Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 197–223, http://dx.doi.org/10.2478/JOS-2020-0010

Fully Bayesian Benchmarking of Small Area Estimation Models

Junni L. Zhang[1] *and John Bryant*[2]

Estimates for small areas defined by social, demographic, and geographic variables are increasingly important for official statistics. To overcome problems of small sample sizes, statisticians usually derive model-based estimates. When aggregated, however, the model-based estimates typically do not agree with aggregate estimates (benchmarks) obtained through more direct methods. This lack of agreement between estimates can be problematic for users of small area estimates. Benchmarking methods have been widely used to enforce agreement. Fully Bayesian benchmarking methods, in the sense of yielding full posterior distributions after benchmarking, can provide coherent measures of uncertainty for all quantities of interest, but research on fully Bayesian benchmarking methods is limited. We present a flexible fully Bayesian approach to benchmarking that allows for a wide range of models and benchmarks. We revise the likelihood by multiplying it by a probability distribution that measures agreement with the benchmarks. We outline Markov chain Monte Carlo methods to generate samples from benchmarked posterior distributions. We present two simulations, and an application to English and Welsh life expectancies.

Key words: Small domain estimation; Bayesian hierarchical model; area-level model; life expectancy.

1. Introduction

Small area estimation is the problem of obtaining estimates for many areas or domains defined by social, demographic and geographic variables where the number of observations in an area can be small. It has many practical applications, from monitoring unemployment to the targeting of anti-poverty programs (Pfeffermann 2013; Rao and Molina 2015), and is increasingly important for official statistics. In the United States, for instance, county-level estimates of poverty rates from the Small Area Income and Poverty Estimates (SAIPE) program are used to allocate federal funding (U.S. Census Bureau 2014). In areas where the number of observations is small, 'direct' methods, such as estimating rates by dividing the number of events in the area by the population at risk, perform poorly. Small area estimation models compensate for small sample sizes by exploiting additional information, such as covariate data or values from similar areas.

Because of their practical importance, small area estimates often receive extensive public scrutiny. This scrutiny typically includes a consistency check: estimates for small areas should agree with aggregate estimates for large areas, which are generally obtained

[1] National School of Development, Center for Statistical Science and Center for Data Science, Peking University, Beijing, 100871 China. Email: junnizhang@163.com
[2] Bayesian Demography Limited, Christchurch, New Zealand. Email: john@bayesiandemography.com

using direct methods. Model-based estimates of the number of poor people in each county, for instance, should add up to direct estimates of the number of poor people in the state. Minor discrepancies may be tolerated, but major discrepancies undermine the credibility of the estimates. Moreover, if estimates are used to allocate funding, discrepancies create grounds for dispute. Many statistical offices and funding bodies accordingly have a "one- figure" policy, whereby estimates in different tables describing the same phenomenon must all agree with each other (De Waal 2016, 232). The U.S. Census Bureau, for instance, adjusts county-level small area estimates to agree with state-level ones as part of the SAIPE program (U.S. Census Bureau 2014). Within the field of small area estimation, the aggregate estimates are referred to as benchmarks, and techniques for forcing small area estimates to agree with the benchmarks are known as benchmarking (Pfeffermann 2013).

Many existing methods for benchmarking treat benchmarks as a type of constraint. The methods differ, however, in the way that the constraints are interpreted, and in the way that the constraints are incorporated into the estimation procedures. Some methods follow a two-step procedure: first estimating the small area models, and then modifying the resulting point estimators to satisfy the benchmarking constraints (You et al. 2004; Datta et al. 2011; Berg and Fuller 2009; Berg et al. 2012; Fabrizi et al. 2012; Steorts and Ghosh 2013; Fabrizi et al. 2014; Ghosh et al. 2015). Some methods treat benchmarks as constraints on the underlying small area parameters and estimate the small area models under these constraints (Pfeffermann and Barnard 1991; Pfeffermann and Tiller 2006; Fabrizi et al. 2012; Pfeffermann et al. 2014). Some methods estimate the small area models in a way that the benchmarking constraints are satisfied for point estimators of the small area parameters (You and Rao 2002, 2003; Wang et al. 2008; You et al. 2013; Bell et al. 2013; Ranalli et al. 2018).

Most methods, including all of the ones cited above, focus on obtaining point estimates of small area parameters and associated uncertainty measures. Some Bayesian benchmarking methods, however, provide probability distributions for small area parameters (Toto and Nandram 2010; Nandram et al. 2011; Nandram and Sayit 2011; Vesper 2013). These methods are fully Bayesian in the sense that they yield a full posterior distribution for all unknown quantities after benchmarking. On this definition of fully Bayesian benchmarking, methods such as those of You et al. (2004) and Datta et al. (2011), which derive posterior distributions without benchmarking but provide point estimators after benchmarking, are not fully Bayesian. The advantage of having a full posterior distribution is that it automatically provides measures of uncertainty for all model parameters, small area parameters, and derived quantities.

In this article, we present an approach to fully Bayesian benchmarking that can be applied to a wide range of small area models. We treat benchmarks as estimates for underlying aggregate parameters. To measure agreement with the benchmarks, we specify a probability distribution for the benchmarks conditional on the aggregate parameters. We revise the likelihood function by multiplying the original likelihood function by the probability distribution for the benchmarks. Multiplying the revised likelihood function by the prior distribution then yields the benchmarked posterior distribution.

In the main body of the article, we focus on 'area-level' models, as opposed to 'unit-level' models (Rao and Molina 2015). Area-level models relate small area direct estimators to area-specific covariates. The Fay-Herriot model (Fay and Herriot 1979), for instance, is a popular area-level model used for the estimation of small area means. Unit-level models relate the unit values of an outcome variable to unit-specific covariates. The World Bank or ELL method (Elbers et al. 2003), for instance, is a widely used method for estimating small area poverty indicators, in which a unit-level model is fitted using survey data, and then applied to census data to obtain values of the outcome for all units. In the online Supplemental data (see Section 5) we discuss how our methods could be extended to unit-level models.

We implement our approach using Markov chain Monte Carlo (MCMC) methods. The methods are designed to work with complicated models that would be difficult to benchmark using previous fully Bayesian benchmarking approaches.

Our approach accommodates multiple benchmarks, and benchmarks that are nonlinearly related to small-area quantities. There is little previous research on nonlinear benchmarks: exceptions are Datta et al. (2011) and Fabrizi et al. (2012). In the application section, we estimate age-specific mortality rates benchmarked to life expectancies, which are nonlinearly related to the age-specific rates.

Our approach also allows control over the degree of agreement between model-based estimates and benchmarks. In some applications, users require exact agreement between small areas estimates and benchmarks, while in others, they may tolerate minor discrepancies. We refer to methods that achieve complete agreement as exact benchmarking, and methods that allow discrepancies as inexact benchmarking. Almost all previous methods have implemented exact benchmarking. Exceptions include Bell et al. (2013, Section 2), Nandram and Sayit (2011), and Vesper (2013).

The rest of the article is organized as follows. Section 2 describes our approach, including an outline of the associated MCMC methods. Section 3 compares our approach with previous approaches. Section 4 uses two simulation studies to illustrate the effect of benchmarking on the performance of small area models. Section 5 applies our methods to the problem of estimating district-level life expectancy in England and Wales. Section 6 summarizes the advantages of our methods.

2. A Fully Bayesian Approach to Benchmarking

2.1. Conceptual Framework

We start with a standard setup for the fully Bayesian estimation of area-level models. The aim is to estimate area-level parameters $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_n\}^\top$, such as means, rates, or probabilities, on the n areas defined by a multiway classification constructed from variables such as age, sex, and region. The data are area-level observations $\boldsymbol{y} = \{y_1, \ldots, y_n\}$. In a hierarchical Bayesian model, the likelihood is $p(\boldsymbol{y} \mid \boldsymbol{\gamma})$, the prior distribution is $p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a vector of hyperparameters, and the posterior distribution is

$$p(\boldsymbol{\gamma}, \boldsymbol{\phi} \mid \boldsymbol{y}) \propto p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})p(\boldsymbol{y} \mid \boldsymbol{\gamma}). \tag{1}$$

The prior may itself have a complicated hierarchical structure. Throughout the article, we use Roman letters to denote data, and use Greek letters to denote parameters.

We extend this setup to incorporate benchmarking. The statistician carrying out the small area estimation is provided with a set of benchmarks $\boldsymbol{m} = \{m_1, \ldots, m_d\}^\top$, with d much less than n. The benchmarks are pre-existing summary statistics at a more aggregate level than $\boldsymbol{y}$. If $\boldsymbol{y}$ is numbers of people in the labor force disaggregated by age, sex, and education level, for example, then $\boldsymbol{m}$ might be estimates of labor force participation rates disaggregated only by sex. If $\boldsymbol{y}$ is death counts disaggregated by age, sex and region, then $\boldsymbol{m}$ might be estimates of life expectancy by sex and region. The statistician is required to make estimates of the area-level parameters $\boldsymbol{\gamma}$ agree with the benchmarks $\boldsymbol{m}$.

The benchmarks could be calculated from $\boldsymbol{y}$, or from other data sources. Within the small area estimation literature, benchmarking where $\boldsymbol{m}$ is calculated from $\boldsymbol{y}$ itself is known as internal benchmarking, and benchmarking where $\boldsymbol{m}$ is calculated from other sources is known as external benchmarking (e.g., Bell et al. 2013).

Decisions on whether to benchmark, on which statistics to benchmark to, on whether to use internal or external benchmarking, and on the degree of agreement required between small area estimates and benchmarks, are typically determined by the institutional setting and the specifics of the application. Statistical agencies often have a policy of using direct methods for aggregate measures where sample sizes are large, and using model-based methods to disaggregate further, with the requirement that model-based estimates agree with aggregate ones (Little 2012). In other words, statistical agencies require statisticians to perform internal benchmarking.

If the small area estimates will be used to allocate funding, then exact benchmarking may be required, to avoid surpluses or shortfalls. In contrast, if the main users of small area estimates are researchers and policy analysts, then some discrepancies between small area estimates and aggregates estimates may be acceptable.

