NISS

A Bayesian Semiparametric Model for Small Domain Estimation, with application to the National Survey of Recent College Graduates

Neung Soo Ha, Alan F. Karr, and Hang J. Kim

Technical Report 195
April 13, 2016

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

A Bayesian Semiparametric Model for Small Domain Estimation, with application to the National Survey of Recent College Graduates

Neung Soo Ha[1], Alan F. Karr[2], and Hang J. Kim[3]

[1]Nielsen Company, Seoul, Korea. , neung.ha@nielsen.com
[2]RTI International, Research Triangle Park, NC. karr@rti.org
[3]University of Cincinnati, Cincinnati, OH. hang.kim@uc.edu

Abstract

When sample sizes are too small to produce reliable direct estimates in survey statistics, the model-based methods are often used to obtain population-level quantities of interest for those small geographical areas or small population subgroup domains. One well-known model for small area/domain level estimates is the Fay-Herriot model, which can be interpreted as a linear mixed effects model in which the true domain-level means are normally distributed. However, it is challenging to verify their distributional assumption since they are not directly observable. In this paper, we formulate a semi-parametric extension of the Fay-Herriot model in which the default normality assumption for the true means is replaced by a nonparametric specification. While we investigate the intercept-only model, which is often used in the absence of the domain-level covariates, we illustrate the robustness of our estimators for domain-level means as well as the distribution of their "ensemble" through simulations under different distributional assumptions. Viability of the approach and the effects are illustrated using the 2008 National Survey of Recent College Graduates to estimate mean salaries for demographic subgroups of interest.

Keywords: Complex survey, Dirichlet process prior, Fay-Herriot model, National Survey of Recent College Graduates, Small area estimation.

1 Introduction

Many government agencies administer large-scale sample surveys to study various population attributes of interest, for which they typically represent large geographical areas, such as the entire country, or large domains, such as country's male population. Subject to cost constraints, these agencies construct survey designs such that the allocated survey samples yield the estimates with desired accuracies. Often, however, when the focus is on small geographical regions or population subdomains, these direct design-based estimates become unreliable, primarily due to small sample sizes. When those instances occur, survey analysts rely on small area estimation methodologies that "borrow strength" via explicit or implicit model-based approaches to optimally estimate the desired population attributes, such as the area means and the corresponding uncertainties. Furthermore, those model-based methods often incorporate supplementary data, like administrative records and other surveys, (for example, the American Community Survey (ACS)) to increase the reliability of the estimators. For a comprehensive review of the small area estimation literature, see Jiang and Lahiri (2001) or Rao and Molina (2015).

As illustrated in Jiang (2007), consider the estimation of true domain-level means, θ_i, where i is a domain index, for a continuous characteristic of population members—for instance, salary, as in Section 5. In this paper, we use the terms "domain" and "area" interchangeably. As discussed in more detail in Section 2, θ_i can be estimated directly using model-based methods even though they are defined from individual-level information. Within this context, one widely used model is the Fay-Herriot (FH) model (Fay and Herriot 1979; You 2008), which assumes normal distributions for both the domain-level direct estimates, $\hat{\theta}_i$, and their corresponding true means, θ_i. While the central limit theorem is often used as a justification for the distribution of $\hat{\theta}_i$ (Rao 2003), making the same assertion for θ_i is often problematic and difficult to verify (Sinharay and Stern 2003). Moreover, misspecification of their distribution can lead to undesired consequences for the estimators, such as incorrect posterior distribution convergence, bias in estimates, and poor posterior variance estimates.

When the assumption of normality for θ_i is not tenable, other parametric models have been proposed. For example, to capture the outliers, Bell and Huang (2006) considered the t-distribution,

while Farrell et al. (1994) considered Laplace priors for their model. More recently, Liu (2009) considered the use of the exponential power distribution. To use asymmetric distributions for θ_i, Fabrizi and Trivisano (2010) used a skewed exponential power distribution and Diallo and Rao (2014) used skewed normal distributions.

In this paper we propose a semiparametric model with nonparametric specification, based on a Dirichlet process prior, for the distribution of the θ_i. As demonstrated in Section 4, this model can successfully be applied when the assumed distribution for θ_i deviates from the normal distribution, for instance, by being heavy-tailed, asymmetric, or mixed. Moreover, the model performs comparably to the FH model even when θ_i are normally distributed, illustrating the robustness of our estimators with respect to the distribution. We pay special attention to not only the point estimates from the models but also their empirical distribution over the domains. To this end, we compare the empirical distributions between the ensemble of posterior estimates from our method to the corresponding estimates from the standard FH model (Fabrizi and Trivisano 2010).

The paper is organized as follows. Section 2 introduces the FH model and discusses its previous extensions. Section 3 describes our proposed approach with the Dirichlet process mixture (DPM) model. In Section 4, we use simulations to compare between the DPM model with the standard FH model under a variety of distributions for θ_i. In Section 5, we apply both the FH model and our DPM model to the 2008 National Survey of Recent College Graduates (NSRCG), which was conducted by the National Center for Science and Engineering Statistics at the U.S. National Science Foundation. We compare the domain-level salaries between the direct estimates, posterior means from the FH model, and those from our approach. Among our findings, the NSRCG salary data exhibit a multimodality captured by the DPM model but not by the FH model. Section 6 contains conclusions and some paths for further research.

2 Previous Model-based Estimation Methods

In this section, we describe the FH model, which was originally developed to obtain model-based county-level income estimates where the housing data and the tax records were used as the

area-level supplementary information. Recently, the U.S. Census Bureau used the FH model that incorporates data from the ACS and the Current Population Survey (CPS) to produce the estimates for the numbers of school-age children in poverty in each state and its counties. See http://www.census.gov/did/www/saipe/.

Let y_{ij} be the response, assumed numerical (in Section 5, the response is salary), of unit/person j in domain i, $i = 1,\ldots,m$. Domains can be defined in terms of geography, demography, or even a cross-classification of several variables. We wish to estimate the true domain means $\theta_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$, where N_i is the population size in domain i. The direct design-based (or Horvitz-Thompson) estimator is defined as $\hat{\theta}_i = \frac{\sum_{j\in s_i} w_{ij} y_{ij}}{\sum_{j\in s_i} w_{ij}}$, where w_{ij} is the survey weight for the unit j, s_i denotes the set of sampled units, and n_i is the number of sampled units in corresponding domain i.

For inference regarding θ_i, Fay and Herriot (1979) introduced the following two-level hierarchical model:

FH model

$$\text{Level 1 (}\textit{Sampling model}\text{)}: \qquad \hat{\theta}_i|\theta_i,\psi_i \overset{ind}{\sim} N(\theta_i,\psi_i) \tag{1}$$

$$\text{Level 2 (}\textit{Prior model}\text{)}: \qquad \theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \nu_i, \nu_i \overset{iid}{\sim} N(0,\sigma_\tau^2) \tag{2}$$

where the ν_i represent domain-level random effects that account for heterogeneity among domains. The variance component in (1), ψ_i, is called sampling variance and assumed to be known. The domain-level covariates are denoted as $\boldsymbol{x}_i$ where they are typically drawn from external sources. The hyperparameters $\boldsymbol{\beta}$ and σ_τ^2 are called model parameters, which are treated as unknowns and they are usually assigned non-informative or weak prior distributions in a hierarchical Bayesian setting.

In practice, when the domain-level covariates are not available (Sinharay and Stern 2003), the *Prior model* in (2) reduces to

$$\theta_i = \mu + \nu_i, \nu_i \overset{iid}{\sim} N(0,\sigma_\tau^2) \tag{3}$$

where it is often referred to as an unbalanced one-way analysis of variance (Jones and Spiegelhalter 2011). In a Bayesian setting, once we appropriately specify the distributions for model parameters

(μ, σ_τ^2), it becomes relatively straightforward to obtain the posterior distribution of θ_i, $f(\theta_i|\hat{\theta})$, via Gibbs sampling in a Markov chain Monte Carlo (MCMC) algorithm (You et al. 2003; Gelman et al. 2004).

Unlike the *Sampling model* in equation (1), where appeal to the central limit theorem can be warranted from the superpopulation perspective, the *Prior model* in (2) or in (3) is challenging to justify. To this end, there have been some efforts, especially for the *Prior model* to guard against using a single parametric distribution or having a normality assumption. For example, Maiti (2001) applied a finite mixture of normal distributions via hierarchical modeling and Articus and Burgard (2014) used the EM algorithm to obtain the inference for θ_i.

3 Dirichlet Process Extension of the Fay-Herriot Model

3.1 The Dirichlet Process

A random probability distribution G is called a Dirichlet process (DP) with base distribution G_0 and concentration parameter α on a space S if for every partition $\{T_1, T_2, \ldots, T_K\}$ of S,

$$(G(T_1), \ldots, G(T_K)) \sim Dir\left(\alpha G_0(T_1), \ldots, \alpha G_0(T_K)\right), \tag{4}$$

where *Dir* denotes the Dirichlet distribution. We denote this by $G \sim DP(\alpha, G_0)$. The larger α, the more G resembles G_0; in all cases, $E(G) = G_0$. In a Bayesian setting, we can model uncertainty in θ by assuming $\theta|G \overset{\text{iid}}{\sim} G$, where $G \sim DP(\alpha, G_0)$, which we call the DP prior.

