamma_0 + \gamma_1 h(y_i)) + j(\gamma_0 + \gamma_1 h(y_i)) - \log(j!)\}]}, \\
&\qquad\qquad n_i = y_i, y_i + 1, \ldots. \qquad (3)
\end{aligned}
$$

While we can in principle extend the truncated Poisson specification to incorporate overdispersion and/or replace the truncated Poisson specification by any model for count data, it is important to keep in mind that we expect P to be small, so it may be difficult to estimate the parameters in such models. If P is so small that it is difficult to estimate both parameters in the Poisson specification, we may choose to fit the null version of the model with $\gamma_1 = 0$ or the no intercept model with $\gamma_0 = 0$.

The sampling procedure entails observing Y_i and N_i for the networks labelled $\{i_1, \ldots, i_m\}$. It is convenient to use the subscript '0' to identify the observed component and '1' the unobserved component, and write $\mathbf{N} = (\mathbf{N}_0, \mathbf{N}_1)$, $\mathbf{Y} = (\mathbf{Y}_0, \mathbf{Y}_1)$, $X =$

$X_0 + X_1$ and $P = P_0 + P_1$ to distinguish between observed and unobserved quantities. Here X_0 is the observed number of nonempty grid cells and P_0 is the observed number of nonempty networks in the sample.

The sampling design satisfies the condition

$$[\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}, \mathbf{N}] = [\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}], \tag{4}$$

so we can write the product of (1), (2) and (3) as

$$\begin{aligned}
[\{i_1, \ldots, i_m\}, X, P, \mathbf{Y}, \mathbf{N}|\alpha, \beta, \boldsymbol{\gamma}] &= [\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}, \mathbf{N}][X, P, \mathbf{Y}, \mathbf{N}] \\
&= [\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}][X, P, \mathbf{Y}, \mathbf{N}|\alpha, \beta, \boldsymbol{\gamma}] \\
&= [\{i_1, \ldots, i_m\}|X, P, \mathbf{Y}][X, P, \mathbf{Z}|\alpha, \beta][\mathbf{N}|X, P, \mathbf{Y}, \boldsymbol{\gamma}] \\
&= [\{i_1, \ldots, i_m\}X, P, \mathbf{Y}|\alpha, \beta][\mathbf{N}|P, \mathbf{Y}, \boldsymbol{\gamma}]. \qquad (5)
\end{aligned}$$

That is, the model factorises into two terms and, at the population level, (α, β) and $\boldsymbol{\gamma}$ are orthogonal parameters. However, as described in Breckling et al. (1994), the likelihood at the sample level is obtained by summing over the quantities which are unknown and not otherwise observed in the selected sample to obtain

$$\begin{aligned}
[X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0|\alpha, \beta, \boldsymbol{\gamma}] &= \sum_{\mathbf{N}_1} \sum_{\mathbf{Y}_1} \sum_{P_1} \sum_{X_1} [\{i_1, \ldots, i_m\}, X, P, \mathbf{Y}, \mathbf{N}|\alpha, \beta, \boldsymbol{\gamma}] \qquad (6) \\
&= \sum_{\mathbf{N}_1} \sum_{\mathbf{Y}_1} \sum_{P_1} \sum_{X_1} [\{i_1, \ldots, i_m\}, X, P, \mathbf{Y}|\alpha, \beta][\mathbf{N}|P, \mathbf{Y}, \boldsymbol{\gamma}].
\end{aligned}$$

Thus, at the population level, the model can be interpreted as a spatial version of the models of Mullahy (1986); Heilbron (1994) or A.H. Welsh and Lindenmayer (1996) for modelling count data with many zeros but the same factorisation does not hold at the sample level. Even if we set $\gamma_1 = 0$ so there is no relationship between $\mathbf{N}$ and $\mathbf{Y}$, we still have $[\mathbf{N}|P, \mathbf{Y}, \boldsymbol{\gamma}] = [\mathbf{N}|P, \boldsymbol{\gamma}]$ in (3) so the factors are linked by the common unobserved P_1.

The sums we need to evaluate to obtain the sample level likelihood are too complicated to be evaluated analytically. We have therefore chosen to use a Markov Chain Monte-Carlo (MCMC) approach to estimate the unknown parameters α, β and $\boldsymbol{\gamma}$ in the model and the unobserved quantities X_1, P_1, $\mathbf{Y}_1$ and $\mathbf{N}_1$. In difficult problems such as adaptive cluster sampling the Bayesian approach has several advantages: i) we can improve estimation of quantities for which there is little sample information by incorporating other information such as the underlying population is sparse; ii) we can propagate the uncertainty in our estimates through to the final predictions; and iii) we can obtain a natural predictive distribution for predicting N. In this problem it is natural to use the prior distributions to incorporate the knowledge about the populations assumed in the sampling design into the analysis and this is one of the key improvements to the original technique. However, it is not the purpose of this paper to address whether Bayesian inference should be used and it can therefore be noted that, in principle, non-informative priors could be chosen to make the inference 'likelihood-like';

there is a cost in that the model is then much more difficult to fit. Whichever approach we use, note that a natural predictor of N is

$$\hat{N} = \mathbf{1}_{P_0}^T \mathbf{N}_0 + \mathbf{1}_{\hat{P}_1}^T \hat{\mathbf{N}}_1$$

and that its distribution can be obtained from the MCMC output.