We distinguish between the benchmarks and the underlying parameters that they estimate. Let $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_d\}^\top$denote the parameters that the benchmarks $\boldsymbol{m}$ estimate. Vector $\boldsymbol{\psi}$ is derived from $\boldsymbol{\gamma}$ through a deterministic benchmarking function $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$, which consists of d components $\psi_j = f_j(\boldsymbol{\gamma})$, $j = 1, \ldots, d$. For each benchmarking parameter ψ_j, let δ_j denote the set of areas i such that γ_i contributes to ψ_j. We require that the δ_j do not overlap, in that each area i belongs to at most one δ_j. This restriction is commonly used in applications of benchmarking.

The components of the benchmarking function are typically linear, so that

$$\psi_j = \sum_{i=1}^{n} b_{ij}\gamma_i, \qquad i = 1, \ldots, n, j = 1, \ldots, d, \tag{2}$$

where the b_{ij} are known constants and $b_{ij} = 0$ for $i \notin \delta_j$. Equivalently, $\boldsymbol{\psi} = \boldsymbol{B}^\top \boldsymbol{\gamma}$ where $\boldsymbol{B}$ is a $n \times d$ matrix of b_{ij}. For example, if $\boldsymbol{\gamma}$ is labor force participation rates by age, sex and education level, and $\boldsymbol{\psi}$ is labor force participation rates by sex, then δ_j consists of all areas associated with sex j, and $b_{ij} = w_i / \sum_{i' \in \delta_j} w_{i'}$, where w_i is the population count for area i. However, the components of the benchmarking function may also be nonlinear. For example, if $\boldsymbol{\gamma}$ is mortality rates by age, sex and region, and $\boldsymbol{\psi}$ is life expectancy by sex and region, then δ_j consists of all areas associated with each combination j of sex and region, and f_j is a nonlinear deterministic function of $\{\gamma_i : i \in \delta_j\}$ (Preston et al. 2001, chap. 3). In the above formulation, we have assumed that there is only one set of benchmarks corresponding to mutually exclusive sets of small areas. In the Supplementary data

(Section 6) we discuss how our approach can be extended to allow for multiple sets of benchmarks, for instance, with one set of benchmarks estimating labor force participation rates by sex, and a second set of benchmarks estimating labor force participation rates by age.

To measure agreement with the benchmarks, we specify a probability distribution for the benchmarks conditional on the aggregate parameters, $p^{[m\,|\,\psi]}(\boldsymbol{m}\,|\,\boldsymbol{\psi}) = p^{[m\,|\,\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$. We then multiply the original likelihood $p(\boldsymbol{y}\,|\,\gamma)$ by this distribution. The modified likelihood $p(\boldsymbol{y}\,|\,\boldsymbol{\gamma})p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$ is a compromise between the original likelihood and the requirement to agree with the benchmarks. The component $p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$ pulls the original likelihood towards the benchmarks. For values of $\boldsymbol{\gamma}$ yielding larger (smaller) values for $p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$, the original likelihood is inflated (deflated).

In the special case of external benchmarking where $\boldsymbol{m}$ comes from completely separate data sources from $\boldsymbol{y}$ and where $p^{[m\,|\,\psi]}$ describes the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, the revised likelihood gives the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{m}$ given the parameters $\boldsymbol{\gamma}$. But in external benchmarking where $p^{[m\,|\,\psi]}$ is not equal to the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, or in internal benchmarking, $p^{[m\,|\,\psi]}$ cannot be interpreted as a standard component of the likelihood, but rather as a device for enforcing the extra requirement to agree with the benchmarks.

With the revised likelihood, the benchmarked posterior distribution is given by

$$p(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{y}, \boldsymbol{m}) \propto p(\boldsymbol{\phi})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\phi})p(\boldsymbol{y}\,|\,\boldsymbol{\gamma})^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma})). \tag{3}$$

In external benchmarking, a possible alternative approach is to incorporate the benchmarks into the prior. Under this approach, conditional on the benchmarks $\boldsymbol{m}$, the parameters $\boldsymbol{\psi}$ are assumed to have a prior distribution $p^{[\psi\,|\,m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})$. There is a second prior $p^*(\boldsymbol{\psi})$, implied by $p(\boldsymbol{\phi})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\phi})$ and $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$. The two priors $p^{[\psi\,|\,m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})$ and $p^*(\boldsymbol{\psi})$ need to be combined. This can be regarded as a special case of Bayesian melding proposed by Poole and Raftery (2000). Poole and Raftery (2000) note that the problem of combining priors is addressed by the literature on combining expert judgements, with a standard method being logarithm pooling, which leads to the pooled prior distribution for $\boldsymbol{\psi}$,

$$\tilde{p}(\boldsymbol{\psi}\,|\,\boldsymbol{m}) \propto \left[p^*(\boldsymbol{\psi})\right]^{\alpha}\left[p^{[\psi\,|\,m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})\right]^{1-\alpha}. \tag{4}$$

for some value $0 < \alpha < 1$. However, Equation (4) needs to be inverted, through a complicated procedure, to the parameter space for $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ to yield a pooled prior distribution $\tilde{p}(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{m})$. Simulating from the corresponding posterior distribution, $\tilde{p}(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{m})p(\boldsymbol{y}\,|\,\gamma)$, can be difficult with complicated models. Furthermore, logarithm pooling has undesirable properties for probability calculations (O'Hagan et al. 2006, Subsection 9.2.2.).

Under external benchmarking, our approach corresponds to treating benchmarks as data and incorporating them into the likelihood. This approach is also related to the literature on combining expert judgements, in particular Morris (1974), Morris (1977), Lindley et al. (1979), Lindley (1983), Roback and Givens (2001), and Albert et al. (2012), who argue for treating expert judgements as data, and for building models of the accuracy of these judgements. This approach avoids the limitations of logarithm pooling.

2.2. Exact Benchmarking

Under exact benchmarking, model-based estimates are required to agree perfectly with the benchmarks. We interpret perfect agreement to mean that

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m}/\boldsymbol{\psi}) = \begin{cases} 1 & \text{if } \boldsymbol{m} = \boldsymbol{\psi}; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This interpretation of exact benchmarking is effectively the same as the one adopted by Pfeffermann and Barnard (1991), Pfeffermann and Tiller (2006), Fabrizi et al. (2012), and Pfeffermann et al. (2014), all of whom take frequentist approaches, and Nandram and Sayit (2011), who take a fully Bayesian approach. These methods all treat the benchmarks as constraints on the small area parameters.

When Equation (5) is plugged into Equation (3), the benchmarked posterior distribution becomes a singular distribution concentrated on the region $\{(\boldsymbol{\gamma},\boldsymbol{\phi}) : \boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}\}$. Every $\boldsymbol{\gamma}$ in the posterior distribution satisfies the restriction $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}$. Therefore, any point estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, such as the posterior mean or posterior median, satisfies $\boldsymbol{f}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{m}$. We show how samples can be generated from the singular posterior distribution in Subsection 2.5.

2.3. Inexact Benchmarking

Under inexact benchmarking, $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$ is a non-degenerate distribution. The statistician can define $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$ so that it operationalizes the definition required by the particular institutional setting. For example, if it is required that most discrepancies are smaller than a given tolerance a, such that $\Pr(\,|m_j - \psi_j| < a \mid \psi_j) \geq q$ for $j = 1, \ldots, d$, then it may be appropriate to specify $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$ as

$$m_j \overset{\text{ind}}{\sim} \mathrm{N}\left(\psi_j, \left(\frac{a}{z_{(1-q)/2}}\right)^2\right), \tag{6}$$

where $\overset{\text{ind}}{\sim}$ indicates independent distributions, and $z_{(1-q)/2}$ is the upper $(1 - q)/2$ quantile of a standard normal distribution.

In some applications, the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, $p^{[\boldsymbol{m}|\boldsymbol{\psi}]}_{\text{sample}}$, is known, and it may be appropriate to incorporate the sampling distribution into the measure of agreement. The measure can be customized by including a discrepancy parameter λ, with smaller values of λ enforcing greater agreement. We illustrate with two examples.

In the first example, the data $\boldsymbol{y}$ are obtained from a survey, and the benchmarks $\boldsymbol{m}$ are direct estimates calculated from $\boldsymbol{y}$, with standard errors $\boldsymbol{s}$. If the survey was implemented well, then $\boldsymbol{m}$ should be unbiased for $\boldsymbol{\psi}$, and $\boldsymbol{s}$ should be approximately correct. If each m_j is derived from a large number of observations, then we can assume that, conditional on the ψ_j, each m_j is independently normally distributed with mean ψ_j and standard deviation s_j. The sampling distribution $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}$ is given by

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \exp\left(-\sum_{j=1}^{d} \frac{(m_j - \psi_j)^2}{2s_j^2}\right). \tag{7}$$

Incorporating a discrepancy parameter $0 < \lambda \leq 1$ into Equation (7) yields

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}\left(\boldsymbol{m} \mid \boldsymbol{\psi}\right) \propto \exp\left(-\sum_{j=1}^{d} \frac{\left(m_j - \psi_j\right)^2}{2\lambda s_j^2}\right). \tag{8}$$

When $\lambda \rightarrow 0$, Equation (8) converges to Equation (5). Hence exact benchmarking is a limiting case of inexact benchmarking.

In the second example, the data are counts of events that follow Poisson distributions. We have $y_i \sim \text{Poisson}(w_i \gamma_i)$, where w_i is the known exposure for area i. Let $v_j = \sum_{i \in \delta_j} w_i$ denote the total exposure associated with δ_j. Then $\psi_j = \sum_{i \in \delta_j} w_i \gamma_i / v_j$, and its estimate is $m_j = \sum_{i \in \delta_j} y_i / v_j$ where $\sum_{i \in \delta_j} y_i \sim \text{Poisson}\left(\sum_{i \in \delta j} w_i \gamma_i\right) \sim \text{Poisson}\left(v_j \psi_j\right)$. The sampling distribution $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}$ is given by

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}\left(\boldsymbol{m} \mid \boldsymbol{\psi}\right) \propto \prod_{j=1}^{d} \text{Poisson}\left(v_j m_j \mid v_j \psi_j\right). \tag{9}$$

Incorporating a discrepancy parameter $0 < \lambda \leq 1$ into (9) yields

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}\left(\mathbf{m} \mid \boldsymbol{\psi}\right) \propto \prod_{j=1}^{d} \text{Poisson}\left(\lambda v_j m_j \mid \lambda v_j \psi_j\right), \tag{10}$$

with convergence to exact benchmaking as $\lambda \rightarrow 0$.

