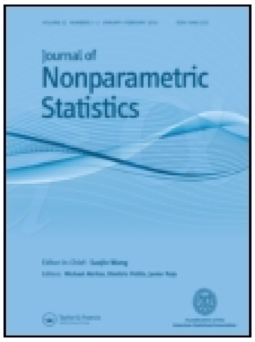

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse

Balgobin Nandram & Jai Won Choi

Nonparametric Statistics
Vol. 16(6), December 2004, pp. 821–839

Taylor & Francis
Taylor & Francis Group

NONPARAMETRIC BAYESIAN ANALYSIS OF A PROPORTION FOR A SMALL AREA UNDER NONIGNORABLE NONRESPONSE

BALGOBIN NANDRAM[a,*] and JAI WON CHOI[b,†]

[a]*Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA;* [b]*National Center for Health Statistics, CDC, Room 915, 6525 Belcrest Road, Hyattsville, MD 20782, USA*

(Received October 2002; Revised August 2003; In final form November 2003)

In small area estimation, it is a standard practice to assume that the area effects are exchangeable. This is obtained by assuming that the area effects have a common parametric distribution, and a Bayesian approach is attractive. The Dirichlet process prior (DPP) has been used to provide a nonparametric version of this approach. The DPP is useful because it makes the procedure more robust, and the Bayesian approach helps to reduce the effect of nonidentifiability prominent in nonignorable nonresponse models. Using the DPP, we develop a Bayesian methodology for the analysis of nonignorable nonresponse binary data from many small areas, and for each area, we estimate the proportion of individuals with a particular characteristic. Our DPP model is centered on a baseline model, a standard parametric model. We use Markov chain Monte Carlo methods to fit the DPP model and the baseline model, and our methodology is illustrated using data on victimization in ten domains from the National Crime Survey. Our comparisons show that it may be preferable to use the nonparametric DPP model over the parametric baseline model for the analysis of these data.

Keywords: Dirichlet process prior; Exchangeability; Identifiability; Griddy Gibbs sampler; Selection model

1 INTRODUCTION

The Dirichlet process prior (DPP) has been used to model clustered data in situations where the cluster effects are exchangeable, thereby providing a nonparametric Bayesian approach to the analysis of clustered data. In addition, there has recently been much activity in the analysis of survey nonresponse, and the response rates in many surveys have been decreasing internationally (Groves and Couper, 1998; De Heer, 1999;). For many of these surveys the responses are binary. For example, the National Crime Survey (NCS) estimates the proportion of households with at least one experience of crime. Using the DPP, we develop a Bayesian methodology for the analysis of *nonignorable* nonresponse binary data from small areas to estimate the proportion of individuals with a particular characteristic for each area.

[*] Corresponding author. E-mail: balnan@wpi.edu
[†] E-mail: jwc7@cdc.gov

ISSN 1048-5252 print; ISSN 1029-0311 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/10485250310001652629

In small area estimation, it is a standard practice to assume that the area effects are exchangeable. This assumption is accommodated by allowing the area effects to have a common parametric distribution, and it facilitates a ‘borrowing of strength’ across the ensemble. However, when there is a preference for a more robust approach, one may prefer a nonparametric Bayesian approach. A natural way to deal with this situation is to use the DPP. Ferguson (1973, 1974), Escobar (1994) and Escobar and West (1995) have used the DPP to perform nonparametric Bayesian analysis on normal data. Additionally, Kong *et al* (1994) have used nonparametric Bayesian analysis for binomial data. The DPP offers some flexibility and robustness to departures from the assumption of a parametric distribution.

There are three types of nonresponse (see Little and Rubin, 1987). When the response indicator is independent of all other variables in the survey, the nonresponse is said to be missing completely at random (MCAR). When the response indicator depends only on the observed values, the nonresponse is said to be missing at random (MAR). When the response indicator depends on the unobserved values, the nonresponse is said to be nonignorable or informative. Models with an MCAR or MAR response mechanism are called ignorable when the parameters for the response model are distinct from the parameters for the data model; otherwise the models are nonignorable (Rubin, 1976). One of the two approaches to model nonignorability is to introduce parameters which control the extent of nonignorability in models for the observed data and considers the sensitivity of quantities of interest to *a priori* realistic changes in these parameters (*e.g.*, Forster and Smith, 1998; Nandram and Choi, 2002a,b).

An ongoing issue in the nonresponse literature is whether the selection or the pattern mixture approaches should be used for modeling of nonresponse. Little and Rubin (1987) and Little (1993) distinguished between these two classes of models for missing data. In the selection approach, the hypothetical complete data are modeled, and a model for the non-response mechanism is added conditional on the hypothetical data. In the pattern mixture approach, the population is stratified into two patterns, respondents and nonrespondents, each being modeled separately, and the final answer is obtained by a probabilistic mixture of these two. The selection approach is more natural and convenient for our application.

In nonignorable nonresponse models, some parameters are not identifiable. In a pattern mixture model, the parameters associated with the nonresponse pattern are not identifiable because there are no data to estimate these parameters, unless a centered model can be constructed. However, in a selection model it is possible to estimate the parameters with reasonable efficiency using a Bayesian approach because the parameters cannot be separated from the data in the likelihood function (false in the pattern mixture model). Although the parameters are still not identifiable, the Bayesian approach helps to reduce this effect. We note that the nonparametric approach using the DPP has nothing to do with this issue, and in the selection model, it is not true that the data do not play a role when these ‘nonidentifiable’ parameters are estimated.

Stasny (1991) used a hierarchical Bayesian model to study victimization in the NCS. Stasny (1991) used the Bayesian selection approach which was developed primarily to study sample selection problems (*e.g.*, Heckman, 1976; Olson, 1980). However, the Stasny Bayes empirical Bayes approach assumes that the hyper-parameters are fixed but unknown, and these parameters are estimated using maximum likelihood methods. This approach has been extended in several directions. (See Nandram and Choi (2002a,b) and Nandram *et al.* (2002) for full Bayesian analyses.) To permit a flexibility in robustness to the prior specifications, we study a nonparametric hierarchical Bayes model that can be used to study nonignorable nonresponse for binary data from many areas. We use a nonparametric Bayesian method to analyze nonignorable nonresponse binary data. But unlike Stasny (1991), Nandram and Choi (2002a,b) and Nandram *et al.* (2002) who assume that the parameters identifying the areas come from a common probability density function, we assume the DPP for these parameters. Thus, in our model with the DPP, the Stasny's nonignorable model is our baseline model. However, unlike Nandram and Choi (2002a,b), we do not express uncertainty about ignorability in this article.

A related literature is on what is now known as uncertain pooling used primarily for experiments, not small areas. The experiments are partitioned and there can be many partitions depending on the number of experiments. The experiments in each partition set are assumed to be similar, and there is uncertainty about which partition is the correct one. This methodology works well for a small number of experiments, but for problems with many experiments (or areas) it may be infeasible. Malec and Sedransk (1992), Consoni and Veronese (1995), Mallick and Walker (1997) and Evans and Sedransk (2001), discussed Bayesian methodology for combining results from several normal or binomial experiments. However, while we are not dealing with the problem of uncertain pooling, the methodology we describe in this article can be used to identify certain groups.

Our main objective is to develop a methodology for the analysis of nonignorable nonresponse binary data from many small areas when a relatively robust approach may be desirable. We use data from the NCS to illustrate our methodology. The rest of the article is organized as follows. In Section 2, we describe the NCS data to further motivate our objectives. In Section 3, we first describe a parametric baseline hierarchical Bayesian model. We describe the nature of nonidentifiability in the baseline model. Then, we describe our DPP model which is centered on the baseline model, and like the baseline model it assumes that the area effects are exchangeable. Both models are constructed for nonignorable nonresponse binary data. We also describe how to fit the models using Markov chain Monte Carlo (MCMC) methods. In Section 4, we illustrate our methodology and provide some empirical results using the NCS data. Finally, Section 5 contains concluding remarks.

2 NATIONAL CRIME SURVEY

The NCS is a large scale household survey conducted by the U.S. Bureau of the Census for the Bureau of Justice Statistics. Stasny (1991) summarized the NCS data and discussed the design of the NCS. Like Stasny (1991), features of the sampling design are not reflected in our modeling except that we assume the data are collected by probability sampling.