3 Model Fitting

Assume that the three unknown parameters α, β and $\boldsymbol{\gamma}$ are a priori independent and denote their marginal prior distributions by $\pi(\alpha)$, $\pi(\beta)$ and $\pi(\boldsymbol{\gamma})$ respectively. The summand in the expression (6) for $[X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0|\alpha, \beta, \boldsymbol{\gamma}]$ is the product of (1), (2) and (3). If we then incorporate the prior information, the modified summand becomes the joint distribution of all the quantities in the model (including the parameters) which is

$$\begin{aligned}
&[X, P, \mathbf{Y}, \mathbf{N}, \alpha, \beta, \boldsymbol{\gamma}] \\
&\qquad = \prod_{j=1}^{m} \frac{z_{i_j} \times g_{i_j,j}}{\sum_{i=1}^{M-x+p} z_i - \sum_{k=0}^{j-1} z_{i_k}} \times \binom{M}{x} \frac{\alpha^x (1-\alpha)^{M-x}}{1-(1-\alpha)^M} \pi(\alpha) \\
&\qquad \times \binom{x}{p} \frac{\beta^p (1-\beta)^{x-p}}{1-(1-\beta)^x} \pi(\beta) \times (x-p)! \prod_{i=1}^{p} \frac{1}{y_i - 1} \left(\frac{1}{p}\right)^{y_i - 1} \\
&\qquad \times \prod_{i=1}^{p} \frac{\exp\{-\exp(\gamma_0 + \gamma_1 y_i) + n_i(\gamma_0 + \gamma_1 y_i)\}}{n_i![1 - \sum_{j=1}^{y_i} \exp\{-\exp(\gamma_0 + \gamma_1 Y_i) + j(\gamma_0 + \gamma_1 Y_i) - \log(j!)\}]} \\
&\qquad \times \pi(\boldsymbol{\gamma}). \qquad (7)
\end{aligned}$$

The steps in the Gibbs sampler which we use to fit (7) are

1) Specify initial values for the unsampled components X_1, P_1 and $\mathbf{Y}_1$.

2) Generate α from the conditional distribution $[\alpha|X, P, \mathbf{Y}, \mathbf{N}, \beta, \boldsymbol{\gamma}] = [\alpha|X]$.

3) Generate β from the conditional distribution $[\beta|X, P, \mathbf{Y}, \mathbf{N}, \alpha, \boldsymbol{\gamma}] = [\beta|X, P]$

4) Generate $\boldsymbol{\gamma}$ from the conditional distribution of $[\boldsymbol{\gamma}|X, P, \mathbf{Y}, \mathbf{N}, \alpha, \beta] = [\boldsymbol{\gamma}|P, \mathbf{Y}]$

5) Generate $(X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1)$ from the conditional distribution $[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1|X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0, \alpha, \boldsymbol{\gamma}]$

6) Iterate from 2).

3.1 Generating α

The conditional distribution of α given everything else in the model is formed from all of the terms in the model (7) in which α appears so

$$[\alpha|X, P, \mathbf{Y}, \mathbf{N}, \beta, \boldsymbol{\gamma}] \propto \frac{\alpha^{x_0+x_1}(1-\alpha)^{M-x_0-x_1}}{1-(1-\alpha)^M}\pi(\alpha).$$

We take the prior distribution $\pi(\alpha)$ to be a conjugate beta(a, b) distribution with $a = 3$ and $b = 15$. The parameters of the Beta distribution are chosen to reflect the fact that α is necessarily small in a sparse, clustered population. The robustness of this prior choice and those following is examined in section 5. The conditional distribution is not truly a beta distribution (due to the truncation term) so a Metropolis-Hastings accept-reject step is used to sample for α. However, the conditional distribution is close to a beta distribution so we will use the beta$(x_0 + x_1 + a - 1, M - x_0 - x_1 + b - 1)$ distribution as the proposal distribution.

3.2 Generating β

The conditional distribution of β given everything else in the model is

$$[\beta|X, P, \mathbf{Y}, \mathbf{N}, \alpha, \boldsymbol{\gamma}] \propto \frac{\beta^{p_0+p_1}(1-\beta)^{x_0+x_1-p_0-p_1}}{1-(1-\beta)^{x_0+x_1}}\pi(\beta).$$

We again take the prior distribution $\pi(\beta)$ to be a beta(a, b) distribution with $a = 1$ and $b = 9$. In this case β is necessarily small, otherwise the population would not be clustered. We use the beta$(p_0 + p_1 + a - 1, x_0 + x_1 - p_0 - p_1 + b - 1)$ distribution as the proposal distribution in a Metropolis-Hastings accept-reject step.

