Forming Post-strata Via Bayesian Treed Capture-Recapture Models

Xinlei Wang, Johan Lim and Lynne Stokes*

September 2004

Abstract

For the problem of dual system estimation, we propose a treed Capture Recapture Model (CRM) to account for heterogeneity of capture probabilities where individual auxiliary information is available. A treed CRM uses a binary tree to partition the covariate space into "homogeneous" regions, within each of which the capture response can be described adequately by a simple model that assumes equal catchability. In this paper, a Bayesian approach is presented to fit and search promising treed CRMs. We compare the performance of estimators based on this approach to those of alternative models in three examples. The attractive features of the proposed models include reduction of correlation bias, robustness, practical flexibility as well as simplicity and interpretability. In addition, they provide a systematic and effective way to form post-strata for the Sekar and Deming estimator of population size.

Keywords: Binary trees; Bayesian model selection; Dual system estimation; Closed population; Heterogeneity; Parallel tempering.

**Xinlei Wang is Assistant Professor and Lynne Stokes is Professor, Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, Texas 75275-0332, swang@mail.smu.edu and lstokes@mail.smu.edu. Johan Lim is Assistant Professor, Department of Statistics, Texas A & M University, johanlim@stat.tamu.edu.*

1 Introduction

The problem we consider is that of estimating the size of a closed population from capture-recapture data when capture probabilities are heterogeneous. We restrict attention to the case of two capture periods, as in Census undercount estimation. In this context, the problem is known as dual system estimation.

Under the classical model for capture-recapture data, the maximum likelihood estimator (MLE) for population size N is $\hat{N} = n_1 n_2/m$ where n_1 and n_2 are the sample sizes in each of the two capture periods and m is the number of previously captured individuals captured in the second period. This estimator is asymptotically unbiased for N under the assumption of equal catchability, but asymptotically biased if capture probabilities vary in both capture periods. The bias is negative when capture probabilities in the first and second capture periods are positively correlated (Alho et al. 1993), as is typically true in real populations. As a result, this bias is sometimes called correlation bias.

The presence of unequal catchability and the problems it causes for estimation of N have been long recognized in both animal and human populations (e.g., Young et al. 1952, Wilbur & Landwehr 1974, Sekar & Deming 1949). Several alternative estimation approaches have been proposed for the two-capture period case. Among the earliest was that of Sekar & Deming (1949), who suggested a method whose basic idea is still used by the U.S. Bureau of the Census today in undercount estimation. Their approach is to divide the captured individuals into post-strata that are believed to be more homogeneous with respect to capture probabilities than the entire population; then make separate estimates of population sizes within the post-strata and sum them up. That is, compute $\hat{N}_{SD} = \sum_i \hat{N}_i$, where $\hat{N}_i$ is the MLE of population size in the i-th stratum. If the post-stratification is perfectly effective so that the capture probabilities are equal within each stratum, then $\hat{N}_{SD}$ would be asymptotically unbiased. However, there is little guidance in the literature about how to choose a post-stratification scheme. Sekar and Deming suggest that "...the population need be divided only to the stage when further division shows no increase in $\hat{N}$...". Though it appears reasonable, their method can result in overstratification because an increase in $\hat{N}$ can always be expected by randomly dividing sampled individuals into more groups (Appendix A).

More recently, there have been several methods proposed that model the selection probabilities as regression functions whose coefficients are estimated from the data of the individuals captured at least once (Alho 1990, Huggins 1991, Pollock et al. 1984, Pollock 2002). The estimated probabilities are then used to form a kind of Horvitz-Thompson estimator of N. These regression methods have been shown to reduce correlation bias, as well as to provide useful information about which individuals are easy and which are difficult to capture (Alho et al. 1993).

In this paper, we propose a new class of models allowing heterogeneous capture probabilities. We refer to our models as Bayesian Treed Capture Recapture Models (BTCRM). As with the approaches using regression functions, our method requires that auxiliary information about each captured individual is available, and that these covariates are potentially related to that individual's capture probabilities. There are two key features that make our models different from past work. First, given a set of covariates X, instead of using a regression setup, our models use binary trees to partition the domain of X, denoted $\mathcal{X}$, into subsets. Within each of the subsets, individuals are assumed to have equal capture rates, so the classical model can adequately describe the distribution of the capture history Y. Such treed models, though simple, are flexible and robust in practice (Breiman et al. 1984, Hastie et al. 2001). This is because a series of binary splits can achieve good approximation to many different forms of functions, including non-smooth relationships that are difficult to describe by a standard regression model. Second, our approach to finding a good tree is a supervised learning process guided by the information in the observed capture history y. Intrinsically, it incorporates Bayesian model selection to account for model uncertainty. Rather than use ad hoc penalty criteria for ranking models, the Bayesian approach coherently ranks trees by the posterior distribution. MCMC is used to search for high posterior trees. Due to its stochastic nature, MCMC can readily find systematic structures that tend to be overlooked by a short-sighted greedy search. The use of the BTCRM could be easily integrated into the Census undercount estimation by using a tree chosen based on the posterior to define the post-strata for $\hat{N}_{SD}$.

The idea of such Bayesian treed models was pioneered by Chipman et al. (1998) (hereafter CGM 1998) and Denison et al. (1998) to find and fit CART models, which partition $\mathcal{X}$

into regions where $E(Y|X)$ is constant. Chipman et al. (2002, 2003) (hereafter CGM 2002 and 2003) extended it to treed regressions and treed GLMs, where linear regression and generalized linear models were used to describe the variation within each subset of the partition, respectively. What distinguishes our work from these earlier treed models is that we consider a capture-recapture model with homogeneous capture probabilities at each terminal node, and this induces considerable differences in prior specification and posterior calculation.

The remainder of the paper is organized as follows. Section 2 introduces BTCRMs in a general mathematical framework. In Section 3, we propose several prior distributions for trees and capture probabilities and address the problem of computing the marginal densities of the data. A Metropolis-Hastings (MH) algorithm with parallel tempering is presented to search for promising trees, and discussed in Section 4. The potential of our proposed BTCRMs is illustrated through three examples in Section 5, where we also compare their predictive performance in estimating N with competing models on an actual dataset from the 1990 Census. Section 6 concludes with discussion.

2 The Bayes Setup for Treed CRMs

We begin by discussing the general structure of a treed model that describes the conditional distribution of a random vector of interest Y given a set of known covariates X. Typically, a binary tree is used to divide the covariate space $\mathcal{X}$ into regions where a single parametric model for $Y|X$ is adequate. Consider a tree T with b terminal nodes denoted $T_1, \ldots, T_b$, corresponding to the distinct regions of a partition of $\mathcal{X}$. Let $(\mathbf{X}_i, \mathbf{Y}_i)$ denote the observations assigned to the $i-$th terminal node T_i, and $\mathbf{Y}_i|\mathbf{X}_i \sim f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta}_i)$ where $\boldsymbol{\theta}_i$ is the parameter vector associated with T_i. Define $\mathbf{X} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_b)'$, $\mathbf{Y} \equiv (\mathbf{Y}_1, \ldots \mathbf{Y}_b)'$ and $\Theta \equiv (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_b)'$. Assuming y values across terminal nodes are independent, the treed model distribution of the data will be of the form

$$f(\mathbf{Y}|\mathbf{X}, \Theta, T) = \prod_{i=1}^{b} f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta}_i). \tag{1}$$

A Bayesian solution to model uncertainty problems for the general treed setup in (1) proceeds as follows. We seek priors of the form

$$\pi(\Theta, T) = \pi(\Theta|T)\pi(T) = \prod_{i=1}^{b} \pi(\boldsymbol{\theta}_i) \cdot \pi(T). \tag{2}$$

so that $\boldsymbol{\theta}_i$'s across terminal nodes are assumed a priori independent. Such prior distributions lead to posterior distributions over T of form

$$f(T|\mathbf{X}, \mathbf{Y}) \propto f(\mathbf{Y}|\mathbf{X}, T)\pi(T), \tag{3}$$

where

$$f(\mathbf{Y}|\mathbf{X}, T) \;=\; \prod_{i=1}^{b} \int f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i \tag{4}$$

is the marginal density of $\mathbf{Y}$ given T. Then an MCMC method is used to simulate a sample from $f(T|\mathbf{X}, \mathbf{Y})$, which tends to gravitate toward trees with high posterior probabilities.