For computational efficiency, one can use a constructive representation of the *DP*, referred to as the stick-breaking representation (Sethuramen 1994), given by

$$\begin{aligned} G(\cdot) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot) \\ \pi_k &= v_k \prod_{g<k} (1 - v_g) \qquad \text{where } \sum_{k=1}^{\infty} \pi_k = 1 \\ v_k &\sim Beta(1, \alpha), \end{aligned} \tag{5}$$

where $\delta_\theta(\cdot)$ is the point mass at θ.

In practice, the infinite sum in (5) is replaced by the finite sum, with a large value of K such that $\sum_{k=K+1}^{\infty}\pi_k$ has a distribution concentrated near zero. Ishwaran and James (2001) suggest the truncation approximation to *DP* such that

$$
\begin{aligned}
G(\cdot) &= \sum_{k=1}^{K}\pi_k\delta_{\theta_k}(\cdot) \qquad (6)\\
\pi_k &= v_k\prod_{g<k}(1-v_g) \ \text{ for } k=1,\ldots,K,\\
v_k &\sim Beta(1,\alpha) \ \text{ for } k=1,\ldots,K-1;\ v_K=1.
\end{aligned}
$$

The truncation is helpful to decrease the computational burden of the MCMC implemented for inference.

3.2 Our Approach Using the DP prior

Our model extends the FH model by relaxing the parametric distributional assumption for θ_i. Specifically, for the *Prior model* in (3), we use a Dirichlet process mixture (DPM), which has been widely used in Bayesian analysis (Escobar 1994; Muller et al. 1996). See Escobar and West (1995), Kim et al. (2014), and Kim et al. (2015) for more details.

To recapitulate notation, let $(\theta_1,\ldots,\theta_m)$ be the true domain-level means and $(\hat{\theta}_1,\ldots,\hat{\theta}_m)$ be the direct, design-based estimates, where m is the total number of domains. In our approach, we assume that each domain i belongs to one of K latent mixture components where $z_i\in\{1,\ldots,K\}$ denote the component index for domain i. The DPM model, then, comprises into three levels, such that:

DPM model

$$
\begin{aligned}
&\text{Level 1 (}\textit{Sampling model}\text{)}: && \hat{\theta}_i|\theta_i,\psi_i \overset{ind}{\sim} N(\theta_i,\psi_i)\\
&\text{Level 2 (}\textit{Prior model I}\text{)}: && \theta_i|z_i \overset{ind}{\sim} N(\mu_{z_i},\tau^2_{z_i})\\
&\text{Level 3 (}\textit{Prior model II}\text{)}: && z_i \overset{iid}{\sim} Categorical(\pi_1,\ldots,\pi_K), \qquad (7)
\end{aligned}
$$

where $(\pi_1, \ldots, \pi_K)$ are interpreted as component weights for the latent class k and follow the truncated *DP* in (6). Marginalized over z_i, the prior model of θ_i reduces to

$$p(\theta_i|\boldsymbol{\mu}, \boldsymbol{\tau}^2, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k N(\theta_i; \mu_k, \tau_k^2). \tag{8}$$

We use the conjugate prior for $\boldsymbol{\mu}, \boldsymbol{\tau}^2$ given by

$$\mu_k|\tau_k^2 \overset{ind}{\sim} N\left(0, \frac{\tau_k^2}{h_0}\right), \tau_k^2 \overset{iid}{\sim} IG(a_\tau, b_\tau), \tag{9}$$

where *IG* denotes the inverse Gamma distribution. Following Wang and Dunson (2011) and Kim et al. (2014), we simply chose the value of K large enough ($K = 25$ for simulations in Section 4 and for data analysis in Section 5) to include all relevant possibilities. In contrast to Ohlssen et al. (2007), where a prior distribution is assigned, we use a fixed value for $\alpha = 1$. For the other hyperparameters, we set fixed values as $a_\tau = 2$ and $b_\tau = 1$, which can be interpreted as small prior sample sizes and $h_0 = 0.1$. With these specifications, we can explicitly derive the posterior distributions for θ_i by the MCMC using a Gibbs sampler. See Appendix A for detail.

4 Simulation Study

4.1 Simulation Structure

We compare the performance of our proposed DPM model and the FH model via simulation studies in which the direct estimates, $\hat{\theta}_i$, follow the *Sampling model* in (1), but the true domain-level means, θ_i, may deviate from the normal distribution in (3). Specifically, we consider the following four cases in which the distributions for θ_i are defined as:

Case 1: Normal distribution. $\theta_i \overset{iid}{\sim} N(4, 1)$.

Case 2: Heavy-tailed distribution. $\theta_i \overset{iid}{\sim} T_2$, a t-distribution with 2 degrees of freedom.

Case 3: Skewed distribution. $\theta_i \overset{iid}{\sim} SN(\xi = 0, \omega = 1, \gamma = 2)$, where $SN(\xi, \omega, \gamma)$ denotes the skewed normal distribution with location parameter ξ, scale parameter ω and slant parameter γ.

Case 4: Mixed distribution. $\theta_i \overset{iid}{\sim} .3N(0,1) + .5N(5,1) + .2N(10,1)$.

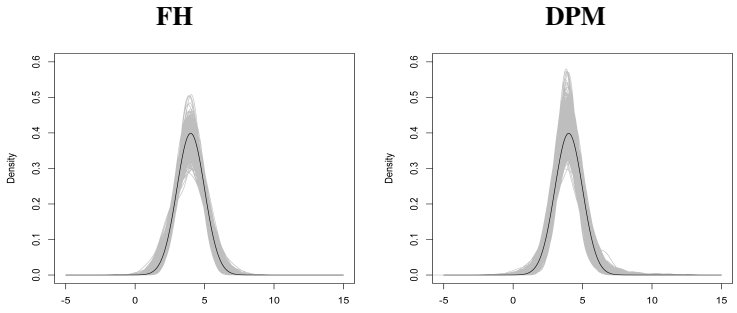
We separately generate a set of sampling variances $\psi_i \sim \Gamma(\text{shape} = 6, \text{rate} = 2)$ to apply in each case. The direct estimates, $\hat{\theta}_i$, are then generated from $\hat{\theta}_i \overset{ind}{\sim} N(\theta_i, \psi_i)$, $i = 1, \ldots, m$, and $m = 500$. We simulated a total of 100 sets of replicated samples for this study.

To implement the DPM approach, we follow the model and the prior distributions described in Section 3.2 and Appendix A. On the other hand, for the FH model we use weak priors for the hyperparameters, $\mu \sim N(0, 1000)$, $\sigma_\tau \sim Unif(0, \infty)$ suggested by Gelman (2006). For both methods, we used 10,000-iteration MCMC runs with 5,000-iteration burn-in periods with 4 thinning steps.

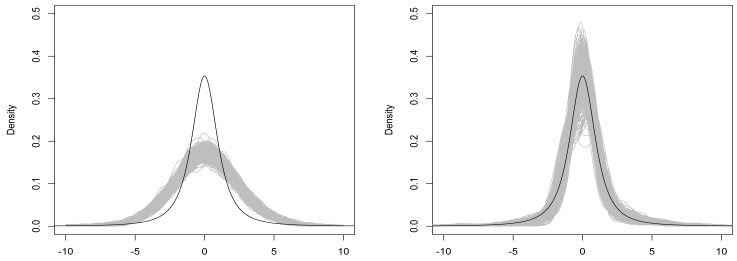
4.2 Simulation Results: One Replication

We first compare between our proposed DPM approach and the FH model based on one replicate sample. Specifically, by plotting them, we examine samples from each of their posterior distributions, $f(\theta_i|\hat{\boldsymbol{\theta}})$, against the true distribution for the four cases listed in Section 4.1. In Figure 1, the results from the FH models are in the left-hand panels, and those from the DPM models are in the right-hand panels, where we use kernel density estimates to represent the distributions from each posterior sample with gray curves and the true distribution with a black curve.

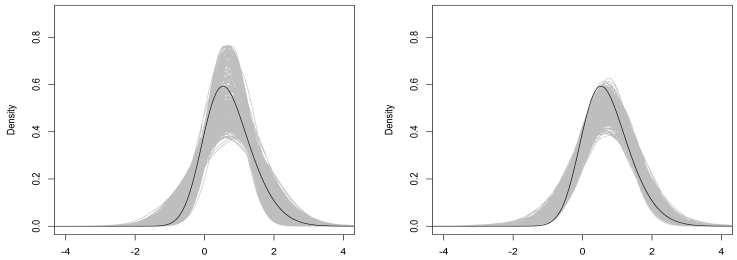
In Case 1, estimated densities of the posterior samples from the FH model are, unsurprisingly, closer to the true density than those from the DPM approach. However, those from the DPM approach also locate the true normal distribution of θ_i, albeit with larger variation. In Case 2, the estimated densities from the FH model are overly dispersed and thus fail to capture the true underlying density. On the other hand, those densities from the DPM approach capture the true density reasonably well, although they may seem too concentrated around the mean. At first glance for Case 3, the FH may seem to perform slightly better than the DPM approach since the density estimates capture the true density reasonably well. However, while the posterior draws from the FH model seem symmetric and lack the skewness, those from the DPM model appear to be more skewed to the left but yet only just capture the true distribution. In Case 4, the estimated densities from the FH model overly smooth the variation of the true distribution. Moreover, they perform very poorly due to the substantial distributional deviation from the true distribution for θ_i. On the other hand,



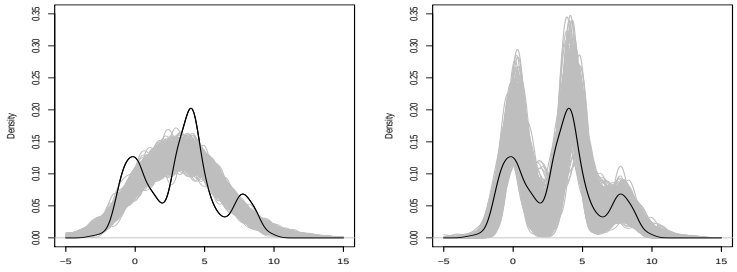
(a) **Case 1**: $\theta_i \sim N(4,1)$



(b) **Case 2**: $\theta_i \sim T_2$



(c) **Case 3**: $\theta_i \sim SN(0,1,2)$



(d) **Case 4**: $\theta_i \sim .3N(0,1) + .5N(5,1) + .2N(10,1)$

Figure 1: Density comparison between the FH model (left panels) and the DPM approach (right panels). The solid black curve represents the true density for θ_i.

the estimated densities from the DPM model capture the variation of true distribution almost perfectly although it does show wide variation around the true distribution. These types of multi-modal distribution characteristics are observed in our analyses of the NSRCG data in Section 5.