In the above examples of $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$, we have assumed conditional independence of m_j's given the underlying parameters ψ_j's, and used simple models for $p(m_j \mid \psi_j)$. This is similar to assuming conditional independence of y_i's given γ_i's and using simple models for $p(y_i \mid \gamma_i)$. Unconditionally, the m_j's can have complicated correlations, such as correlations between neighbouring time points or age groups. Such correlations are captured by the prior model on the underlying benchmarking parameters $\boldsymbol{\psi}$, which is implied by the prior model on $\boldsymbol{\gamma}$, $p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})$, and the equality $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$. The prior model on $\boldsymbol{\gamma}$ typically uses a complicated hierarchical structure to model relationship between the underlying parameters, such as similarities between neighbouring time points or age groups.

The appropriate value for the discrepancy parameter λ in any particular application depends on the sizes of discrepancies between model-based estimates and benchmarks that can be tolerated in that application. As we discuss in Subsection 2.6, the effects of benchmarking on performance measures such as accuracy are difficult to predict. One possible approach to setting λ is to fit a model several times with alternative values for λ, and use the highest value that gives acceptable levels of discrepancy.

2.4. An Illustrative Analytical Example

To illustrate the benchmarked posterior distribution, we present an example in which the distribution can be derived in closed form. The data $\boldsymbol{y} = \{y_i, \ldots, y_{\text{n}}\}^\top$ are generated and modelled using

$$y_i \overset{\text{ind}}{\sim} \text{N}\left(\gamma_i, \sigma^2\right) \tag{11}$$

$$\gamma_i \overset{\text{ind}}{\sim} \text{N}\left(\mu_0, \tau^2\right), \tag{12}$$

where μ_0, σ^2 and τ^2 are known. There is a single benchmark $m = \sum_{i=1}^{n} w_i y_i$ estimating benchmarking parameter $\psi = \sum_{i=1}^{n} w_i \gamma_i$, where the w_i's are a set of weights satisfying $\sum_{i=1}^{n} w_i = 1$. Let $\boldsymbol{w} = (w_1, \ldots, w_\mathrm{n})^\top$ denote the vector of weights, and $\mathbf{1}_n$ a vector of n ones. Then $\boldsymbol{w}^\top \boldsymbol{y} = m$, $\boldsymbol{w}^\top \boldsymbol{\gamma} = \psi$, and $\boldsymbol{w}^\top \mathbf{1}_n = 1$. Under exact benchmarking, $p^{[\mathbf{m} \mid \psi]}$ is given by Equation (5). Under inexact benchmarking, the sampling distribution $p_{\text{sample}}^{[m \mid \psi]}$ is given by

$$p_{\text{sample}}^{[m \mid \psi]}(m \mid \psi) \sim \mathrm{N}\left(\psi, \left(\boldsymbol{w}^\top \boldsymbol{w}\right)\sigma^2\right). \tag{13}$$

We incorporate a discrepancy parameter λ into (13) and arrive at

$$p^{[m \mid \psi]}(m \mid \psi) \sim \mathrm{N}\left(\psi, \lambda\left(\boldsymbol{w}^\top \boldsymbol{w}\right)\sigma^2\right), \tag{14}$$

where $0 < \lambda \leq 1$.

Let $\boldsymbol{I}_\mathrm{n}$ be the $n \times n$ identity matrix. As shown in the Supplemental data (Section 1), Equations (5), (11), (12) and (14) yield posterior distributions for $\boldsymbol{\gamma}$ that are multivariate normal under no benchmarking (NB), exact benchmarking (EB), and inexact benchmarking (IB), with means and variances

$$\boldsymbol{\mu}^{\mathrm{NB}} = -\frac{\sigma^2}{\sigma^2 + \tau^2} \mathbf{1}_n \mu_0 + \frac{\tau^2}{\sigma^2 + \tau^2} \boldsymbol{y}, \tag{15}$$

$$\boldsymbol{\Sigma}^{\mathrm{NB}} = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \boldsymbol{I}_n, \tag{16}$$

$$\boldsymbol{\mu}^{\mathrm{EB}} = \frac{\sigma^2}{\sigma^2 + \tau^2} \left[\mathbf{1}_n - \frac{1}{\boldsymbol{w}^\top \boldsymbol{w}} \boldsymbol{w}\right] \mu_0 + \frac{\sigma^2}{\sigma^2 + \tau^2} \frac{1}{\boldsymbol{w}^\top \boldsymbol{w}} \boldsymbol{w} m + \frac{\tau^2}{\sigma^2 + \tau^2} \boldsymbol{y}, \tag{17}$$

$$\boldsymbol{\Sigma}^{\mathrm{EB}} = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \left[\boldsymbol{I}_n - \frac{1}{\boldsymbol{w}^\top \boldsymbol{w}} \boldsymbol{w} \boldsymbol{w}^\top\right], \tag{18}$$

$$\boldsymbol{\mu}^{\mathrm{IB}} = \left[1 - \frac{\tau^2}{\lambda\sigma^2 + (\lambda + 1)\tau^2}\right] \boldsymbol{\mu}^{\mathrm{NB}} + \frac{\tau^2}{\lambda\sigma^2 + (\lambda + 1)\tau^2} \boldsymbol{\mu}^{\mathrm{EB}}, \tag{19}$$

$$\boldsymbol{\Sigma}^{\mathrm{IB}} = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \left[\boldsymbol{I}_n - \frac{\tau^2}{\left(\lambda\sigma^2 + (\lambda + 1)\tau^2\right) \boldsymbol{w}^\top \boldsymbol{w}} \boldsymbol{w} \boldsymbol{w}^\top\right]. \tag{20}$$

With no benchmarking, the posterior mean for γ_i equals the observation y_i shrunk towards the prior mean μ_0. With exact benchmarking, the posterior mean is instead shrunk towards a linear combination of the prior mean μ_0 and the benchmark m. With inexact benchmarking, the posterior mean is a compromise between the means under no benchmarking and exact benchmarking. Benchmarking reduces posterior variance in this setting, with exact benchmarking leading to larger reductions than inexact benchmarking.

2.5. *A General MCMC Approach to Sampling from a Benchmarked Posterior Distribution*

In practical applications, closed form expressions for the benchmarked posterior distribution Equation (3) are seldom available, and posterior inference must be carried out via simulation. We outline a general MCMC strategy for sampling from Equation (3).

We first discuss the case where the components of the benchmarking function are linear. Under exact benchmarking, Equation (3) is a singular distribution concentrated on the region $\{(\boldsymbol{\gamma}, \boldsymbol{\phi})\colon \boldsymbol{B}^\top \boldsymbol{\gamma} = \boldsymbol{m}\}$. We obtain draws from this singular distribution by choosing an initial value $\boldsymbol{\gamma}^{(0)}$ that satisfies $\boldsymbol{B}^\top \boldsymbol{\gamma}^{(0)} = \boldsymbol{m}$, and then repeatedly iterating through the following steps:

E1. Update $\boldsymbol{\gamma}^{(t)} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y}$ subject to $\boldsymbol{B}^\top \boldsymbol{\gamma}^{(t)} = \boldsymbol{m}$.

E2. Update $\boldsymbol{\phi}^{(t)} \mid \boldsymbol{\gamma}^{(t)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{y}$.

Step E2 can be done using standard methods. Step E1 ensure that the constraint $\boldsymbol{B}^\top \boldsymbol{\gamma} = \boldsymbol{m}$ continues to be satisfied. It is carried out as follows.

An area i_1 is randomly selected from $\{1, \ldots, n\}$. If area i_1 does not belong to any δ_j, so that γ_{i_1} is not subject to any benchmarking constraint, then γ_{i_1} is updated using standard methods. If area i_1 is the only area in an δ_j, then γ_{i_1} is fully determined by the benchmark m_j and is not updated. Otherwise, γ_{i_1} is updated through a Metropolis-Hastings step. A proposal $\gamma^*_{i_1}$ is generated from $J\left(\gamma^*_{i_1} \mid \gamma^{(t-1)}_{i_1}\right)$. Then another area i_2 from the same δ_j that i_1 belongs to is randomly selected, and $\gamma^*_{i_2}$ is obtained as $\gamma^*_{i_2} = \gamma^{(t-1)}_{i_2} + b_{i_1 j}/b_{i_2 j}\left(\gamma^{(t-1)}_{i_1} - \gamma^*_{i_1}\right)$, which ensures that $b_{i_1 j}\gamma^*_{i_1} + b_{i_2 j}\gamma^*_{i_2} = b_{i_1 j}\gamma^{(t-1)}_{i_1} + b_{i_2 j}\gamma^{(t-1)}_{i_2}$. Setting $\gamma^*_i = \gamma^{(t-1)}_i$ for $i \notin \{i_1, i_2\}$ yields a proposed value γ^* for which $\boldsymbol{B}^\top \boldsymbol{\gamma}^* = \boldsymbol{m}$ continues to be satisfied.

To calculate the joint proposal density $J\left(\left(\gamma^*_{i_1}, \gamma^*_{i_2}\right) \mid \left(\gamma^{(t-1)}_{i_1}, \gamma^{(t-1)}_{i_2}\right)\right)$, we need to take account of the fact that we could have arrived at $\left(\gamma^*_{i_1}, \gamma^*_{i_2}\right)$ in one of two ways: by drawing $\gamma^*_{i_1}$ and then calculating $\gamma^*_{i_2}$ or by drawing $\gamma^*_{i_2}$ and then calculating $\gamma^*_{i_1}$. The resulting Metropolis-Hastings ratio is

$$r = \left[\frac{p(\boldsymbol{\gamma}^* \mid \boldsymbol{\phi}) p(\boldsymbol{y} \mid \boldsymbol{\gamma}^*)}{p(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{\phi}) p(\boldsymbol{y} \mid \boldsymbol{\gamma}^{(t-1)})}\right] \times \frac{J\left(\gamma^{(t-1)}_{i_1} \mid \gamma^*_{i_1}\right) + \left| b_{i_1 j}/b_{i_2 j}\right| J\left(\gamma^{(t-1)}_{i_2} \mid \gamma^*_{i_2}\right)}{J\left(\gamma^*_{i_1} \mid \gamma^{(t-1)}_{i_1}\right) + \left| b_{i_1 j}/b_{i_2 j}\right| J\left(\gamma^*_{i_2} \mid \gamma^{(t-1)}_{i_2}\right)}. \tag{21}$$

Since the benchmarked posterior distribution Equation (3) under exact benchmarking is a singular distribution, it is not immediately obvious that the above Metropolis-Hastings algorithm has the desired convergence property. The Supplemental data (Subsection 2.1) provides a proof that under exact benchmarking, the stationary distribution of chains produced by the above algorithm is indeed Equation (3).

