Survey Methodology, December 2012
Vol. 38, No. 2, pp. 203-214
Statistics Canada, Catalogue No. 12-001-X

Bayesian inference for finite population quantiles from unequal probability samples

Qixuan Chen, Michael R. Elliott and Roderick J.A. Little [1]

Abstract

This paper develops two Bayesian methods for inference about finite population quantiles of continuous survey variables from unequal probability sampling. The first method estimates cumulative distribution functions of the continuous survey variable by fitting a number of probit penalized spline regression models on the inclusion probabilities. The finite population quantiles are then obtained by inverting the estimated distribution function. This method is quite computationally demanding. The second method predicts non-sampled values by assuming a smoothly-varying relationship between the continuous survey variable and the probability of inclusion, by modeling both the mean function and the variance function using splines. The two Bayesian spline-model-based estimators yield a desirable balance between robustness and efficiency. Simulation studies show that both methods yield smaller root mean squared errors than the sample-weighted estimator and the ratio and difference estimators described by Rao, Kovar, and Mantel (RKM 1990), and are more robust to model misspecification than the regression through the origin model-based estimator described in Chambers and Dunstan (1986). When the sample size is small, the 95% credible intervals of the two new methods have closer to nominal confidence coverage than the sample-weighted estimator.

Key Words: Bayesian analysis; Cumulative distribution function; Heteroscedastic errors; Penalized spline regression; Survey samples.

1. Introduction

We consider inference for finite population quantiles of a continuous variable from a sample survey with unequal inclusion probabilities. The finite-population quantiles are usually estimated by the sample-weighted quantiles, a Horvitz-Thompson type estimator. Often in sample surveys the design variable (here, the inclusion probability) or a correlated auxiliary variable is measured on the non-sampled units, and this information can be used to improve the efficiency of the sample-weighted estimators (Zheng and Little 2003; Chen, Elliott, and Little 2010).

Methods for using auxiliary information in estimating finite-population distribution functions have been extensively studied. Chambers and Dunstan (1986) proposed a model-based method, illustrating their approach for a zero intercept linear regression superpopulation model. We refer to this estimator from now on as the CD estimator. Dorfman and Hall (1993) applied the CD approach, replacing the linear regression model with a non-parametric model. Lombardía, González-Manteiga, and Prada-Sánchez (2003, 2004) proposed a bootstrap approximation to these estimators based on resampling a smoothed version of the empirical distribution of the residuals. Kuk and Welsh (2001) also modified the CD approach to address departures from the model by estimating the conditional distribution of residuals as a function of the auxiliary variable. Rao, Kovar, and Mantel (RKM 1990) demonstrated advantages of design-based ratio and difference estimators over the CD estimator when the model is misspecified. Wang and Dorfman (1996) suggested a weighted average of the CD and the RKM estimators. Kuk (1993) proposed a kernel-based estimator that combines the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the value of the auxiliary variable. Chambers, Dorfman, and Wehrly (1993) proposed a kernel-smoothed model-based estimator, and Wu and Sitter (2001) and Harms and Duchesne (2006) proposed calibration type estimators.

Research on using auxiliary information for inference about finite population quantiles (defined as the inverse of the distribution function) is more limited. Chambers and Dunstan (1986) discussed estimation by inverting the CD estimator of the distribution function, but did not compare the performance of this quantile estimator with alternatives. Rao *et al.* (1990) proposed simple ratio and difference quantile estimators that were considerably more efficient than the sample-weighted estimator when the survey outcome was approximately proportional to the auxiliary variable.

We assume here unequal probability sampling with inclusion probabilities that are known for all the units in the population. We develop two Bayesian spline-model-based estimators of finite population quantiles that incorporate the inclusion probabilities. The first method is to estimate the distribution function at a number of sample values using

1. Qixuan Chen is Assistant Professor, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. E-mail: qc2138@columbia.edu; Michael R. Elliott and Roderick J.A. Little are professors, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu and rlittle@umich.edu.

Bayesian penalized spline predictive estimators (Chen *et al.* 2010). The finite population quantiles are then estimated by inverting the predictive distribution function. The second method is a Bayesian two-moment penalized spline predictive estimator, which predicts the values of non-sampled units based on a normal model, with mean and variance both modeled with penalized splines on the inclusion probabilities. We compare the performance of these two new methods with the sample-weighted estimator, the CD estimator, and the RKM's ratio and difference estimators, using simulation studies on artificially generated data and farm survey data.

2. Estimators of the quantiles

Let s denote an unequal probability random sample of size n, drawn from the finite population of N identifiable units according to inclusion probabilities $\{\pi_i, i = 1, ..., N\}$, which are assumed to be known for all the units before a sample is drawn. Let Y denote a continuous survey variable, with values $\{y_1, y_2, ..., y_n\}$ observed in the random sample s. The finite-population α-quantile of Y is defined as:

$$\theta(\alpha) = \inf\left\{t;\ N^{-1}\sum_{i=1}^{N} \Delta(t - y_i) \geq \alpha\right\}, \quad (1)$$

where $\Delta(u) = 1$ when $u \geq 0$ and $\Delta(u) = 0$ elsewhere. The $\theta(\alpha)$ is often estimated using the sample-weighted α-quantile $\hat{\theta}(\alpha) = \inf\{t, \hat{F}_w(t) \geq \alpha\}$, where $\hat{F}_w(t)$ is the sample-weighted distribution function given by

$$\hat{F}_w(t) = \sum_{i \in s} \pi_i^{-1} \Delta(t - y_i) \Big/ \sum_{i \in s} \pi_i^{-1}.$$

Woodruff (1952) proposed a method of calculating confidence limits for the sample weighted α-quantile. First, a pseudo-population is obtained by weighting each sample item by its sampling weight; the standard deviation of the percentage of items less than the estimated α-quantile is estimated; and the estimated standard deviation is multiplied by the appropriate z percentile and is added to and subtracted from α to construct the confidence limits for the percentage of items less than the estimated α-quantile. Finally, the values of the survey variable corresponding to the confidence limits of the percentage of items less than the estimated α-quantile are read-off the weighted pseudo-population arrayed in order of size. Variance estimation of the percentage of items in the pseudo-population less than the estimated α-quantile is discussed in Woodruff (1952). Sitter and Wu (2001) showed that the Woodruff intervals perform well even in moderate to extreme tail regions of the distribution function. An alternative variance estimate was derived by Francisco and Fuller (1991) using a smoothed version of the large-sample test inversion.