4.3 Simulation Results: Multiple Replications

Based on our analysis in the previous section, we have shown that our DPM approach is more effective when underlying true distribution deviates from a normal distribution. In this section, we extend our analysis by comparing the domain-level estimates based on the posterior means, $\hat{\theta}_i^* = E(\theta_i|\hat{\theta})$, from each model and their ensembles $\{\hat{\theta}_i^*\}_{i=1}^m$ by using the following summary measures on multiple replicates.

***1. Root Average Squared Bias*, (RASB)**

$$RASB = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i^* - \theta_i)^2},$$

which is an aggregate of differences between the posterior mean, $\hat{\theta}_i^*$, from each model of each simulated replicate and the true domain-level mean θ_i.

***2. Root Integrated Squared Error Loss*, (RISEL)**

$$RISEL = \sqrt{\int \left(F_n(t) - \tilde{F}_n^*(t)\right)^2 dt},$$

(Shen and Louis 1998), where $F_n(t)$ is the empirical distribution function (EDF) of the ensemble of true θ_i's, $\theta = \{\theta_i\}_{i=1}^m$:

$$F_n(t) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(\theta_i \leq t),$$

and $\tilde{F}_n^*(t)$ is the corresponding EDF for the estimators, $\hat{\theta}^* = \{\hat{\theta}_i^*\}_{i=1}^m$, from each replicate. RISEL measures the average difference between two empirical distributions.

Table 1: Averages of summary statistics from posterior means over all 100 replicated samples. The numbers in the parenthesis shows the 5% and 95% quantiles.

	RASB		RISEL		K-S	
	FH	DPM	FH	DPM	FH	DPM
Case 1: Normal dist'n	1.210 (1.139, 1.295)	1.201 (1.139, 1.273)	0.029 (0.022, 0.045)	0.031 (0.024, 0.050)	0.081 (0.065, 0.111)	0.090 (0.073, 0.137)
Case 2: t-dist'n	2.141 (2.081, 2.202)	1.723 (1.665, 1.792)	0.030 (0.027,0.032)	0.010 (0.008,0.013)	0.158 (0.146, 0.172)	0.081 (0.067, 0.102)
Case 3: Skewed normal	0.867 (0.741, 0.993)	0.880 (0.826, 0.958)	0.056 (0.035, 0.123)	0.042 (0.031, 0.062)	0.141 (0.089, 0.331)	0.107 (0.084, 0.159)
Case 4: Mixture of normals	4.495 (4.352, 4.642)	2.032 (1.961, 2.105)	0.047 (0.044, 0.053)	0.032 (0.026, 0.043)	0.103 (0.093, 0.111)	0.087 (0.073, 0.118)

3. Kolmogorov-Smirnov Statistic, (K-S)

$$D(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^*) = \sup_t |F_n(t) - \tilde{F}_n^*(t)|,$$

which measures the maximum distance between the EDFs from the estimators and the true domain-level means.

Table 1 shows the averages of each summary statistic; the numbers in the parenthesis represent the 5% and 95% quantiles over 100 replicates. For Case 1, the DPM and FH results are essentially identical, showing that the DPM model suffers no performance degradation when θ_i are in fact normally distributed, i.e., when the assumptions for underlying model of the FH are satisfied. For Case 2, the DPM approach outperforms the FH model rather dramatically for all three measures, especially for RISEL and K-S. For Case 3, DPM appears to be marginally superior to FH in terms of means, but it is more so for RISEL and K-S than for RASB. This is expected from the result in the previous section since the DPM approach is more successful at capturing the skewness of the distribution. Additionally, the DPM approach yields shorter inter-quantile distances than FH, suggesting that its estimates show less variation. Lastly in Case 4, there is a factor-of-two difference in RASB, a 50% difference in RISEL and approximately a 20% difference in K-S, all in favor of the DPM although the DPM produced wider inter-quantile distances.

5 Application to Salary Estimation for the NSRCG

In this section, we illustrate our DPM approach by analyzing salary data from the 2008 National Survey of Recent College Graduates (NSRCG). We cannot, of course, know the "ground truth", but our results suggest a mixed (multi-modal) distribution of the domain-level means, implying that higher credence could be placed in the results from the DPM model as demonstrated in Section 4.

5.1 The NSRCG

Conducted by the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation, the NSRCG provides information about recent recipients of bachelor's and master's degrees in science, engineering, and health (SEH) fields from U.S. academic institutions. The NSRCG was a biennial, cross-sectional survey carried out from 1973 to 2010, and collected demographic, educational and employment information from respondents. According to the NCSES website (http://www.nsf.gov/statistics/srvyrecentgrads/), NSRCG data "help users understand and predict trends in education, employment opportunities, and salaries of recent (SEH) graduates."

In particular, we analyze the 2008 NSRCG. For the survey, the eligible participants were under the age 76, non-institutionalized, and obtained their degrees during the 2006 and 2007 academic years (that is, between July 1, 2005 and June 30, 2007). The survey design follows a two-phase sampling scheme. In the first phase, a sample of 288 institutions was selected with probability proportional to size (PPS) from a list of 2,027 eligible institutions, and in the second phase approximately 18,000 graduates of those institutions who are in the U.S. were selected. For more information about the survey design, see http://www.nsf.gov/statistics/nsf12328/content.cfm?pub_id=4169&id=3.

The 2008 NSRCG survey was designed to provide reliable national estimates for domains defined by degree type (bachelor's or master's), major field of degree, race/ethnicity, and gender. However, it was *not* designed to provide estimates for a cross-classification of those domains, and thus constructing such small-domain estimates was the original motivation for this study.

5.2 Analysis of Salaries

Building on the work of Carrillo and Karr (2013) regarding estimation of salaries of Ph.D. recipients using data from NCSES Survey of Doctorate Recipients, we restrict the data to respondents who are employed full time, work more than 35 weeks, and report salary income exceeding $5,000 per year. The reduced data set contains approximately 8,300 respondents.

With these criteria, we estimate mean salaries for small domains constituting a cross-classification of five variables—gender, race, degree level, field of degree, and Carnegie code of the degree-granting institution. Although more detail is available in the data, these variables were re-coded to the levels shown in Table 2. When fully crossed, the total number from the five variables resulted in 190 domains since two of these domains are empty in our case. The distribution of the sample sizes, on the square-root scale, is shown in Figure 2, which shows that more than one-half of the domains have sample sizes less than 19.

Table 2: The five variables used in the analysis of mean salaries for a total of 190 domains of interest.

Variable	Level	Notation
Gender	Female	F
	Male	M
Race	Asian only	A
	Black only	B
	White only	W
	All other races, including mixed	O
Degree level	Bachelor's	B
	Graduate	G
Field of degree	Biological and environmental sciences	Bio_Env
	Computational and mathematical sciences	Comp_Sci
	Engineering	Eng
	Physical sciences	Physical
	Social sciences	Soc
	Other SEH-related fields	Rel
Carnegie code	Research and doctorate-granting	R_U
	Other	Other

To calculate the domain-level estimates, let y_{ij} be the annual salary of individual j in domain i.

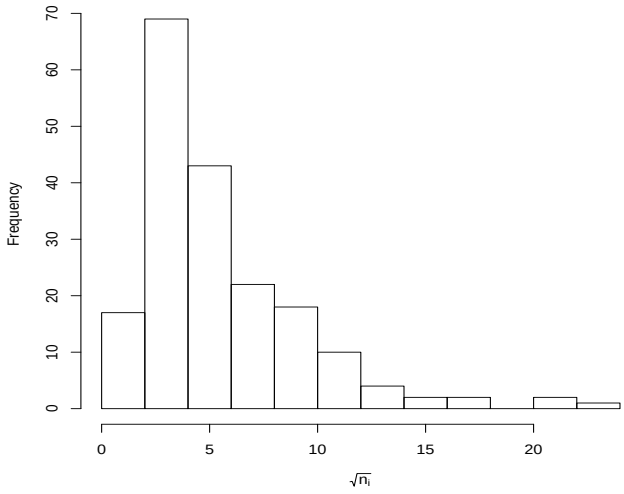


Figure 2: Distribution of the sample sizes, in square root scale, of the 190 non-empty domains defined by gender, race, degree level, field of degree, and Carnegie code of the degree-granting institution.