Under this general framework, there have previously been three classes of Bayesian treed models proposed. The difference in these models lies in what kind of terminal node distribution they consider. For example, for CART models, the conditional distribution of $Y|X$ under the terminal node T_i is given by a normal distribution with constant mean and variance, i.e., $\mathbf{N}(\mu_i, \sigma_i^2)$; for treed regression models, it is given by $\mathbf{N}(x\boldsymbol{\beta}_i, \sigma_i^2)$, where $\boldsymbol{\beta}_i$ is the vector of regression coefficients associated with T_i. In this paper, we introduce a new class of treed models that are Bayesian Treed Capture Recapture models, by considering a conditional multinomial as the terminal node distribution. Once a tree is given, each individual in the population under consideration is assigned to a terminal node. Let N_i be the unknown number of individuals assigned to T_i in the population, $i = 1, \ldots, b$. For the $k-$th individual assigned to T_i, define indicator variables u_{ijk} and m_{ik} for $k = 1, \ldots, N_i$,

$$u_{ijk} \;=\; \begin{cases} 1, & \text{if individual } k \text{ is captured on occasion } j \text{ only, } j = 1, 2; \\ 0, & \text{otherwise;} \end{cases}$$

$$m_{ik} \;=\; \begin{cases} 1, & \text{if individual } k \text{ is captured twice;} \\ 0, & \text{otherwise.} \end{cases}$$

Define $M_{ik} = u_{i1k} + u_{i2k} + m_{ik}$, $u_{ij} = \sum_k u_{ijk}$, $m_i = \sum_k m_{ik}$, $n_{ij} = \sum_k n_{ijk}$, and $M_i = u_{i1} + u_{i2} + m_i$. Define p_{ij} as the capture probability on the $j-$th occasion for all individuals assigned to T_i. Then we have the following generalized Bernoulli (GB) model for each

individual under T_i:

$$(u_{i1k}, u_{i2k}, m_{ik}, 1 - M_{ik}) \sim \text{GB}\,(1, p_{i1}(1-p_{i2}), (1-p_{i1})p_{i2}, p_{i1}p_{i2}, (1-p_{i1})(1-p_{i2})) \tag{5}$$

Assuming that the captures of different individuals are independent, the *conditional likelihood* given the observed individuals under T_i (i.e., those with $M_{ik} = 1$) is

$$f(\mathbf{Y}_i|\boldsymbol{\theta}_i) = \frac{p_{i1}^{u_{i1}+m_i}(1-p_{i1})^{u_{i2}}p_{i2}^{u_{i2}+m_i}(1-p_{i2})^{u_{i1}}}{[1-(1-p_{i1})(1-p_{i2})]^{u_{i1}+u_{i2}+m_i}} \tag{6}$$

Here, $\mathbf{Y}_i \equiv (u_{i1}, u_{i2}, m_i)$ and $\boldsymbol{\theta}_i \equiv (p_{i1}, p_{i2})$. As in the CART model, (6) does not depend on $\mathbf{X}_i$. This reduces $f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta}_i)$ to $f(\mathbf{Y}_i|\boldsymbol{\theta}_i)$, $f(\mathbf{Y}|\mathbf{X}, T)$ to $f(\mathbf{Y}|T)$, and $f(T|\mathbf{X}, \mathbf{Y})$ to $f(T|\mathbf{Y})$.

The practical value of the introduced BTCRMs relies heavily on the flexibility of the prior distributions and the ease of posterior computation. In the next section, we address both of these issues for various prior settings on Θ.

3 Prior Specifications and Marginal Densities

For the tree prior $\pi(T)$, we note the specification in CGMs (i.e., CGM 1998, 2002 and 2003) is independent of the terminal node models. Hence, it is sufficiently general for all kinds of tree structures. In this paper, we adopt their specification because of its flexibility and ease of implementation. For completeness, a brief description is provided below.

Unlike classical prior specifications, the CGM version of $\pi(T)$ has no closed-form density. Instead, it is implicitly defined by a tree-generating stochastic process which is controlled by two functions $f_{split}(\eta, T)$ and $f_{rule}(\rho|\eta, T)$. $f_{split}(\eta, T)$ specifies the probability that terminal node η of tree T is split and $f_{rule}(\rho|\eta, T)$ specifies the probability of assigning splitting rule ρ to η when the tree is split. A general form for $f_{split}(\eta, T)$ is given by $\alpha(1+d_\eta)^{-\beta}$, where d_η is the depth of the node η and $\beta \geq 0$. This allows for controlling the size and the shape of the generated trees through (α, β): a large α will tend to grow larger trees and a large β will make deeper nodes less likely to split. A default choice of $f_{rule}(\rho|\eta, T)$, which will be used throughout this paper, is a prior distribution that is uniform on all available X variables and within a given variable, uniform on all possible splits for that variable. For a discussion about a variety of $f_{rule}(\rho|\eta, T)$ specifications or more details about $\pi(T)$, see CGM 1998.

We now proceed to choose the conditional prior $\pi(\Theta|T)$ on the parameter space and calculate the marginal density of the data $f(\mathbf{Y}|T)$. As in CGMs, one simplifying and reasonable assumption we take is that the components $\boldsymbol{\theta}_i = (p_{i1}, p_{i2})$ of Θ are a priori independently and identically distributed (IID). This reduces our consideration of $\pi(\Theta|T)$ to a unified $\pi(p_{i1}, p_{i2})$ for every terminal node. Such consideration must confront the difficulty that the integrations in (4) are often analytically intractable. In what follows, we discuss three different priors on capture probabilities p_{ij} for a given terminal node T_i to accommodate different needs in practice; for each of the priors, we derive the corresponding approximate representation for the marginal $f(\mathbf{Y}|T)$, based on the Laplace method or its modification, which leads to analytical tractability and computational simplicity.

Jeffreys prior for p_{ij}

When there is no real prior information about p_{ij}, as will frequently be the case in practice, a noninformative prior is needed. Also in some applications, especially politically charged ones such as Census undercount estimation, one might prefer objective and automatic methods that require as little human decision-making as possible, such as setting of tuning parameters. In this case as well, a noninformative prior would be desirable. A widely used method to derive a noninformative prior is that of Jeffreys (1961), which is to choose $\pi(\boldsymbol{\theta}) \propto [\det \mathbf{I}(\boldsymbol{\theta})]^{1/2}$, where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix $-E_{\boldsymbol{\theta}}[\partial^2 \log f(\mathbf{Y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2]$. As noted in Berger (1985), an attractive feature of the Jeffreys prior is that, when dealing with restricted parameter spaces, one of the situations where noninformative priors are useful, it is not affected by the restriction. From the terminal node distribution (6) of T_i, it is straightforward to obtain $\mathbf{I}(\boldsymbol{\theta}_i) \propto [I_{kl}]_{2\times 2}$ where

$$\begin{aligned}
I_{11} &= 1/p_{i1} + p_{i2}/(1-p_{i1}) - (1-p_{i2})^2/(p_{i1}+p_{i2}-p_{i1}p_{i2}),\\
I_{12} &= I_{21} = -1/(p_{i1}+p_{i2}-p_{i1}p_{i2}),\\
I_{22} &= 1/p_{i2} + p_{i1}/(1-p_{i2}) - (1-p_{i1})^2/(p_{i1}+p_{i2}-p_{i1}p_{i2}).
\end{aligned}$$

So the Jeffreys prior of $\boldsymbol{\theta}_i = (p_{i1}, p_{i2})$, denoted by π^J, has the form

$$\pi^J(p_{i1}, p_{i2}) \propto (1-p_{i1})^{-1/2}(1-p_{i2})^{-1/2}[1-(1-p_{i1})(1-p_{i2})]^{-1/2}. \tag{7}$$

With some calculus, we can show π^J is proper so that it is well suited for our Bayesian treed model selection. Then the marginal prior density on p_{ij} is

$$\pi^J(p_{ij}) = \frac{1}{\pi \log 2} \frac{\arcsin \sqrt{1-p_{ij}}}{1-p_{ij}}, \quad \text{for} \quad j = 1, 2$$

and the normalizing constant for π^J is indeed $2\pi \log 2$. This constant cannot be ignored when comparing trees with different numbers of terminal nodes.