The incorporation of these prior distributions allows coherence through the inference, reflecting the same data in the sampling as in the modelling. However, the specific parameters for the beta can be varied under the assumptions and the robustness of the inference to this is explored in section 4.

3.3 Generating $\boldsymbol{\gamma}$

The conditional distribution of $\boldsymbol{\gamma}$ given everything else in the model is

$$[\boldsymbol{\gamma}|X, P, \mathbf{Y}, \mathbf{N}] \propto \prod_{i=1}^{p_0+p_1} \frac{\exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + n_i(\gamma_0+\gamma_1 h(y_i))\}}{1-\sum_{j=1}^{y_i}\exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + j(\gamma_0+\gamma_1 h(y_i)) - \log(j!)\}}\pi(\boldsymbol{\gamma}).$$

We take the prior distribution $\pi(\boldsymbol{\gamma})$ to be the product of two Gamma distributions (say Gamma$(2, 7)$ or Gamma$(2, 5)$) and use the normal distribution centred at the posterior mode with variance 16 as the proposal distribution in a Metropolis-Hastings accept-reject step. Simulations (not presented here) show that the inference is not sensitive to the prior choice for γ, this reinforces the fact that the distribution for the number of objects in a cluster is independent of the model set-up and can be chosen as desired.

3.4 Generating X_1, P_1, $\mathbf{Y}_1$ and $\mathbf{N}_1$

The conditional distributions of X_1, P_1, $\mathbf{Y}_1$ and $\mathbf{N}_1$ given everything else in the model can be obtained from (7) but the expressions are complicated and do not suggest simple methods of generating the required random variables. The joint conditional distribution of $(X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1)$ given everything else in the model is

$$
\begin{aligned}
&[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1 | X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0, \alpha, \beta, \boldsymbol{\gamma}] \\
&\propto \prod_{j=1}^{m} \frac{g_{i_j,j}}{\sum_{i=1}^{M-x_0-x_1+p_0+p_1} z_i - \sum_{k=0}^{j-1} z_{i_k}} \times \frac{\alpha^{x_1}(1-\alpha)^{-x_1}}{(M-x_0-x_1)!} \\
&\quad \times \frac{1}{(p_0+p_1)!} \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{1-(1-\beta)^{x_0+x_1}} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!} \left(\frac{1}{p_0+p_1}\right)^{\sum_{i=1}^{p_0+p_1}(y_i-1)} \\
&\quad \times \prod_{i \notin s}^{p_1} \frac{\exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + n_i \gamma_1 h(y_)i\}}{n_i![1-\sum_{j=1}^{y_i} \exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + j(\gamma_0+\gamma_1 h(y_i)) - \log(j!)\}]} \qquad (8)
\end{aligned}
$$

which is also complicated to sample from but can be used in a Metropolis-Hastings algorithm for sampling $(X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1)$ jointly; we construct a proposal distribution from which it is straightforward to sample $(X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1)$ and jointly accept or reject these values using (8) as the target distribution.

It is convenient to generate X_1 using a discrete random walk; i) given the past value X^* for X, at each step to generate a new X by sampling from the discrete uniform distribution centred at X^* with support $\{X^* \pm k : k = 1, \ldots, 5\}$ and a reflecting boundary at X_0 and M to ensure that X stays within bounds $X_1 = X - X_0$, and ii) we set $X_1 = X - X_0$. Next, given X_1 and β, we note that P_0 is the number of observed nonempty networks formed out of X_0 observed non-empty grid cells and P_1 is the additional number of non-empty networks formed out of the remaining X_1 nonempty grid cells, so we generate P_1 by sampling from the truncated binomial(X_1, β) distribution. Given X, P_1 and Y_0, we note that $\mathbf{Y}_1$ is the number of nonempty grid cells in each of the P_1 additional networks, so we generate $\mathbf{Y}_1$ from the distribution of $\mathbf{Y}_1 | \mathbf{Y}_0, X, P$ which is the $\mathbf{1}_{P_1}$ + multinomial$(X_1 - P_1, \frac{1}{P_1}\mathbf{1}_{P_1})$ distribution. Finally, given P_1, $\mathbf{Y}_1$ and $\boldsymbol{\gamma}$, we generate $\mathbf{N}_1$ as independent truncated Poisson$(\exp\{\gamma_0 + \gamma_1 h(y_i)\})$ random variables. The proposal distribution is therefore

$$
\begin{aligned}
&[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1]_{prop} \\
&= \frac{1}{10} \times \frac{x_1!}{p_1!} \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{1-(1-\beta)^{x_1}} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!} \left(\frac{1}{p_1}\right)^{\sum_{i \notin s}^{p_1}(y_i-1)} \\
&\quad \times \prod_{i=1}^{P} \frac{\exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + n_i(\gamma_0+\gamma_1 h(y_i))\}}{n_i![1-\sum_{j=1}^{y_i} \exp\{-\exp(\gamma_0+\gamma_1 h(y_i)) + j(\gamma_0+\gamma_1 h(y_i)) - \log(j!)\}]}.
\end{aligned}
$$

By construction, it is straightforward to sample from the proposal distribution.