Data from the NCS are used to produce quarterly estimates of victimization rate and yearly estimates of the prevalence of crimes. Individuals interviewed for the NCS are asked about crimes (*e.g.*, assault, auto or motor vehicle theft, burglary, larceny, rape and robbery) committed against them or against their property in the previous six months.

We used the data created by Stasny (1991), who took a random start at the record for the eighth household (ordered on the original longitudinal file) in the full data set and then every 15th record after that. The data are poststratified into domains according to three neighborhood characteristics: (i) urban (U) and rural (R), (ii) central city (C), other incorporated place (I) and unincorporated or not a place (N) and (iii) low poverty level (L) (9% or fewer of families below poverty level) and high poverty level (H) (10% or more of families below poverty level). Since it is practically impossible for a rural area to be a central city, as observed by Stasny (1991), this poststratification results in 10 domains.

We define a binary variable to be 1 if there is at least one crime in a household (*i.e.*, household is victimized), and 0 otherwise. Let

$$y_{ij} = \begin{cases} 1, & \text{if households } j \text{ in area } i \text{ is victimized} \\ 0, & \text{if household } j \text{ in area } i \text{ is not victimized} \end{cases}$$

and

$$r_{ij} = \begin{cases} 1, & \text{if households } j \text{ in area } i \text{ is a respondent} \\ 0, & \text{if household } j \text{ in area } i \text{ is not a respondent,} \end{cases}$$

$i = 1, \dots, \ell$, $j = 1, \dots, n_i$. Essentially our models start with the y_{ij} and r_{ij}. We define $y_i = \sum_{j=1}^{r_i} y_{ij}$ and $r_i = \sum_{j=1}^{n_i} r_{ij}$. That is, y_i is the number of successes (*i.e.*, households with crimes in the NCS), r_i is the number of respondents and n_i is the number of households sampled in the ith domain or area), $i = 1, \dots, \ell$ where $\ell = 10$ domains.

It is convenient to present the data for the NCS in Table I. The nonresponse rate in these domains ranges from 9.4% to 16.9%, and one reason for nonresponse is that a woman may be embarrassed to report a rape committed by an attacker. The domains UNH, RIL and RIH can be considered small areas because there are relatively few households in these domains, and they have relatively high observed nonresponse rates. But note that UNL has the highest nonresponse rate.

Let $\hat{p}_i$ be the observed proportion of households with at least one crime in the ith domain. Inspection of the $\hat{p}_i$ seems to reveal two groups of domains, which are (UCL, UCH, UIL, UIH, UNL, UNH, RIL) and (RIH, RNL, RNH). For the first group, the average (standard deviation) of the $\hat{p}_i$ is 0.230 (0.022) and for the second group 0.120 (0.038). Thus, it appears that there are two groups of data, but making two different assumptions for these two groups of data will definitely lose efficiency since one group has seven domains and the other has three domains. It is possible to still assume that the p_i are exchangeable, but one can use more robust assumption. The DPP is an appropriate candidate.

Stasny (1991) suggested that nonresponse does not occur at random with respect to victimization status (see also Saphire, 1984; Stasny, 1990). For the analysis of this data set, Nandram and Choi (2002a) made two key contributions: (a) discern whether nonresponse is ignorable or not and (b) introduce a new model in which the degree of ignorability may vary from one area to another. Here, our contribution is to perform a nonparametric Bayesian analysis of these data by providing a prior that makes our procedure more robust. This can help to reduce overshrinkage, a possible nuisance in small area estimation.

Finally, we note that individuals who are victimized tend to respond less frequently than individuals who are not victimized. This information is useful when the DPP model is fit; it permits stability in computation, helps to improve precision, and it reduces the effects of

TABLE I NCS Data (Stasny, 1991).

Domain	y	$r - y$	$n - r$	$\hat{p}$	$\hat{\delta}$
UCL	156	555	104	0.219	0.872
UCH	95	364	73	0.207	0.862
UIL	162	557	101	0.225	0.877
UIH	72	262	36	0.216	0.903
UNL	92	297	79	0.237	0.831
UNH	15	40	9	0.273	0.859
RIL	11	36	7	0.234	0.870
RIH	10	105	20	0.087	0.852
RNL	35	274	32	0.113	0.906
RNH	79	413	64	0.161	0.885

Note: Crimes were committed during January–June, 1975. The response is binary: 0 for no crimes and 1 for at least one crime; $\hat{p}$ is the observed proportion of households with at least one crime; $\hat{\delta}$ is the observed proportion of respondents.

nonidentifiability. We note that this information was not incorporated into the models of Stasny (1991) and Nandram and Choi (2002a).

3 HIERARCHICAL BAYES NONRESPONSE MODELS

In Section 3.1, we describe the baseline nonignorable nonresponse model and the nonparametric Bayesian model. In Section 3.2, we discuss the issue of identifiability for the baseline nonignorable model. We describe the DPP model in Section 3.3 and its computations in Section 3.4.

3.1 Baseline Nonresponse Model

Our baseline model is a nonignorable nonresponse model, and is given by

$$y_{ij} \mid p_i \overset{\text{iid}}{\sim} \text{Bernoulli}\ (p_i), \quad j = 1, \ldots, n_i, \ \ i = 1, \ldots, \ell,$$

$$r_{ij} \mid y_{ij} = s, \ \pi_{is} \overset{\text{iid}}{\sim} \text{Bernoulli}\ (\pi_{is}), \ \ s = 0, 1. \tag{1}$$

$$p_i \mid \mu_1, \tau_1 \overset{\text{iid}}{\sim} \text{Beta}\ (\mu_1\tau_1, (1 - \mu_1)\tau_1), \tag{2}$$

$$\pi_{is} \mid \mu_{s+2}, \tau_{s+2} \overset{\text{iid}}{\sim} \text{Beta}\ (\mu_{s+2}\tau_{s+2}, (1 - \mu_{s+2})\tau_{s+2}), \quad s = 0, 1. \tag{3}$$

Assumptions (2) and (3) express similarity among the states. This similarity helps when the weakly identified parameters like π_{i0} and π_{i1} are estimated, but may encourage too much pooling. Therefore, to restrict the pooling one may use a more robust prior specification.

We complete the prior specification by taking μ_k, $k = 1, 2, 3$ and τ_k, $k = 1, 2, 3$ to be independent. Specifically, for the μ_k

$$\mu_1 \sim \text{Uniform}(0, 1), \quad \mu_2 \sim \text{Uniform}(\mu_3, 1) \quad \text{and} \quad \mu_3 \sim \text{Uniform}(0, 1). \tag{4}$$

To incorporate the information that non-victimized households tend to respond more frequently than victimized households, in Eq. (4) we take $\mu_2 \geq \mu_3$. In other situations, with expert opinion one might prefer to reverse this inequality.

For the τ_k, we take

$$\tau_k \overset{\text{iid}}{\sim} S(1), \quad k = 1, 2, 3, \tag{5}$$

where the notation $X \overset{\text{iid}}{\sim} S(a)$ means that $p(x) = a/(a + x)^2$, $x \geq 0$ and $a \geq 0$. This is the shrinkage prior density, used to keep X away from the boundary of the parameter space (*i.e.*, near $x = 0$).

3.2 Discussion of Identifiability

To illustrate the issue of identifiability, we consider the baseline model for a single area, momentarily dropping the notation for the single area.

The likelihood function is

$$L(p, \pi_0, \pi_1 \mid y, r) = (\pi_1 p)^y (\pi_0(1 - p))^{r-y} (1 - \pi_1 p - \pi_0(1 - p))^{n-r}.$$

Now consider the transformation $\alpha = \pi_1 p$ and $\beta = \pi_0(1 - p)$. Then, the likelihood function becomes

$$L(\alpha, \beta \mid y, r) = \alpha^y \beta^{r-y}(1 - \alpha - \beta)^{n-r},$$

which is a function only of two parameters, α and β. Thus, p, π_0 and π_1 are not identifiable, and so they cannot be estimated.