Under inexact benchmarking, the algorithm for sampling from Equation (3) is

I1. Update $\boldsymbol{\gamma}^{(t-\frac{1}{2})} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y}$ subject to the constraint $\boldsymbol{B}^\top \boldsymbol{\gamma}^{(t-1)} = \boldsymbol{\psi}^{(t-1)}$, where $\boldsymbol{\psi}^{(t-1)} = \boldsymbol{B}^\top \boldsymbol{\gamma}^{(t-1)}$.

I2. Update $\boldsymbol{\gamma}^{(t)} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-\frac{1}{2})}, \boldsymbol{y}$ with no constraint.

I3. Update $\boldsymbol{\phi}^{(t)} \mid \boldsymbol{\gamma}^{(\mathrm{t})}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{y}$.

Step I1 is similar to step E1 with exact benchmarking, and step I3 is the same as step E2 with exact benchmarking. Step I2 can be carried out using Metropolis-Hastings updates similar to those for an unbenchmarked model, except that the density $p^{[\boldsymbol{m}/\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{B}^\top \boldsymbol{\gamma})$ needs to be accounted for in the Metropolis-Hastings ratio. Step I1 is not strictly necessary, but speeds up the exploration of the parameter space when $p^{[\boldsymbol{m}/\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{B}^\top \boldsymbol{\gamma})$ is tightly concentrated around the hyperplane defined by $\boldsymbol{m} \mid \boldsymbol{B}^\top \boldsymbol{\gamma}$. The Supplemental data (Subsection 2.2) provides a proof that under inexact benchmarking, the stationary distribution of chains produced by the above algorithm is the benchmarked posterior distribution in Equation (3).

When the components of the benchmarking function are nonlinear, under inexact benchmarking Equation (3) is a singular distribution concentrated on the region $\{(\boldsymbol{\gamma},\boldsymbol{\phi}) : \boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}\}$. There is generally no efficient way to implement a step similar to E1 or I1 which ensures that the constraint $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}$ or $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{\psi}^{(t-1)}$ continues to be satisfied. Instead we use steps I2 and I3 for inexact benchmarking, and approximate exact benchmarking by using inexact benchmarking with discrepancy parameter λ close to zero.

We have implemented our general MCMC approaches with a specific family of area-level hierarchical models:

$$y_i \mid \gamma_i, \sigma^2 \overset{\text{ind}}{\sim} G(\gamma_i, w_i, \sigma^2), \tag{22}$$

$$g(\gamma_i) \mid \boldsymbol{\beta}, \tau^2 \overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \tau^2). \tag{23}$$

In Equation (22), y_i is an observation for area i within a multiway classification, G denotes the normal, Poisson or binomial distribution, w_i is a known weight, exposure or number of trials, and σ^2 is a variance, used only with the normal distribution. In Equation (23), g is the identity, log or logit link function. The transformed values $g(\gamma_i)$ are modelled using a structure similar to analysis of variance. Vector $\boldsymbol{\beta}$ contains batches of coefficients representing main effects and interactions formed from the cross-classifying dimensions. Vector $\boldsymbol{x}_i$ is a vector consisting of ones and zeros indicating which main effects and interactions are associated with each area i. We place no restrictions on the prior for $\boldsymbol{\beta}$, and it will typically have a complicated hierarchical structure. The Supplemental data (Section 3) gives details of the specific MCMC samplers.

We have written R packages implementing the models, which can be obtained from github.com/statisticsnz/R. The family of models included in the packages can accommodate a wide range of real applications, and the packages are user-friendly, making it easy for practitioners to implement the fully Bayesian benchmarking approached presented in this article.

2.6. *The Effects of External and Internal Benchmarking on Model Performance*

External benchmarking allows information from the external data sources to be incorporated into the analysis. When $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$ is constructed from a correctly specified $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}$, external benchmarking should, on average, improve model performance as measured by criteria such as accuracy and coverage.

The effect of internal benchmarking on accuracy and coverage is more ambiguous. Internal benchmarking entails using data $\boldsymbol{y}$ twice: once when calculating benchmarks $\boldsymbol{m}$, and again in $p(\boldsymbol{y} \mid \boldsymbol{\gamma})$. If a correctly-specified model is subject to internal benchmarking, then it is no longer correctly specified. Performance on accuracy and coverage can be expected to suffer. Previous studies with frequentist and empirical Bayes approaches have confirmed that this is indeed the case: when the unbenchmarked model is correctly specified, benchmarking typically reduces accuracy and coverage (Pfeffermann and Barnard 1991; Wang et al. 2008; Datta et al. 2011; Bell et al. 2013).

When the unbenchmarked model is correctly specified, methods of benchmarking that enforce stronger forms of agreement with the benchmarks can be expected to perform worse in terms of accuracy and coverage. For example, exact benchmarking can be expected to have poorer accuracy and coverage than inexact benchmarking.

In real applications, however, the model is almost always misspecified. When the model is misspecified, the effects of internal benchmarking on model performance are uncertain. Previous simulation studies suggest that, depending on the details of the data and model, internal benchmarking can sometimes improve performance (Pfeffermann and Tiller 2006; Nandram et al. 2011; Pfeffermann 2013; Vesper 2013; Ranalli et al. 2018).

Given the uncertainty about the effect of benchmarking on model performance, we suggest that benchmarking not be seen as a method for protecting against model misspecification. Instead, analysts should use standard model-checking tools such as posterior predictive checks (Gelman et al. 2014, chap. 6) to detect possible problems with their models, and adjust the models accordingly. Benchmarking should, rather, be seen as a method for achieving agreement between model-based estimates and benchmarks.

3. Comparison with Previous Approaches

3.1. An Alternative Interpretation of Exact Benchmarking

In our interpretation of exact benchmarking, set out in Equation (5), the entire posterior distribution must agree with the benchmarks. Under this approach, any standard point estimate derived from the posterior distribution, such as the posterior mean or posterior median, automatically agrees with the benchmarks.

Most previous approaches interpret exact benchmarking less strictly. Instead of working with full distributions, they work only with point estimates. They require a specific point estimate, $\hat{\gamma}^{\text{Spe}}$, to agree with the benchmarks,

$$\boldsymbol{f}\left(\hat{\gamma}^{\text{Spe}}\right) = \boldsymbol{m}. \tag{24}$$

You et al. (2004), Datta et al. (2011), and Ghosh et al. (2015), for example, obtain point estimates $\hat{\gamma}^{\text{FB}}$ from a fully Bayesian model, and then adjust them to obtain a new set of estimates $\hat{\gamma}^{\text{Spe}}$ that satisfy the benchmarking constraint. When the benchmarks are linear, one such estimator is the raked or ratio-adjusted estimator,

$$\hat{\gamma}_i^{\text{Spe}} = \hat{\gamma}_i^{\text{FB}} \frac{m_j}{\sum_{i'=1}^{n} b_{i'j} \hat{\gamma}_{i'}^{\text{FB}}}. \tag{25}$$

The raked estimator is easy to implement, and is widely used in practice, but has been characterised as ad hoc (Ghosh et al. 2015). Datta et al. (2011) (henceforth DGSM) instead propose an estimator that minimizes the expected posterior loss based on a weighted squared error loss function $L(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \rho_i (\hat{\gamma}_i - \gamma_i)^2$, where ρ_i are known weights and $\hat{\gamma}$ satisfies $\boldsymbol{f}(\hat{\gamma}) = \boldsymbol{m}$. As DGSM point out, with the appropriate choice of ρ_i, the raked estimator can be derived as a special case of their estimator. Ghosh et al. (2015) argue that the squared error loss function may not be appropriate for estimating positive quantities such as income, and propose an estimator that minimizes the expected posterior loss based on a variant of the Kullback-Leibler loss function.

Our approach to exact benchmarking enforces stronger forms of agreement with the benchmarks. As discussed in Subsection 2.6, this can lead to poorer performance on criteria such as accuracy and coverage, when the unbenchmarked model is correctly specified. However, when the model is misspecified, depending on the details of the data and model, enforcing stronger forms of agreement may sometimes improve performance, as we illustrate in Subsection 4.1.

Requiring that the entire posterior distribution agree with the benchmarks, as we do with exact benchmarking, means that the statistician does not have to choose a particular loss function. Moreover, with the entire posterior distribution available, the statistician can obtain a posterior distribution for any function of the small area parameters (Gelman et al. 2014, 261–262). For instance, given a posterior distribution for county-level income levels, the statistician can derive a posterior distribution for county income rankings.

3.2. *Previous Fully Bayesian Benchmarking Approaches*

Toto and Nandram (2010) and Nandram et al. (2011) use fully Bayesian benchmarking on a model where the posterior distribution is multivariate normal, and where there is a single benchmark and a linear benchmarking function. Their model is specified at the unit level rather than the area level. We discuss unit models further in the Supplemental data, Section 4.

Nandram and Sayit (2011) benchmark an area-level beta-binomial hierarchical Bayesian model,

$$y_i \mid \gamma_i \overset{\text{ind}}{\sim} \text{Binomial}(w_i, \gamma_i), \tag{26}$$

$$\gamma_i \mid \mu, \tau \overset{\text{ind}}{\sim} \text{Beta}\big(\mu\tau, (1-\mu)\tau\big), \tag{27}$$

$$p(\mu, \tau) \propto \left(1+\tau^2\right)^{-1}, \quad 0 < \mu < 1, \quad \tau \geq 0. \tag{28}$$

Here w_i is a known number of trials for area i. The authors work with a single benchmarking parameter $\psi = \sum_{i=1}^{n} b_i \gamma_i$, where $b_i = w_i / \sum_{i'=1} w_{i'}$. Instead of incorporating the benchmark into the likelihood, Nandram and Sayit (2011) incorporate it into a prior distribution for ψ, $p(\psi) \sim \text{Beta}(m\tau_0, (1-m)\tau_0)$. The authors consider three scenarios for $p(\psi)$: (1) exact benchmarking, with $\tau_0 \to \infty$ and $p(\psi)$ a point mass at m; (2) inexact benchmarking, with τ_0 specified by the user; and (3) inexact benchmarking, with $m = 1/2$ and $\tau_0 = 2$, so that $p(\psi) \sim \text{Uniform}[0, 1]$.