Letting C denote the proportionality constant, the logged ratio of the target distribution over the proposal distribution is given by

$$\begin{aligned}
&\log\left(\frac{[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1 | X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0, \alpha, \beta, \boldsymbol{\gamma}]}{[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1]_{prop}}\right) \\
&= \log(10C) + \sum_{j=1}^{m} \log\left(\frac{g_{i_j,j}}{\sum_{i=1}^{M-x_0-x_1+p_0+p_1} z_i - \sum_{k=0}^{j-1} z_{i_k}}\right) \\
&\quad - \log\left(\frac{(M-x_0-x_1)!(p_0+p_1)!x_1!}{p_1!}\right) + x_1 \log\left(\frac{\alpha}{1-\alpha}\right) \\
&\quad + \log\left(\frac{1-(1-\beta)^{x_1}}{1-(1-\beta)^{x_0+x_1}}\right) - \sum_{i\in s}^{p_0} (y_i - 1)\log(p_0+p_1) \\
&\quad + \sum_{i\notin s}^{p_1} (y_i - 1)\log\left(\frac{p_1}{p_0+p_1}\right)
\end{aligned}$$

so defining X_1', P_1', $\mathbf{Y}_1'$ and $\mathbf{N}'$ to be the new proposed variables respectively, the Metropolis-Hastings test criterion is

$$\log\left(\frac{[X_1', P_1', \mathbf{Y}_1', \mathbf{N}_1' | X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0, \alpha, \beta, \boldsymbol{\gamma}][X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1]_{prop}}{[X_1', P_1', \mathbf{Y}_1', \mathbf{N}_1']_{prop}[X_1, P_1, \mathbf{Y}_1, \mathbf{N}_1 | X_0, P_0, \mathbf{Y}_0, \mathbf{N}_0, \alpha, \beta, \boldsymbol{\gamma}]}\right).$$

Note that in this expression, the $g_{i_j j}$ are recalculated at each step as they are functions of $\mathbf{Y}_1$.

4 Results

To examine the performance of the Bayesian estimator, we sampled several simulated clustered populations and estimated the population totals from the samples. The population estimates were then compared with the true underlying population totals to evaluate the performance of the model.

Each population was created in a 20×20 grid of $M = 400$ cells to be directly comparable to the populations considered in Thompson (1990). The populations were generated using the assumed model for 16 sets of values (α, β) with $\alpha, \beta \in \{0.05, 0.1, 0.15, 0.2\}$ and $\gamma = (1.0, 0.3)$. Using these fixed parameter values means that the prior distributions were not used to generate the data and allows a limited exploration of the sensitivity of the prior. For each parameter setting, 30 populations were simulated and each was sampled from 5 times by adaptive cluster sampling with the first stage a simple random sample with replacement of size $n = 15$ giving 90 estimates for each parameter setting and $16 \times 150 = 2400$ estimates in total.

Figure 1 shows the difference between the actual and estimated population totals computed using a Beta$(3, 15)$ prior for α, a Beta$(1, 9)$ prior for β and a Gamma$(2, 7)$ prior for γ plotted against against the range of underlying parameter settings. (The horizontal axis in Figure 1 is actually the run number but the ordering is the same as in Figure 2.)

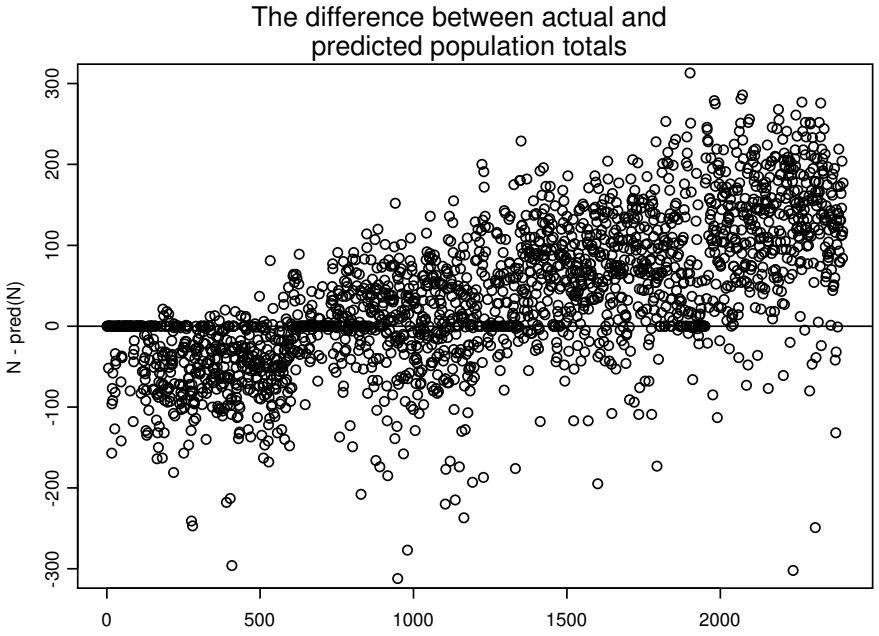

Figure 1: Plot showing the best overall fit under the $\pi(\alpha)$ = Beta$(3, 15)$, $\pi(\beta)$ = Beta$(1, 9)$ and $\pi(\gamma)$ = Gamma$(2, 7)$ priors. The labels on the x-axis give the run number.