Letting $\gamma = \pi_1/\pi_0$, observe that $\alpha/\beta = \gamma(p/(1 - p))$. Thus, if γ is known, then p is identifiable. For example, if $\gamma = 1$ (*i.e.*, $\pi_0 = \pi_1$, ignorable nonresponse model), then $p = \alpha/(\alpha + \beta)$. Thus, once γ is unknown (*i.e.*, the relation between π_0 and π_1 is unknown), p, π_0 and π_1 are all nonidentifiable. The failure of the non-Bayesian method is primarily due to the lack of information about p, π_0 and π_1. If there is some knowledge about the relation between π_0 and π_1, there will be an improvement in inference; see Nandram and Choi (2002a,b) for an approach via an expansion model when γ is unknown.

Let z denote the number of households with at least one crime among the nonrespondents, and note that z is a latent variable. The entire nonresponse problem is solved once z becomes known. Suppose we consider proper noninformative prior densities for p, π_0 and π_1 with $p, \pi_0, \pi_1 \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$. Then the joint posterior density of the parameters z, p, π_0, π_1, given y and r is

$$f(p, \pi_0, \pi_1, z \mid y, r) \propto \begin{pmatrix} n - r \\ z \end{pmatrix} p^{y+z}(1 - p)^{n-y-z}\pi_0^{r-y}(1 - \pi_0)^{n-r-z}(\pi_1)^y(1 - \pi_1)^z,$$

where $0 < p, \pi_0, \pi_1 < 1$, and $z = 0, 1, \ldots, n - r$. Now, letting $B(u, v) = \Gamma(u)\Gamma(v)/\Gamma(u + v)$ ($\Gamma(u)$ is the gamma function) denote the beta function, the normalization constant is

$$\sum_{z=0}^{n-r} \begin{pmatrix} n - r \\ z \end{pmatrix} B(y + z + 1, n - y - z + 1)B(r - y + 1, n - r - z + 1)B(y + 1, z + 1)$$

and obviously this latter quantity is finite. That is, as indicated above, the joint posterior density of $p, \pi_0, \pi_1 \mid y, r$ is proper. Thus, p, π_0, π_1 are all identifiable, *albeit* wealdy.

Note that we have incorporated virtually no information through the uniform priors. Thus, it is simply because the parameters p, π_0, π_1 and z are bounded that help them to become identifiable. This is a strength of the Bayesian paradigm.

Inference about p, π_0 and π_1 can be easily obtained because the joint posterior density of p, π_0 and π_1 is

$$f(p, \pi_0, \pi_1 \mid y, r) = \sum_{z=0}^{n-r} f(p, \pi_0, \pi_1 \mid Z = z, y, r) \Pr(Z = z \mid y, r),$$

where

$$\Pr(Z = z \mid y, r) = \frac{\omega_z}{\sum\limits_{z'=0}^{n-r} \omega_{z'}} = \tilde{\omega}_z, \quad z = 0, \ldots, n - r,$$

and

$$\omega_z = \begin{pmatrix} n - r \\ z \end{pmatrix} B(y + z + 1, n - y - z + 1)B(r - y + 1, n - r - z + 1)B(y + 1, z + 1).$$

Also, given z, y and r, it is clear that p, π_0 and π_1 are independent with

$$p \mid z, y, r \sim \text{Beta}(y + z + 1, n - y - z + 1),$$
$$\pi_0 \mid z, y, r \sim \text{Beta}(r - y + 1, n - r - z + 1) \quad \text{and}$$
$$\pi_1 \mid z, y, r \sim \text{Beta}(y + 1, z + 1).$$

Thus, *a posteriori* inference about p, π_0 and π_1 can be made by drawing samples from $p, \pi_0, \pi_1, z \mid y, r$ using the composition method. In particular, we note that while *a priori* inference about p, π_0 and π_1 is based on uniform distributions, *a posteriori* inference about p, π_0 and π_1 is based on mixtures of beta distributions. Specifically, $p \mid y, r \sim \sum_{z=0}^{n-r} \tilde{\omega}_z \text{Beta}(y + z + 1, n - y - z + 1)$, $\pi_0 \mid y, r \sim \sum_{z=0}^{n-r} \tilde{\omega}_z \text{Beta}(r - y + 1, n - r - z + 1)$, and $\pi_1 \mid y, r \sim \sum_{z=0}^{n-r} \tilde{\omega}_z \text{Beta}(y + 1, z + 1)$. Then, obviously there must be improved inference about p, π_0 and π_1 *a posteriori*.

3.3 Model with Dirichlet Process Prior

We maintain the structure in Eq. (1),

$$y_{ij} | p_i \overset{\text{iid}}{\sim} \text{Bernoulli } (p_i),$$
$$r_{ij} | y_{ij} = s, \ \pi_{is} \overset{\text{iid}}{\sim} \text{Bernoulli } (\pi_{is}), \quad s = 0, 1.$$

Instead of the prior densities in Eqs. (2) and (3), we use the DPP.

Letting $\boldsymbol{\theta}_i = (p_i, \pi_{i0}, \pi_{i1})$, we assume that, given a cumulative distribution function, G say, that

$$\boldsymbol{\theta}_i | G \overset{\text{iid}}{\sim} G(\cdot).$$

To express uncertainty about $G(\cdot)$, we assume that given α and $G_0(\cdot)$,

$$G(\cdot) \sim \text{Dirichlet}\{\alpha G_0(\cdot)\},$$

a Dirichlet process defined by α, a positive real number, and $G_0(\cdot)$, the prior specification of $G(\cdot)$. In fact, $E(G(\boldsymbol{\theta})) = G_0(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ and α is a precision parameter, determining the concentration of the prior distribution for $G(\cdot)$ around $G_0(\cdot)$. Here α is assumed unknown, and $G_0(\cdot)$ has a specified form with its parameters unknown.

This is the standard assumption of exchangeability for $\boldsymbol{\theta}_i$, $i = 1, \ldots, \ell$, within a Bayesian nonparametric framework, in which $\boldsymbol{\theta}_i$, $i = 1, \ldots, \ell$ are the first ℓ realizations from a general Polya-urn scheme (Blackwell and MacQueen, 1973). That is, conditional on G_0 and α, when G is integrated over its prior distribution, the sequence of $\boldsymbol{\theta}_i$s follows the general urn scheme

$$\boldsymbol{\theta}_1 \sim G_0,$$
$$\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1} \begin{cases} = \boldsymbol{\theta}_j & \text{with probability } \dfrac{1}{\alpha + i - 1} \\ \sim G_0 & \text{with probability } \dfrac{\alpha}{\alpha + i - 1}, \end{cases}$$
$$j = 1, \ldots, i - 1, \quad i = 2, \ldots, \ell.$$

A key feature of DPP is associated with the discreteness of $G(\cdot)$ under the Dirichlet process assumption (Ferguson, 1973). In any sample, $\boldsymbol{\theta}_i, i, = 1, \ldots, \ell$, from $G(\cdot)$, there is a positive probability that some of these $\boldsymbol{\theta}_i$ coincide. That is, there are $k(1 \leq k \leq \ell)$, parameters that describe the ℓ areas. The structure is such that the posterior distribution will strongly support common values of individual parameters, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_{i'}$ for data points (y_i, r_i) and $(y_{i'}, r_{i'})$ that are close. Thus, we can combine information locally in the sample space to estimate the local structure. This is where the nonparametric structure in our formulation arises.

In our application, we take $G_0(\cdot)$ corresponding to $\boldsymbol{\theta}_i$ as

$$G_0(p_i, \pi_{i0}, \pi_{i1}) = G_{01}(p_i) G_{02}(\pi_{i0}) G_{03}(\pi_{i1}), \tag{6}$$

where the prior densities for p_i, π_{i0} and π_{i1} are assumed to be independent as in the baseline model and specified as

$$p_i | \mu_1, \tau_1 \overset{\text{iid}}{\sim} \text{Beta}(\mu_1 \tau_1, (1 - \mu_1)\tau_1),$$
$$\pi_{is} | \mu_{s+2}, \tau_{s+2} \overset{\text{iid}}{\sim} \text{Beta}(\mu_{s+2}\tau_{s+2}, (1 - \mu_{s+2})\tau_{s+2}), \quad s = 0, 1.$$

As in the baseline model, *a priori* we specify $\mu_k, k = 1, 2, 3$ and $\tau_k, k = 1, 2, 3$ exactly as in Eqs. (4) and (5), respectively. The specification $\mu_2 \geq \mu_3$ is specifically useful when the DPP model is fit. In the NCS we have only 10 areas, so that when the DPP is fit, the number of clusters (k) of areas is smaller than 10 and can obviously be as small as 1. When the number of clusters is small (*e.g.*, 2 or 3), obviously there is a difficulty in running the Gibbs sampler to fit the conditional posterior density of the hyperparameters.