2.1 Bayesian model-based approach, inverting the estimated CDF

The finite population quantile function is the inverse of the finite population cumulative distribution function (CDF), defined as $F(t) = N^{-1}\sum_{i=1}^{N} \Delta(t - y_i)$, where $\Delta(x) = 1$ when $x \geq 0$ and $\Delta(x) = 0$ elsewhere. We can estimate the finite population quantiles by first building a continuous and strictly monotonic predictive estimate of $F(t)$, by treating $\Delta(t - y)$ as a binary outcome variable and applying methods for estimating finite population proportions.

In particular, Chen *et al.* (2010) proposed a Bayesian penalized spline predictive (BPSP) estimator for finite population proportions in unequal probability sampling. They regress the binary survey variable z on the inclusion probabilities in the sample, using the following probit penalized spline regression model (2) with m pre-selected fixed knots:

$$\Phi^{-1}(E(z_i \mid \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l (\pi_i - k_l)_+^p,$$
$$b_l \sim N(0, \tau^2). \quad (2)$$

Self-representing units are included by setting $\pi_i = 1$. Assuming non-informative prior distributions for β and τ^2, they simulated draws of z for the non-sampled units from their posterior predictive distribution. A draw from the posterior distribution of the finite population proportion is then obtained by averaging the observed sample units and the draws of the non-sample units. This is repeated many times to simulate the posterior distribution of the finite population proportion. Simulation studies indicated that the BPSP estimator is more efficient than the sample-weighted and generalized regression estimators of the finite population proportion, with confidence coverage closer to nominal levels.

We employ the BPSP approach n times to estimate $F(t)$ at each of the sampled values of y, $t = \{y_1, y_2, ..., y_n\}$. This estimator does not take into account the fact that we are estimating a whole distribution function, and is not necessarily a monotonic function. In addition, linear interpolation of the n estimated distribution functions may lead to a poorly-estimated CDF. To overcome these two problems, we fit a smooth cubic regression curve to the n estimated distribution functions with monotonicity constraints (Wood 1994). We denote the resulting estimated distribution function as $\hat{F}(t)$. The Bayesian model-based estimator of $\theta(\alpha)$, obtained by inverting the estimated CDF, is then defined as follows:

$$\hat{\theta}_{\text{inv-CDF}}(\alpha) = \inf\{t;\ \hat{F}(t) \geq \alpha\}. \quad (3)$$

We also fit two other monotonic smooth regression curves to the upper and lower limits of the 95% credible intervals (CI) of these estimated distribution functions, denoted as $\hat{F}_U(t)$ and $\hat{F}_L(t)$. To reduce computation time in our simulation studies, we only estimate the CDF at $k < n$ pre-selected sample points.

The basic idea behind this approach is shown graphically in Figure 1. Suppose a sample of size 100 is drawn from a finite population. We pick 20 observations from the sample and estimate their corresponding distribution functions and associated 95% CI using the BPSP estimator. In Figure 1(a) we plot the BPSP estimates of these 20 points with black dots and the upper and lower limits of 95% CI with "-" signs, and connect the upper and lower limits with solid lines. In Figure 1(b) we add three monotonic smooth predictive curves using black solid curve for the point estimate and black dash curves for the upper and lower limits of the 95% CI.

In Figure 1(c) we draw a horizontal line across the graph with α as the y-axis value. We read x_A, x, and x_B respectively from the x-axis such that $\hat{F}_L(x_A) = \alpha$, $\hat{F}(x) = \alpha$, and $\hat{F}_U(x_B) = \alpha$. Then x is the inverse-CDF Bayesian estimate of $\theta(\alpha)$. If the 95% CI of the distribution function $F(\cdot)$ is formed by splitting the tail areas of the posterior distribution equally, the interval formed by x_A and x_B is a 95% CI of $\theta(\alpha)$. The proof is as follows: If α is the lower limit of the 95% CI of $F(x_A)$, only 2.5 percent of the draws of $F(x_A)$ in the posterior distribution are smaller than α. That is,

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0.025.$$

Similarly with α as the upper limit of the 95% CI of $F(x_B)$, $\Pr(\theta(\alpha) < x_B) = 0.975$. Therefore, there is 95% probability that $\theta(\alpha)$ is within x_A and x_B in the posterior distribution, given the sample.

This inverse-CDF Bayesian model-based approach avoids strong modeling assumptions, and can be applied to normal or skewed distributions. Estimating the distribution function at all n sample units makes full use of the sample information, but is computationally intensive; estimating the distribution function at $k < n$ values reduces computation time at the expense of some loss of efficiency. In the traditional approach, the population quantiles are estimated by inverting the unsmoothed empirical CDF. We recommend fitting a smooth cubic regression curve to the estimated distribution functions before inverting the estimated CDF. The resulting quantile estimates are more efficient, because the smooth curve exploits information from all the data. Simulations not shown here suggest that the estimated CDF distribution function curve estimated based on a well-chosen subset of the k sample units is similar to the curve estimated based on all sample units, but the computation time is significantly reduced.

We suggest choosing the subset of k data points at evenly spaced intervals in the middle of the distribution, and more frequent intervals in the extremes to improve the estimate of the CDF in the tails. For instance, in our simulation study with a sample size of 100, we estimated the distribution functions at 20 points: the 3 smallest, the 3 largest, and 14 other equally spaced points in the middle of the ordered sample.