The population domain-level means are given by

$$\theta_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}, \tag{10}$$

where N_i is the population size in domain i. The direct design-based, or Horvitz-Thompson estimator of θ_i is defined as

$$\hat{\theta}_i = \frac{\sum_{j \in s_i} w_{ij} y_{ij}}{\sum_{j \in s_i} w_{ij}}, \tag{11}$$

where s_i is the set of sampled individuals and w_{ij} is the survey weight for person j in domain i. The weights reflect the complex sample design in the NSRCG. Direct use of design-based sampling variance, ψ_i, for all domains is not a viable option due to the small sample sizes. Thus, we apply an adjustment method similar to that in Ha et al. (2014). For a detailed description of our method, see Appendix B.

5.3 Result for Domain-level Salaries

The complete set of domain-level direct design-based, FH, and DPM estimates appears in the table in Appendix D. Here we compare the results in terms of five domains with the highest and

Table 3: The five domains with highest estimated salaries (top) and five domains with lowest estimated salaries (bottom). The underlined numbers represent the common domains between the FH and the DPM.

Five Domains with Highest Salaries

Gender	Race	Degree Level	Field	Carnegie	n_i	Salary ($)
Design-based estimates						
M	A	G	Rel	Other	29	107,648
M	O	G	Rel	Other	1	103,000
M	A	G	Rel	R_U	12	97,485
M	O	G	Rel	R_U	4	92,701
M	A	G	Rel	Other	6	92,133
FH estimates						
M	W	G	Rel	R_U	42	89,475
F	A	G	Eng	Other	24	77,173
M	W	G	Comp_Sci	R_U	96	77,144
F	B	G	Comp_Sci	R_U	27	75,938
M	W	G	Eng	Other	105	74,603
DPM estimates						
F	A	G	Eng	Other	24	77,173
M	W	G	Comp_Sci	R_U	96	77,144
F	B	G	Comp_Sci	R_U	27	75,938
M	W	G	Eng	Other	105	74,603
M	A	G	Comp_Sci	R_U	42	74,511

Five Domains with Lowest Salaries

Gender	Race	Degree Level	Field	Carnegie	n_i	Salary ($)
Design-based estimates						
F	A	B	Physical	Other	7	23,776
M	B	G	Physical	R_U	9	26,755
M	B	B	Bio_Envir	Other	5	28,990
M	B	B	Physical	Other	14	29,222
M	O	B	Bio_Envir	Other	1	30,000
FH estimates						
F	A	B	Physical	Other	7	23,776
M	B	B	Physical	Other	14	29,222
F	W	B	Social	Other	203	32,482
F	B	B	Physical	Other	15	33,197
F	W	B	Bio_Envir	Other	73	33,320
DPM estimates						
M	B	B	Physical	Other	14	29,222
F	W	B	Social	Other	203	32,482
F	A	B	Bio_Envir	R_U	17	32,641
F	W	B	Bio_Envir	Other	73	33,320
F	B	B	Physical	Other	15	33,197

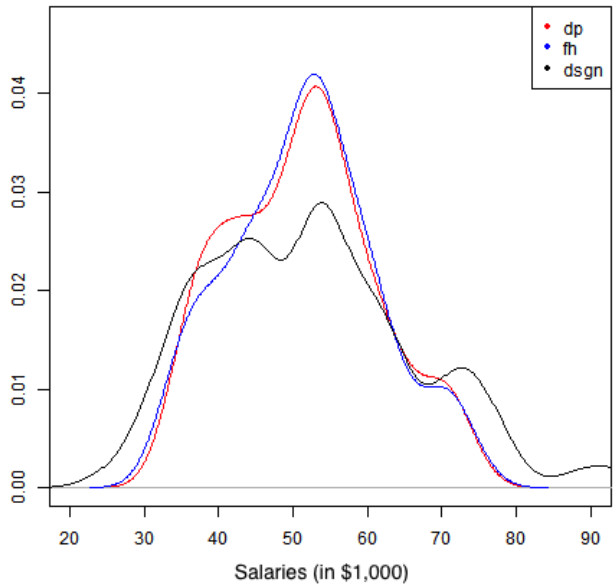


Figure 3: Kernel densities for posterior means of salary estimates.

lowest estimated salaries, which appear in Table 3. This table shows that the posterior means from the FH and the DPM models are quite similar for most domains, which is expected from the posterior model checking with the Bayesian p-value in the Appendix C. Evidently, the choice of prior distribution did not make a meaningful difference when we examined only the posterior means from both models.

Figure 3 shows the estimated densities of $\hat{\theta}_i$'s from all three (DPM, FH, and design-based) estimates. As expected, there is a large difference between the design-based estimates and the FH and DPM estimates. However, there is much less difference between posterior estimates from the FH and the DPM models overall while the distributional difference of posterior means for the FH and the DPM model are more apparent when the salary estimates are between $35,000 and $45,000. Table 4 shows the groups and posterior means for each model. However, when we examined estimates that fall within that range, we found no systematic patterns of differences for the domains.

Table 4: Comparison of FH and DPM salary estimates lying between $35,000 and $45,000.

Gender	Race	Degree Level	Field	Carnegie	FH	DPM
F	B	B	Comp_Sci	R_U	45,161	44,041
M	W	G	Physical	R_U	45,464	44,413
M	A	B	Social	R_U	45,284	44,240
F	O	G	Social	R_U	34,793	36,623
F	B	B	Bio_Envir	Other	34,045	35,243
F	A	B	Physical	Other	33,836	36,677
M	B	B	Eng	Other	46,086	44,445

5.4 Distribution of Subdomain Salaries

Finally, we compare the models in terms of how they differentiate salaries by type of degree. Figure 4 contains Box plots of the salary distributions for recipients of bachelor's degrees (left) and graduate degrees (right). Within each pair of Box plots, FH estimates are on the left, and DPM estimates are on the right. For both degree levels, estimates from the DPM model have slightly less variation than those from the FH model, but the differences are not substantial, and the median estimates from the FH model slightly exceed those from the DPM model.

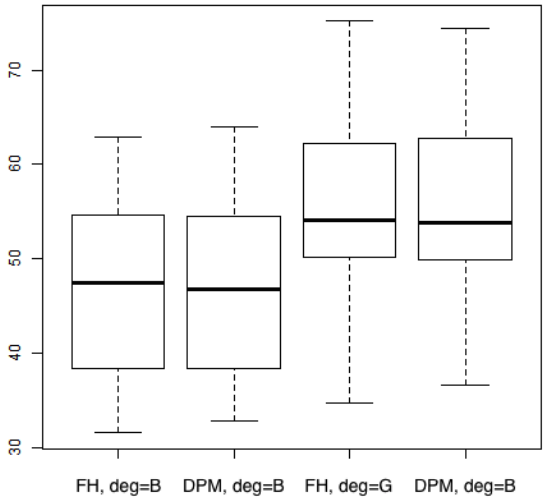


Figure 4: Box plots comparing densities of subdomain salaries by degree level. Left: FH estimates compared to DPM estimates for bachelors degree recipients. Right: FH estimates compared to DPM estimates for graduate degree recipients.

Figure 5, however, tells a more nuanced story. In the figure, we see that the density estimates

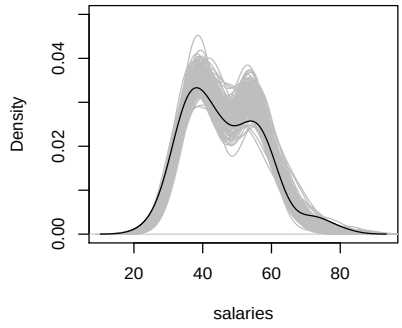
from both methods are cross-classified with the degree types: The DPM density estimates are on the left and the FH density estimates of the right; bachelor's degrees are at the top, and graduate degrees are at the bottom. The grey lines show posterior density estimates of θ_i from both models, and the black lines show the distribution of design-based estimates. As expected from the Bayesian p-value analysis, the distribution of the design-based estimates lies within the distributions from the two models; see Appendix C.

We observe substantial differences between distributions from the DPM and FH models when their shapes are compared. For bachelor's degrees (upper plots in Figure 5), the posterior distribution from the DPM model clearly shows a multi-modality, whereas the posterior distribution from the FH model does not, i.e., it is flatter and with no clear multi-modality. For graduate degrees (lower plots in Figure 5), the results are similar: the posterior density estimates of the DPM model demonstrate a mixture of distributions with different modes, whereas those of the FH model appears smoother in comparison. Based on our simulation results in Section 4, we believe that the flexible distributional assumption of the DPM approach allows it to conform better to the true underlying distribution, which appears to be multi-modal.

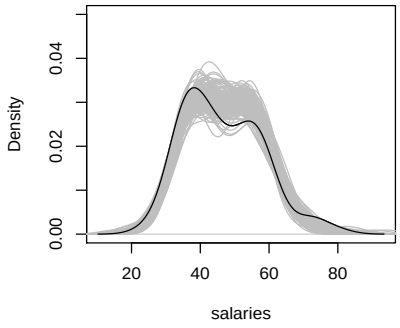
6 Concluding Remarks

In this paper, we have introduced an extension of the classical FH model by employing a DP prior for the distribution of the true domain-level means. Using simulation studies, we demonstrated that our approach with the DPM model outperforms the FH model when the true distribution deviates from normality—heavy tails, skewness and mixtures, and with no loss of performance when the normality assumption is satisfied.