Finally, we consider the prior specification for α, and we denote the prior density for α by $p(\alpha)$. In particular, Escobar and West (1995) derived an elegant form for the conditional posterior density of γ when the prior is $\alpha \sim \text{Gamma}(a, b)$; see Appendix A for a review of the discussion. In fact, for the astronomy data of Roeder (1990), Escobar and West (1995) took $a = 2$ and $b = 4$. Nowadays, it is a standard specification to take $a = 0.001$ and $b = 0.001$ for a proper diffuse prior for α in practical Bayesian work as documented in readily available software (see Spiegelhalter *et al.* 1996). We have found that inference is sensitive to the choice of a and b so that this simple prior density may not work so well. To avoid this problem, one of the recently recommended priors for variance components is the shrinkage prior (see Daniels, 1999; Natarajan and Kass, 2000) which is also proper and the median of this prior density must be specified. From Appendix A we have

$$p(\alpha | k) \propto p(\alpha) \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + \ell)}, \quad 0 < \alpha < \infty.$$

We use the shrinkage prior

$$p(\alpha) = \frac{\kappa_0}{(\kappa_0 + \alpha)^2}, \quad 0 < \alpha < \infty$$

where κ_0 is the median of $p(\alpha)$ and is to be chosen.

3.4 Computations for the Model with Dirichlet Process Prior

We use the griddy Gibbs sampler to obtain samples from the joint posterior density of all the parameters. See Ritter and Tanner (1992) for an introduction to the griddy Gibbs sampler and

Tanner (1993) for a more elaborate pedagogy. We can draw samples from $f(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})$ as follows. We note that $f(\boldsymbol{\theta}_i, z_i|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau}) = g_1(\boldsymbol{\theta}_i|z_i, \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})\ p(z_i|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})$, where $g_1(\boldsymbol{\theta}_i|z_i, \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})$ can be easily written down.

Note that, given z_i, y_i, r_i, $\boldsymbol{\mu}$, $\boldsymbol{\tau}$, the parameters p_i, π_{i0} and π_{i1} are independent with

$$\begin{aligned}
p_i|z_i, y_i, r_i\mu_1, \tau_1 &\overset{\text{ind}}{\sim} \text{Beta}(z_i + y_i + \mu_1\tau_1, n_i - z_i - y_i + (1-\mu_1)\tau_1),\\
\pi_{i0}|z_i, y_i, r_i, \mu_2, \tau_2 &\overset{\text{ind}}{\sim} \text{Beta}(r_i - y_i + \mu_2\tau_2, n_i - r_i - z_i + (1-\mu_2)\tau_2),\\
\pi_{i1}|z_i, y_i, r_i, \mu_3, \tau_3 &\overset{\text{ind}}{\sim} \text{Beta}(y_i + \mu_3\tau_3, z_i + (1-\mu_3)\tau_3).
\end{aligned}$$

The posterior distribution of z_i is as follows. Letting

$$p(Z_i = z_i|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau}) \propto \frac{\omega_{z_i}}{\sum_{z_i'=0}^{n_i-r_i} \omega_{z_i'}}$$

where

$$\begin{aligned}
\omega_{z_i} = \binom{n_i - r_i}{z_i} & B(z_i + y_i + \mu_1\tau_1, n_i - z_i - y_i + (1-\mu_1)\tau_1)\\
&\times B(r_i - y_i + \mu_2\tau_2, n_i - r_i - z_i + (1-\mu_2)\tau_2) B(y_i + \mu_3\tau_3, z_i + (1-\mu_3)\tau_3),
\end{aligned}$$

$z_i = 0, \ldots, n_i - r_i$ and $i = 1, \ldots, \ell$. Thus, we draw z_i from $p(Z_i = z_i|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})$, and with this z_i, we draw p_i, π_{i0} and π_{i1} independently.

Next, we describe the conditional posterior densities for $\theta_1, \ldots, \theta_\ell$. Let $\boldsymbol{\theta}_{(i)}$ denote the vector of all $\boldsymbol{\theta}_i$ except the ith one. That is, $\boldsymbol{\theta}_{(i)} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_\ell)'$, $i = 1, \ldots, \ell$. It is pertinent to describe the conditional posterior density of $\boldsymbol{\theta}_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}$. First, we describe two important components of this distribution.

Under the baseline model, the likelihood function is

$$\begin{aligned}
p(y_i, r_i|\boldsymbol{\theta}_i) = \binom{n_i}{r_i}\binom{r_i}{y_i} & (\pi_{i1}p_i)^{y_i}(\pi_{i0}(1-p_i))^{r_i-y_i}\\
&\times \{(1-\pi_{i1})p_i + (1-\pi_{i0})(1-p_i)\}^{n_i-y_i}, \qquad (7)
\end{aligned}$$

$y_i = 0, \ldots, r_i$ and $r_i = y_i, y_{i+1}, \ldots, n_i$, with independence over i, $i = 1, \ldots, \ell$. Since the number of households in the ith domain with at least one crime for the nonrespondents is unknown, we denote it by the latent variable z_i, and the number of households with no victimization among the nonrespondents is $n_i - r_i - z_i$. The introduction of z_i into the analysis simplifies the computations. Then, incorporating the latent variables into our model, we obtain the augmented likelihood function

$$\begin{aligned}
p(y_i, r_i, z_i|\,\boldsymbol{\theta}_i) = \binom{n_i}{r_i}\binom{r_i}{y_i} & (\pi_{i1}p_i)^{y_i}(\pi_{i0}(1-p_i))^{r_i-y_i}\\
&\times \{(1-\pi_{i1})p_i\}^{z_i}\{(1-\pi_{i0})(1-p_i)\}^{n_i-r_i-z_i},
\end{aligned}$$

$i = 1, \ldots, \ell$. By marginalizing over $\boldsymbol{\theta}_i$, we have

$$A(y_i, r_i) = \binom{n_i}{r_i}\binom{r_i}{y_i}\sum_{z_i=0}^{n_i-r_i}\binom{n_i-r_i}{z_i}\frac{B(y_i+z_i+\mu_1\tau_1, n_i-y_i-z_i+(1-\mu_1)\tau_1)}{B(\mu_1\tau_1,(1-\mu_1)\tau_1)}$$
$$\times\frac{B(r_i-y_i+\mu_2\tau_2, n_i-r_i-z_i+(1-\mu_2)\tau_2)}{B(\mu_2\tau_2,(1-\mu_2)\tau_2)}$$
$$\times\frac{B(y_i+\mu_3\tau_3, z_i+(1-\mu_3)\tau_3)}{B(\mu_3\tau_3,(1-\mu_3)\tau_3)}, \quad i=1,\ldots,\ell. \tag{8}$$

The second quantity is the posterior density of $\boldsymbol{\theta}_i$ under the baseline model which is

$$f(p_i, \pi_{0i}, \pi_{1i}|\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\tau})$$
$$\propto \sum_{z_i=0}^{n_i-r_i}\left\{\binom{n_i-r_i}{z_i}B(y_i+z_i+\mu_1\tau_1, n_i-y_i-z_i+(1-\mu_1)\tau_1)\right.$$
$$\times B(r_i-y_i+\mu_2\tau_2, n_i-r_i-z_i+(1-\mu_2)\tau_2)B(y_i+\mu_3\tau_3, z_i+(1-\mu_3)\tau_3)$$
$$\times\frac{p_i^{y_i+z_i+\mu_1\tau_1-1}(1-p_i)^{n_i-y_i-z_i+(1-\mu_1)\tau_1-1}}{B(y_i+z_i+\mu_1\tau_1, n_i-y_i-z_i+(1-\mu_1)\tau_1)}$$
$$\times\frac{\pi_{i0}^{r_i-y_i+\mu_2\tau_2-1}(1-\pi_{i0})^{n_i-r_i-z_i+(1-\mu_2)\tau_2-1}}{B(r_i-y_i+\mu_2\tau_2, n_i-r_i-z_i+(1-\mu_2)\tau_2)}$$
$$\left.\times\frac{\pi_{i1}^{y_i+\mu_3\tau_3-1}(1-\pi_{i1})^{z_i+(1-\mu_3)\tau_3-1}}{B(y_i+\mu_3\tau_3, z_i+(1-\mu_3)\tau_3)}\right\}, \quad i=1,\ldots,\ell. \tag{9}$$

Now, letting Q_j denote the probability that θ_i is the same as θ_j conditional on α, we have

$$Q_j = \frac{p(y_i, r_i|\boldsymbol{\theta}_j)}{\alpha A(y_i, r_i) + \sum_{j=1, j\neq i}^{\ell} p(y_i, r_i|\boldsymbol{\theta}_j)},$$

where $p(y_i, r_i|\boldsymbol{\theta}_j)$ is given by Eq. (7) with $\boldsymbol{\theta}_i$ begin replaced by $\boldsymbol{\theta}_j$ and $A(y_i, r_i)$ is given by Eq. (8).