Let $p_{\text{NB}}(\gamma_1, \ldots, \gamma_{n-1}, \gamma_n, \mu, \tau \mid \boldsymbol{y})$ denote the unbenchmarked posterior distribution for (γ, μ, τ). By using the identity $\gamma_n = \left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right) / b_n$, Nandram and Sayit (2011) are able to work with $(\gamma_1, \ldots, \gamma_{n-1}, \psi)$ instead of $(\gamma_1, \ldots, \gamma_{n-1}, \gamma_n)$, and derive the benchmarked posterior distribution for ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$) as

$$\begin{aligned} & p(\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau \mid \boldsymbol{y}) \\ & \propto p(\psi) p_{\mathrm{NB}}\left(\gamma_1, \ldots, \gamma_{n-1}, \frac{1}{b_n}\left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right), \mu, \tau \mid \boldsymbol{y}\right). \end{aligned} \tag{29}$$

Nandram and Sayit (2011) use a Gibbs sampling algorithm to sample ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$) from Equation (29). The full conditional distributions for γ_i ($i = 1, \ldots, n - 1$) and ψ are both proportional to the product of two density functions, one being a truncated beta density and the other being a generalized beta density. Specialized algorithms are used to draw samples from these distributions. After obtaining samples for ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$), samples for γ_{n} are then obtained using the identity $\gamma_n = \left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right)/b_n$.

Implementation of the approach used by Nandram and Sayit (2011) depends on the choice of which small area is labeled as area n and left out. This choice affects the specific posterior distribution derived in Equation (29) and hence affects the computational efficiency of the MCMC algorithms. Nandram and Sayit (2011) sort the areas in ascending order of y_i, with area n having the largest value of y_i. When there are multiple benchmarks, with this approach, one area needs to be left out in each δ_j. Poor choices may lead to poor computational efficiency. In contrast, our MCMC approach in Subsection 2.5 does not depend on the labeling of areas.

The approach of Nandram and Sayit (2011) is also difficult to generalize to nonlinear benchmarking functions. Even with a single benchmarking parameter $\psi = f(\gamma_1, \ldots, \gamma_n)$ where f is nonlinear, it can be difficult to write γ_n analytically as a function of $\gamma_1, \ldots, \gamma_{n-1}, \psi$. Therefore, it may not be possible to write out a benchmarked posterior distribution similar to Equation (29), or to draw samples from it. In contrast, our approach can accommodate nonlinear benchmarking functions.

Vesper (2013) works with the Fay-Herriot model (Fay and Herriot 1979):

$$y_i \mid \gamma_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\gamma_i, \sigma_i^2\right), \tag{30}$$

$$\gamma_i \mid \boldsymbol{\beta}, \tau^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \tau^2\right), \tag{31}$$

$$p(\boldsymbol{\beta}, \tau^2) \propto \left(\tau^2\right)^{-a-1} e^{-b/\tau^2}, \tag{32}$$

where σ_i^2 is the variance of y_i and is assumed known, $\boldsymbol{x}_i$ is an observed vector of covariates for area i, and a and b are known constants. There is a single benchmarking parameter $\psi = \sum_{i=1}^n b_i \gamma_i$, and $p(\psi) \sim N\left(m, \sum_{i=1}^n b_i^2 \sigma_i^2\right)$. The benchmarked posterior distribution is similar to Equation (29) from Nandram and Sayit (2011). This approach is subject to the same implementation limitations as that of Nandram and Sayit (2011).

4. Two Simulation Studies

4.1. Estimation of Fertility Rates from Registration Data

We use simulated data on births to examine how benchmarking affects accuracy, coverage, and agreement between model-based estimates and benchmarks. We examine performance with correctly specified models and with misspecified models. All code and data for the simulation are available at github.com/bayesiandemography/fertsim.

The simulated data consist of counts of births and reproductive-age women. With the baseline ‘no change’ data sets, counts of births are generated using the model

$$y_{art} \overset{\text{ind}}{\sim} \text{Poisson}(w_{art}\gamma_{art}^{\text{Tr}}), \tag{33}$$

$$\log \gamma_{art}^{\text{Tr}} \overset{\text{ind}}{\sim} \text{N}(\beta_a^{\text{age-std}} + \beta_r^{\text{reg-Tr}} + \beta_t^{\text{time-Tr}}, \sigma_{\text{Tr}}^2), \tag{34}$$

where y_{art} is the number of births to women in age group $a \in \{15-19, \ldots, 40-44\}$ in region $r \in \{1, \ldots, 30\}$ during period $t \in \{1, 2, 3\}$, w_{art} is the corresponding person-years of exposure, and γ_{art}^{Tr} is the true underlying birth rate. We set $w_{art} = 300$ for all a, r, t, and set $\sigma_{\text{Tr}} = 0.1$. Age effects $\beta_a^{\text{age-std}}$ are taken from the ‘standard’ fertility schedule in Table 1; region effects have distribution $\beta_r^{\text{reg-Tr}} \overset{\text{ind}}{\sim} \text{N}(0, 0.1^2)$; and time effects have distribution $\beta_t^{\text{time-Tr}} \overset{\text{ind}}{\sim} \text{N}(0, 0.1^2)$.

To explore how benchmarking affects performance under model misspecification, we also construct ‘change in level’ and ‘change in distribution’ data sets by perturbing the ‘no change’ data set. The ‘change in level’ data set represents a sudden change in the level of fertility in a subset of regions. Birth rates and counts are identical to the ‘no change’ data set, except for areas in regions 26–30 during period 3, which we refer to as being ‘nonstandard’ areas. Log rates for the nonstandard areas are generated by adding log 0.2, to the existing log rates. Counts for nonstandard areas are obtained by drawing new values from Equation (33). The ‘change in distribution’ data set represents a sudden change in the age-pattern, rather than the level, of fertility. Birth rates and counts are again identical to the ‘no change’ data set except in regions 26–30 during period 3. Log rates for nonstandard areas are generated by replacing the ‘standard’ age effects from Table 1 with the ‘nonstandard’ age effects, and leaving region effects and time effects the same. Counts for nonstandard areas are obtained by drawing new values from Equation (33).

We simulate an analysis seeking to estimate γ_{art}^{Tr}. The simulated analysis model has likelihood

$$y_{art} \overset{\text{ind}}{\sim} \text{Poisson}(w_{art}\gamma_{art}), \tag{35}$$

and assumes that

$$\log \gamma_{art} \overset{\text{ind}}{\sim} \text{N}(\beta^0 + \beta_a^{\text{age}} + \beta_r^{\text{reg}} + \beta_t^{\text{time}}, \sigma^2). \tag{36}$$

Table 1. Age effects (exponentiated).

	Standard	Nonstandard
15–19	0.0288	0.0695
20–24	0.0713	0.1225
25–29	0.1083	0.0936
30–34	0.1210	0.0726
35–39	0.0653	0.0394
40–44	0.0127	0.0098
Total	0.4074	0.4074

Equation (36) is correctly specified for the ‘no change’ data sets. However, it is misspecified for the ‘change in level’ and ‘change in distribution’ data sets, since the true data-generating processes for these data sets contain interactions.

The intercept term in the simulated analysis model has a proper but diffuse prior, $\beta^0 \sim \mathrm{N}(0, 10^2)$. The region effect has a normal prior $\beta_r^{\mathrm{reg}} \sim \mathrm{N}(0, \tau_{\mathrm{reg}}^2)$, with a weakly informative half-t prior with seven degrees of freedom on the standard deviation (Gelman et al. 2008), $\tau_{\mathrm{reg}} \sim t_7^+(0, 0.25^2)$. The age effect has a ‘random walk with noise’ (Prado and West 2010, 119–120) prior,

$$\beta_a^{\mathrm{age}} \overset{\mathrm{ind}}{\sim} \mathrm{N}(\eta_a^{\mathrm{age}} \tau_{\mathrm{age}}^2) \tag{37}$$

$$\eta_a^{\mathrm{age}} \overset{\mathrm{ind}}{\sim} \mathrm{N}(\eta_{a-1}^{\mathrm{age}} \omega^2) \tag{38}$$

with $\eta_0^{\mathrm{age}} \sim \mathrm{N}(0, 10^2)$, $\tau_{\mathrm{age}} \sim t^+(0, 1)$, and $\omega \sim t_7^+(0, 1)$. The random walk with noise prior recognizes the tendency for neighbouring age groups to have similar values. The standard deviation parameter from Equation (36) has a weakly informative half-t prior, $\sigma \sim t_7^+(0, 0.25^2)$. The time effect has the same prior as the region effect.

The analysis model is fitted with (i) no benchmarking, (ii) exact benchmarking, and (iii) inexact benchmarking. Let $v_{rt} = \sum_a w_{art} = 1,800$. The benchmarks are region-time means $m_{rt} = \sum_a y_{art}/v_{rt}$, which estimate benchmarking parameters $\psi_{rt} = \sum_a w_{art}\gamma_{art}/v_{rt}$. Under exact benchmarking, $p^{[\mathbf{m}\,|\,\boldsymbol{\psi}]}(\mathbf{m} \mid \boldsymbol{\psi})$ is given by Equation (5). Under inexact benchmarking, $p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi})$ is given by Equation (10), which becomes

$$p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \prod_{r=1}^{30} \prod_{t=1}^{3} \mathrm{Poisson}(\lambda v_{rt} m_{rt} \mid \lambda v_{rt} \psi_{rt}). \tag{39}$$

We consider the case where $\lambda = 1$, which allows discrepancies between model-based estimates and benchmarks to vary in line with Poisson variation in birth counts. As discussed in Subsection 2.3, lower values for λ would lead to smaller discrepancies. We use the posterior means of γ_{art} and ψ_{rt} as point estimators.