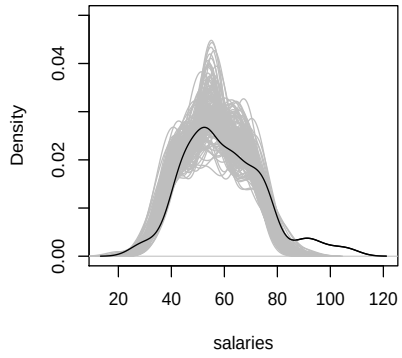
For the 2008 NSRCG, the DPM model and the FH model were both applied to estimate the salaries for 190 domains comprising from a full cross-classification of gender, race, degree-level, field of degree, and Carnegie classification. The differences between the estimates produced by two models are subtle but meaningful. Based on our findings from the simulations studies, we claim that the underlying distribution for domain-level salary averages may exhibit a multi-modality and



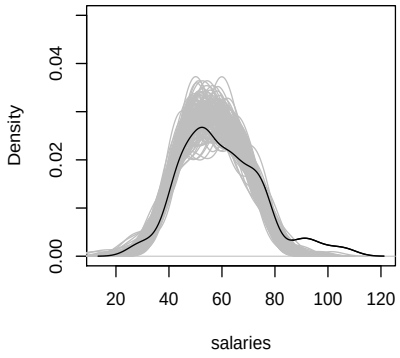
(a) DPM, deg=B



(b) FH, deg=B



(c) DPM, deg=G



(d) FH, deg=G

Figure 5: Density estimates comparing subdomain salaries by degree level. *Left:* DPM estimates, with bachelor's degrees above and graduate degrees below. *Right:* FH estimates, with bachelor's degrees above and graduate degrees below.

that the DPM model could be more successful with capturing that characteristic.

Similar to other model-based estimation methods, our approach is still subject to possible model failure. One possible solution could be using a benchmarking method similar to that in Datta et al. (2009). Benchmarking is a popular method in small area estimation because the method provides techniques such that the sum of the estimates for small domains becomes equivalent to that of the corresponding larger domains.

Our analysis focused on the intercept-only model for the prior distribution since the NSRCG does not contain relevant covariates for the domains of our interest. However, for many cases in small area estimation problems, there are available domain/area-level covariates, and our DPM

approach can be extended to those general cases.

References

Articus, C. and J. P. Burgard. 2014. *A finite mixture Fay Herriot-type model for estimating regional rental prices in Germany*. Research Papers in Economics. Universitat Trier.

Bell, W. R. and E. T. Huang. 2006. "Dealing with influential observationa and outliers in small area estimation". In XXIII International Symposium on Methodological Issues. Ottawa, Canada.

Carrillo, I. and A. F. Karr. 2013. "Combining cohorts in longitudinal surveys". *Survey Methodology* 39, 149–182.

Datta, G. S., M. Ghosh, R. Steorts, and J. Maples. 2009. *Bayesian benchmarking with applications to small area estimation*. Research Report Series. The U.S. Census Bureau.

Diallo, A. and J. N. K. Rao. 2014. "Small area estimation of complex parameters under unit-level models with skew-normal errors". In Proceedings of the Survey Research Methods Section, ASA, 970–984.

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

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

Fabrizi, E. and C. Trivisano. 2010. "Robust linear mixed models for small area estimation". *Journal of Statistical Planning and Inference* 140, 433–443.

Farrell, P. J., B. MacGibbon, and T. J. Tomberlin. 1994. "Protection against outliers in empirical Bayes estimation". *Canadian Journal of Statistics* 22, 365–376.

Fay, R. E. and R. A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data". *Journal of the American Statistical Association* 74, 269–277.

Fay, R. E. and G. Train. 1995. "Aspects of survey and Model-based postcensal estimation of income and poverty characteristics and states and counties". In Proceedings of the Government Statistics Section, ASA, 154–159.

Gelman, A. 2006. "Prior distributions for variance parameters in hierarchical models". *Bayesian Analysis* 1, 515–533.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. *Bayesian Data Analysis*. New York, NY.: Chapman and Hall/CRC.

Ha, N. S., P. Lahiri, and P. Parsons. 2014. "Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey". *Statistics in Medicine* 33, 3932–3945.

Ishwaran, H. and L. F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors". *Journal of the American Statistical Association* 96, 161–173.

Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York, NY.: Springer.

Jiang, J. and P. Lahiri. 2001. "Empirical best prediction for small area inference with binary data". *Annals of Institute of Statistical Mathematics* 53, 217–243.

Jones, H. E. and D. Spiegelhalter. 2011. "The identification of "unusual" health-care providers from a hierarchical model". *The American Statistician* 65, 154–163.

Kim, H., J. P. Reiter, Q. Wang, L. H. Cox, and A. F. Karr. 2014. "Multiple imputation of missing or faulty values under linear constraints". *Journal of Business and Economic Statistics* 32, 375–386.

Kim, H. J., L. H. Cox, A. F. Karr, J. P. Reiter, and Q. Wang. 2015. "Simultaneous Edit-Imputation for Continuous Microdata". *Journal of the American Statistical Association* 110(511), 987–999. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.2015.1040881.

Liu, B. 2009. "Adaptive hierarchical Bayes estimation of small-area proportions". In Proceedings of the Survey Research Methods Section, ASA, 3785–3799.

Maiti, T. 2001. "Generalized linear mixed models for small area estimation". *Journal of Statistical Planning and Inference* 98, 225–238.

Muller, P., A. Erkanli, and M. West. 1996. "Bayesian curve fitting using multivariate normal mixtures". *Biometrika* 83, 67–79.

Ohlssen, D. I., L. D. Sharples, and D. J. Spiegelhalter. 2007. "Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons". *Statistics in Medicine* 26, 2088–2112.

Rao, J. N. K. 2003. *Small Area Estimation*. Hoboken, NJ.: Wiley Series in Survey Methodology.

Rao, J. N. K. and I. Molina. 2015. *Small Area Estimation*. 2nd. New York, NY.: Wiley.

Sethuramen, J. 1994. "A constructive definition of Dirichlet priors". *Statistica Sinica* 4, 639–650.

Shao, J. 1996. "Resampling methods in sample surveys". *Statistics* 27, 203–254.

Shen, W. and T. Louis. 1998. "Triple-goal estimates in two-stage hierarchical models". *Journal of Royal Statistical Society, Series B* 60, 455–471.

Sinharay, S. and H. S. Stern. 2003. "Posterior predictive model checking in hierarchical models". *Journal of Statistical Planning and Inference* 111, 209–221.

Wang, L. and D. Dunson. 2011. "Fast Bayesian inference in Dirichlet process mixture models". *Journal of Computational and Graphical Statistics* 20, 196–216.

Wolter, Kirk. 1985. *Introduction to Variance Estimation*. New York, NY.: Springer.

You, Y. 2008. "An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada". *Survey Methodology* 34, 19–27.

You, Y., J. N. K. Rao, and J. Gambino. 2003. "Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach". *Survey Methodology* 29, 25–32.

Appendix A Posterior Inference via MCMC

We can construct the posterior distribution for the DPM model via MCMC with the following Gibbs sampling algorithm.

Step 1. For each $i = 1, \ldots, m$, draw $\theta_i \sim N(\mu_i^*, \tau_i^*)$, where

$$\mu_i^* = \frac{\sigma_i^2 \mu_{z_i} + \tau_{z_i}^2 \hat{\theta}_i}{\sigma_i^2 + \tau_{z_i}^2}, \text{ and } \tau_i^* = \frac{\sigma_i^2 \tau_{z_i}^2}{\sigma_i^2 + \tau_{z_i}^2}.$$

Step 2. For each $k = 1, \ldots, K$, draw $\sigma_k^2 \sim \text{IG}(a_k^*, b_k^*)$, and then draw $\mu_k \sim \text{N}\left(\mu_k^*, \sigma_k^2/h_k^*\right)$, where

$$
\begin{aligned}
n_k &= \sum_{\{i:z_i=k\}} 1, & \bar{\theta}_k &= \frac{1}{n_k} \sum_{\{i:z_i=k\}} \theta_i, \\
h_k^* &= n_k + h_0, & \mu_k^* &= \frac{1}{h_k}\left(n_k \bar{\theta}_k + h_0 \mu_0\right), \\
a_k^* &= a_\tau + \frac{n_k}{2}, & b_k^* &= b_\tau + \frac{\sum_{\{i:z_i=k\}} (\theta_i - \bar{\theta}_k)^2}{2} + \frac{(\bar{\theta}_k - \mu_0)^2}{2(1/n_k + 1/h_0)}.
\end{aligned}
$$

Step 3. For each $k = 1, \ldots, K-1$, draw $v_k \sim \text{Beta}\left(1 + n_k, \alpha + \sum_{g>k} n_g\right)$ and let $v_K = 1$. Then calculate the mixture component weights $\pi_k = v_k \prod_{g<k}(1 - v_g)$ for $k = 1, \ldots, K$.

Step 4. For each $i = 1, \ldots, n$, draw $z_i \sim \text{Categorical}(\pi_{i1}^*, \ldots, \pi_{iK}^*)$, where

$$
\pi_{ik}^* = \frac{\pi_k \text{N}(\theta_i; \mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \text{N}(\theta_i; \mu_{k'}, \sigma_{k'}^2)}
$$

for $k = 1, \ldots, K$.