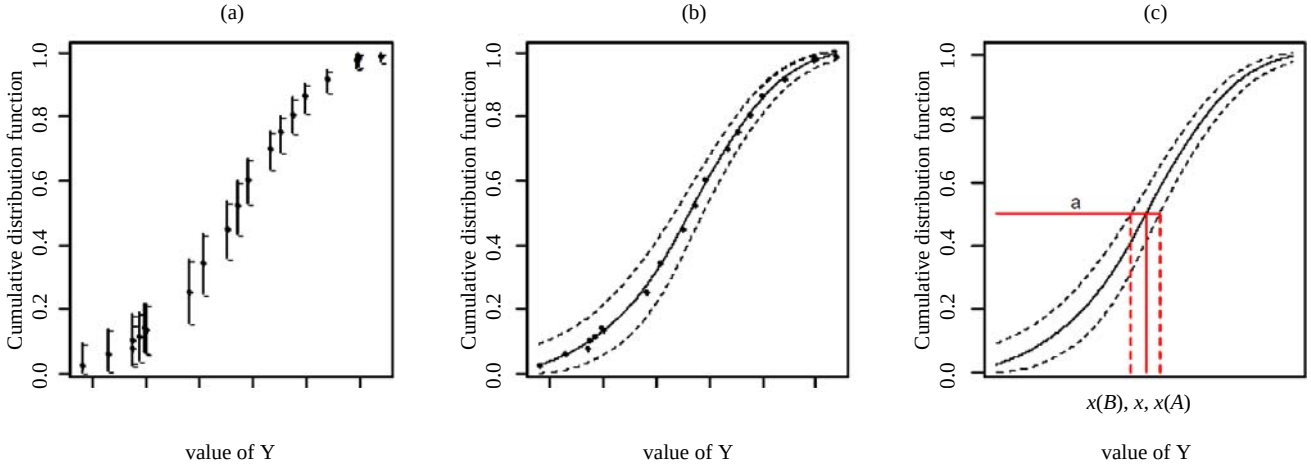

Figure 1 Inverse-CDF Bayesian model-based approach in estimating finite population distribution functions and associated quantiles illustrated using a sample of size 100 drawn from a finite population. (a) BPSP method is used to estimate the finite population distribution functions at 20 sample points; the dots denote BPSP estimators and the minus signs denote the upper and lower limits of the 95% CI. (b) Three monotonic smooth cubic regression models are fit on the BPSP estimators, upper limits, and lower limits; the solid curve is the predictive continuous distribution functions and the two dash curves are the 95% CI of the distribution functions. (c) The point estimate and 95% CI of population α-quantile are obtained by inverting the estimated CDF; x is the point estimate, and x(B) and x(A) are the lower and upper limits of the 95% CI

2.2 Bayesian two-moment penalized spline predictive approach

We consider alternative estimators of finite population quantiles of the form:

$$\tilde{\theta}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t - y_i) + \sum_{j\notin s}\Delta(t - \hat{y}_j)\right) \geq \alpha\right\}, \quad (4)$$

where $\hat{y}_j$ is the predicted value of the j^{th} non-sample unit based on a regression on the inclusion probabilities $\{\pi_i\}$. A basic normal model for a continuous outcome assumes a mean function that is linear in $\{\pi_i\}$, that is:

$$Y_i \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1\pi_i, c_i\sigma^2), \quad (5)$$

with known constants c_i to model non-constant variance. This leads to a biased estimate of $\theta(\alpha)$ when the relationship is not linear. For estimating finite population totals, Zheng and Little (2003, 2005) replaced the linear mean function in (5) with a penalized spline, and assumed $c_i = \pi_i^{2k}$ with some known value of k. Simulations suggested that their model-based estimator of the finite population total outperforms the sample-weighted estimator, even when the variance structure is misspecified.

For estimation of quantiles rather than the total, correct specification of the variance structure is important in order to avoid bias. Therefore, we extend the penalized spline model in Zheng and Little (2003) by modeling both the mean and the variance using penalized splines. The two-moment penalized spline model can be written as (Ruppert, Wand, and Carroll 2003, page 264):

$$\begin{aligned}
&Y_i \overset{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i, k), \exp(\text{SPL}_2(\pi_i, k'))),\\
&\text{SPL}_1(\pi_i, k) = \beta_0 + \sum_{k=1}^{p}\beta_k\pi_i^k + \sum_{l=1}^{m_1} b_l(\pi_i - k_l)_+^p,\\
&\qquad b_l \overset{\text{iid}}{\sim} N(0, \tau_b^2),\\
&\text{SPL}_2(\pi_i, k') = \alpha_0 + \sum_{k=1}^{p}\alpha_k\pi_i^k + \sum_{l=1}^{m_2} \nu_l(\pi_i - k_l')_+^p,\\
&\qquad \nu_l \overset{\text{iid}}{\sim} N(0, \tau_\nu^2). \quad (6)
\end{aligned}$$

In (6), the mean and the logarithm of the variance are modeled as penalized splines (SPL_1) and (SPL_2) on $\{\pi_i\}$. Modeling the logarithm of the variance ensures positive estimates of the variance. We allow different numbers (m_1, m_2) and locations (k, k') of the knots for the two splines.

Ruppert *et al.* (2003) suggested an iterative approach to estimate the parameters in (6). They first assumed that SPL_2 was known and fitted a linear mixed model to estimate the parameters in SPL_1. They calculated the square of the difference between Y and SPL_1, which followed a Gamma distribution with the shape parameter as ½ and the scale parameter of 2SPL_2. They then fitted a generalized linear mixed model for the squared differences to estimate the parameters in SPL_2. They iterated the above procedures until the parameter estimates converged. This iterative approach is simple to implement. However, our goal here is not to estimate the parameters but to obtain Bayesian predictions of Y for the non-sample units so that we can use (4) to estimate the quantiles.

Crainiceanu, Ruppert, Carroll, Joshi, and Goodner (2007) developed Bayesian inferential methodology for (6). They noted that the implementation of MCMC using multivariate Metropolis-Hastings steps is unstable with poor mixing properties. They suggested adding error terms to the second spline to make computations feasible, replacing sampling from complex full conditionals by simple univariate Metropolis-Hastings steps. This idea can be expressed as

$$\begin{aligned}
&Y_i \overset{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i, k), \sigma_\varepsilon^2(\pi_i)),\\
&\log(\sigma_\varepsilon^2(\pi_i)) \overset{\text{iid}}{\sim} N(\text{SPL}_2(\pi_i, k'), \sigma_A^2).
\end{aligned}$$

We used a prior distribution $N(0, 10^6)$ for the fixed effects parameters β and α, and a proper inverse-gamma prior distribution $\text{IGamma}(10^{-6}, 10^{-6})$ for the variance components τ_b^2 and τ_ν^2. We fixed the values of $\sigma_A^2 = 0.1$. The full conditionals of the posterior are detailed in Crainiceanu *et al.* (2007).