We also adjust the posterior means of γ_{art} from the ‘no benchmarking’ case to obtain the raked estimator in Equation (25), and the DGSM estimator based on a weighted squared error loss function. Following Datta et al. (2011, 580) and Wang et al. (2008), we set ρ_{art}, the weight in the weighted squared error loss function, equal to the inverse of the estimated variance of the direct estimate y_{art}/w_{art}. Since these two estimators achieve exact benchmarking, the corresponding point estimators of ψ_{rt} are equal to m_{rt}.

We apply four performance measures. The first is

$$D_{rt} = E\left(\frac{|\,\hat{\psi}_{rt} - m_{rt}\,|}{m_{rt}}\right), \tag{40}$$

where $\hat{\psi}_{rt}$ is the point estimator of ψ_{rt}. This measure captures discrepancies (i.e., levels of disagreement) between the model-based estimates and benchmarks. With exact benchmarking under our approach, the raked estimator, and the DGSM estimator, D_{rt} always equals 0.

The second performance measure is the mean squared error from using the point estimator $\hat{\gamma}_{rt}$ to estimate $\gamma_{art}^{\mathrm{Tr}}$,

$$\mathrm{MSE}_{art} = \mathrm{E}(\hat{\gamma}_{art} - \gamma_{art}^{\mathrm{Tr}})^2. \tag{41}$$

This measure captures the accuracy of the point estimator.

The third measure, W_i^q, is the expected width of a $(1 - q) \times 100\%$ credible interval for γ_{art}. Let $\gamma_{art}^{q/2}$ and $\gamma_{art}^{1-q/2}$ be the $q/2$ and $1 - q/2$ quantiles for the posterior distribution of γ_{art}. Then

$$W_{art}^q = \mathrm{E}\left(\gamma_{art}^{1-q/2} - \gamma_{art}^{q/2}\right). \tag{42}$$

Values for W_i^q cannot be calculated for the raked and DGSM estimators, since these estimators do not come with measures of uncertainty.

The fourth measure, C_{art}^q, is the expected coverage rate of a $(1 - q) \times 100\%$ credible interval for $\gamma_{art}^{\mathrm{Tr}}$,

$$C_{art}^q = \Pr\left(\gamma_{art}^{q/2} \leq \gamma_{art}^{\mathrm{Tr}} \leq \gamma_{art}^{1-q/2}\right). \tag{43}$$

Again, values for C_{art}^q cannot be calculated for the raked and DGSM estimators.

We use $K = 100$ simulation replicates. As discussed in the Supplementary data, 100 replicates is enough to obtain stable estimates for the performance indicators we are interested in. At each replicate, results for no benchmarking, exact benchmarking, inexact benchmarking, raked estimators, and DGSM estimators are obtained for each of the ‘no change’, ‘change in distribution’, and ‘change in level’ data sets, yielding $5 \times 3 = 15$ sets of results. With the unbenchmarked model, the Gibbs sampler is run with four independent chains, each with 20,000 iterations. Every 40th draw from the final 10,000 iterations of each chain is recorded, yielding a combined total of 2,000 draws from the posterior distribution. With the benchmarked models, which converge more quickly, the number of iterations and thinning ratios are both reduced by a factor of five.

When calculating performance measures D_{rt}, MSE_{art}, W_{art}^q, and C_{art}^q, we use means across K replicates to approximate $\mathrm{E}(\cdot)$ or $\Pr(\cdot)$ in (40)–(43). These measures are calculated separately for each rt or art. Figure 1 summarizes the resulting distributions across rt or art using boxplots, where W_{art}^q and C_{art}^q are calculated for $q = 0.95$. The median for each distribution is printed above the corresponding notch in the boxplot.

The top row of Figure 1 gives results for the ‘no change’ data sets, where the analysis model is correctly specified. The model without benchmarking departs furthest from the benchmarks, but has the lowest MSE. The version of our model with exact benchmarking has the opposite strengths and weaknesses, agreeing exactly with the benchmarks (by construction), but having the highest MSE. The raked and DGSM estimators also obtain complete agreement, but with median MSE that is approximately 3–4% lower than the model with exact benchmarking. The model with inexact benchmarking is in an intermediate position, with moderate agreement and moderate MSE.

The models with no benchmarking, inexact benchmarking, and exact benchmarking have coverage rates close to the nominal 95%, but the model without benchmarking achieves this with the narrowest credible intervals. As noted above, the raked and DGSM estimators do not have uncertainty measures and therefore do not have coverage rates.

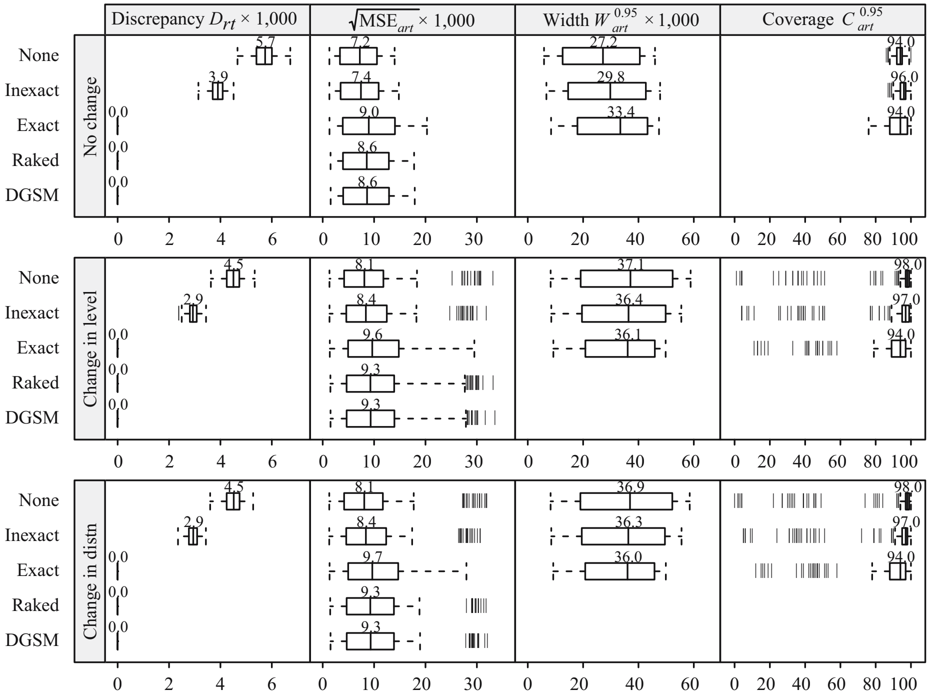


Fig. 1. Performance of models of fertility rates, by type of benchmarking and data set. The results are based on K = 100 replicates.

The second row of Figure 1 shows results for the 'change in levels' data sets. The rank order of the discrepancy measures and MSEs is preserved. The rank order of the width of credible intervals has changed, with the model without benchmarking having the widest credible intervals. In general, credible intervals are wider than in the 'no change' case, but coverage rates for a subset of areas are poor, with and without benchmarks. Similar results are obtained with the 'change in distribution' data sets.

Figure 2 shows results for the 'change in levels' data sets, but distinguishing between standard and non-standard areas. In the standard areas, performance is similar to the overall picture in Figure 1. In the nonstandard areas, our benchmarked models have smaller MSE than the non-benchmarked model, and the raked and DGSM estimators. Coverage rates for the nonstandard areas are poor for all three versions of our model, but the model with exact benchmarking has better coverage rates than the other two. As can be seen in Figure 3, a similar pattern is found with 'change in distribution' data sets.

4.2. Estimation of Smoking Prevalence from Survey Data

In the second simulation we compare the performance of benchmarked and non-benchmarked models when estimating finite-population quantities.

Ideally, the distinction between finite-population and super-population quantities should be reflected in the benchmarking procedures, so that, for instance, agreement with $\boldsymbol{m}$ is measured using the finite-population equivalent of $\boldsymbol{\psi}$. We have not done so, on pragmatic grounds. Using super-population quantities simplifies the MCMC computations, and when

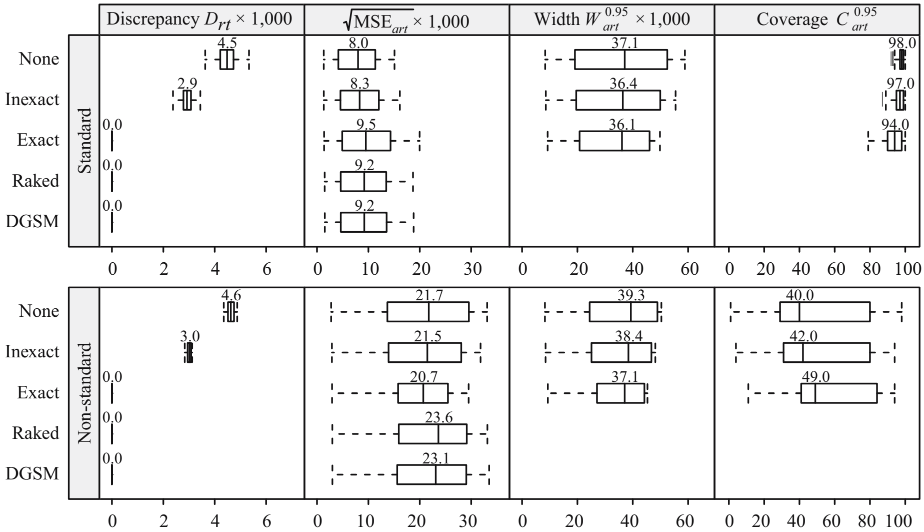


Fig. 2. Performance of model of fertility rates when applied to 'change in level' data sets, distinguishing between standard and nonstandard areas. The results are based on K = 100 replicates.

sample sizes are large, as they typically are for aggregate quantities such as $\boldsymbol{m}$ and $\boldsymbol{\psi}$, superpopulation quantities closely approximate finite-population ones.

The simulation uses artificial surveys of smoking prevalence constructed from real data from the 2013 New Zealand population census. The artificial surveys are generated by randomly drawing records from a file containing unit-level census data on the population aged 15 and over. The file contains information on age, sex, region within the country, income level, and whether the respondent currently smokes. The file excludes the 8% of people who did not answer the smoking question, leaving a total of 3.07 million records.