Under the prior (7) and the likelihood (6) for T_i, we express the integral in (4) as

$$f(\mathbf{Y}|T) = \prod_{i=1}^{b} \left\{ K \cdot \int_0^1 \int_0^1 \exp\left[w_i(p_{i1}, p_{i2})\right] \, \mathrm{d}p_{i1} \mathrm{d}p_{i2} \right\} \tag{8}$$

where $K = 1/(2\pi \log 2)$, and $w_i(p_{i1}, p_{i2})$ is given by

$$w_i = A_{i1} \log p_{i1} + B_{i1} \log(1-p_{i1}) + A_{i2} \log p_{i2} + B_{i2} \log(1-p_{i2}) - C_i \log(p_{i1} + p_{i2} - p_{i1}p_{i2}) \tag{9}$$

and $A_{i1} = n_{i1}$, $B_{i1} = u_{i2} - 1/2$, $A_{i2} = n_{i2}$, $B_{i2} = u_{i1} - 1/2$, and $C_i = M_i + 1/2$. To solve these analytically intractable integrations in (8), we resort to the Laplace method to approximate $f(\mathbf{Y}|T)$ (Tierney & Kadane 1986, Smith 1991, Kass & Raftery 1995, CGM 2002, etc.):

$$\tilde{f}(\mathbf{Y}|T) = \prod_{i=1}^{b} \left\{ 2\pi K \cdot \left| -\widetilde{\mathbf{H}}_i \right|^{-1/2} \cdot \exp\left[w_i(\tilde{p}_{i1}, \tilde{p}_{i2})\right] \right\} \tag{10}$$

and

$$\tilde{f}(\mathbf{Y}|T) = f(\mathbf{Y}|T) \cdot [1 + O(1/\min_i M_i)]. \tag{11}$$

In (10), $\mathbf{H}_i = w_i''(p_{i1}, p_{i2})$ and $\widetilde{\mathbf{H}}_i$ is $\mathbf{H}_i$ evaluated at $(\tilde{p}_{i1}, \tilde{p}_{i2})$ where w_i peaks; $(\tilde{p}_{i1}, \tilde{p}_{i2})$ satisfies $\partial w_i / \partial p_{ij} = 0, \; j = 1, 2$, which can be simplified to the following two equations:

$$\begin{aligned} (A_{i1} + B_{i1} - C_i)(A_{i2} + D_i)p_{i1}^2 - [(A_{i1} + A_{i2})(A_{i1} + B_{i1} - C_i) & \\ + A_{i1}A_{i2} + D_i(A_{i1} - C_i)\,]\, p_{i1} + A_{i1}(A_{i1} + A_{i2} - C_i) & = 0 \qquad (12) \\ A_{i1}/p_{i1} - A_{i2}/p_{i2} - D_i & = 0 \qquad (13) \end{aligned}$$

where $D_i = (A_{i1} + B_{i1}) - (A_{i2} + B_{i2})$ is reduced to zero here. Obviously, (12) and (13) can be analytically solved. Further, it is not hard to show the following results for any T_i:

1. For any $u_{i1} \geq 0$, $u_{i2} \geq 0$ and $m_i \geq 0$, $\int_0^1 \int_0^1 \exp\left[w_i(p_{i1}, p_{i2})\right] \, \mathbf{d}p_{i1} \mathbf{d}p_{i2}$ exists.

2. If $u_{i1} > 0$, $u_{i2} > 0$ and $m_i > 0$, (12) and (13) have a single pair of roots in the region $[0,1]^2$, and w_i achieves the global maximum at these roots.
3. If $M_i > 0$ but $u_{i1}\, u_{i2}\, m_i = 0$, then the global maximum of w_i is $+\infty$ achieved at the boundary of the region $[0,1]^2$. In this case, the Laplace approximation (10) cannot be applied, so a 2-dimension numerical integration need to be performed.

It is notable that the computation required for the marginal via (10) is indeed minimum and the results are often surprisingly accurate.

Beta prior for p_{ij}

There exist situations where prior information about capture probabilities is available from direct knowledge or previous surveys. For example, field biologists may know some species are hard to catch so want to concentrate the prior on small capture probabilities. Such information can be incorporated into the analysis by using an informative prior. Following Castledine (1981) and George & Robert (1992), we consider the beta prior, denoted π^B, in which p_{i1} and p_{i2} are a priori independent, namely

$$p_{ij} \sim^{\text{IID}} \text{Beta}(a_j, b_j), \text{ for } i = 1, \ldots b;\ j = 1, 2. \tag{14}$$

This reduces the choice to two priors $\text{Beta}(a_1, b_1)$ and $\text{Beta}(a_2, b_2)$. Both Castledine (1981) and George & Robert (1992) assume the priors of capture probabilities for different capture occasions are exchangeable so that only one prior is needed. The reason that we allow separate priors for the two capture occasions is the experimenter sometimes expends more sampling effort on one occasion than the other. A typical example is U.S. Census undercount estimation, where the first occasion is the well-publicized Census operation and the second is a less extensive sampling operation.

Computing the marginal $f(\mathbf{Y}|T)$ under π^B using the Laplace approximation is essentially the same as that under π^J. Formulas (8), (11), (12) and (13) hold for π^B if we redefine the related quantities in this way: $K = \Gamma(a_1 + b_1)\Gamma(a_2 + b_2)/\{\Gamma(a_1)\Gamma(b_1)\Gamma(a_2)\Gamma(b_2)\}$, $A_{i1} = n_{i1} + a_1 - 1$, $B_{i1} = u_{i2} + b_1 - 1$, $A_{i2} = n_{i2} + a_2 - 1$, $B_{i2} = u_{i1} + b_2 - 1$, and $C_i = M_i$. Similarly, we have the following results for T_i under π^B:

1. For any $u_{ij} \geq 0$, $m_i \geq 0$ and $a_j > 0$, $b_j > 0$, $j = 1, 2$, $\int_0^1 \int_0^1 \exp(w_i)\ \mathrm{d}p_{i1}\mathrm{d}p_{i2}$ exists.
2. If $A_{i1} > 0$, $B_{i1} > 0$, $A_{i2} > 0$, $B_{i2} > 0$ and $m_i + a_1 + a_2 > 2$, (12) and (13) have a single pair of roots in the region $[0, 1]^2$, where w_i achieves the global maximum. In most cases, these conditions hold so the Laplace approximation (10) can be applied easily.
3. If any of A_{i1}, B_{i1}, A_{i2} or $B_{i2} \leq 0$ or $m_i + a_1 + a_2 \leq 2$, then the global maximum of w_i is $+\infty$ and achieved at the boundary of the region $[0, 1]^2$. In this case, a numerical integration or a sampling-based integration is necessary.

The hyperparameters (a_j, b_j) $(j = 1, 2)$ can be flexibly chosen to incorporate available subjective prior information into our treed models. A special case is $a_j = b_j = 1$ where (14) becomes a flat prior so can be used as an alternative to noninformative priors. On the other hand, one can obtain these hyperparameters from prior predictions if historical data exist (Meyer & Laud, 2002). At this stage, the treed CRM is not necessarily used. For example, one might build a conditional logistic regression model (Alho, 1990) on historical data using the common covariates in both studies, and use this model to predict the capture probabilities of the individuals observed in the current study; then a natural way of specifying (a_j, b_j) is that, for each j, let the prior mean of p_{ij} be the sample mean of the predicted probabilities at the $j-$th occasion, let the prior variance be the corresponding sample variance, and solve for a_j and b_j. For historical data containing no covariates, an analysis based on the homogeneous assumption can still guide the choice of (a_j, b_j). In this case, it may be reasonable to simply set the prior mean of p_{ij} at $\hat{p}_j$ estimated from historical data because the capture probabilities in different strata may balance so that $\hat{p}_j$ still provides a rough estimate of the prior mean; however, one should select the prior variance larger than the variance estimated under the homogeneous assumption to allow a reasonable spread in the interval [0,1].