Appendix B Sampling Variance Estimation

In the Fay-Herriot model, the sampling variances ψ_i are assumed to be known; however, in practice their estimated values are used. There are two commonly used methods to obtain estimated sampling variances. The first method uses a jackknife replication method to develop replicate weights (Fay and Train 1995), and this method requires construction of replicate subsamples by using the survey design information, such as strata or the primary sampling units, (Shao 1996). The second method employs the Generalized Variance Function (GVF, Wolter 1985), for approximating variances. The GVF method is a model-based method in which the model describes the relationship between the relative variance of a survey estimator and its expectation.

For our variance estimates, we primarily used the jackknife method, based on replicate weights available in the NSRCG datasets. However, due to the small sample sizes in many domains, some of the variance estimates were either undefined or small. Especially, the design effects (ratios of the variance under the survey design and variance under simple random sampling) for those domains

were less than one, which is uncommon under complex survey designs. Thus, we have made an adjustment in the following manner, with assumptions similar to those in Ha et al. 2014.

First, we examined the variance estimates at larger domains defined by gender $\times$ race $\times$ degree level. We made a conservative assumption, for those subdomains with design effect less than one, that the variation for all subdomains would be similar to that of the corresponding larger domains. Let ψ_i be a variance estimate for small domain i and ψ_j be the corresponding variance estimate where the domain i is a subdomain within the domain j. Then, we assume that $\psi_i/n_i \approx \psi_j/n_j$, where n_i and n_j are sample sizes for domains i and j. Finally, we replaced the ψ_i with ψ_i^* as defined by

$$\psi_i^* = \psi_j \times \frac{n_i}{n_j} \tag{12}$$

for those domains in which the design effect was less than one.

Appendix C Model Fit via Posterior Predictive Model Checking

Sinharay and Stern 2003 and Gelman et al. 2004 and discussed using the Bayesian p-value for checking the adequacy of a model. Let $D(\hat{\theta}^{\text{obs}}, \theta)$ be a test statistic on the observation $\hat{\theta}^{\text{obs}}$ and the parameter θ. Let $\tilde{\theta}$ represent a sample draw from the posterior distribution, $f(\theta|\hat{\theta}^{\text{obs}})$ and let $\tilde{\theta}^*$ represent a draw from $f(\hat{\theta}|\tilde{\theta})$ as in (1). Then marginally, $\tilde{\theta}^*$ is a sample drawn from the posterior predictive distribution $f(\hat{\theta}|\hat{\theta}^{\text{obs}})$. Subsequently, we can define our test statistic D as a χ^2-type discrepancy measure such that,

$$D(\hat{\theta}, \theta) = \sum_{i=1}^{m} \frac{(\hat{\theta}_i - \theta_i)^2}{\psi_i}, \tag{13}$$

where ψ_i denotes the sampling variance for domain i.

By using this discrepancy measure, D, the posterior predictive $p-$value is defined as

$$p_B = \text{Prob}\{D(\tilde{\theta}^*, \theta) \geq D(\hat{\theta}^{\text{obs}}, \theta)|\hat{\theta}^{\text{obs}}\}, \tag{14}$$

Using the posterior simulated samples from the Gibbs sampling, the posterior predictive $p-$value

(14) can be approximated easily. For each iteration ℓ and simulated value $\tilde{\theta}^{\ell}$, we can draw $\tilde{\theta}^{*,\ell}$, and consequently compute $D(\hat{\theta}^{\text{obs}}, \tilde{\theta}^{\ell})$ and $D(\tilde{\theta}^{*,\ell}, \tilde{\theta}^{\ell})$. Then equation (14) can be approximated as:

$$p_B \approx B^{-1} \sum_{\ell=1}^{B} \mathbb{I}\{D(\hat{\boldsymbol{\theta}}^{*,\ell}, \boldsymbol{\theta}^{\ell}) \geq D(\hat{\boldsymbol{\theta}}^{\text{obs}}, \boldsymbol{\theta}^{\ell})\}, \tag{15}$$

where B denotes the total number of Gibbs samplers and $\mathbb{I}$ is an indicator function. An extreme value (near 0 or 1) of p_B indicates possible lack-of-fit, whereas the value near 0.5 reflects adequacy for a given model. For our study with the 2008 NSRCG, the Bayesian p-value under the FH model is 0.399 and under the DPM model is 0.425, and thus, both models are adequate for explaining the data.

Appendix D Salary ($1K) Estimates for the NSRCG

Gender	Race	Degree Level	Field of Degree	Carnegie Code	Salary	Standard Error	Design Effect	Variance	n_i	Fay-Herriot	Dirichlet Process	Difference (DPM - FH)	N_i
F	W	B	Comp_Sci	R_U	51.117	2.301	0.952	20.038	38	51.110	51.740	0.630	4929.369
M	W	B	Comp_Sci	R_U	58.513	2.717	1.103	7.384	69	58.066	57.547	0.519	18622.677
F	A	B	Comp_Sci	R_U	52.812	3.730	0.765	37.548	14	52.458	52.303	0.156	2256.267
M	A	B	Comp_Sci	R_U	63.172	2.414	0.783	10.754	36	62.335	62.849	0.515	8340.328
F	B	B	Comp_Sci	R_U	44.712	3.345	1.093	11.192	14	45.162	44.042	1.120	822.380
M	B	B	Comp_Sci	R_U	72.592	16.913	2.144	286.062	10	58.262	58.129	0.133	1545.942
F	O	B	Comp_Sci	R_U	44.881	3.088	0.701	732.356	6	50.326	50.006	0.321	387.907
M	O	B	Comp_Sci	R_U	59.935	5.088	1.067	25.886	7	58.581	58.343	0.239	1439.888
F	W	G	Comp_Sci	R_U	63.211	3.876	1.827	15.022	75	62.064	62.474	0.409	2461.580
M	W	G	Comp_Sci	R_U	77.144	3.437	1.576	11.812	96	75.174	74.463	0.710	9237.450
F	A	G	Comp_Sci	R_U	62.185	2.726	1.255	7.430	28	61.744	62.195	0.451	3128.425
M	A	G	Comp_Sci	R_U	74.511	4.147	1.169	17.194	42	72.089	71.815	0.274	9197.577
F	B	G	Comp_Sci	R_U	75.939	4.072	1.167	16.578	27	73.331	72.875	0.456	316.010
M	B	G	Comp_Sci	R_U	71.221	3.720	0.750	98.605	20	62.706	63.588	0.882	585.757
F	O	G	Comp_Sci	R_U	77.438	15.752	2.038	248.111	8	60.428	60.916	0.489	280.011
M	O	G	Comp_Sci	R_U	62.382	13.626	0.499	2748.865	2	52.244	51.682	0.562	201.950
F	W	B	Bio_Envir	R_U	34.222	1.656	1.444	2.742	71	34.484	34.888	0.405	13908.531
M	W	B	Bio_Envir	R_U	35.531	2.057	1.581	4.230	79	35.985	36.262	0.277	16405.749
F	A	B	Bio_Envir	R_U	32.642	3.046	1.264	9.277	17	33.875	34.882	1.007	3895.783
M	A	B	Bio_Envir	R_U	36.891	3.742	1.483	14.003	15	38.346	38.219	0.127	3210.508
F	B	B	Bio_Envir	R_U	34.443	3.737	1.319	13.966	14	36.231	36.735	0.504	1421.714
M	B	B	Bio_Envir	R_U	41.019	6.016	1.463	36.192	7	43.093	41.896	1.198	614.300
F	O	B	Bio_Envir	R_U	35.865	2.853	1.898	8.139	19	36.705	37.146	0.440	2149.156
M	O	B	Bio_Envir	R_U	40.515	7.154	1.096	51.177	4	43.554	42.261	1.293	1008.661
F	W	G	Bio_Envir	R_U	42.638	3.256	1.509	10.604	56	43.235	42.175	1.060	4092.723
M	W	G	Bio_Envir	R_U	45.825	4.257	1.802	18.123	53	46.572	45.895	0.678	5646.711
F	A	G	Bio_Envir	R_U	55.552	4.718	1.132	22.261	17	54.959	54.832	0.127	1457.720
M	A	G	Bio_Envir	R_U	54.522	10.120	1.025	102.411	16	53.131	53.099	0.033	1318.987
F	B	G	Bio_Envir	R_U	62.351	8.633	1.468	74.523	17	58.803	58.453	0.350	283.225
M	B	G	Bio_Envir	R_U	52.954	2.555	0.659	657.367	3	52.329	51.051	1.278	61.968
F	O	G	Bio_Envir	R_U	49.152	12.373	1.443	153.082	6	50.845	49.934	0.911	305.575
M	O	G	Bio_Envir	R_U	31.564	4.803	0.554	1832.577	3	50.184	49.878	0.306	165.403
F	W	B	Physical	R_U	36.921	2.011	1.709	4.045	79	37.375	37.509	0.134	3311.844
M	W	B	Physical	R_U	37.680	2.471	1.481	6.105	74	38.343	38.388	0.045	5429.125
F	A	B	Physical	R_U	40.126	3.575	1.567	12.782	26	41.115	40.366	0.748	703.643
M	A	B	Physical	R_U	40.326	3.402	0.886	27.653	14	42.152	41.076	1.077	643.205
F	B	B	Physical	R_U	40.309	4.124	2.418	17.011	24	41.385	40.593	0.792	302.035
M	B	B	Physical	R_U	45.710	8.496	2.335	72.187	10	47.315	46.444	0.871	312.763
F	O	B	Physical	R_U	35.289	2.911	0.802	292.942	15	46.401	45.595	0.806	290.104
M	O	B	Physical	R_U	39.052	5.056	1.450	25.561	10	41.003	40.203	0.800	331.449
F	W	G	Physical	R_U	53.472	4.281	1.365	18.330	53	53.154	53.452	0.298	1431.638
M	W	G	Physical	R_U	45.014	3.501	1.128	12.257	57	45.465	44.413	1.051	3642.088
F	A	G	Physical	R_U	40.726	7.332	0.961	193.366	8	47.099	46.269	0.830	422.466
M	A	G	Physical	R_U	46.376	7.009	1.041	49.133	18	47.827	46.705	1.122	968.925
F	B	G	Physical	R_U	45.005	4.561	0.923	365.284	12	49.422	48.875	0.547	47.981
M	B	G	Physical	R_U	26.756	7.849	0.795	219.122	9	41.913	42.084	0.171	433.696
F	O	G	Physical	R_U	41.606	6.901	1.394	47.630	7	44.211	42.981	1.229	145.972
M	O	G	Physical	R_U	59.600	2.896	0.446	916.288	6	52.639	52.183	0.455	241.486
F	W	B	Soc	R_U	36.845	1.690	1.114	2.857	219	37.153	37.191	0.038	58476.158
M	W	B	Soc	R_U	43.458	2.296	1.369	5.272	147	43.737	42.982	0.755	39793.151
F	A	B	Soc	R_U	42.360	1.520	0.421	9.222	57	42.995	41.970	1.026	12402.901
M	A	B	Soc	R_U	44.966	2.637	1.060	6.955	35	45.284	44.240	1.044	7920.778
F	B	B	Soc	R_U	36.721	1.526	0.995	20.477	37	38.510	38.648	0.138	5887.567
M	B	B	Soc	R_U	50.354	20.915	1.180	437.450	27	50.950	50.246	0.704	3670.130
F	O	B	Soc	R_U	51.778	14.483	1.004	209.759	29	51.687	51.300	0.388	6668.851
M	O	B	Soc	R_U	43.190	3.641	1.340	13.256	26	43.943	42.659	1.284	5432.709
F	W	G	Soc	R_U	46.073	1.813	2.086	3.287	268	46.179	45.587	0.593	13651.393
M	W	G	Soc	R_U	53.059	2.998	1.417	8.987	126	52.882	53.562	0.680	7874.127
F	A	G	Soc	R_U	52.175	4.332	1.038	18.763	32	52.114	52.568	0.453	2169.035
M	A	G	Soc	R_U	52.216	6.115	1.092	37.398	19	52.137	52.067	0.070	1180.318
F	B	G	Soc	R_U	44.267	3.027	2.487	9.162	70	44.806	43.424	1.381	1359.590
M	B	G	Soc	R_U	58.346	10.060	3.983	101.197	30	55.236	55.397	0.161	692.505