The positive gradient in Figures 1 and 2 shows that the performance of the estimator is dependent on the underlying values of α and β (which are increasing from left to right in the plots). For small α and β the population total is overestimated while for large α and β the population total is increasingly underestimated. This makes intuitive sense; we are assuming that we are modelling sparse clustered data and if α and β are large then we have many small clusters which does not accord well with this assumption. (In this case we have a much more evenly distributed population; with larger grid cells the same population could be thought of as randomly scattered single units rather than clusters as the clusters will be of a similar size.) In addition, the steeper the slope, the more sensitive the estimates to the choice of prior. As the variances do not change much with the values of α and β, it is clear that these parameters mainly affect the bias of

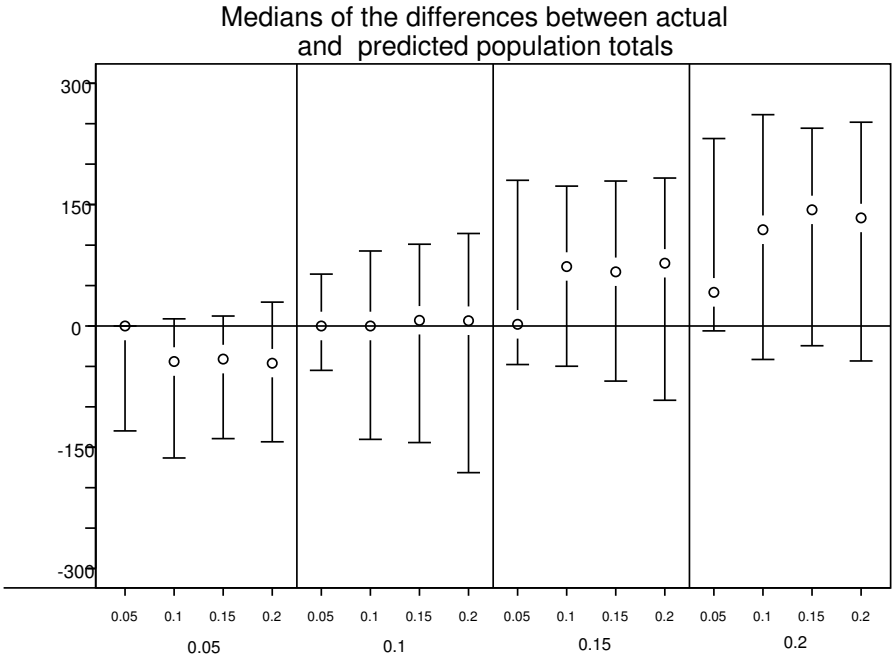

Figure 2: Plot showing the medians of the best overall fit under the $\pi(\alpha) = \text{Beta}(3, 15)$, $\pi(\beta) = \text{Beta}(1, 9)$ and $\pi(\gamma) = \text{Gamma}(2, 7)$ priors. The labels on the x-axis in (a) give the run number and in (b) the top row is α and the bottom row is β. The interval bars represent the 95% confidence interval.

the estimator.

Figure 1 also gives insight into the effect of changing the grid size. The important quantity is the relationship between the size of the cells and the size of the clusters so the effect of increasing the grid size while holding the cluster size fixed is similar to that of holding the grid size fixed and decreasing the cluster size. Letting α and β increase as we have done shows the effect of increasing the grid size; the amount by which the estimates underestimate the total increases as the grid size increases.

5 Prior Choice and Convergence

As we noted in the Introduction, it is generally difficult to carry out adaptive cluster sampling successfully in the complete absence of any prior information about the population of interest because this information is used implicitly in the sampling design. In particular, we need to know that the population is sparse and clustered to justify

using adaptive cluster sampling – if the population is not actually sparse or clustered, then neither our Bayesian approach nor the frequentist Horvitz-Thompson approach described in the next section will out-perform standard non-clustered methods – and we need knowledge of the size of objects, the size and the likely number of clusters of these objects to determine the size, layout (and hence number) of grid cells, a choice which is critical to the success of the approach. The Bayesian analysis we have developed goes further by incorporating this information explicitly into the analysis as well.

The priors used in the next Section are informative, but they are based on information which we feel we already need to ensure the design is meaningful and our choices are relatively robust over the ranges of the underlying parameters α and β we have considered. This means that they can be used even if the only knowledge that can be obtained about the underlying population is that it is sparse and clustered. If a clearer idea of the type or structure of the population can be obtained (i.e. a rough estimate of the expected number of clusters in an area) then better estimates can be achieved for specific values. A full list of the priors explored and when they can be applied can be found in Rapley (2004).