Normal prior for logit(p_{ij})

In many practical cases, capture probabilities in the first and second capture periods are correlated and people may have prior information about the correlation structure of (p_{i1}, p_{i2}), obtained from previous studies or expert knowledge. Whenever available, such information should be used to help reduce the correlation bias in estimating N. Although flexible in

shape, the beta prior assumes independence between p_{i1} and p_{i2}. That's why we further consider the normal prior based on the logit transformation (Castledine 1981), which can easily incorporate a correlation structure. Define $\psi_{ij} = \log[p_{ij}/(1-p_{ij})]$ and $\boldsymbol{\psi}_i = (\psi_{i1}, \psi_{i2})^T$, and consider the bivariate normal prior on $\boldsymbol{\psi}_i$, denoted π^N, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Under this normal prior, we express the integral in (4) as

$$f(\mathbf{Y}|T) = \prod_{i=1}^{b} \left\{ L \cdot \int_{R^2} \exp\left[v_i(\boldsymbol{\psi}_i) + \lambda_i(\boldsymbol{\psi}_i)\right] \, \mathrm{d}\boldsymbol{\psi}_i \right\}, \tag{15}$$

where $L = 1/(2\pi|\boldsymbol{\Sigma}|^{1/2})$ and

$$\begin{aligned} v_i(\boldsymbol{\psi}_i) &= n_{i1}\psi_{i1} + n_{i2}\psi_{i2} - M_i \log\left[e^{\psi_{i1}} + e^{\psi_{i2}} + e^{\psi_{i1}+\psi_{i2}}\right], \\ \lambda_i(\boldsymbol{\psi}_i) &= -(\boldsymbol{\psi}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\psi}_i - \boldsymbol{\mu})/2. \end{aligned}$$

Applying the Laplace method to (15) yields

$$\tilde{f}(\mathbf{Y}|T) = \prod_{i=1}^{b} \left\{ 2\pi L \cdot \left| -\widetilde{\mathbf{G}}_i \right|^{-1/2} \cdot \exp\left[v_i(\widetilde{\boldsymbol{\psi}}_i) + \lambda_i(\widetilde{\boldsymbol{\psi}}_i) \right] \right\}, \tag{16}$$

where $\widetilde{\boldsymbol{\psi}}_i$ is the posterior mode that maximizes $v_i + \lambda_i$, $\mathbf{G}_i = v_i'' + \lambda_i''$ and $\widetilde{\mathbf{G}}_i$ is $\mathbf{G}_i$ evaluated at $\widetilde{\boldsymbol{\psi}}_i$. However, unlike $(\tilde{p}_{i1}, \tilde{p}_{i2})$ under π^J or π^B, $\widetilde{\boldsymbol{\psi}}_i$ cannot be analytically solved from the first-order conditions. To avoid the need for calculating $\widetilde{\boldsymbol{\psi}}_i$ via an optimization algorithm, we take an approach similar to that of Raftery (1996), which substantially reduces the computational burden of tree posterior exploration. First apply the one-step Newton's method to obtain

$$\widetilde{\boldsymbol{\psi}}_i \approx \widehat{\boldsymbol{\psi}}_i + \widehat{\mathbf{G}}_i^{-1} \boldsymbol{\Sigma}^{-1} \cdot (\widehat{\boldsymbol{\psi}}_i - \boldsymbol{\mu}), \tag{17}$$

where $\widehat{\boldsymbol{\psi}}_i$ is the MLE of $\boldsymbol{\psi}_i$ and $\widehat{\mathbf{G}}_i$ is $\mathbf{G}_i$ evaluated at $\widehat{\boldsymbol{\psi}}_i$, namely

$$\widehat{\boldsymbol{\psi}}_i = \begin{pmatrix} \log(m_i/u_{i2}) \\ \log(m_i/u_{i1}) \end{pmatrix}, \quad \widehat{\mathbf{G}}_i = \begin{pmatrix} -u_{i2}n_{i1}/M_i & u_{i1}u_{i2}/M_i \\ u_{i1}u_{i2}/M_i & -u_{i1}n_{i2}/M_i \end{pmatrix} - \boldsymbol{\Sigma}^{-1} \ .$$

Then noting $\widetilde{\mathbf{G}}_i \approx \widehat{\mathbf{G}}_i$ and inserting this and (17) in (16) yields an approximation for $f(\mathbf{Y}|T)$

$$\tilde{f}(\mathbf{Y}|T) \approx \prod_{i=1}^{b} \left\{ 2\pi L \cdot \left| -\widehat{\mathbf{G}}_i \right|^{-1/2} \cdot \exp\left[v_i(\widehat{\widetilde{\boldsymbol{\psi}}}_i) + \lambda_i(\widehat{\widetilde{\boldsymbol{\psi}}}_i) \right] \right\}, \tag{18}$$

where $\widehat{\widetilde{\boldsymbol{\psi}}}_i$ is the right hand side of (17). Because of the Newton's step, (18) appears less accurate than the Laplace approximation (16). But the error remains $O(1/\min_i M_i)$ and (18) works equivalently well as (16) especially when $\min_i M_i$ is reasonably large.

As described for π^B, when historical data exist, specifying the hyperparameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be easily done through prior predictions for capture probabilities, combined with the moment estimation method based on the logit transformation.

4 Posterior Exploration with Parallel Tempering

Due to the huge size of the tree space, it is infeasible to calculate the posterior distribution over all possible trees. CGMs proposed an MH algorithm to stochastically search for high posterior trees, which iteratively simulates a Markov chain with limiting distribution $f(T|\mathbf{Y}) \propto f(\mathbf{Y}|T)\,\pi(T)$. A drawback of this algorithm, as mentioned in CGMs, is that the simulated chains tend to quickly gravitate towards a region where the tree posterior is large and then stabilize, move locally in that region for a long time. To reduce the time for the chains to move away from local maxima, different methods have been proposed in the existing literature, such as search with multiple starting points (CGM 1998 and 2002), parallel tempering (Geyer & Thompson 1995), evolutionary MCMC (Liang & Wong, 2000), dynamic weighting (Liu et al. 2001), and many others. In this paper, we adopt parallel tempering (PT) to our BTCRMs due to its better performance than using multiple starts and its simplicity. The basic idea of the PT method is, instead of using a single long run, it simulates a set of Markov chains in parallel, and updates them by both within-chain and across-chain operations in each iteration. An important feature of such parallel chains is that each of them uses a different temperature; a high temperature can make the limiting distribution with sharp peaks become flat so help a chain escape from local maxima, while a low temperature can make a chain quickly move to peaks nearby. Moreover, instead of letting chains run independently, the system can be substantially mixed by passing useful information among chains via exchange operations. Under the context of the tree models, a new MH sampler with implementation of PT is outlined by the following steps:

1. Initialize a set of chains $\mathbf{T}^0 = \{T_1^0, \ldots, T_R^0\}$ with the null tree (i.e., the single node

tree), and specify a temperature ladder $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_R\}$ where $1 = \tau_1 < \cdots < \tau_R$ and τ_r is associated with the $r-$th chain, $r = 1, \ldots, R$.

2. For each member of the population $\mathbf{T}^i$ at the $i-$th iteration (say member r), run the following MH algorithm to generate a sample T_r^{i+1}.

   (a) Generate a candidate value T_r^* from T_r^i with probability distribution $q(T_r^i, T_r^*)$ that randomly chooses among four actions: GROW, PRUNE, CHANGE, and SWAP (for a detailed description of these actions, see CGM1998).

   (b) Set $T_r^{i+1} = T_r^*$ with probability

$$\phi(T_r^i, T_r^*) = \min\left\{\left[\frac{\tilde{f}\left(\mathbf{Y}|T_r^*\right)\pi(T_r^*)}{\tilde{f}\left(\mathbf{Y}|T_r^i\right)\pi(T_r^i)}\right]^{1/\tau_r} \frac{q(T_r^*, T_r^i)}{q(T_r^i, T_r^*)}, 1\right\}. \tag{19}$$

   Otherwise, set $T_r^{i+1} = T_r^i$.

3. Exchange T_l^{i+1} with T_k^{i+1} for R pairs of (l, k) with probability

$$\phi_E(T_l^{i+1}, T_k^{i+1}) = \min\left\{\left[\frac{\tilde{f}\left(\mathbf{Y}|T_k^{i+1}\right)\pi(T_k^{i+1})}{\tilde{f}\left(\mathbf{Y}|T_l^{i+1}\right)\pi(T_l^{i+1})}\right]^{1/\tau_l - 1/\tau_k}, 1\right\} \tag{20}$$

   where l is sampled uniformly on $\{1, \cdots, R\}$; for $1 < l < R$, $k = l \pm 1$ with probability 0.5, for l of 1, $k = 2$, and for l of R, $k = R - 1$.

4. Repeat step 2 and 3 for the $(i+1)$-th iteration until the chains converge.

To actually estimate the total population size N, we must choose a specific tree from all of those visited by the parallel chains. It is natural to use posterior probabilities to rank trees. However, this suffers from the dilution phenomenon discussed in CGM 1998, so is not a good selection criterion. Instead, the "best" tree can be determined by choosing the tree with the highest marginal likelihood $f(\mathbf{Y}|T)$. This is equivalent to using Bayes factors as the selection criterion. Here, we should note that trees visited under different tree priors are comparable with regard to their marginal likelihoods. So in practice, it may be appropriate to explore the tree space under several (α, β) choices to accommodate various reasonable beliefs about tree size, then find a single "best" tree over all the choices.