The posterior distribution of the finite population α-quantile is simulated by generating a large number D of draws and using the predictive estimator form

$$\tilde{\theta}^{(d)}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t - y_i) + \sum_{j\notin s}\Delta(t - \hat{y}_j^{(d)})\right) \geq \alpha\right\},$$

where $\hat{y}_j^{(d)}$ is a draw from the posterior predictive distribution of the j^{th} non-sampled unit of the continuous outcome. The average of these draws simulates the Bayesian two-moment penalized spline predictive (B2PSP) estimator of the finite population α- quantile,

$$\hat{\theta}_{\text{B2PSP}}(\alpha) = D^{-1}\sum_{d=1}^{D}\tilde{\theta}^{(d)}(\alpha).$$

The Bayesian 95% credible interval for the population α-quantile in the simulations is formed by splitting the tail area equally between the upper and lower endpoints.

3. Simulation study

3.1 Simulation study with artificial data

We first simulated a super-population of size M = 20,000. The size variable X in the super-population takes 20,000 consecutive integer values from 710 to 20,709. A finite population of size N = 2,000 was then selected from the super-population using systematic probability proportional to size (pps) sampling with the probability proportional to the inverse of the size variable. Consequently, the size variable in the finite population has a right skewed distribution. The survey outcome Y was drawn from a normal distribution with mean $f(\pi)$ and error variance equal to 0.04 (homoscedastic error) or π (heteroscedastic error). Three different mean structures $f(\pi)$ were simulated: no association between Y and π (NULL) $f(\pi)$ = 0.5, a linear association (LINUP) $f(\pi) = 6\pi$, and a nonlinear association (EXP) $f(\pi) = \exp(-4.64 + 52\pi)$. For each of the six simulation conditions, one thousand replicate finite populations were generated, and a systematic pps sample (n = 100) was drawn from each population with x as the size variable; thus $\pi_i = nx_i / \sum_{j=1}^{N} x_j$. Scatter plots of Y versus π for these six populations are displayed in Figure 2.

We compared the performance of the Bayesian inverse-CDF and the B2PSP estimators with five alternative approaches:

a) SW, the sample-weighted estimator defined by inverting $\hat{F}_w$.
b) Smooth-SW, the smooth sample-weighted estimator. A smooth cubic regression curve was fit to $\hat{F}_w$, and denoted as $\tilde{F}_w$. The smooth sample-weighted estimator is then defined as $\tilde{\theta}_w = \inf\{t; \tilde{F}_w \geq \alpha\}$.
c) CD, the Chambers and Dunstan estimator (1986), by assuming the following model: $Y_i = \beta\pi_i + \sqrt{\pi_i} U_i$, where U_i is an independent and identically distributed random variable with zero mean.
d) Ratio, the RKM's ratio estimator (1990) given by $\{\hat{\theta}_y(\alpha) / \hat{\theta}_x(\alpha)\} \times \theta_x(\alpha)$, where $\hat{\theta}_y(\alpha)$ and $\hat{\theta}_x(\alpha)$ denotes respectively the sample-weighted estimates for Y and the size variable X, and $\theta_x(\alpha)$ is the known population quantile of X.
e) Diff, the RKM's difference estimator (1990) given by $\hat{\theta}_y(\alpha) + \hat{R} \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$, where $\hat{R}$ is the sample-weighted estimate of Y/X.

The seven estimators for the finite-population 10th, 25th, 50th, 75th, and 90th percentiles were compared in terms of empirical bias and root mean squared error (RMSE). Because of the complexity in the variance estimation for the CD and RKM's estimators, we only compared the average width and the non-coverage rate of the 95% confidence/credible interval (CI) for the two Bayesian model-based estimators and the sample-weighted estimator. For the 95% CI, we used Woodruff's method for the sample-weighted estimator, the method illustrated in Figure 1(c) for the inverse-CDF Bayesian estimator, and the 95% posterior probability of the quantile with equal tails for the B2PSP estimator. We used cubic splines with 15 equally spaced knots.

Tables 1 and 2 show the empirical bias and RMSE for the three normal distributions with homoscedastic errors and with heteroscedastic errors, respectively. Overall, the empirical bias in estimating the five quantiles is similar using the two Bayesian estimators, the two sample-weighted estimators, and the RKM's two design-based estimators. In contrast, the CD estimator has large bias and RMSE in all scenarios except for LINUP with heteroscedastic error, where its underlying model is correctly specified. The two Bayesian model-based estimators yield smaller root mean squared errors than the other estimators, and this improvement in efficiency is substantial in some scenarios, especially using the B2PSP estimator. By applying a smooth cubic regression curve on the estimated empirical sample-weighted CDF, the smooth-sample-weighted estimator gains some efficiency over the conventional sample-weighted estimators, but the RMSE is still larger than the Bayesian Inverse-CDF estimator. Comparisons of the three design-based estimators suggest that none of the three estimators uniformly dominates the other two. Specifically, the sample-weighted estimator has smaller RMSE than the RKM difference and ratio estimators for all five quantiles in the NULL and for the lower quantiles in the LINUP and EXP populations; on the other hand, the RKM estimators have smaller RMSE at the upper quantiles in the LINUP and EXP populations.

Table 3 shows the average width and non-coverage rate of 95% CI for the two Bayesian model-based estimators and the sample-weighted estimator. Overall, the two Bayesian model-based estimators yield shorter average 95% CI widths than the sample-weighted estimator. The coverage rate of the 95% CI is similar among the three estimators, except that when α is equal to 0.1, where the 95% CI of the B2PSP estimator has the shortest average width and very good coverage, while the sample-weighted estimator has serious under-coverage. This happens because the Woodruff method for estimating the variance of the sample-weighted estimator is based on a large sample assumption, but here the pps sampling leads to only a small number of cases being sampled in the lower tail.