On the other hand, completely non-informative priors under-perform in comparison to the priors we suggest, largely because the kind of sparse, clustered data we are trying to model does not contain much information about the parameters, particularly α. One perspective on varying the prior for α is that it allows us to quantify just how much extra information we have to provide to make the analysis practically useful.

Throughout this work the convergence of the Gibbs samplers was assessed using CUSUM plots Yu and Mykland (1998) as this method can be run using the sample outputs, thereby limiting the additional calculations needed. The 'burn-in' length in all runs was taken to be 2000 as in all cases examined convergence had been reached within 500 samples.

We present two of the plots obtained in our calculations: Figure 3 shows the CUSUM for α and Figure 4 shows the CUSUM for β. The plots are compared to a benchmark plot of independent and identically distributed random normal variates with the mean and variance of the estimated parameters.

The plot comparisons indicate that the sampler has converged. The more quantitative measure of convergence proposed by Brooks (1996) was also calculated. This measure takes values in $[0, 1]$ with a value of 0.5 indicating 'perfect' convergence. In our example the measure for α in 0.501 and for β is 0.536. This again indicates that the samples have converged.

6 Comparison with Thompson (1990)

The idea of this paper is to improve on the population estimates obtained by Thompson through the use of Bayesian model-based inference. To assess the effectiveness of our methodology, we compare results for our approach to those obtained in Thompson (1990) for the design-based approach. We use a reduced version of the setting described

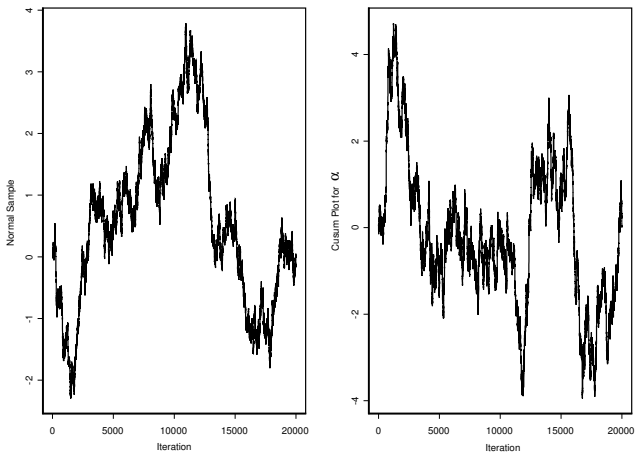

Figure 3: An example of a CUSUM plot for α compared with a plot generated from independent and identically distributed normal variates with the mean and variance estimated from α.

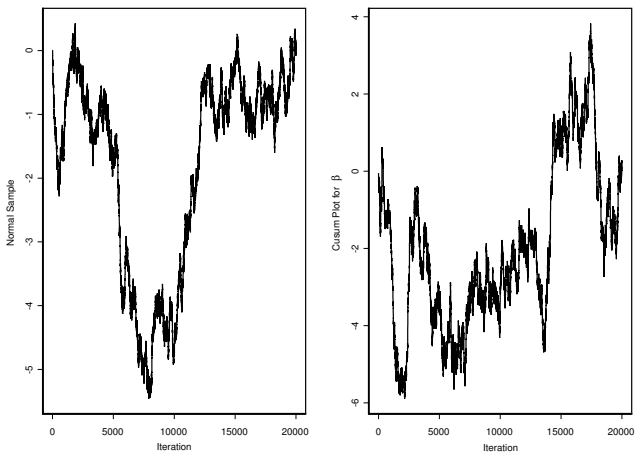

Figure 4: An example of a CUSUM plot for β compared with a plot generated from independent and identically distributed normal variates with the mean and variance estimated from β.

in Section 4 with 16 parameter settings each generating 15 populations from which we draw 2 samples giving 480 different runs. We used the same set of values for α and β but now set $\gamma = (0, 5)$ to obtain larger populations which are more like those used by Thompson (1990). Note that again the prior distributions were not used to generate the populations. We considered an additional factor in that we allowed the first sample (chosen by simple random sampling without replacement) to have a sample size of $n = 10$ or $n = 15$; the second sample was chosen by adaptive sampling as described previously in this paper. We also tried larger initial sample sizes but found that then we frequently observed all the clusters leaving nothing to estimate.

The results in Table 1 below are obtained using the $\pi(\alpha)$ =Beta$(3, 15)$, $\pi(\beta)$ =Beta$(1, 9)$, $\pi(\gamma)$ =Gamma$(2, 5)$ prior distributions. As stated above these priors are not those that produce the best estimates for particular scenarios but are the priors we recommend as useful for a range of typical adaptive cluster sampling situations. The results are averaged over the range of settings.