Then, using Theorem 1 of Escobar (1994), we have

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}\begin{cases} = \boldsymbol{\theta}_j & \text{with probability } Q_j, i \neq j \\ \sim f(\boldsymbol{\theta}_i|\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\tau}) & \text{with probability } (1 - \sum_{j=1, j\neq i}^{\ell} Q_j), \end{cases}$$

where $f(\boldsymbol{\theta}_i|\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\tau})$ is the probability density under the baseline model in Eq. (9). Note that the original ℓ areas replaced by k areas ($k \leq \ell$). That is, this conditional posterior density describes how the discreteness arises, and $\sum_{j=1, j\neq i}^{\ell} Q_j$ is the probability $\boldsymbol{\theta}_i$ is one of the other $\boldsymbol{\theta}_{i'}$, $i' \neq i$ conditional to α.

We use grids to obtain samples from the conditional posterior density of $(\boldsymbol{\mu}, \boldsymbol{\tau})$, given $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$ be the k distinct values, where $\boldsymbol{\theta}_i^* = (p_i^*, \pi_{i0}^*, \pi_{i1}^*)$. Then,

$$p(\boldsymbol{\mu}, \boldsymbol{\tau}|\boldsymbol{\theta}^{(*)}, k)$$
$$\propto \prod_{i=1}^{k}\left\{\frac{p_i^{*(\mu_1\tau_1-1)}(1-p_i^*)^{(1-\mu_1)r_1-1}}{B(\mu_1\tau_1,(1-\mu_1)\tau_1)}\frac{\pi_{i0}^{*(\mu_2\tau_2-1)}(1-\pi_{i0}^*)^{(1-\mu_2)\tau_2-1}}{B(\mu_2\tau_2,(1-\mu_2)\tau_2)}\right.$$
$$\left.\times\frac{\pi_{i1}^{*(\mu_3\tau_3-1)}(1-\pi_{i1}^*)^{(1-\mu_3)\tau_3-1}}{B(\mu_3\tau_3,(1-\mu_3)\tau_3)}\right\}\prod_{r=1}^{3}\frac{1}{(1+\tau_r)^2}.$$

For example, letting $a = \prod_{i=1}^{k} p_i^*$ and $b = \prod_{i=1}^{k}(1 - p_i^*)$, for (μ_1, τ_1) the joint conditional posterior density is

$$p(\mu_1, \tau_1|\boldsymbol{\theta}^*, k) \propto \frac{1}{(1+\tau_1)^2} \frac{a^{\mu_1\tau_1 - 1} b^{(1-\mu_1)\tau_1 - 1}}{B(\mu_1\tau_1, (1-\mu_1)\tau_1)}$$

Then, for μ_1

$$p(\mu_1|\tau_1, \boldsymbol{\theta}^*, k) \propto \frac{a^{(\mu_1\tau_1 - 1)} b^{(1-\mu_1)\tau_1 - 1}}{B(\mu_1\tau_1, (1-\mu_1)\tau_1)}.$$

For τ_1, we make the transformation $\tau_1 = \nu_1/(\mu_1 - \nu_1)$, with $0 < \nu_1 < \mu_1$, to have

$$p(\nu_1|\mu_1, \boldsymbol{\theta}^*, k) \propto \frac{a^{\mu_1(\nu_1/(\mu_1 - \nu 1)) - 1} b^{(1-\mu_1)(\nu_1/(\mu_1 - \nu_1)) - 1}}{B(\mu_1(\nu_1/(\mu_1 - \nu_1)), (1-\mu_1)(\nu_1/(\mu_1 - \nu_1)))}, \quad 0 < \nu_1 < \mu_1.$$

The conditional posterior density of α is only related to k. Escobar and West (1995) show how to get samples from the conditional posterior density, and their argument shows that it is same for all DPPs of the general three-stage hierarchical model. Specifically, the conditional posterior density for α is

$$p(\alpha|k) \propto (\kappa_0 + \alpha)^{-2} \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + \ell)}, \quad 0 < \alpha < \infty.$$

Transforming α to $\rho = \alpha/(\alpha + 1)$, $0 < \rho < 1$, the conditional posterior density for ρ is

$$p(\rho|k) \propto (1-\rho)^{-2} \left\{\frac{\rho}{1-\rho}\right\}^k \left\{\kappa_0 + \frac{\rho}{1-\rho}\right\}^{-2} \frac{\Gamma\{\rho/(1-\rho)\}}{\Gamma\{\ell + \rho/(1-\rho)\}}, \quad 0 < \rho < 1$$

where κ_0 is to be specified.

For μ_1, ν_1 and ρ, the grid procedure is the same. Bounded intervals improve the grid method. We stratify the range into a large number of grids (*e.g.*, 100) to approximate the probability density function by a probability mass function.

We have used the griddy Gibbs sampler to fit both the baseline and the DPP models. The baseline model was fit using the Metropolis-Hastings sampler in our previous work. However in this work for a more appropriate comparison we use the griddy Gibbs sampler. We drew 11,000 iterates, threw out the first 1000, and took every tenth. This is very conservative because convergence is very rapid.

4 NUMERICAL RESULTS

First, we consider the baseline model to assess the impact of the data on the prior distributions of the π_{i0} and π_{i1}. For the prior distributions, we fit the NCS data to the baseline model to get estimates of (μ_2, μ_3) and (τ_2, τ_3), and we sample the beta distributions $\pi_{is-2} \sim$ iid Beta$(\mu_s\tau_s/10, (1-\mu_s)\tau_s/10)$, $s = 2, 3$ to construct prior data. The griddy Gibbs sampler provide data for the baseline model from the corresponding posterior densities. Note that the prior distributions for the π_{i0} or the π_{i1} are the same, but the corresponding posterior densities are different. The comparisons are presented in Figure 1 for the π_{i0} and in Figure 2

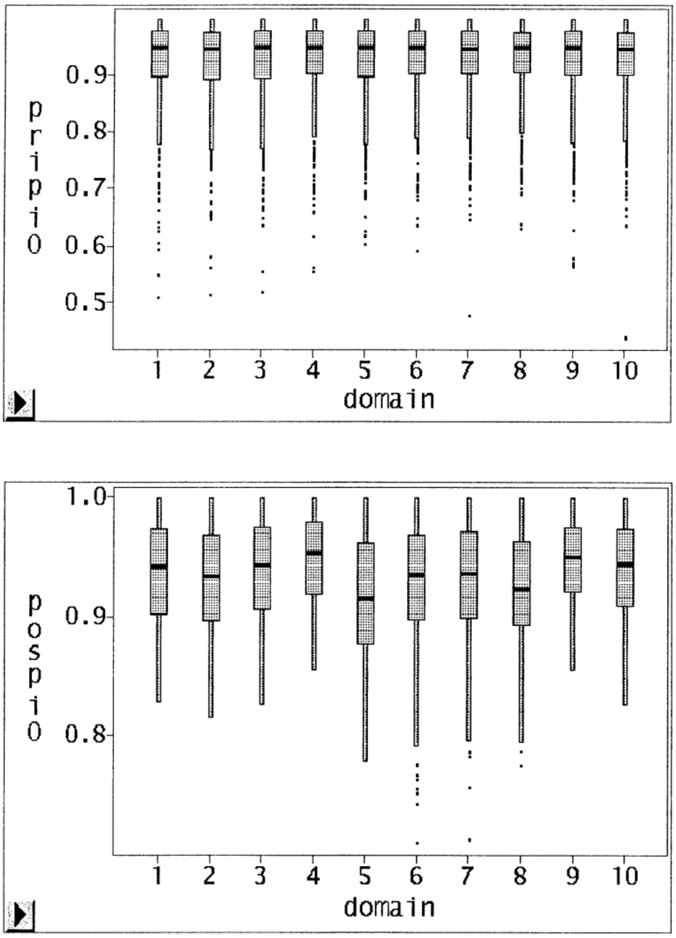

FIGURE 1 Box plots of the prior densities (top panel) and the posterior densities (bottom panel) of π_{i0} by domain.

for the π_{i1}. In both cases, the posterior distributions are less variable than the prior distributions, and the posterior distributions are different across the domains showing the impact of the data on the prior distributions.