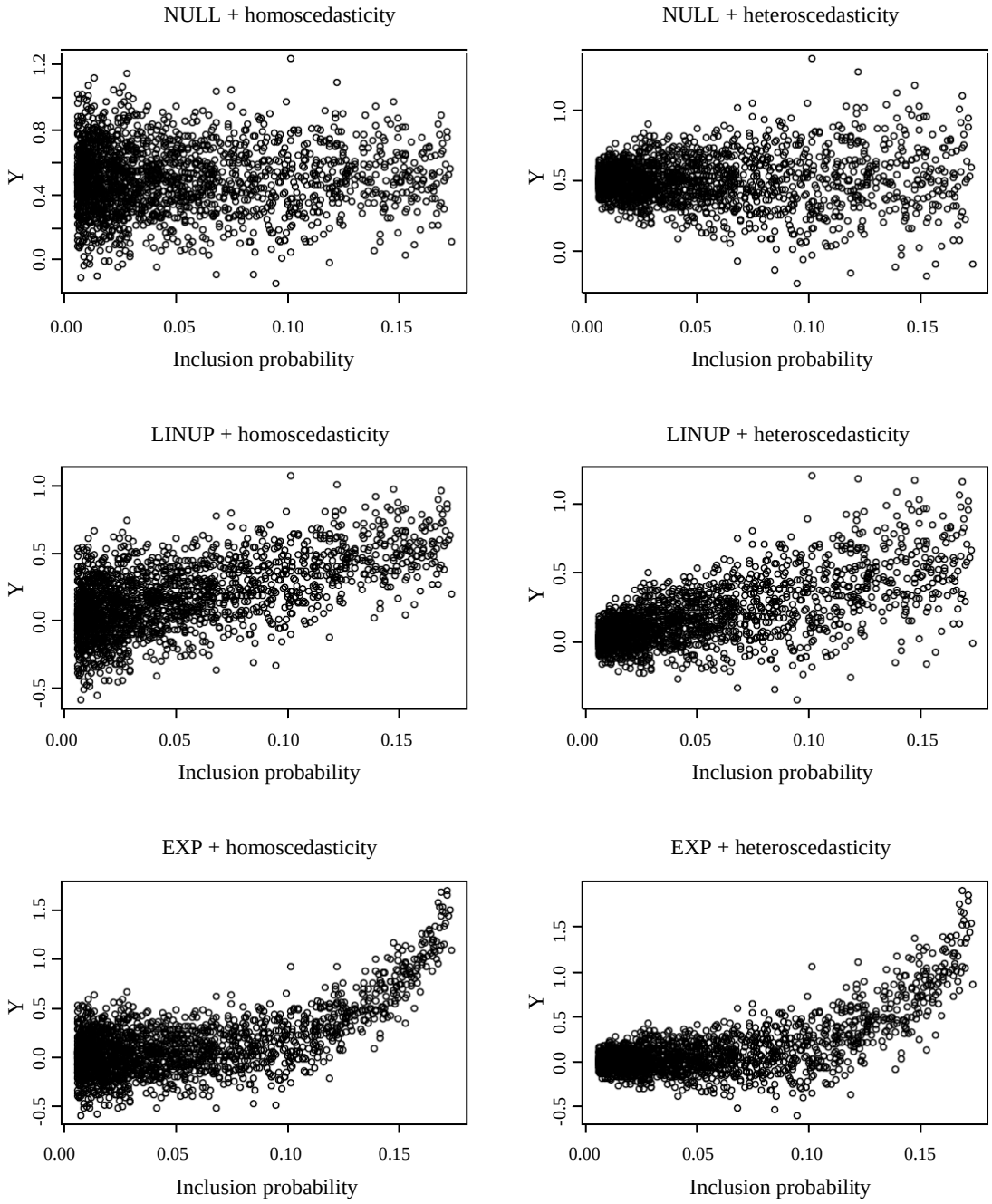

Figure 2 Scatter plots of Y versus the inclusion probabilities for the six artificial finite populations of size equal to 2,000

Although the sample-weighted estimator performs similarly with the two Bayesian spline-model-based estimators in terms of overall empirical bias, the conditional bias of estimates varies largely as the sample mean of the inclusion probability increases. Following Royall and Cumberland (1981), the estimates from the 1,000 samples were ordered according to the sample mean of the inclusion probabilities and were split into 20 groups of 50 each, and then the empirical bias was calculated for each group. Figure 3 displays the conditional bias of the two Bayesian estimators and the sample-weighted estimator for the 90th percentile in the "EXP + homoscedastic error" case. Figure 3 shows that there is a linear trend for the bias in the sample-weighted estimator as the sample mean of the inclusion probabilities increases, while the grouped bias of the two Bayesian spline-model-based estimators is less affected by the sample mean of inclusion probabilities. Similar findings are also seen in other scenarios.

Table 1
Comparisons of empirical bias and root mean squared errors × 10^3 of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9: Scenarios with homoscedastic errors

	Empirical bias					Empirical RMSE				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
NULL										
Inverse-CDF	-6	-3	-1	-1	-5	46	37	36	37	45
B2PSP	-5	-1	1	2	6	41	33	31	34	42
SW	-5	-3	-1	-4	-6	54	41	39	41	50
Smooth-SW	-7	-4	-1	-2	-5	50	39	37	38	47
CD	-197	-272	-265	-108	168	203	274	266	115	189
RKM's Ratio	3	25	33	16	6	77	125	159	112	79
RKM's Diff	-5	-1	6	14	14	58	58	94	122	113
LINUP										
Inverse-CDF	-15	-3	-2	-1	-2	70	49	39	34	33
B2PSP	-3	-1	1	4	7	56	43	35	31	29
SW	-15	-3	-3	-2	-6	77	57	48	44	42
Smooth-SW	-14	-5	-2	-1	-4	72	53	45	42	41
CD	101	35	-37	-49	1	104	38	39	53	31
RKM's Ratio	-23	-9	2	5	-0.2	95	67	53	51	40
RKM's Diff	-15	-4	-4	-0.2	-2	77	55	45	43	38
EXP										
Inverse-CDF	-8	0.4	4	7	4	60	45	41	43	49
B2PSP	-10	-6	-3	0.3	13	52	40	35	36	36
SW	-9	-3	-2	-2	-8	65	49	46	50	72
Smooth-SW	-12	-5	-2	-1	-2	62	47	43	46	68
CD	92	54	14	19	61	96	57	21	31	75
RKM's Ratio	-17	-11	1	3	-5	87	65	50	53	55
RKM's Diff	-9	-4	-2	-2	-7	65	49	47	47	59