Once a tree is selected, it is straightforward to use the Sekar and Deming estimator $\hat{N}_{SD}$ for estimating N. For a final tree with b terminal nodes, this is given by $\hat{N}_{SD} =$

$\sum_{i=1}^{b} n_{i1}n_{i2}/m_i$ and an estimate for the asymptotic variance is $\hat{V}(\hat{N}_{SD}) = \sum_{i=1}^{b} n_{i1}n_{i2}u_{i1}u_{i2}/m_i^3$ (Sekar & Deming 1949). The advantage of this classical approach is that it is computationally simple, and at least for the Census undercount application, a well-understood estimator. The tree model can be viewed as simply a rational and data-based post-stratification mechanism. This is the estimator we examine in the paper.

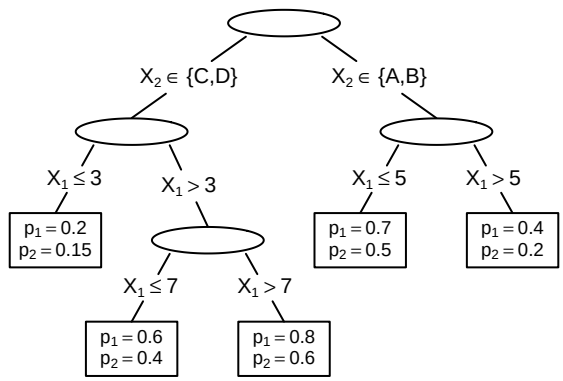
In other applications, however, one might be interested in a Bayesian analysis for each N_i that can incorporate useful prior information into the data. In this case, calculations leading to the posterior distribution $\pi(N_i|\mathbf{Y}_i)$ are necessary so that $\hat{N}_i$ can be the posterior mean or mode and the posterior variance of $N|\mathbf{Y}$ can be given by $\sum_{i=1}^{b} V(N_i|\mathbf{Y}_i)$. Note under the homogeneity assumption, there have been comprehensive Bayesian inferences about population size (e.g., George & Robert 1992, Castledine 1981, Smith 1991).

Before we end this section, one implementation detail of the search algorithm is worth noting. A tree having at least one node with m_i equal to zero is not permitted in a run. In contrast, a tree having one or more nodes with u_{i1} or u_{i2} equal to zero can be visited during a run. However, once such trees are visited, none of the nodes with u_{i1} or u_{i2} equal to zero is allowed to split for two reasons. First, splitting such nodes would not help us to estimate N because $\hat{N}_{SD}$ would remain unchanged by doing so. Second, such operations cannot avoid numerical or sampling-based integration to obtain the marginals, which slows down the algorithm greatly. In many applications, people may be able to set reasonable and realistic constraints on any of m_i, u_{i1}, u_{i2}, n_{i1}, n_{i2} when splitting a terminal node. Such constraints, when available, are recommended to use for avoiding trivial cases and achieving fast computing speed.

5 Examples

In this section, we illustrate Bayesian treed CRMs and evaluate their performance in estimating total population sizes on three examples. The first example uses data simulated from a true tree structure, so we expect BTCRMs to do well on it. We use this example to show that using parallel tempering improves the effectiveness of stochastic search, compared to the strategy of using multiple starts for the MCMC algorithm. We also consider various

Figure 1: A capture-recapture tree



choices of hyperparameters and investigate their effects on estimation. Although such effects could vary from one dataset to another, our hope is that the results can shed lights on typical cases. The second example deals with data simulated from logistic regression models, where we show that BTCRMs with high posteriors can provide reasonable approximations to the true model and identify important predictors to produce good estimates. The third example demonstrates how BTCRMs perform under the null case, where the population under consideration is homogeneous.

5.1 First Simulated Example

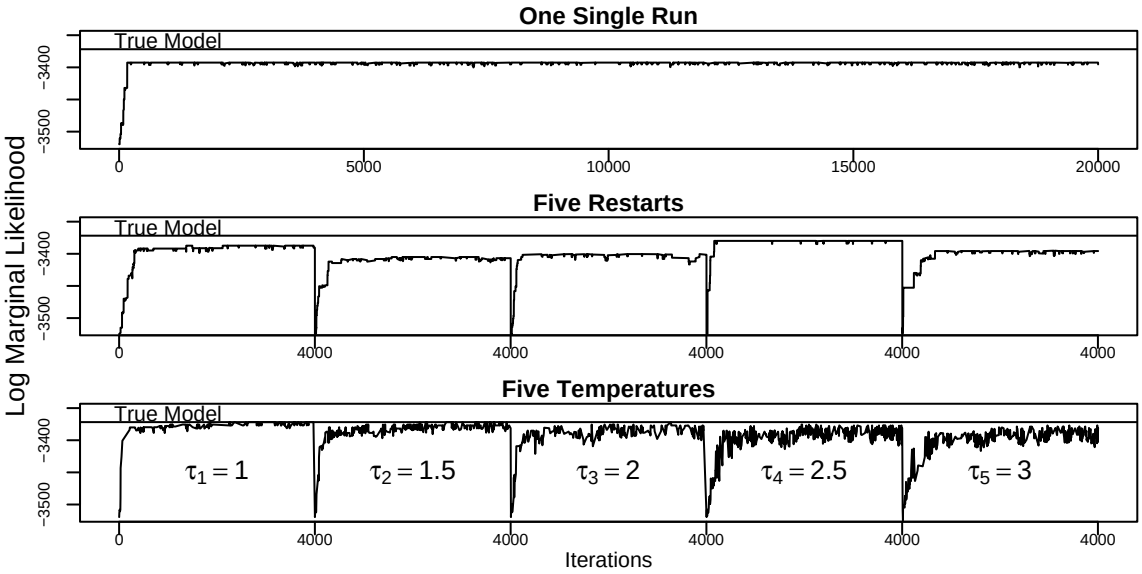
In this experiment, we extended aspects of the simulation setup in Section 6 of CGM 1998 for BTCRMs. First, we simulated a population of size $N = 5000$ with capture probabilities following a tree structure depicted by Figure 1. For the $k-$th individual in the population, the covariates x_{k1} and x_{k2} were simulated uniformly from $\{1, 2, \ldots, 10\}$ and $\{A, B, C, D\}$, respectively. Then the indicators of capture status, z_{kj}, $j = 1, 2$, were generated from Bernoulli$(1, p_{kj})$ independently. Note this tree has the same splitting rules as the one in CGM 1998 and tends to elude identification by some "short-sighted" greedy algorithms.

We begin by applying our approach to the "visible" capture data where $z_{k1} + z_{k2} \neq 0$. To illustrate the behavior of the PT algorithm for our treed setup, we performed a comparison of three strategies: (1) one long run with 20,000 iterations, (2) 5 restarts with 4000 iterations for each restart, and (3) PT with 5 parallel chains having a temperature ladder $\boldsymbol{\tau} = \{1, 1.5, 2, 2.5, 3\}$ and 4000 iterations for each chain. These temperatures were

chosen so that we can control the PT sampler to yield reasonable acceptance rates: the overall local updating rate was about 0.25, and the exchange rate was about 0.4. All the strategies are based on the noninformative Jeffreys prior on capture probabilities, and the tree prior $\pi(T)$ with $\alpha = 0.95$ and $\beta = 0.5$, which assigns a prior mass of approximately 0.80 to trees with less than 10 terminal nodes. Each panel of Figure 2 displays the log marginal likelihoods for visited trees under one of the strategies, with a line drawn at the log marginal likelihood of the true tree. As Figure 2 shows, the strategy of using a long run was worst; the algorithm quickly got to a region of high posterior probabilities, and then was trapped in that region for a long time. The strategy of using multiple starts was much better than that of using a single run; the fourth restart luckily found trees with likelihoods only a bit lower than the "optimal" value. The PT strategy performed best in this example and correctly identified trees with the "optimal" likelihood. Its second chain with temperature 1.5 first visited such trees, and passed these via the exchange operations to the first chain with temperature 1 that reflects the true posterior. The other three chains were also helpful in sense that they accepted trees with less rigid standards and passed good candidates through several exchange operations to the second chain that led to finding the "optimal" trees. Overall, it appears that the PT strategy improves the effectiveness of the MH sampler, compared to using multiple starts, due to the attractive features of sampling along a temperature ladder and exchange of useful information among chains.