Next, we have used the NCS data to compare inference about p_i and $\delta_i = (1-p)\pi_{i0} + p\pi_{i1}$; π_{i0} and π_{i1} are nuisance parameters.

We present the results in Tables II and III. In each table, we have presented the posterior means (PM), the posterior standard deviations (PSD), the numerical standard errors (NSE) and the 95% credible intervals (CI) for the baseline and the DPP models. To obtain 95% interval, we ordered the 1000 iterates of each parameter from the smallest to the largest, and obtain the 25th and 975th values. The NSE is a measure of the changes that are expected when the computations are repeated in exactly the same manner; we use the batch means method with 40 batches of length 25. Obviously, we need the NSE to be as small as possible.

First, consider the p_i in Table II. The PMs are very similar for the two models except perhaps for UIH, UNL and UNH. As is expected, the PSDs are larger for the DPP model except for UNH and RIL. The NSEs are larger for the DPP model, but observe how close they are for RIH. Consequently, the 95% CI for the DPP model are generally wider with overlaps for some domains.

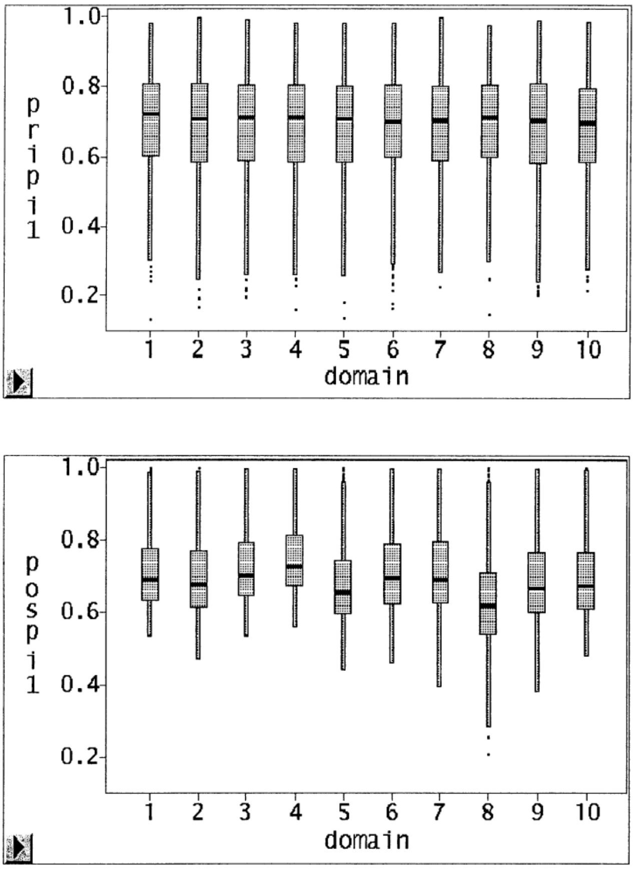

FIGURE 2 Box plots of the prior densities (top panel) and the posterior densities (bottom panel) of π_{i1} by domain.

We note one important feature. For the baseline model, the average (standard deviation) of the PM for the first seven domains is 0.273 (0.015) and for the last three domains 0.186 (0.024). For the DPP model, the average (standard deviation) of the PM for the first seven domains is 0.274 (0.003) and for the last three domains 0.197 (0.022). Thus, the two groups of domains are more distinct when the DPP model is fit rather than the baseline model.

Second, we consider the response probabilities δ_i in Table III. The PM for the two models are generally very similar with some differences. The largest difference occurs for UNL; compare 0.838 for the baseline model with 0.863 for the DPP model. The NSE are very similar and very small reflecting good performance of the griddy Gibbs sampler for the δ_i. It is surprising that the PSD for the DPP model are much smaller than the corresponding PSD for the baseline model (*e.g.*, the ratio of the PSD for the DPP versus the baseline model for UNH is 0.43). This makes the 95% CI for the DPP model shorter than those of the baseline model. This seems very interesting.

Another quantity of interest is the posterior probability mass function of k. Since the conditional posterior density of k is $p(k|\alpha) = s_\ell(k)\alpha^k\Gamma(\alpha)/\Gamma(\alpha+\ell)$, $k = 1, \ldots, \ell$, a Rao–Balckwellized (RB) estimator of $p(k|\boldsymbol{y}, \boldsymbol{r})$ is

$$\hat{p}_k^{(rb)} = M^{-1}\sum_{h=1}^{M} s_\ell(k)\alpha_{(h)}^k \frac{\Gamma(\alpha_{(h)})}{\Gamma(\alpha_{(h)}+\ell)}, \quad k = 1, \ldots, \ell,$$

TABLE II Comparison of the PM, PSD, NSE and 95% CI for p from the Baseline and DPP Models.

Domain	*PM*	*PSD*	*NSE*	*CI*
(a) Baseline model				
UCL	0.269	0.036	0.017	(0.200, 0.329)
UCH	0.262	0.038	0.016	(0.190, 0.328)
UIL	0.273	0.037	0.018	(0.198, 0.332)
UIH	0.254	0.034	0.014	(0.187, 0.313)
UNL	0.295	0.046	0.019	(0.205, 0.374)
UNH	0.291	0.056	0.020	(0.186, 0.408)
RIL	0.269	0.058	0.019	(0.164, 0.389)
RIH	0.178	0.049	0.020	(0.087, 0.274)
RNL	0.168	0.034	0.013	(0.105, 0.234)
RNH	0.213	0.034	0.015	(0.150, 0.275)
(b) DPP model				
UCL	0.274	0.045	0.041	(0.196, 0.327)
UCH	0.274	0.045	0.041	(0.196, 0.327)
UIL	0.275	0.045	0.041	(0.196, 0.329)
UIH	0.272	0.045	0.040	(0.186, 0.327)
UNL	0.280	0.049	0.041	(0.197, 0.365)
UNH	0.276	0.048	0.042	(0.196, 0.342)
RIL	0.270	0.049	0.039	(0.185, 0.331)
RIH	0.186	0.049	0.022	(0.084, 0.265)
RNL	0.183	0.042	0.020	(0.096, 0.248)
RNH	0.223	0.051	0.030	(0.126, 0.312)

Note: $p = \Pr(y = 1|p)$ where $y = 1$ for a victimized household and 0 otherwise.

where $\alpha_{(h)}$, $h = 1, \ldots, M$ are $M = 1000$ iterates from the griddy Gibbs sampler. It is known that the RB estimator has smaller mean square error than the corresponding empirical estimator.

We found an interesting feature in estimating the posterior mass function of k. Even if $p(k|\boldsymbol{y}, \boldsymbol{r})$ is negligible, this estimator can give a substantial probability to this value. For instance, at $\kappa_0 = 1$, of the 1000 iterates from the griddy Gibbs sampler virtually none has $k = 1$, yet the estimate of this probability is 0.225 when the RB estimator $\hat{p}_k^{(\boldsymbol{rb})}$ is used. Essentially the RB estimator splits the probability at $k = 2$ into $k = 1$ and $k = 2$, allocating a substantial probability to $k = 1$. This does not show that the RB estimator performs poorly. In fact, we have found that this behavior is associated with the estimation of α. We have computed the posterior probabilities at $\kappa_0 = 1$ for $k = 1, 2$ at the lower quartile (0.36), median (0.74) and upper quartile (1.35) of the posterior density of α; at $k = 2$ the probabilities are 0.39, 0.17, 0.05 and at $k = 3$ they are 0.40, 0.35, 0.20. It appears that the problem is caused by the fact that the posterior distribution tends to concentrate near zero (*i.e.*, α tends to be too small). This problem can be resolved if there is improved prior information about α which can be obtained through prior elicitation.