Table 2
Comparisons of empirical bias and root mean squared errors × 10^3 of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9: Scenarios with heteroscedastic errors

	Empirical bias					Empirical RMSE				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
NULL										
Inverse-CDF	-9	-8	-2	4	1	30	24	22	24	31
B2PSP	-6	-6	1	7	7	25	21	19	23	27
SW	-4	-3	-2	-1	-5	34	26	23	26	35
Smooth-SW	-4	-5	-2	1	-4	34	26	23	26	35
CD	-298	-325	-253	-46	270	302	327	255	60	288
RKM's Ratio	8	31	32	16	5	81	143	154	94	57
RKM's Diff	-5	-1	6	17	16	44	54	87	113	97
LINUP										
Inverse-CDF	-11	-1	5	2	-3	32	24	24	29	35
B2PSP	-10	-1	7	3	1	29	22	22	24	30
SW	-5	-1	-0.1	-1	-4	31	28	33	45	51
Smooth-SW	-11	-3	2	-0.4	-5	32	26	30	44	50
CD	10	7	6	7	11	20	13	13	20	32
RKM's Ratio	-7	-3	2	3	1	36	29	30	35	41
RKM's Diff	-5	-2	-1	1	-0.2	32	27	28	33	41
EXP										
Inverse-CDF	-8	-3	5	7	-3	30	23	23	30	48
B2PSP	-11	-7	2	6	7	28	23	20	25	36
SW	-3	-3	-2	1	-2	30	26	26	41	84
Smooth-SW	-8	-5	1	2	-5	30	23	24	39	86
CD	18	16	35	84	68	27	21	38	88	81
RKM's Ratio	-5	-6	-1	2	-0.1	36	31	27	32	62
RKM's Diff	-3	-3	-2	1	-0.1	32	28	28	31	67

Table 3
Comparisons of average width and non-coverage rate of 95% CI × 10^3 of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9

	Average width of 95% CI					Non-coverage rate of 95% CI				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
					Homoscedastic errors					
NULL										
Inverse-CDF	199	156	141	152	184	46	35	44	38	67
B2PSP	178	134	118	134	177	52	55	61	59	50
SW	195	164	151	167	237	112	65	46	40	38
LINUP										
Inverse-CDF	257	207	157	139	141	61	45	37	46	52
B2PSP	230	167	134	123	121	58	54	44	57	59
SW	248	231	188	179	187	119	60	42	41	39
EXP										
Inverse-CDF	234	184	163	177	234	59	44	47	40	42
B2PSP	217	157	132	144	156	54	59	55	53	60
SW	231	199	175	210	402	106	64	47	40	40
					Heteroscedastic errors					
NULL										
Inverse-CDF	146	104	90	101	137	42	43	38	38	47
B2PSP	107	89	79	89	107	38	49	37	68	65
SW	146	101	91	113	169	80	60	51	37	42
LINUP										
Inverse-CDF	131	107	104	124	154	70	31	36	42	40
B2PSP	125	97	87	93	116	47	35	50	58	52
SW	141	110	133	184	219	138	69	41	50	42
EXP										
Inverse-CDF	131	99	99	134	242	63	49	34	40	41
B2PSP	116	92	84	98	139	57	55	40	63	59
SW	135	100	106	186	378	111	65	46	45	34

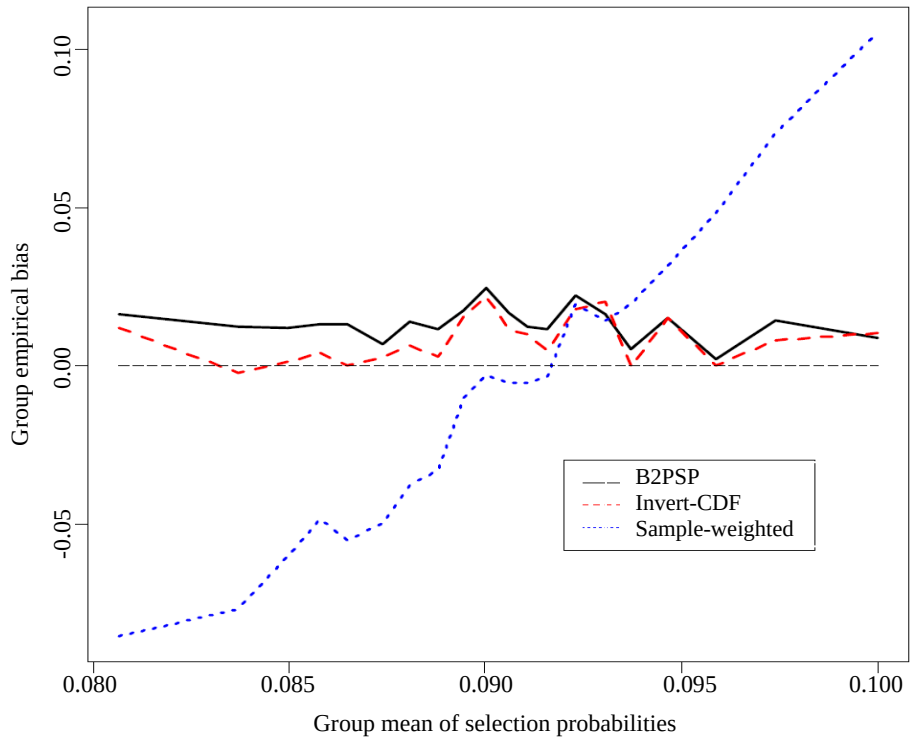

Figure 3 Variation of empirical bias of the three estimators for 90th percentile from the "EXP + homoscedasticity" case

3.2 Simulation study with the broadacre farm survey data

The B2PSP estimator assumes the outcome has a normal distribution, after conditioning on the inclusion probabilities. Since the inverse-CDF Bayesian model-based approach does not assume normality, we might expect it to out-perform the B2PSP when the normality assumption is violated. This motivates a comparison of the sample-weighted and the inverse-CDF Bayesian estimators for non-normal data.