Now we report results from a sensitivity analysis of various prior choices for capture probabilities. What we considered in the analysis are the Jeffreys prior, the flat prior, and informative beta priors including Beta$(0.5, 0.5)$, Beta$(1, 2)$, Beta$(1.5, 1.5)$ and Beta$(2, 1)$ for both p_{i1} and p_{i2}, or for one of p_{i1} and p_{i2} while the other uses the flat prior Beta$(1, 1)$. Note these beta priors represent very different subjective prior information: Beta$(0.5, 0.5)$ is symmetric convex, Beta$(1.5, 1.5)$ is symmetric concave, Beta$(2, 1)$ is a line with a positive slope, and Beta$(1, 2)$ is a line with a negative slope, each of which is in favor of different probability values in [0,1]. In total, there were 14 prior distributions on (p_{i1}, p_{i2}) tested. For each of these priors, we considered three sets of the hyperparameters (α, β) of the tree prior $\pi(T)$: (i) $(\alpha, \beta) = (0.95, 1.5)$, (ii) $(\alpha, \beta) = (0.95, 0.5)$ and (iii) $(\alpha, \beta) = (0.95, 0.25)$. This gives $14 \times 3 = 42$ possible prior settings. Here, (i) expresses the prior belief that small

Figure 2: Comparison of the strategy PT with no restart and multiple restarts



trees should yield adequate fits to the data, giving 2.8 expected terminal nodes; (ii) is in favor of medium trees, giving 6.5 expected terminal nodes; and (iii) induces a loose prior that puts a lot of weight on large trees, giving 11.8 expected terminal nodes. Note in this example α is fixed at 0.95 so that the root node has a large chance to split, which reflects the prior belief that the null tree would not fit the data adequately. For each of the 42 prior settings, to search for promising trees over the induced posterior, we ran the PT sampler with the temperature ladder $\boldsymbol{\tau} = \{1, 1.5, 2, 2.5, 3\}$ for 4000 iterations; then we selected the tree with the highest marginal likelihood as the "best" and report this in Table 1 along with the null tree and the true tree. Finally, for each of the 14 prior distributions on (p_{i1}, p_{i2}), we selected the most likely tree denoted $\hat{T}$ over the small, medium and large tree priors, and its characteristics are displayed in Figure 3. The left panel shows log Bayes factors for the true tree and each $\hat{T}$, the middle panel is for size of $\hat{T}$ with a line drawn at the size of the true tree, and the right panel shows predictive loss of $\hat{T}$ with a line drawn at the predictive loss of the true tree. Here, the Bayes factor for tree T is defined as $\mathrm{BF}_T = f(\mathbf{Y}|T)/f(\mathbf{Y}|T_0)$ where T_0 is the null tree; the predictive loss is defined as $L = \left|\hat{N}_{SD} - N\right|/N$ where $N = 5000$ in this example.

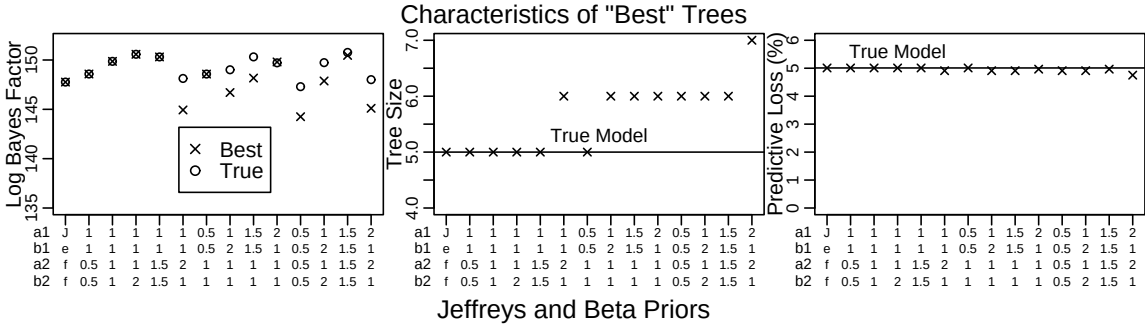
From Table 1 and Figure 3, we can see that the marginal likelihoods (or Bayes factors) for

Table 1: Results from Various Priors on Capture Probabilities

Prior	Log Marginal Likehood				
$\pi(p_{i1}, p_{i2})$	Null Tree	True Tree	Best (S)	Best (M)	Best (L)
Jeff	-3519.6	-3371.8	-3371.8	-3371.8	-3371.8
Beta(a_1, b_1, a_2, b_2)					
(1.0,1.0,1.0,1.0)	-3519.0	-3369.2	-3373.3	-3369.2	-3371.5
(1.0,1.0,0.5,0.5)	-3519.5	-3370.9	-3374.6	-3370.9	-3373.5
(1.0,1.0,1.0,2.0)	-3518.9	-3368.3	-3368.3	-3370.2	-3370.2
(1.0,1.0,1.5,1.5)	-3518.8	-3368.5	-3368.5	-3368.7	-3368.5
(1.0,1.0,2.0,1.0)	-3519.2	-3371.0	-3375.5	-3377.5	-3374.2
(0.5,0.5,1.0,1.0)	-3519.4	-3370.9	-3370.9	-3373.6	-3373.6
(1.0,2.0,1.0,1.0)	-3519.4	-3370.4	-3377.1	-3372.7	-3375.1
(1.5,1.5,1.0,1.0)	-3518.8	-3368.5	-3372.2	-3371.7	-3370.7
(2.0,1.0,1.0,1.0)	-3518.8	-3369.1	-3371.4	-3375.4	-3369.0
(0.5,0.5,0.5,0.5)	-3519.9	-3372.6	-3377.9	-3377.9	-3375.6
(1.0,2.0,1.0,2.0)	-3519.3	-3369.5	-3374.0	-3374.6	-3371.4
(1.5,1.5,1.5,1.5)	-3518.6	-3367.8	-3369.9	-3368.1	-3369.9
(2.0,1.0,2.0,1.0)	-3518.9	-3370.9	-3375.2	-3373.8	-3375.2

Note: S, M and L stand for the small, medium and large tree prior, respectively.

Figure 3: Sensitivity Analysis of Various Priors on Capture Probabilities



the null tree, the true tree or the "best" trees are not sensitive to the prior choice; instead, they are quite resistant to the prior change. As shown in the left panel of Figure 3, the true tree has the highest likelihood under every prior setting. In about half of cases including the Jeffreys and flat priors, the true tree was selected as the "best" tree while in the other cases with only one exception, trees with one more terminal nodes were identified as the "best". However, the predictive losses of $\hat{N}$ from the "best" trees under all prior settings are very close to that of the true tree (5.0%), much better than that of the null tree (15.8%). This is because the "best" trees selected by this Bayesian approach often have subtrees with the same splitting rules of the true tree, so their predictive performance is as good as that of the true tree, as indicated in the right panel of Figure 3.

In this example (the sample size $M = 3373$), each run of 4000 iterations using the PT sampler with five temperatures (i.e., total 20,000 iterations) took about 40 seconds (1.8GHz Xeon processor and 1GB of RAM).

5.2 Second Simulated Example

Treed models are often powerful and robust as they can provide convenient but reasonable approximation to reality in a wide range of applications. However, whether BTCRMs, as a new class of treed models, can achieve this remains an open question. To address this, we deliberately simulated data from populations in which each individual has capture probabilities following logistic regression models instead of a binary tree. And we tested how well BTCRMs would perform even when the underlying assumptions were not satisfied.

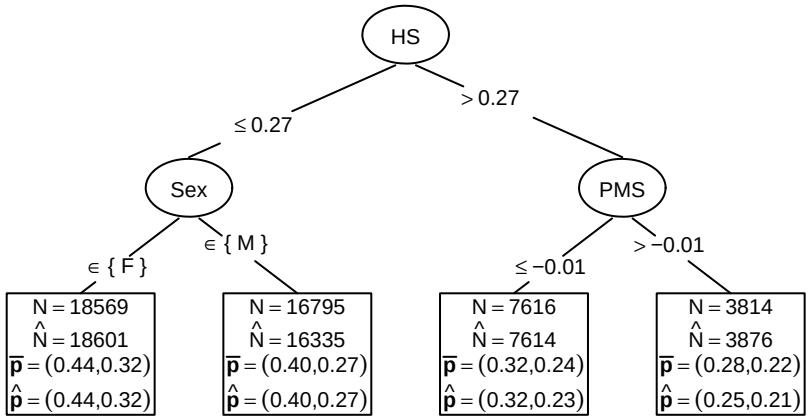
Following Caples (2000), we used actual data for a particular post-stratum (minority home-owner households) from the 1990 U.S. Census and Post-Enumeration Survey to construct populations. The data contains the following descriptive variables for 46,794 individuals: Age, Sex, Marital Status (MS), Household size (HS), Percent Non-Owner in block (PNO), Percent Multiunit Structure in block (PMS), Vacancy Rate in block (VR), and two interaction terms Age*Sex and PNO*PMS. To generate capture rates, we used the same model as in Caples (2000), given by the equations

$$\begin{aligned} &\text{Capture:} \quad \text{logit}(p_{k1}) = \eta_1 + 0.1528 \text{ Sex} - 0.3027 \text{ HS} - 0.0967 \text{ PMS} \\ &\text{Recapture:} \quad \text{logit}(p_{k2}) = \eta_2 + 0.2036 \text{ Sex} - 0.1998 \text{ HS} - 0.0475 \text{ PMS.} \end{aligned} \tag{21}$$