Sample Size	$\overline{(r_{HT*})}$	$\overline{(r_B)}$	$med(r_B)$	$var(r_{HT*})$	$var(r_B)$	$\widehat{var}(t_{HT*})$	$\widehat{var}(t_B)$
10	-86.53	44.79	35.00	16884.48	8617.66	40510.35	7801.35
15	-47.39	29.86	17.50	13750.54	7264.52	21442.58	8231.73

Table 1: Comparison between design-based and model-based estimates with priors $\pi(\alpha)$ =Beta$(3, 15)$, $\pi(\beta)$ =Beta$(1, 9)$ and $\pi(\gamma)$ =Gamma$(2, 5)$. Here B denotes the Bayesian estimator and $HT*$ the modified Horvitz-Thompson estimator; r denotes the average difference between the true and predicted population totals. Both the mean and median for B are presented to show that the posterior distribution is not symmetric. Also, var is the variance of the difference between the true values and the estimator over all populations and $\hat{var}$ is the average variance of the estimator calculated from each sample.

The model-based method produces estimates with smaller actual variances than the design-based method of Thompson (1990). The estimator is also on average closer to the true values in these particular cases so has smaller mean squared error than the design-based estimator. We have not presented the efficiency statistic given in Thompson (1990) as the variances are large in comparison to the very small variance of the expansion estimator. The expansion estimator gives small variances in all cases but the population estimates are so biased as to make the method useless under adaptive cluster sampling. The average mean squared error of the expansion estimator of the population total is 48866, while the average variance is 1.458.

We can also compare the coverage properties of confidence intervals based on the Horvitz-Thompson estimates and credibility intervals based on the model-based estimates. In both cases, we use normal approximations to set the intervals and report the empirical coverage, namely the proportion of times the population total lies within the

interval. This is the standard method of constructing design-based confidence intervals based on the Horvitz-Thompson estimator but it is not the recommended method of constructing Bayesian credibility intervals; using the MCMC output to estimate posterior quantiles directly. However, due to the requirement for a direct comparison to Horvitz-Thompson and the computationally intensive nature of the calculations for multiple populations, we feel the normal approximation is justified for this comparison. Our results for the Bayesian credibility intervals are therefore only suggestive and should be interpreted cautiously.

n		10		15	
Standard Deviation	Nominal Level	$HT*$	B	$HT*$	B
1	0.67	0.87	0.52	0.79	0.63
2	0.95	1.00	0.81	0.97	0.86
3	0.99	1.00	0.89	1.00	0.91

Table 2: Coverage properties of the approximate confidence interval based on the Horvitz-Thompson estimator and the approximate Bayesian credibility interval. Both intervals are based on normal approximations. The estimated coverage is the proportion of times the population total lies within an interval of half-length the stated number of standard deviations.

Table 2 shows that the Horvitz-Thompson confidence interval has much higher coverage than the nominal level, whereas the Bayesian credibility interval has lower coverage. The Horvitz-Thompson intervals are on average 2.11 times larger than the Bayes intervals when $n = 10$ and 1.32 times larger when $n = 15$. The Horvitz-Thompson estimator has large variance but the design-based estimates of the variance tend to be quite conservative (particularly with small n) so the intervals are larger than they need be and therefore have higher than nominal coverage levels. On the other hand, the Bayesian estimator has smaller variance which is estimated quite well. This suggests that the observed lower than nominal coverage levels are largely due to the weakness of the normal approximation to the posterior distribution and that more sophisticated credibility intervals would perform better.

For balance, we also include a comparison of our estimator with the Horvitz-Thompson estimator when a prior which is not robust to parameter changes is used; in other words, a prior which gives poor results for some parameter combinations. The Beta$(1, 25)$ prior for α puts more weight on small values of α than our recommended Beta$(3, 15)$ prior. The effect of this change is to increase the bias in the Bayesian estimates, leave their variability unchanged but decrease the estimate of the variance so that it is over-optimistic. The properties of the Horvitz-Thompson have changed only slightly showing that the inter-simulation variability is small. While the Horvitz-

Thompson estimator on average produces less-biased estimates of the population total, it still has a larger variance than our procedure.

Sample Size	$\overline{(r_{HT*})}$	$\overline{(r_B)}$	$med(r_B)$	$var(r_{HT*})$	$var(r_B)$	$\widehat{var}(t_{HT*})$	$\widehat{var}(t_B)$
15	-44.58	89.61	77.50	21480.4	1314.2	14080.44	8257.69

Table 3: Comparison between design-based and model-based estimates with priors $\pi(\alpha)$ =Beta$(1, 25)$ and $\pi(\beta)$ =Beta$(1, 9)$. Here B denotes the Bayesian estimator and $HT*$ the modified Horvitz-Thompson estimator; r denotes the average difference between the true and predicted population totals. Both the mean and median of B are presented to show that the posterior distribution is asymmetric. Also, var is the variance of the difference between the true values and the estimator over all populations and $\hat{var}$ is the average variance calculated from each sample.