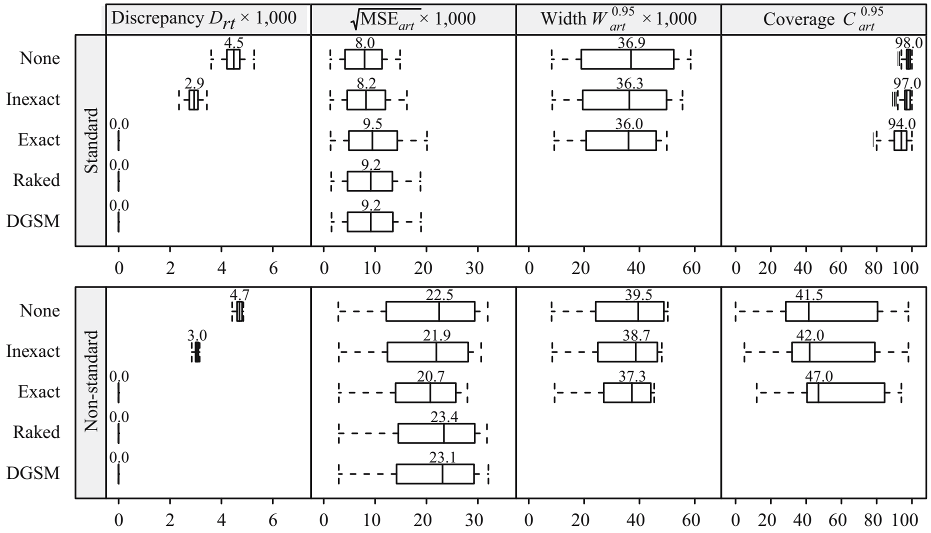


Fig. 3. Performance of model of fertility rates when applied to 'change in distribution' data sets, distinguishing between standard and nonstandard areas. The results are based on K = 100 replicates.

At each replicate of the simulation, we construct a simulated survey data set by drawing a sample of 60,000 records from the file. The sample is stratified by region, with simple random sampling within each region. There are 16 regions in total, with populations ranging from 23,000 to just over 1 million. Regional sample sizes are proportional to the square root of regional population size.

Let N_{aslr} be the number of people in age-group $a \in \{15–24, 25–34, \ldots, 55–64, 65+\}$, sex s, income level $l \in \{$No income or loss, NZD1–NZD20,000, NZD20,001–NZD40,000, NZD40,001–NZD60,000, NZD60,001–NZD100,000, NZD100,001+$\}$, and region $r \in \{1, \ldots, 16\}$. Let Y_{aslr} be the number of people who smoke. We treat the census file as the true finite population. Within the simulated analysis, N_{aslr} is known but Y_{aslr} is not. The aim of the simulated analysis is to estimate finite-population smoking prevalence by age, sex, and income, $p_{asl} = \sum_r Y_{aslr} / \sum_r N_{aslr}$.

Let y_{aslr} and n_{aslr} be the sample equivalents of Y_{aslr} and N_{aslr}. The model

$$y_{aslr} \overset{\text{ind}}{\sim} \text{Binomial}(n_{aslr}, \gamma_{aslr}) \tag{44}$$

$$\text{logit}(\gamma_{aslr}) \overset{\text{ind}}{\sim} \text{N}\Big(\beta^0 + \beta_a^{\text{age}} + \beta_s^{\text{age}} + \beta_l^{\text{income}} + \beta_r^{\text{reg}} + \beta_{al}^{\text{age:income}}, \sigma^2\Big) \tag{45}$$

is fitted to the artificial survey data. Region is included in the model to account for the stratified sample design. The age effects, income effects, region effects, and age-income interaction all have normal priors with mean 0. However, following the approach that Little (2011) suggests for statistical agencies that are reluctant to adopt informative priors, we use improper uniform priors over the set of positive real numbers for the standard deviation terms. We use an improper uniform prior for the sex effect.

The model is fitted without benchmarking, and with exact and inexact benchmarking. The benchmarks are estimated mean smoking prevalence by income level

$$m_l = \frac{\sum_{a,s,r} N_{aslr} y_{aslr} / n_{aslr}}{\sum_{a,s,r} N_{aslr}}.$$

The corresponding super-population benchmarking parameters are

$$\psi_l = \sum_{a,s,r} N_{aslr} \gamma_{aslr} / \sum_{a,s,r} N_{aslr}.$$

The finite-population equivalent of ψ_l is $\psi_l^{\text{fin}} = \sum_{a,s,r} Y_{aslr} / \sum_{a,s,r} N_{aslr}$.

Under exact benchmarking, $p^{[m \mid \psi]}(m \mid \psi)$ is given by Equation (5). Under inexact benchmarking, $p^{[\mathbf{m} \mid \psi]}(\mathbf{m} \mid \psi)$ is given by Equation (8), which becomes

$$p^{[m \mid \psi]}(m \mid \psi) \propto \exp\left(-\sum_l \frac{(m_l - \psi_l)^2}{2\lambda s_l^2} \right), \tag{46}$$

where s_l is the standard error of using m_l to estimate ψ_l. We examine the cases where $\lambda = 1$ and where $\lambda = 0.5$. The $\lambda = 1$ case allows discrepancies between model-based estimates and benchmarks to vary in line with sampling variation, while the $\lambda = 0.5$ case allows smaller discrepancies.

Benchmarks m_l and standard errors s_l are calculated using function `svymean` from *R* package `survey` (Lumley 2004). Calculating standard errors that properly account for the stratified sample design is complicated; function `svymean` uses replicate weights (Lumley 2011, 32).

Performance measures D_l, MSE_{as1}, W^q_{asl}, and C^q_{asl} are calculated for finite-population smoking prevalence p_{asl}. Here D_l is defined as

$$D_l = \text{E}\left(\frac{|\,\bar{\psi}_l^{\text{fin}} - m_l\,|}{s_l}\right), \tag{47}$$

where $\bar{\psi}_l^{\text{fin}}$ is the posterior mean of ψ_l^{fin}. This measures discrepancies in units of standard errors. To estimate p_{asl} and ψ_l^{fin}, it is necessary to estimate $Y_{asl} = \sum_r (y_{aslr} + y_{aslr}^{\text{non}})$, where y_{aslr}^{non} is the number of non-sampled people in area *aslr* who smoke. Draws from the posterior distribution of y_{aslr}^{non} can be generated using $y_{aslr}^{\text{non}(t)} \sim \text{Binomial}(N_{aslr} - n_{aslr}, \gamma_{aslr}^{(t)})$, where $\gamma_{aslr}^{(t)}$ is the *t*th draw from the posterior sample for γ_{aslr}.

As with the fertility simulation, we use $K = 100$ replicates. The Gibbs sampler is run with six independent chains, each with 100,000 iterations. Every 250th draw from the final 50,000 iterations of each chain is recorded, yielding a combined total of 1,200 draws from the posterior distribution.

The results from the simulation are summarized in Figure 4, with $q = 0.95$. Benchmarking improves agreement between the model-based estimates and the benchmarks, with exact benchmarking giving the largest improvement, and inexact benchmarking with $\lambda = 1$ the smallest. Exact benchmarking does not achieve complete agreement, since the benchmarks are applied to super-population prevalences, rather than finite population ones. However, the median absolute difference between model-based estimates and benchmarks is only 0.01 standard errors.

Benchmarking degrades overall accuracy and coverage. However, the most striking feature of the distributions of MSE_{asl}, $W^{0.95}_{asl}$, and $C^{0.95}_{asl}$ in Figure 4 is the long tails. These long tails result from a small number of outliers, notably people aged 15–24 with incomes of NZD100,000 or higher. This group has high smoking prevalence, even though youth and high incomes are, in general, associated with low prevalence. With these particular data, rather than providing robustness to outliers, benchmarking decreases robustness.

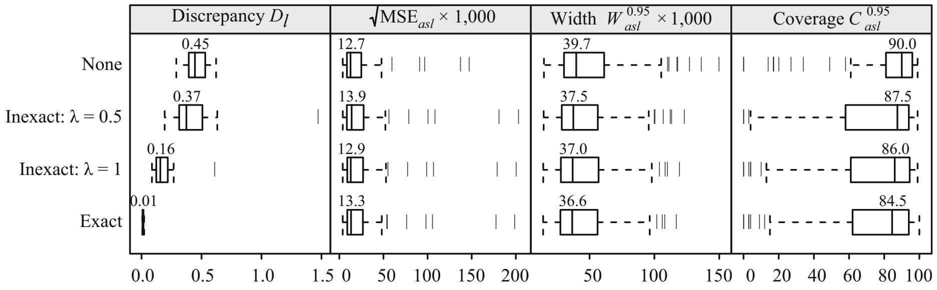


Fig. 4. Performance of model of smoking prevalence, by type of benchmarking. The results are based on K = 100 replicates.

5. Application

We now apply fully Bayesian benchmarking to a real data set. We estimate age-sex-specific mortality rates for local authority districts in England and Wales, using as benchmarks sex-specific life expectancies for regions. All code and data for the application can be obtained from github.com/bayesiandemography/britmort.

Our data are counts of deaths and populations at risk in 2014, disaggregated into 20 age groups, 2 sexes, and 348 local authority districts. The total number of deaths is 500,314, and the total population at risk is 57,408,654. The median number of deaths per area is 8, and 16% of areas have 0 deaths. Direct estimates of mortality rates (i.e., counts of deaths for each area, divided by the corresponding population at risk) for five randomly-selected districts are shown in Figure 5. Because the graphs are drawn on a log scale, estimates for which the count of deaths and direct estimate are 0 are omitted. As is apparent from the graphs, direct estimates of age-sex-specific mortality rates at the district level are unstable below age 60.