Gender	Race	Degree Level	Field of Degree	Carnegie Code	Salary	Standard Error	Design Effect	Variance	n_i	Fay-Herriot	Dirichlet Process	Difference (DPM - FH)	N_i
F	O	G	Soc	R_U	30.046	6.244	4.244	38.993	31	34.794	36.623	1.830	1225.449
M	O	G	Soc	R_U	62.654	4.698	0.643	274.886	20	55.435	55.070	0.365	685.995
F	W	B	Eng	R_U	54.624	1.012	1.772	1.024	272	54.637	54.334	0.303	7851.723
M	W	B	Eng	R_U	57.143	0.874	1.247	0.764	428	57.134	56.925	0.209	41757.955
F	A	B	Eng	R_U	57.414	2.044	1.390	4.176	80	57.223	56.561	0.661	3017.704
M	A	B	Eng	R_U	58.710	1.732	1.499	3.001	115	58.514	58.440	0.074	12210.722
F	B	B	Eng	R_U	55.372	1.414	1.204	1.998	76	55.349	54.815	0.535	843.120
M	B	B	Eng	R_U	55.216	2.376	1.962	5.646	54	55.097	54.533	0.564	2146.541
F	O	B	Eng	R_U	54.500	2.734	1.222	7.476	40	54.342	54.236	0.106	735.317
M	O	B	Eng	R_U	57.300	4.552	2.500	20.718	49	56.655	56.078	0.577	3220.609
F	W	G	Eng	R_U	64.386	1.668	2.332	2.781	406	64.151	64.353	0.202	4400.171
M	W	G	Eng	R_U	69.049	1.586	2.041	2.517	506	68.711	68.859	0.148	17121.917
F	A	G	Eng	R_U	61.960	2.792	1.787	7.794	119	61.357	61.641	0.284	5164.136
M	A	G	Eng	R_U	69.778	2.312	1.153	5.347	175	69.131	69.193	0.062	13508.570
F	B	G	Eng	R_U	62.512	2.244	1.883	5.034	115	62.118	62.454	0.336	320.147
M	B	G	Eng	R_U	73.131	3.884	2.390	15.085	79	70.858	71.008	0.150	865.719
F	O	G	Eng	R_U	73.878	10.268	2.802	105.442	41	64.163	64.595	0.432	233.432
M	O	G	Eng	R_U	62.124	7.804	3.875	60.898	52	58.612	58.838	0.226	1969.797
F	W	B	Rel	R_U	48.223	2.086	1.107	4.353	68	48.298	48.092	0.206	22960.482
M	W	B	Rel	R_U	55.148	2.682	0.899	35.134	21	54.593	54.602	0.009	5213.612
F	A	B	Rel	R_U	54.473	4.311	0.954	43.806	12	53.354	53.810	0.456	3323.187
M	A	B	Rel	R_U	43.119	3.122	0.715	129.049	3	47.186	46.358	0.828	400.617
F	B	B	Rel	R_U	50.193	3.392	0.908	63.137	12	50.874	50.187	0.687	2846.446
M	B	B	Rel	R_U	79.920	7.473	0.408	1102.055	5	54.461	54.614	0.153	606.607
F	O	B	Rel	R_U	39.990	5.821	0.486	2197.067	2	50.999	50.827	0.172	648.793
M	O	B	Rel	R_U	70.905	0.701	0.500	321.118	2	57.504	57.053	0.451	209.352
F	W	G	Rel	R_U	61.666	1.934	0.796	32.977	119	59.776	59.509	0.268	27775.236
M	W	G	Rel	R_U	89.475	10.623	1.413	112.838	42	72.434	71.379	1.055	8483.949
F	A	G	Rel	R_U	69.784	15.529	0.995	154.692	10	60.330	60.846	0.516	2250.846
M	A	G	Rel	R_U	97.485	17.736	1.329	314.575	12	65.250	65.353	0.103	1311.377
F	B	G	Rel	R_U	67.718	8.804	0.839	208.734	21	57.950	57.452	0.498	2089.641
M	B	G	Rel	R_U	53.237	9.734	0.876	394.420	5	51.920	51.523	0.397	318.903
F	O	G	Rel	R_U	55.702	5.132	1.387	26.340	10	54.968	55.019	0.051	1387.232
M	O	G	Rel	R_U	92.701	29.269	1.198	856.662	4	57.426	57.523	0.097	768.869
F	W	B	Comp_Sci	Other	39.218	1.712	0.894	13.359	57	40.263	39.807	0.456	8225.145
M	W	B	Comp_Sci	Other	48.942	1.756	1.080	3.085	102	49.044	49.226	0.182	28962.298
F	A	B	Comp_Sci	Other	45.341	8.382	0.478	262.835	2	49.273	48.561	0.713	270.952
M	A	B	Comp_Sci	Other	60.467	5.092	0.894	35.195	11	58.577	58.600	0.024	3253.467
F	B	B	Comp_Sci	Other	58.168	9.820	2.036	96.426	16	55.421	55.784	0.363	1192.103
M	B	B	Comp_Sci	Other	40.325	6.248	1.561	39.043	16	43.030	41.706	1.324	2474.899
F	O	B	Comp_Sci	Other	45.312	7.145	1.067	51.048	6	47.000	45.740	1.260	433.861
M	O	B	Comp_Sci	Other	57.484	8.201	0.844	91.748	7	55.174	55.273	0.099	1652.306
F	W	G	Comp_Sci	Other	56.600	10.288	3.452	105.850	39	54.473	53.949	0.524	1517.866
M	W	G	Comp_Sci	Other	86.938	10.724	1.385	115.004	41	70.764	70.183	0.581	2542.505
F	A	G	Comp_Sci	Other	53.768	5.904	1.054	34.854	9	53.421	53.474	0.053	1869.551
M	A	G	Comp_Sci	Other	63.583	8.835	1.098	78.050	7	59.350	59.157	0.193	1713.134
F	B	G	Comp_Sci	Other	59.055	6.132	1.222	37.597	23	57.328	57.559	0.232	233.512
M	B	G	Comp_Sci	Other	72.276	10.464	1.516	109.490	8	63.168	63.648	0.480	230.120
F	O	G	Comp_Sci	Other	50.844	1.966	0.527	232.660	6	51.047	51.064	0.017	110.494
M	O	G	Comp_Sci	Other	78.357	14.400	1.199	207.365	5	62.222	62.323	0.101	232.416
F	W	B	Bio_Envir	Other	33.321	1.776	1.117	3.153	73	33.750	34.170	0.420	15863.275
M	W	B	Bio_Envir	Other	35.362	2.404	1.429	5.779	52	35.952	36.474	0.522	10476.930
F	A	B	Bio_Envir	Other	33.619	5.870	1.586	34.462	11	37.427	37.811	0.384	2097.115
M	A	B	Bio_Envir	Other	39.040	7.132	0.982	64.525	6	42.957	42.342	0.615	1758.101
F	B	B	Bio_Envir	Other	32.749	3.354	2.269	11.247	19	34.046	35.243	1.197	1459.954
M	B	B	Bio_Envir	Other	28.991	2.619	0.498	1102.055	5	49.050	48.534	0.516	419.428
F	O	B	Bio_Envir	Other	37.045	3.544	1.068	12.560	10	38.226	38.161	0.065	776.384
M	O	B	Bio_Envir	Other	30.000	0.000		642.236	1	47.665	47.737	0.072	319.505
F	W	G	Bio_Envir	Other	39.021	5.443	2.384	29.631	36	40.931	40.257	0.674	2331.408
M	W	G	Bio_Envir	Other	37.283	5.726	1.272	32.784	18	39.784	39.403	0.381	1888.476
F	A	G	Bio_Envir	Other	54.302	2.638	0.706	515.641	3	51.664	51.904	0.240	336.627
M	A	G	Bio_Envir	Other	45.532	3.740	0.895	235.212	5	49.546	48.898	0.649	353.043
F	B	G	Bio_Envir	Other	40.099	3.572	1.089	12.759	7	41.079	40.418	0.661	106.845
M	B	G	Bio_Envir	Other	46.189	10.230	1.307	104.649	5	48.631	47.662	0.969	113.463
F	O	G	Bio_Envir	Other	66.099	3.315	0.254	348.989	4	55.996	54.209	1.787	242.537
F	W	B	Physical	Other	34.810	2.017	1.746	4.069	89	35.307	35.682	0.376	4181.038
M	W	B	Physical	Other	42.379	6.596	1.922	43.505	78	44.582	43.196	1.386	5453.701