It is sensible to consider the alternative simple empirical estimator, based on the Gibbs iterates,

$$\hat{p}_k^{(e)} = \frac{N_k}{M}, \quad k = 1, \ldots, \ell$$

where N_k is the number of iterates out of M with k distinct groups. In Table IV, we present the $\hat{p}_k^{(e)}$ and $\hat{p}_k^{(\boldsymbol{rb})}$ for $\kappa_0 = 1$. (We will consider the entire table later.) The $\hat{p}_k^{(e)}$ indicate that $k = 2$

TABLE III Comparison of the PM, PSD, NSE and 95% CI for δ from the Baseline and DPP Models.

Domain	*PM*	*PSD*	*NSE*	*CI*
(a) Baseline model				
UCL	0.872	0.011	0.002	(0.851, 0.892)
UCH	0.864	0.014	0.003	(0.836, 0.889)
UIL	0.875	0.011	0.002	(0.853, 0.896)
UIH	0.893	0.014	0.003	(0.864, 0.920)
UNL	0.838	0.016	0.004	(0.807, 0.867)
UNH	0.861	0.028	0.005	(0.797, 0.907)
RIL	0.867	0.030	0.006	(0.801, 0.919)
RIH	0.866	0.026	0.006	(0.804, 0.909)
RNL	0.900	0.015	0.003	(0.867, 0.927)
RNH	0.884	0.012	0.003	(0.858, 0.907)
(b) DPP model				
UCL	0.870	0.006	0.003	(0.857, 0.883)
UCH	0.870	0.007	0.003	(0.855, 0.882)
UIL	0.870	0.006	0.003	(0.857, 0.881)
UIH	0.872	0.010	0.003	(0.858, 0.900)
UNL	0.863	0.015	0.004	(0.821, 0.881)
UNH	0.868	0.012	0.004	(0.837, 0.885)
RIL	0.870	0.013	0.003	(0.840, 0.898)
RIH	0.882	0.020	0.005	(0.826, 0.913)
RNL	0.891	0.014	0.003	(0.866, 0.922)
RNH	0.882	0.012	0.003	(0.862, 0.906)

Note: $\delta = \Pr(r = 1|p, \pi_0, \pi_1) = (1 - p)\pi_0 + p\pi_1$.

has estimated probability 0.524 as compared with 0.266 for the $\hat{p}_k^{(\boldsymbol{rb})}$. Notice that $\hat{p}_1^{(e)} = 0.000$ while $\hat{p}_1^{(\boldsymbol{rb})} = 0.225$, a substantial difference, as we remarked above.

We have looked at the clustering of the domains further. Of the 1000 iterates from the griddy Gibbs sampler, 524 indicate that there are two clusters of domains. Of these, 294 have two clusters: the first cluster is (UCL, UCH, UIL, UIH, UNL, UNH) and the second cluster is (RIH, RNL, RNH). We have considered posterior inference for p_1^* and p_2^* (*i.e.*, the proportion of households with at least one crime) conditional on these two clusters. For p_1^* and p_2^* respectively, the PM (standard deviations) are 0.269 (0.047) and 0.193 (0.041), and the 95% credible intervals are (0.205, 0.327) and (0.112, 0.249). See the PM in Table II. We also consider the three clusters. The number of iterates in which there are three clusters is 270; of these only 60 iterates have the clusters (UCL, UCH, UIL, UIH, UNL, UNH), (RIH, RNL) and (RNH). Thus, this is a potentially useful methodology that can be used to cluster the domains.

We have considered sensitivity to inference for the specification of κ_0 in the prior density for α. We consider five choices of κ_0 (*i.e.*, $\kappa_0 = 0.001, 0.01, 1.00, 100, 1000$). The response probabilities δ_i are clearly nonsensitive to the choice of κ_0; see Table V(b). There is some sensitivity to inference about the p_i, but this is reasonably small (see Tab. V(a)). But in Table IV, inference about k is sensitive to the choice of κ_0. For small values of κ_0, small values of k are dominant, and for large values of κ_0, large values of k are dominant. Also, note the disparity between the $p_k^{(e)}$ and $p_k^{(\boldsymbol{rb})}$ for different choices of κ_0; but note that $\kappa_0 = 0.001$ and $\kappa_0 = 1000$ are two extreme cases.

TABLE IV Comparison of Probability Mass Functions of k for Various Choices of κ_0.

	$\kappa_0 = 0.001$		$\kappa_0 = 0.01$		$\kappa_0 = 1$		$\kappa_0 = 100$		$\kappa_0 = 1000$	
k	$\hat{p}^{(e)}$	$\hat{p}^{(rb)}$	$\hat{p}^{(e)}$	$\hat{p}^{(rb)}$	$\hat{p}^{(e)}$	$\hat{p}^{(rb)}$	$\hat{p}^{(e)}$	$\hat{p}^{(rb)}$	$\hat{p}^{(e)}$	$\hat{p}^{(rb)}$
1	0.037	0.785	0.021	0.668	0.000	0.225	0.000	0.042	0.000	0.020
2	0.866	0.137	0.815	0.202	0.524	0.266	0.157	0.074	0.066	0.033
3	0.072	0.046	0.125	0.075	0.270	0.215	0.133	0.089	0.047	0.038
4	0.020	0.019	0.030	0.031	0.123	0.142	0.105	0.091	0.045	0.038
5	0.004	0.008	0.008	0.013	0.058	0.082	0.100	0.089	0.043	0.037
6	0.001	0.003	0.000	0.006	0.014	0.042	0.080	0.085	0.053	0.038
7	0.000	0.001	0.001	0.003	0.006	0.019	0.064	0.086	0.043	0.044
8	0.000	0.001	0.000	0.001	0.003	0.007	0.083	0.098	0.072	0.063
9	0.000	0.000	0.000	0.000	0.000	0.003	0.124	0.133	0.168	0.149
10	0.000	0.000	0.000	0.000	0.002	0.002	0.154	0.213	0.463	0.539

Note: $\hat{p}^{(e)}$ is the empirical estimator based on the Gibbs iterates and $\hat{p}^{(rb)}$ is the RB estimator.

As a summary, there are differences between the DPP and the baseline models. The PM for the p_i and their PSD under the DPP model are generally larger. It is surprising that inference for the δ_i is at least as good for the DPP model as for the baseline model. With respect to the choice of κ_0, inference about the p_i is mildly sensitive, inference about the δ_i is not sensitive, but inference about k is somewhat sensitive.

TABLE V Comparison of 95% CI of p and δ for Various Choices of κ_0.