The regression coefficients in (21) are the MLEs fitted for the above model based on the entire post-stratum and actual capture status data. But the constant terms were modified to create two populations having capture rates similar to ones associated with model I and II in Alho (1990). For population I, $\eta_1 = 0.3688$ and $\eta_2 = 1.3619$, which yield p_{k1}'s with mean 0.608, SD 0.075, and p_{k2}'s with mean 0.810, SD 0.038. For population II, $\eta_1 = -0.5312$ and $\eta_2 = -1.082$, which yield p_{k1}'s with mean 0.392, SD 0.070, and p_{k2}'s with mean 0.275, SD 0.042. The size of both populations is 46,794. For each population, to simulate capture status indicators, individuals were sampled according to (p_{k1}, p_{k2}) calculated from (21).

In this example, preliminary runs indicated that the "best" trees selected under different

Figure 4: A Selected Tree for Population II



Note: $\bar{\mathbf{p}}$ is the average of true probabilities and $\hat{\mathbf{p}}$ is the classical MLE under a terminal node. Covariates were standardized in this example. The predictive loss of this tree is 0.79%.

tree priors yield similar log marginal likelihoods and predictive losses, and they performed equally well in estimating N. As a result, many different trees can provide good approximation to the true logistic models. Usually, such trees contain splitting rules for the right covariates SEX, HS or PMS or others correlated to these three, an example of which is shown in Figure 4 for population II.

To further investigate the predictive performance of BTCRMs, we generated 50 samples from each population and repeated our procedure for each sample under the noninformative Jeffreys prior for (p_{i1}, p_{i2}) and the tree prior of $(\alpha, \beta) = (0.95, 0.5)$. In a run, we employed the PT sampler with the temperature ladder $\boldsymbol{\tau} = \{1, 2, 3, 4, 5\}$ for 4000 iterations. With these temperatures, the overall acceptance rate of local updating was about 0.3, and the overall exchange rate was about 0.15 for Population I and 0.3 for Population II. For comparison, we also applied conditional logistic regression models (Alho, 1990) for each dataset. Table 2 compares the performance of BTCRMs with classical CRMs and Alho's logistic regression models. The predictive losses show that the BTCRMs performed well. They are substantially better than classical CRMs and close to Alho's logistic models. Note that we cannot expect BTCRMs to outperform Alho's logistic models since they were the models from which the data were simulated. Also, the performance of Alho's models reported in Table 2 is optimistic because we fitted the model with the correct variables directly instead of using any variable

Table 2: Comparison of Bayesian Treed CRM with Classical CRM and Alho's Logistic Regression based on 50 Samples of Simulated Census Data

(N=46794)	Pop I ($\eta_1 = 0.3688, \eta_2 = 1.3619$)			Pop II ($\eta_1 = -0.5312, \eta_2 = -1.082$)		
	Alho Logistic	Null Tree	Best Tree	Alho Logistic	Null Tree	Best Tree
AVG(Size)	–	1	9.9	–	1	5.2
AVG(Log BF)	–	0	362.0	–	0	86.2
AVG(Pred. Loss) (%)	0.17	0.60	0.18	0.71	2.79	0.78
AVG($\hat{N}$)	46785.6	46514.3	46759.7	46771.6	45489.2	46526.8
SD($\hat{N}$)	96.3	87.3	97.6	402.7	370.8	397.8
AVG($\hat{\sigma}(\hat{N})$)	91.2	81.7	92.4	478.3	412.2	468.0

Note: the averages and SDs were calculated based on 50 samples from each population.

selection procedure. In contrast, our BTCRMs automatically involved variable selection among the 9 variables.

In this example, each run of 4000 iterations for population I using the PT sampler with five temperatures (the average sample size $\bar{M}$ is about 43,000) took about 24 minutes, and for population II ($\bar{M}$ is about 26,000) it took about 14 minutes (1.8GHz Xeon processor and 1GB of RAM).

5.3 The Null Case

Like previous tree-based models, a treed CRM is subject to the potential criticism that it may overfit and find complicated structures when there is none. We now examine how our Bayesian search for treed CRMs will perform when the true model is indeed homogeneous. To do this, we first constructed a dataset with 1000 records and six possible explanatory variables to represent a population of size 1000 in a capture-recapture experiment. The pairs (x_{1k}, x_{2k}) were evenly spaced on a grid over (0,1)×(0,1), the triples (x_{3k}, x_{4k}, x_{5k}) were generated from a multivariate normal with mean $(0, 0, 0)^T$ and covariance matrix $[(1.0, 0.2, -0.4), (0.2, 1.0, 0.7), (-0.4, 0.7, 1.0)]^T$, and x_{6k} were generated as independent Bernoulli with $p = 0.5$. For this population, we considered three pairs of capture probabilities: (i) small-small combination: $(p_1, p_2) = (0.2, 0.1)$; (ii) large-small: $(p_1, p_2) = (0.8, 0.2)$; (iii) large-large: $(p_1, p_2) = (0.9, 0.75)$. For each pair (p_1, p_2), we generated 100 samples of capture data from the population, then for each sample we ran the PT sampler with the same setting as in the first example. As usual, the "best" tree was selected

Table 3: Simulation Study for The Null Case

Cap. Prob	p_1=0.2 p_2=0.1			p_1=0.8 p_2=0.2			p_1=0.9 p_2=0.75		
(N=1000)	Null Tree	Best (M1)	Best (M2)	Null Tree	Best (M1)	Best (M2)	Null Tree	Best (M1)	Best (M2)
% Hits	–	23	98	–	11	100	–	5	100
AVG(Size)	1	1.8	1.0	1	1.9	1	1	2.1	1
AVG(Log BF)	0	1.5	0.1	0	1.7	0	0	2.2	0
AVG(Pred. Loss) (%)	17.2	21.0	17.1	2.6	2.8	2.6	0.5	0.5	0.5
AVG($\hat{N}_{SD}$)	1049.0	1109.4	1047.6	996.7	1002.9	996.7	999.7	1000.1	999.7
SD($\hat{N}_{SD}$)	225.6	288.6	224.5	31.8	34.9	31.8	6.0	6.3	6.0
AVG($\hat{\sigma}(\hat{N}_{SD})$)	210.6	245.2	210.4	31.2	35.8	31.2	6.1	6.4	6.1

Note: the averages and SDs were calculated based on 100 samples for each pair of (p_1, p_2).

as the one with the highest log marginal likelihood. Here, we also considered another selection procedure suggested in CGM 2003: first identify the most frequently visited tree size, then for this size choose the tree with the highest log marginal likelihood. We refer to the former method as "**M1**" and the latter as "**M2**". Turning to the choice of the priors, we used the Jeffreys prior on (p_1, p_2) and the tree prior with $(\alpha, \beta) = (0.95, 0.5)$. As mentioned before, this tree prior indicates that the null tree would not fit the data adequately, as will be typically what people believe when they try BTCRMs.

Table 3 shows the proportion of correct $\hat{T} = T_0$ hits, the average size and the average Bayes factors for $\hat{T}$ selected by **M1** and **M2** respectively over the 100 samples for each pair of (p_1, p_2). It also compares $\hat{T}$ with the null tree T_0 for each case. Our findings, shown in Table 3, are that the BTCRMs performed well in all cases. The selection procedure **M2** gave nearly 100% correct identification of T_0 in all cases. In contrast, the **M1** procedure tended to incorrectly select trees with two nodes, especially when p_1 and p_2 are both large. Such selected trees have (usually slightly) higher marginal likelihoods than that of T_0. Even so, except for small p_1 and p_2, $\hat{T}$ selected by **M1** achieved similar values for predictive loss, sample mean and SD of $\hat{N}_{SD}$ as T_0 did. So overfitting would only become a problem for estimation when small capture probabilities are involved. In this case, $\hat{N}_{SD}$ calculated from a tree with more than one node tends to overestimate N, as shown in Appendix A theoretically.