The population considered here is defined by 398 broadacre farms (farms involved in the production of cereal crops, beef, sheep and wool) with 6,000 or less hectares that participated in the 1982 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics (ABARE 2003). The Y variable is the total farm cash receipts. One thousand systematic pps samples of size equal to 100 were drawn with the farm area, X, as the size variable, that is, larger farms are more likely to be selected into the sample. Figure 4 is the scatter plot of Y versus the size variable X for these farms, with filled circles representing a selected pps sample. This shows that the variation of Y increases as X increases. Moreover, Y is right-skewed given X. A simulation study using this broadacre farms data was conducted to compare the two Bayesian spline-model-based estimators with the sample-weighted estimator.

Table 4 shows the simulation results. The inverse-CDF Bayesian approach yields smaller empirical bias and RMSE, and shorter average length of 95% CI than the sample-weighted estimator in general. The 95% CI of the inverse-CDF Bayesian approach also have closer to nominal level confidence coverage than the sample-weighted estimator when α is 0.1 and 0.25. However, in the upper tail with $\alpha = 0.90$, the non-coverage rate of the inverse-CDF Bayesian approach is higher than the nominal level 0.05, while the Woodruff CI of the sample-weighted estimator does well. This is consistent with the findings of Sitter and Wu (2001) that the Woodruff intervals perform well even in the moderate to extreme tail regions of the distribution function. Since the conditional normality assumption is not reasonable here, the B2PSP estimator is biased and the 95% CI has poor confidence coverage.

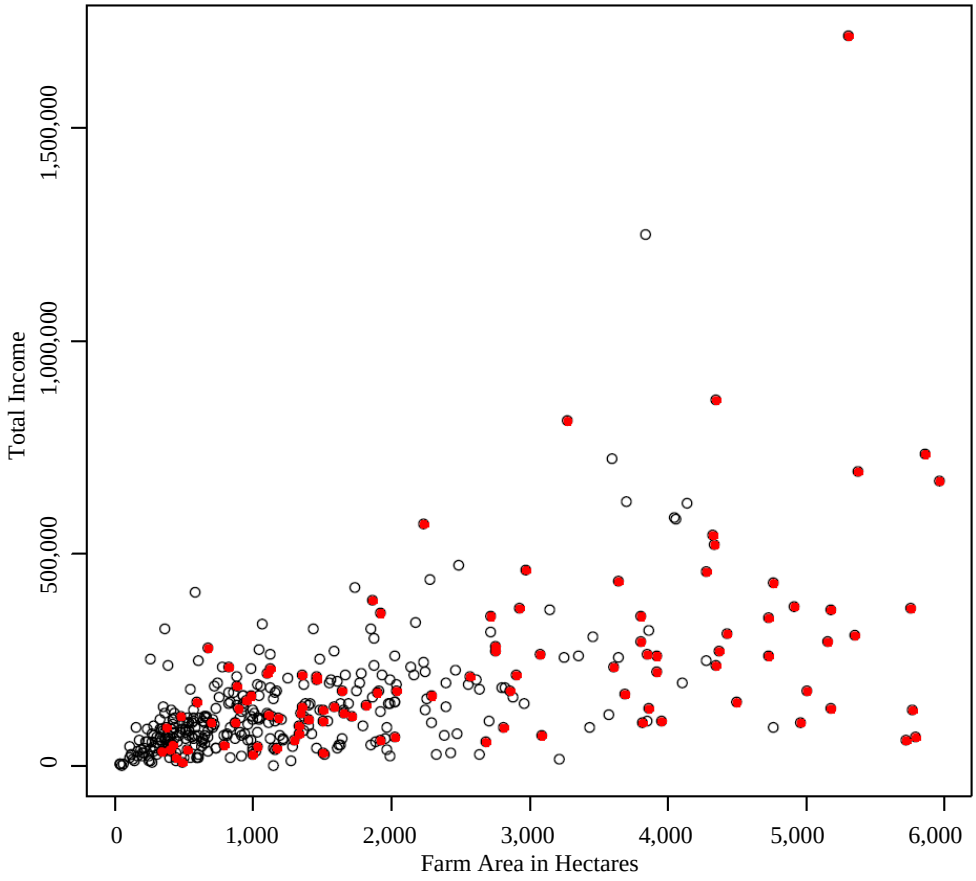

Figure 4 Scatter plot of the broadacre farm data with the filled circles representing a pps sample

Table 4
Empirical bias × 10^{-2}, root mean squared errors × 10^{-2}, average width of 95% CI × 10^{-2}, and non-coverage rate of 95% CI × 10^{3} of $\theta(\alpha)$ for α = 0.1, 0.25, 0.5, 0.75, and 0.9: The broadacre farm data

	0.1	0.25	0.5	0.75	0.9
			Empirical bias		
Inverse-CDF	8	14	10	-22	-60
B2PSP	-110	-125	-63	-12	88
SW	20	-19	-17	-21	-61
			Empirical RMSE		
Inverse-CDF	117	117	108	164	256
B2PSP	113	141	124	140	206
SW	132	173	167	226	350
			Average width of 95% CI		
Inverse-CDF	402	443	501	697	906
B2PSP	170	327	539	726	964
SW	285	468	615	864	1,589
			Non-coverage rate of 95% CI		
Inverse-CDF	96	53	26	52	90
B2PSP	670	258	42	8	17
SW	220	121	68	42	44

4. Discussion

Sample-weighted estimators for finite population quantiles are widely used in survey practice. Although the sample-weighted estimators with Woodruff's confidence intervals are easy to compute and can provide valid large-sample inferences, they may be inefficient and confidence coverage can be poor in small-to-moderate-sized samples. Model-based estimators can improve the efficiency of the estimates when the model is correctly specified, but lead to biased estimates when the model is misspecified. To achieve the balance between robustness and efficiency, we considered spline-model-based estimators. For the quantile estimation of a continuous survey variable, we can either estimate the model-based distribution functions and invert the distribution functions to obtain quantiles, or model the survey outcome on the inclusion probabilities directly. In this paper, we proposed two Bayesian spline-model-based quantile estimators. The first method is the Bayesian inverse-CDF estimator, obtained by inverting the spline-model-based estimates of distribution functions. The second method is the B2PSP estimator, estimated by assuming a normal distribution for the continuous survey outcome, with the mean function and the variance function both modeled using splines.