Gender	Race	Degree Level	Field of Degree	Carnegie Code	Salary	Standard Error	Design Effect	Variance	n_i	Fay-Herriot	Dirichlet Process	Difference (DPM - FH)	N_i
F	A	B	Physical	Other	23.776	1.778	0.545	75.096	7	33.836	36.677	2.841	226.073
M	A	B	Physical	Other	53.469	3.251	0.724	64.525	6	52.720	52.829	0.109	387.747
F	B	B	Physical	Other	33.198	2.229	1.140	4.970	15	33.785	34.442	0.657	239.609
M	B	B	Physical	Other	29.223	4.170	1.817	17.389	14	31.613	34.024	2.411	204.678
F	O	B	Physical	Other	34.939	3.868	1.894	14.962	14	36.636	37.130	0.495	227.879
M	O	B	Physical	Other	52.070	7.141	1.011	50.993	6	51.852	51.818	0.034	181.435
F	W	G	Physical	Other	48.365	6.909	1.395	47.738	20	49.289	48.714	0.575	536.007
M	W	G	Physical	Other	44.339	6.395	1.501	40.897	16	45.875	45.174	0.702	1178.005
F	A	G	Physical	Other	50.680	0.346	0.017	515.641	3	51.230	51.243	0.013	22.779
M	A	G	Physical	Other	52.259	4.388	0.662	235.212	5	52.021	51.498	0.524	163.300
F	B	G	Physical	Other	43.187	4.648	1.223	21.600	7	44.168	42.997	1.171	30.819
M	B	G	Physical	Other	58.763	3.158	1.090	9.971	5	58.234	57.453	0.780	83.341
M	O	G	Physical	Other	52.500	0.000		5497.730	1	51.929	51.277	0.652	4.921
F	W	B	Soc	Other	32.482	1.110	1.674	1.233	203	32.655	32.885	0.230	60785.668
M	W	B	Soc	Other	38.018	1.527	1.399	2.332	129	38.221	38.233	0.012	38075.362
F	A	B	Soc	Other	36.447	2.397	0.778	18.774	28	38.163	38.256	0.093	6263.777
M	A	B	Soc	Other	46.641	4.012	0.737	32.262	12	47.544	47.056	0.488	2470.130
F	B	B	Soc	Other	36.925	2.971	1.061	8.826	58	37.810	37.931	0.121	10103.688
M	B	B	Soc	Other	36.622	4.229	1.079	17.887	14	38.307	38.552	0.245	2497.635
F	O	B	Soc	Other	31.077	2.122	0.776	175.765	25	42.454	42.635	0.181	5019.505
M	O	B	Soc	Other	36.413	3.118	1.104	9.722	12	37.363	37.646	0.284	2658.069
F	W	G	Soc	Other	44.594	1.719	1.545	2.954	135	44.742	44.243	0.498	13696.746
M	W	G	Soc	Other	53.632	3.211	1.052	10.309	46	53.515	53.696	0.181	3221.371
F	A	G	Soc	Other	54.530	6.711	1.465	45.034	10	53.622	53.851	0.229	959.774
M	A	G	Soc	Other	46.731	9.420	0.845	235.212	5	50.139	48.522	1.617	400.851
F	B	G	Soc	Other	41.989	3.267	2.621	10.672	70	42.747	41.717	1.030	2256.816
M	B	G	Soc	Other	52.165	3.716	0.447	103.795	19	51.948	51.314	0.633	450.944
F	O	G	Soc	Other	46.349	3.015	1.158	9.089	18	46.680	45.852	0.828	1401.111
M	O	G	Soc	Other	50.741	7.057	2.471	49.796	13	51.005	50.878	0.126	842.644
F	W	B	Eng	Other	51.866	1.763	1.498	3.109	87	51.828	52.697	0.870	1900.255
M	W	B	Eng	Other	54.903	1.248	1.589	1.557	171	54.866	54.455	0.412	16912.362
F	A	B	Eng	Other	61.180	4.951	1.972	24.514	7	59.696	59.525	0.170	285.852
M	A	B	Eng	Other	61.401	2.922	1.195	8.541	25	60.785	60.981	0.196	2680.325
F	B	B	Eng	Other	43.996	2.710	1.003	7.346	30	44.381	43.312	1.069	422.907
M	B	B	Eng	Other	44.284	6.288	1.847	39.534	18	46.086	44.446	1.640	738.986
F	O	B	Eng	Other	74.194	15.068	2.152	227.048	19	59.810	59.715	0.096	300.160
M	O	B	Eng	Other	59.491	6.216	1.685	38.642	20	57.672	57.887	0.214	1199.190
F	W	G	Eng	Other	67.694	3.887	2.347	15.112	66	66.087	66.695	0.608	704.764
M	W	G	Eng	Other	74.603	3.284	1.859	10.783	105	72.871	72.508	0.363	4906.013
F	A	G	Eng	Other	77.174	5.767	1.903	33.253	24	72.104	71.675	0.429	1180.027
M	A	G	Eng	Other	72.699	4.061	1.921	16.489	31	70.276	70.365	0.089	3292.968
F	B	G	Eng	Other	64.165	4.847	1.973	23.489	26	62.225	62.805	0.581	100.100
M	B	G	Eng	Other	76.003	6.282	2.854	39.468	24	70.699	70.535	0.164	359.816
F	O	G	Eng	Other	65.275	3.719	0.525	107.381	13	59.073	59.620	0.547	126.910
M	O	G	Eng	Other	73.760	5.076	1.005	25.769	7	70.107	70.356	0.249	71.938
F	W	B	Rel	Other	51.698	1.459	1.236	2.129	190	51.681	52.497	0.815	70634.199
M	W	B	Rel	Other	53.893	4.271	1.463	18.243	43	53.651	53.822	0.171	9884.133
F	A	B	Rel	Other	55.596	4.067	0.860	26.283	20	54.890	54.855	0.035	5910.623
M	A	B	Rel	Other	70.748	7.977	0.682	96.787	4	62.975	64.033	1.058	617.298
F	B	B	Rel	Other	47.253	2.812	0.990	21.046	36	47.697	47.674	0.023	8837.781
M	B	B	Rel	Other	46.985	6.270	1.381	39.316	9	47.999	47.145	0.853	1111.491
F	O	B	Rel	Other	55.396	4.487	1.118	20.129	13	54.811	54.708	0.103	3862.865
M	O	B	Rel	Other	60.825	6.061	1.175	36.733	8	58.900	58.502	0.397	1904.469
F	W	G	Rel	Other	70.289	4.123	1.506	17.003	136	68.185	68.708	0.523	33233.953
M	W	G	Rel	Other	107.648	17.306	2.612	299.493	29	69.356	68.348	1.007	7306.351
F	A	G	Rel	Other	72.615	15.552	0.502	773.462	2	54.983	54.489	0.494	380.398
M	A	G	Rel	Other	92.134	10.319	0.765	196.010	6	68.083	67.686	0.397	926.224
F	B	G	Rel	Other	64.897	4.270	1.378	18.235	39	63.505	63.968	0.463	3803.613
M	B	G	Rel	Other	55.761	6.151	0.983	179.282	11	53.776	53.189	0.586	751.088
F	O	G	Rel	Other	52.642	3.792	1.087	14.378	11	52.636	53.114	0.478	1775.584
M	O	G	Rel	Other	103.000	0.000		5497.730	1	52.648	51.778	0.870	96.716