	$\kappa_0 = 0.001$	$\kappa_0 = 0.01$	$\kappa_0 = 1$	$\kappa_0 = 100$	$\kappa_0 = 1000$
(a) Proportion p					
UCL	(0.190, 0.329)	(0.207, 0.331)	(0.196, 0.327)	(0.190, 0.333)	(0.191, 0.327)
UCH	(0.190, 0.329)	(0.207, 0.331)	(0.196, 0.327)	(0.189, 0.333)	(0.184, 0.328)
UIL	(0.190, 0.329)	(0.207, 0.331)	(0.196, 0.329)	(0.196, 0.333)	(0.196, 0.330)
UIH	(0.190, 0.329)	(0.207, 0.331)	(0.186, 0.327)	(0.189, 0.327)	(0.184, 0.320)
UNL	(0.190, 0.329)	(0.207, 0.338)	(0.197, 0.365)	(0.196, 0.375)	(0.199, 0.372)
UNH	(0.190, 0.329)	(0.207, 0.331)	(0.196, 0.342)	(0.189, 0.403)	(0.185, 0.402)
RIL	(0.167, 0.329)	(0.188, 0.331)	(0.185, 0.331)	(0.179, 0.380)	(0.155, 0.388)
RIH	(0.092, 0.304)	(0.115, 0.298)	(0.084, 0.265)	(0.089, 0.276)	(0.081, 0.265)
RNL	(0.095, 0.265)	(0.116, 0.264)	(0.096, 0.248)	(0.103, 0.244)	(0.103, 0.234)
RNH	(0.124, 0.311)	(0.129, 0.317)	(0.126, 0.312)	(0.131, 0.302)	(0.144, 0.289)
(b) Response rate δ					
UCL	(0.857, 0.884)	(0.859, 0.882)	(0.857, 0.883)	(0.852, 0.889)	(0.851, 0.890)
UCH	(0.857, 0.884)	(0.859, 0.882)	(0.855, 0.882)	(0.841, 0.886)	(0.839, 0.888)
UIL	(0.857, 0.884)	(0.859, 0.882)	(0.857, 0.881)	(0.855, 0.892)	(0.853, 0.894)
UIH	(0.860, 0.887)	(0.859, 0.887)	(0.858, 0.900)	(0.859, 0.916)	(0.862, 0.918)
UNL	(0.845, 0.884)	(0.833, 0.882)	(0.821, 0.881)	(0.808, 0.876)	(0.806, 0.876)
UNH	(0.857, 0.887)	(0.850, 0.882)	(0.837, 0.885)	(0.801, 0.901)	(0.798, 0.907)
RIL	(0.857, 0.890)	(0.852, 0.891)	(0.840, 0.898)	(0.812, 0.910)	(0.804, 0.922)
RIH	(0.851, 0.912)	(0.852, 0.911)	(0.826, 0.913)	(0.808, 0.911)	(0.807, 0.909)
RNL	(0.863, 0.915)	(0.869, 0.917)	(0.866, 0.922)	(0.867, 0.927)	(0.869, 0.926)
RNH	(0.862, 0.903)	(0.866, 0.904)	(0.862, 0.906)	(0.860, 0.908)	(0.859, 0.907)

Note: $p = \Pr(y = 1|p)$ where $y = 1$ for a victimized household and 0 otherwise; $\delta = \Pr(r = 1 \mid p, \pi_0, \pi_1) = (1 - p)\pi_0 + p\pi_1$.

5 CONCLUDING REMARKS

Our methodology on nonignorable nonresponse has been motivated by the NCS data which appear to have two groups of domains. To this end, we have made several contributions. First, we have developed the appropriate methodology for nonignorable nonresponse binary data under the DPP. This is good for survey samplers who generally assume that small area effects follow a common parametric distribution. Second, we have discussed the identifiability issue in nonignorable nonresponse models. Third, we have shown how to fit the model using MCMC methods. Fourth, we have shown that an alternative prior specification for α can be used. Finally, we have shown that there are benefits to be gained by using the DPP model over its baseline counterpart.

For the NCS data, the DPP model can help to identify the two groups of domains. With the baseline model, this is not so clear; there is no mechanism to tell how many groups there are. We found that there are differences between these models for estimating the p_i and δ_i. As expected, there is an increased variability when the DPP model is used over the baseline model, although surprisingly the DPP model is at least as precise for estimating the δ_i.

Also, we have found that the prior density for α (*i.e.*, $\alpha \sim \text{Gamma}(a, b)$), used by Escobar and West (1995), is sensitive to the choice of a and b. This prior is attractive because of its simplicity (*i.e.*, it is straight forward to draw a sample from it). Instead, we have used the shrinkage prior which has no moments (*i.e.*, almost noniformative) but is proper. However, the conditional posterior density for α is not simple, and we have shown how to draw a sample from it using a grid method. We have used the shrinkage prior for α (*i.e.*, $p(\alpha) = \kappa_0/(\kappa_0 + \alpha)^2$, $\kappa_0, \alpha > 0$). An important issue about this prior density is that one needs to choose κ_0. We have found that inference about k is sensitive to the choice of κ_0, inference for p is mildly sensitive and inference about δ is not sensitive.

One interesting problem can be addressed in which there may be prior information about the number of groups of domains. With the present theory on the DPP, this is not a simple matter. However, it may be possible to control the number of groups that are formed using appropriate prior information about the parameter α of the DPP (see the discussion in Escobar and West, 1995). One can easily guess k, say $k = k_0$, by inspection of the data or preferably by prior elicitation, and use this information to improve the prior specification of α instead of perturbing the DPP directly. This is a possible avenue for future research.

In small area estimation, our nonparametric Bayesian approach to nonignorable nonresponse is novel. Our procedure will be more beneficial in applications with many areas.

Acknowledgements

The authors are grateful to the referee whose comments help to improve the quality of the presentation. This work was supported by a contract from the National Center for Health Statistics.

References

Antoniak, C. E. (1974). Mixture of Dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya-urn schemes. *The Annals of Statistics*, **1**, 353–355.

Consoni, G. and Veronese, P. (1995). A Bayesian method for combining the results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935–944.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, **27**, 569–580.

De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, **15**, 129–142.

Escobar, M. D. (1994). Estimating normal means with Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–277.

Escobars, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Evans, R. and Sedransk, J. (2001). Combining data from experiments that may be similar. *Biometrika*, **88**, 643–656.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Ferguson, T. S. (1974). Prior distributions on space of probability measures. *The Annals of Statistics*, **2**, 615–629.

Forste, J. J. and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society, Series B*, **60**, 57–70.

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. Wiley, New York.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475–492.

Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, **89**, 278–288.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Malec, D. and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, **79**, 593–601.

Mallick, B. K. and Walker, S. G. (1997). Combining information from several experiments with nonparametric priors. *Biometrika*, **84**, 697–706.

Nandram, B. and Choi, J. W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, **97**, 381–388.

Nandram, B. and Choi, J. W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, **21**, 1189–1212.

Nandram, B., Han, G. and Choi, J. W. (2002). A hierarchical Bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, **28**, 145–156.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, **95**, 227–237.

Olson, R. L. (1980). A least squares correction for selectivity bias. *Econometrica*, **48**, 1815–1820.

Ritter, C. and Tanner, M. A. (1992). The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association*, **87**, 861–868.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–590.

Saphire, D. G. (1984). *Estimation of Victimization Prevalance Using Data from the National Crime Survey*, Lecture Notes in Statistics, 23. Springer-Verlag, New York.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Wilks, W. R. (1996). Bugs: Bayesian inference using Gibbs sampling, version 0.50. *Technical Report*, MRC Biostatistics Unit, Cambridge, UK.

Stasny, E. A. (1990). Symmetry in flows among reported victimization classifications with nonresponse. *Survey Methodology*, **16**, 305–330.

Stasny, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse. An example from the National Crime Survey. *Journal of the American Statistical Association*, **86**, 296–303.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed. Springer-Verlag, New York.

APPENDIX A

Conditional Posterior Density of $p(\alpha|k)$ (Escobar and West, 1995)

For n areas using results in Antoniak (1974), Escobar and West (1995) presented the probability mass function

$$p(k|\alpha) = s_n(k)\frac{\alpha^k\Gamma(\alpha)}{\Gamma(\alpha+n)}, \quad k = 1, \ldots, n$$

where $s_n(k)$, not involving α, are the absolute values of the Sterling numbers of the first kind. These Sterling numbers are given by the coefficients of the $(n-1)$ degree polynomial $\prod_{k=0}^{n-1}(x-k)$ where $\prod_{k=0}^{n-1}(x-k) \equiv \sum_{k=0}^{n} s_n(k)x^k$.

Letting D_n represent a configuration of the data into k groups, Escobar and West (1995) argued that

$$p(\alpha \mid k, \boldsymbol{\theta}, D_n) = p(\alpha | k),$$

where clearly $p(\alpha|k) \propto p(\alpha)p(k|\alpha)$ and $p(\alpha)$ is the prior density for α. They took $\alpha \sim G(a, b)$ and specified $a = 2$ and $b = 4$ for the astronomy data studied by Roeder (1990).

Finally, introducing the latent variable γ, where

$$\gamma \mid \alpha, k \sim \text{Beta}(\alpha + 1, n).$$

Escobar and West (1995) showed that

$$\alpha \mid \gamma, k \sim \lambda_{\gamma,k} G\{a + k, b - \log(\gamma)\} + (1 - \lambda_{\gamma,k}) G\{a + k - 1, b - \log(\gamma)\},$$

where $\lambda_{\gamma,k} = (a + k - 1)/\{a + k - 1 + n(b - \log(\gamma))\}$.