The simulations suggest that the two Bayesian spline-model-based estimators outperform the sample-weighted estimator, the design-based ratio and difference estimators, as well as the CD model-based estimator when its assumed model is incorrect. Both new methods yield smaller root mean squared errors whether there is no association, a linear association, or a nonlinear association between the survey outcome and the inclusion probability. In some scenarios, the improvement in efficiency using the two Bayesian methods is substantial. When the normality assumption of the survey outcome given the inclusion probabilities is true, the B2PSP estimator has smaller RMSE and shorter credible interval than the inverse-CDF approach. Moreover, the two Bayesian model-based estimators are robust to the mis-specification in both the mean and variance functions. In contrast, the CD model-based estimator is biased and inefficient when either the mean function or the variance function is misspecified. Finally, the Bayesian model-based methods have the advantage of easier calculation of the 95% CI and inference based on the posterior distributions of parameters. This is appealing, because variance estimation for the alternative design-based estimators can be complicated. Woodruff's variance estimation method for sample-weighted estimator performs well when a large fraction of the data is selected from the finite population, even in the moderate to extreme tail regions of the distribution function. However, when data from the population is sparse, the Woodruff's method tends to underestimate the confidence coverage, whereas both Bayesian methods have closer to nominal level confidence coverages.

All the three design-based estimators have comparable overall empirical bias to the two Bayesian spline-model-based estimators. However, there is a linear trend in the variation of bias for the sample-weighted estimator as the sample mean of inclusion probabilities increases. When

there is no association between the survey outcome and the inclusion probability, the ratio and difference estimators have relatively larger bias and RMSE than the sample-weighted estimator. However, in some simulation scenarios, the ratio and difference estimators achieve smaller RMSE than the sample-weighted estimator. The comparison between the conventional sample-weighted estimator and the smooth sample-weighted estimator suggests that fitting a smooth cubic curve to the sample-weighted CDF can improve the efficiency, but the smooth sample-weighted estimator still has larger RMSE than the Bayesian inverse-CDF estimator.

For normally distributed data, we recommend the use of the B2PSP estimator over the other estimators, because of smaller bias, smaller RMSE, and better confidence coverage with shorter interval length. The B2PSP estimator and its 95% posterior probability interval are easy to obtain using the algorithm proposed by Crainiceanu *et al.* (2007), which also has the advantage of relatively short computation time.

The B2PSP estimator is potentially biased when the conditional normal assumption does not hold. One possibility here is to transform the survey outcome to make the conditional normality assumption more reasonable. The B2PSP estimator can be applied to the transformed data, and the draws from the posterior distributions of the non-sampled units are transformed back to the original scale before estimating the quantiles of interest.

In our simulations with non-normal data, the inverse-CDF Bayesian approach was still more efficient than the sample-weighted estimator. Improvement in the confidence coverage was restricted to situations where the sample size is small, with Woodruff's CI method performing well when the large sample assumption holds. Thus for non-normal data where there no clear transformation to improve normality, we do not recommend the inverse-CDF Bayesian approach when the sample size is large. Given the good properties of the B2PSP estimator in the normal setting, one extension for future work is to relax the normality assumption in our proposed approaches.

We use the probability of inclusion as the auxiliary variable here. When there is only one relevant auxiliary variable, it does not matter whether the inclusion probability or the auxiliary variable is modeled. However, if there is more than one relevant auxiliary variable, the inclusion probability is the key auxiliary variable that needs to be modeled corrected, since misspecification of the model relating the survey outcome to the inclusion probability leads to bias. When other auxiliary variables are observed for all the units in the finite population, both of our Bayesian estimators can be easily extended to include additional auxiliary covariates by adding linear terms for these variables in the corresponding penalized spline model.

One reviewer suggested an alternative weighted Dirichlet approach, which is simple to calculate but it does not utilize the known auxiliary variables in the non-sampled units. Another possibility is to re-define the CD estimator by using the spline model we have used to define the B2PSP. Specifically, instead of assuming a regression model through the origin, a spline model is fitted to the first and second order moments of the conditional distribution of survey outcome given the inclusion probability. The spline-based CD estimator should perform similarly to the B2PSP estimator, and its variance can be estimated using resampling methods.

In the official statistics context, the methods in this article illustrate the potential benefits of a paradigm shift from design-based methods towards Bayesian modeling that is geared to yielding inferences with good frequentist properties. Design-based statistical colleagues raise two principal objections to this viewpoint.

First, the idea of an overtly model-based - even worse, Bayesian - approach to probability surveys is not well received, although our emphasis here is on Bayesian methods with good randomization properties. We believe that classical design-based methods do not provide the comprehensive approach needed for the complex problems that increasingly arise in official statistics. Judicious choices of well-calibrated models are needed to tackle such problems. Attention to design features and objective priors can yield Bayesian inferences that avoid subjectivity, and modeling assumptions are explicit, and hence capable of criticism and refinement. See Little (2004, 2012) for more discussion of these points.

The second objection is that Bayesian methods are too complex computationally for the official statistics world, where large number of routine statistics need to be computed correctly and created in a timely fashion. It is true that current Bayesian computation may seem forbidding to statisticians familiar with simple weighted statistics and replicate variance methods. Sedransk (2008), in an article strongly supportive of Bayesian approaches, points to the practical computational challenges as an inhibiting feature. We agree that work remains to meet this objection, but we do not view it insuperable. Research on Bayesian computation methods has exploded in recent decades, as have our computational capabilities. Bayesian models have been fitted to very large and complex problems, in some cases much more complex than those typically faced in the official statistics world.

Acknowledgements

We thank Dr. Philip Kokic in the Commonwealth Scientific and Industrial Research Organisation for providing us the broadacre form data. We also thank an associate editor

and referees for their helpful comments on the original version of this paper.

References

ABARE (2003). Australian farm surveys report 2003. Canberra.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34.

Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. and Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.

Dorfman, H., and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 1, 37-52.

Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.

Kuk, A.Y.C., and Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society*, Series B, 63, 277-292.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467. {70}, 99, 546-556.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 309-334.

Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.

Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Wang, S., and Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.

Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.

Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.