A simple way to avoid the problem of overfitting for **M1** in the null case is to check the Bayes factor for $\hat{T}$ vs. T_0. Based on the guideline provided in Kass & Raftery (1995), the evidence against T_0 is strong if $\log(BF) > 3$ and Table 3 indicates no strong evidence at all.

6 Discussion

In this paper, we have introduced a new class of treed models, BTCRM, to account for population heterogeneity in a two-period recapture analysis. A Bayesian model selection approach has been developed for finding BTCRMs that fit the data well. We have proposed a variety of prior distributions on the parameter space and for each of them, we have presented an approximation to the corresponding marginal density of the data. Under most nontrivial cases, all these approximations can be obtained as explicit functions of the data (thus no numerical optimization is required) and they are accurate, which reduces the computational burden of the overall approach. Also, the use of parallel tempering in the stochastic search greatly improves the mixing behaviour of MCMC.

The proposed BTCRMs are illustrated with several simulated examples here, one of which attempts to capture features of Census undercount estimation, the application which motivated this work. The following advantages of BTCRMs are especially important in that application.

1. BTCRMs are simple in structure so have a meaningful and easily understandable interpretation. They are easy to use with the Sekar-Deming estimator.
2. To approximate "nonstandard" relationships (e.g., nonlinear, nonsmooth, nonmonotone), methods using regression functions require great human effort in choosing a transformation, creating interaction terms, and determining binning cut points. By contrast, a BTCRM can be grown completely without human intervention, and still provide an adequate description through a series of binary splits. This is especially important when many models are required, such as for all the states in Census undercount estimation.
3. The performance of BTCRMs in estimating population size is robust in practice. They can be applied in situations in which even the underlying treed assumptions are seriously violated, as shown in our second example.
4. The availability of various prior distributions for BTCRMs not only allows for both objective and subjective Bayesian inferences, but also gives practitioners flexibility in choosing either a default and automatic procedure or a procedure with incorporation of real prior information.

5. There often exist situations where plausible or irrelevant covariates are present. Our approach to treed modeling automatically takes model uncertainty into account while competing models such as Alho's logistic regression require model selection beforehand. Also, our approach can produce a set of good trees on which Bayesian model averaging can be based.

6. With BTCRMs, one can easily incorporate additional rules when splitting a terminal node. An example of when this might be useful is to implement a minimum size for each post-stratum.

In summary, BTCRMs provide a potentially useful method for dual system estimation. We end this paper by pointing out two limitations that we are aware of about BTCRMs. First, like regression methods, BTCRMs can only model heterogeneity that can be explained by observable covariates. Second, BTCRMs are based on the assumption of a closed population. This may be restrictive in applications where birth/death or immigration/emigration occur. Therefore, it would be of interest to consider treed models for open populations.

A Biases From Over-stratification

We show below that if one incorrectly treats a homogeneous population as heterogeneous with a strata, the Sekar-Deming estimator would be expected to increase when increasing a.

Suppose the population is partitioned into a subgroups labeled by $\{1, 2, \ldots, a\}$. Let $\underline{N} = (N_1, N_2, \ldots, N_a)$ where N_l is the number of subjects assigned to group l, $l = 1, 2, \ldots, a$. Define n_{1l}, n_{2l} and m_l as the number of subjects in group l captured in the first occasion, the second occasion and both occasions respectively. Let $u_{1l} = n_{1l} - m_l$, $u_{2l} = n_{2l} - m_l$, $\hat{q}_l = m_l/m$ and $q_l = N_l/N$. Also, for notational simplicity, let $\mathbf{A}$ be the event $\{N, n_1, n_2, m, m_l > 0, \forall l\}$. For each l, we have that $u_{1l}|N, N_l, u_1 \sim$ Hypergeometric(N, N_l, u_1), $u_{2l}|N, N_l, u_2 \sim$ Hypergeometric (N, N_l, u_2), and $m_l|N, N_l, m \sim$ Hypergeometric (N, N_l, m). So $\mathrm{E}(u_{1l}|N, N_l, u_1) = u_1 q_l$, $\mathrm{E}(u_{1l}|N, N_l, u_1) = u_2 q_l$ and

$$\begin{aligned}
\mathrm{E}(\frac{1}{\hat{q}_l}|N, N_l, m, m_l > 0) &= \frac{1}{q_l} + \frac{\mathrm{E}\left((\hat{q}_l - q_l)^2 |N, N_l, m\right)}{q_l^3} + O\left(m^{-3/2}\right) \\
&= \frac{1}{q_l} + \frac{(N - N_l)(N - m)}{mN(N-1)} \cdot \frac{1}{q_l^2} + O\left(m^{-3/2}\right)
\end{aligned}$$

Also notice that for the stratified MLE , denoted by $\widehat{N}_a$, we have

$$\begin{aligned}
\mathrm{E}\left(\widehat{N}_a|\mathbf{A},\underline{N}\right) &= \mathrm{E}\left(\sum_{l=1}^{a}\frac{n_{1l}n_{2l}}{m_l}|\mathbf{A},\underline{N}\right) = n_1+n_2-m+\mathrm{E}\left(\sum_{l=1}^{a}\frac{u_{1l}u_{2l}}{m_l}|\mathbf{A},\underline{N}\right) \\
&= n_1+n_2-m+\sum_{l=1}^{a}\mathrm{E}(u_{1l}|N,N_l,u_1)\mathrm{E}(u_{2l}|N,N_l,u_2)\mathrm{E}(\frac{1}{m_l}|N,N_l,m,m_l>0)
\end{aligned}$$

Therefore

$$\mathrm{E}\left(\widehat{N}_a|\mathbf{A}\right) = \frac{n_1n_2}{m}+\frac{u_1u_2}{m}\left[\frac{a-1}{m}\cdot\frac{N-m}{N-1}+O\left(m^{-3/2}\right)\right] \tag{22}$$

Similarly, we can show if we do post-stratification (i.e., the observed individuals are randomly assigned into a subgroups), then

$$\mathrm{E}\left(\widehat{N}_a|\mathbf{A}\right) = \frac{n_1n_2}{m}+\frac{u_1u_2}{m}\left[\frac{a-1}{m}+O\left(m^{-3/2}\right)\right] \tag{23}$$

Equation (22) and (23) together show that for a homogeneous population, no matter whether we use pre- or post-stratification, the Sekar-Deming estimate would be expected to increase if we use more strata.

References

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623–635.

Alho, J. M., Mulry, M. H., Wurdeman, K., & Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88(423), 1130–1136.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, second edition.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Caples, J. (2000). *Variance Reduction and Variable Selection Methods for Alho's Logistic Capture Recapture Model with Applications to Census Data*. Ph. D. Dissertation, Department of MSIS, University of Texas at Austin.

Castledine, B. J. (1981). A bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 68, 197–210.

Chipman, H., George, E. I., & McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443), 935–948.

Chipman, H., George, E. I., & McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning*, 48, 299–320.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2003). Bayesian treed generalized linear models. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smoth, & M. West (Eds.), *Bayesian Statistics*, 7 (pp. 85–104). Oxford Univeristy Press: Clarendon Press.

Denison, D. G. T., Mallick, B. K., & Smith, A. F. M. (1998). A bayesian cart algorithm. *Biometrika*, 85(2), 363–377.

George, E. I. & Robert, C. P. (1992). Capture-recapture estimation via gibbs sampling. *Biometrika*, 79(4), 677–683.

Geyer, C. J. & Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431), 909–920.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47(2), 725–732.

Jeffreys, H. (1961). *Theory of Probability*. Oxford Unversity Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Liang, F. & Wong, W. H. (2000). Evolutionary monte carlo: Applications to cp model sampling and change point problem. *Statistica Sinica*, 10, 317–342.

Liu, J. S., Liang, F., & Wong, W. H. (2001). A theory for dynamic weighting in monte carlo. *Journal of the American Statistical Association*, 96, 561–573.

Meyer, M. C. & Laud, P. W. (2002). Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, 97(459), 859–871.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modeling : An overview. *Journal of Applied Statistics*, 29(1), 85–102.

Pollock, K. H., Hines, J. E., & Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329–340.

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83, 251–266.

Sekar, C. C. & Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245), 101–115.

Smith, P. J. (1991). Bayesian analysis for a multiple capture-recapture model. *Biometrika*, 78(2), 399–407.

Tierney, L. & Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

Wilbur, H. M. & Landwehr, J. M. (1974). The estimation of population size with equal and unequal risks of capture. *Ecology*, 55, 1339–1348.

Young, H., Neess, J., & Emlen, J. T. J. (1952). Heterogeneity of trap response in a population of house mice. *Journal of Wildlife Management*, 16, 169–180.