Other calculations (not reported here) show that the modified Horvitz-Thompson estimator improves its performance when used with populations in which there are more small networks. This makes sense as it is based on simple random sampling so the closer the population comes to being randomly distributed rather than clustered the better it will perform. This is in contrast to our methodology which assumes that there are a few networks and so does better in clustered cases.

Summarizing our empirical results, our estimator out-performs the modified Horvitz-Thompson estimator using a general prior and can significantly out-perform with more specific priors. However, if a poor choice of prior is made the results can under-perform in comparison to Horvitz-Thompson. Thus, if we have an accurate idea of the size and structure of the population which we can incorporate into our priors then we can predict very well. If we do not have this information then we can seriously underestimate the population total and it would be better to restrict our inference to that using the general prior described above. Note that the choice of the appropriate grid size to place on the region also requires knowledge of the population so neither procedure can be expected to perform well in the absence of such prior knowledge. Even with the worst choice of prior our estimates are still as good (in terms of mean squared error) as those used previously to estimate population totals from this kind of data.

7 Conclusion

We have considered the problem of estimating the number of individuals in a rare, clustered population in a specified region. We impose a regular grid on the region and enumerate the number of individuals within grid cells selected by adaptive cluster sampling as described by Thompson (1990). Our approach is to model the observed counts in the selected grid cells and then use a model-based analysis based on this model. The resulting likelihood involves integrals which we cannot evaluate explicitly

so we used Markov Chain Monte Carlo (MCMC) methods to obtain the estimators. The methodology is compared to the design-based modified Horvitz-Thompson estimator of Thompson (1990) by means of a small simulation study. The performance of our estimator can be substantially better than that of the modified Horvitz-Thompson estimator when good prior information is available.

References

A.H. Welsh, C. D., R.B. Cunningham and Lindenmayer, D. (1996). "Modelling the abundance of rare species - statistical models for counts with extra zeros." *Ecological Modelling*, 88: 297–308. 719, 723

Breckling, J., Chambers, R., Dorfman, A., Tang, S., and Welsh, A. (1994). "Maximum likelihood inference from sample survey data." *International Statistical Reviews*, 62: 349–363. 719, 720, 723

Brewer, K. (1963). "Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process." *Australian Journal of Statistics*, 5: 93–105. 718

Brooks, S. (1996). "Quantitative convergence diagnostics for MCMC via CUSUMS." *Technical Report, University of Bristol*. 730

Brus, D. and de Gruijter, J. (1997). "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion)." *Geoderma*, 80: 1–44. 718

G. Kaufman, Y. B. and Kruyt, D. (1975). "A probabilistic model of oil and gas discovery." In *Estimating the Volume of Undiscovered Oil and Gas Resources*, 109–117. Tulsa, Oklahoma: American Association of Petroleum Geologists. 719

Heilbron, D. (1994). "Zero-altered and other models for count data with added zeros." *Biometrical Journal*, 36: 531–547. 719, 723

Mullahy, J. (1986). "Specification and testing of some modified count data models." *Journal of Econometrics*, 33: 341–365. 719, 723

Nair, V. and Wang, P. (1989). "Maximum likelihood estimation under a successive sampling discovery model." *Technometrics*, 31: 423–436. 719, 721

R. Valliant, A. D. and Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons. 718

Rapley, V. (2004). *Model-Based Adaptive Cluster Sampling, PhD Thesis*. Univerity of Southampton. 730

Royall, R. (1976). "Current advances in sampling theory:Implications for human observational studies." *American Journal of Epidemiology*, 104: 463–477. 718

Tanner, M. (1996). *Tools for Statistical Inference*. New York: Springer, 3 edition. 719

Thompson, S. (1990). "Adaptive cluster sampling." *Journal of the American Statistical Association*, 85: 1050–1059. 717, 718, 719, 720, 727, 730, 732, 734, 735

— (1991a). "Adaptive cluster sampling: Designs with primary and secondary units." *Biometrics*, 47: 1103–1115. 717

— (1991b). "Stratified adaptive cluster sampling." *Biometrika*, 78: 389–397. 717

— (1992). *Sampling*. New York: John Wiley & Sons. 717

— (2002). "On sampling and experiments." *Environmetrics*, 13: 429–436. 717, 718

Thompson, S. and Seber, G. (1996). *Adaptive Sampling*. New York: John Wiley & Sons. 717

West, M. (1996). "Inference in successive sampling discovery models." *Journal of Econometrics*, 75: 217–238. 719, 721

Yu, B. and Mykland, P. (1998). "Looking at Markov samplers through CUSUM path plots: a simple diagnostic idea." *Statistics and Computing*, 8: 275–286. 730

Acknowledgments

We are grateful to The Editor and Referees for their helpful suggestions. We are also grateful to Jon Forster for very helpful discussions on modelling and the use of MCMC; in particular, for pointing out the effect of the discreteness of network counts on modelling the sampling process. We are also grateful to the Southampton Statistical Sciences Research Institute (S3RI) and School of Mathematics for providing funding to allow VER to visit the Centre for Mathematics and its Applications (CMA) to complete the writing of this paper and to the CMA for providing accommodation.