Let y_{asd} be the count of deaths for age group a, sex s and district d, and let γ_{asd} and w_{asd} be the corresponding mortality rate and population at risk. We apply the model

$$y_{asd} \overset{\text{ind}}{\sim} \text{Poisson}(w_{asd}\gamma_{asd}) \tag{48}$$

$$\log\ \gamma_{asd} \overset{\text{ind}}{\sim} \text{N}(\beta^0 + \beta_a^{\text{age}} + \beta_s^{\text{sex}} + \beta_d^{\text{dis}} + \beta_{as}^{\text{age:sex}}, \sigma^2). \tag{49}$$

Age effects are assumed to follow a random walk with drift,

$$\beta_a^{\text{age}} \sim t_4(\eta_a^{\text{age}}, \tau_{\text{age}}^2) \tag{50}$$

$$\eta_0^{\text{age}} \sim \text{N}(0, 10^2) \tag{51}$$

$$\eta_a^{\text{age}} \sim \text{N}(\eta_{a-1}^{\text{age}} + \delta_{a-1}^{\text{age}}, \omega^2), \quad a > 0 \tag{52}$$

$$\delta_0^{\text{age}} \sim \text{N}(0, 1) \tag{53}$$

$$\delta_a^{\text{age}} \sim \text{N}(\delta_{a-1}^{\text{age}}, \varphi^2), \quad a > 0. \tag{54}$$

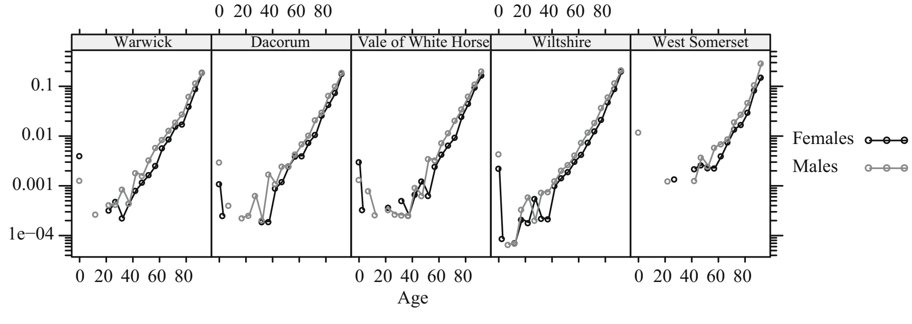


Fig. 5. Direct estimates of mortality rates, by age and sex, in five randomly-selected local authority districts in England and Wales, 2014. For some combinations of age, sex, and district, the counts of deaths, and hence direct estimates, are 0. Estimates for these combinations are not shown, since the graph is on a log scale.

The drift term accounts for the fact that log-mortality rates rise linearly over much of the age range. The use of a t_4 distribution in Equation (50) allows for occasional large departures from trend, as occurs at age 0. The sex effect has a normal prior $\beta_s^{\text{sex}} \sim \text{N}(0, 1)$. The district effect has a normal prior $\beta_d^{\text{dis}} \sim \text{N}(0, \tau_{\text{dis}}^2)$, with a weakly informative half-t prior on the standard deviation, $\tau_{\text{dis}} \sim t_7^+(0, 1)$. The interaction has a normal prior $\beta_{as}^{\text{age:sex}} \sim \text{N}(0, \tau_{\text{age:sex}}^2)$, with a weakly informative half-t prior on the standard deviation, $\tau_{\text{age:sex}} \sim t_7^+(0, 0.5^2)$. We use a smaller scale for the interaction than for the main effect on the principle that interactions are typically smaller in size than main effects (Gelman et al. 2008). Standard deviations terms σ, τ_{age}, ω and φ all have $t_7^+(0, 1)$ priors.

We benchmark the estimates to sex-specific life expectancies for regions. The region is an administrative unit further up the English geographical hierarchy than the local authority district. Counting Wales as one region, there were ten regions in England and Wales in 2014. Life expectancy is the mean number of years a newborn baby would live if prevailing mortality rates were to continue indefinitely. The procedure for calculating life expectancy is given in Preston et al. (2001, chap. 3), but for our purposes, the key point is that life expectancy is a nonlinear deterministic function of age-specific mortality rates.

Let

$$\zeta_{asr} = \sum_{d \in \Delta_r} w_{asd} \gamma_{asd} / \sum_{d \in \Delta_r} w_{asd} \tag{55}$$

be the mortality rate in age group a, sex s and region r, where Δ_r is the set of d such that district d belongs to region r. Life expectancy for sex s in region r is

$$\psi_{sr} = f_{\text{life}}(\zeta_{1sr}, \ . \ . \ . \ , \zeta_{Asr}), \tag{56}$$

where f_{life} is the nonlinear function for calculating life expectancy from age-specific mortality rates, and $A = 20$ is the number of age groups. Similarly, let

$$z_{asr} = \sum_{d \in \Delta_r} y_{asd} / \sum_{d \in \Delta_r} w_{asd} \tag{57}$$

be the direct estimate of the mortality rate in age group a, sex s, and region r. The benchmark for sex s and region r is then

$$m_{sr} = f_{\text{life}}(z_{1sr}, \ . \ . \ . \ , z_{Asr}). \tag{58}$$

Since life expectancies are ordinarily reported to at most two decimal places, most users can tolerate discrepancy of size 0.01. We specify agreement with the benchmarks as

$$m_{sr} \overset{\text{ind}}{\sim} \text{N}(\psi_{sr}, 0.005^2). \tag{59}$$

We fit our model with and without benchmarks, using four independent chains, each with 80,000 iterations. Every 100th draw from the final 40,000 iterations of each chain is recorded, yielding a combined total of 1,600 draws from the posterior distribution.

Benchmarking improves agreement between the modelled life expectancies by sex and region and the benchmarks. Figure 6 compares benchmarks with point estimates

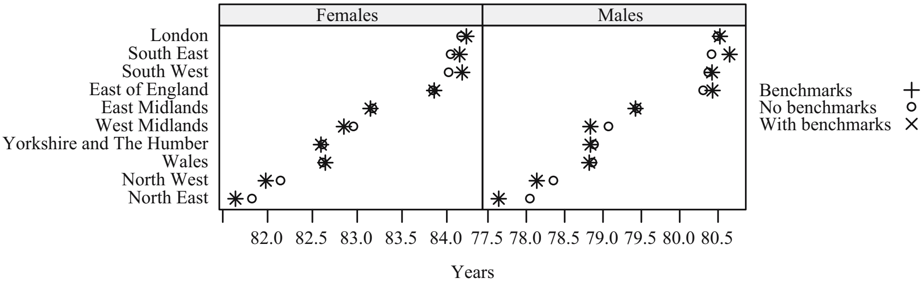
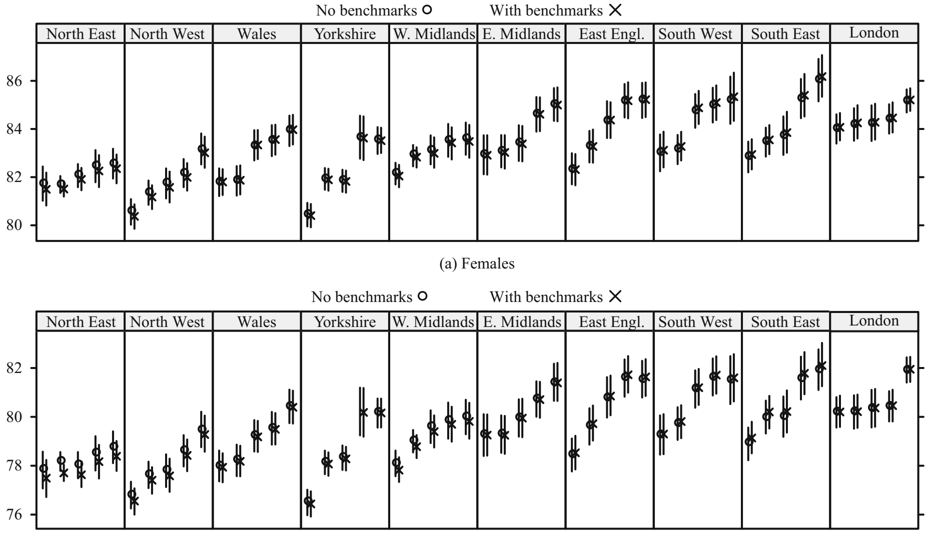


Fig. 6. Point estimates of life expectancy by sex and region: benchmarks versus posterior medians from models.

(posterior medians) from models with and without benchmarking. Without benchmarking, the model-based estimates are noticeably different from the benchmarks, especially for males. With benchmarking, the model-based estimates and benchmarks are indistinguishable.

Figure 7 shows life expectancies by sex and district, with and without benchmarking. Benchmarking shifts most posterior medians. The shifts are larger in some regions, such as the North East, than in others, such as London. Benchmarking changes the width of credible intervals, but only very slightly. The mean width of credible intervals for all age-sex-district-specific log-mortality rates increases from 0.33 to 0.34, and the mean width of credible intervals for sex-district-specific life expectancies decreases from 1.33 to 1.32 (results not shown).

Figure 8 illustrates how benchmarking affects age-sex-specific mortality rates at the district level. The percent differences between posterior medians of mortality rates from benchmarked models and those from non-benchmarked models are all below 4%.



(a) Females

(b) Males

Fig. 7. Life expectancies for 50 local authority districts, with five randomly-selected districts fromeach region. The vertical lines are 95% credible intervals.

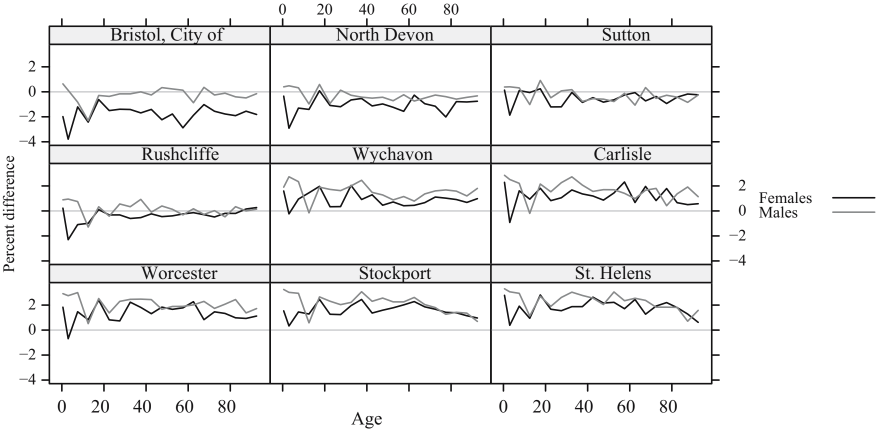


Fig. 8. Percent differences between posterior medians from benchmarked models and posterior medians from non-benchmarked models, for age-sex-specic mortality rates in nine randomly-selected local authority districts.

6. Discussion

We conclude by summarizing the advantages of the fully Bayesian benchmarking methods described in this article.

Our benchmarking methods allow full posterior distributions to be generated for a wide range of models. With full posterior distributions, uncertainty measures can be automatically produced for all unknown quantities in the small area models, as well as derived quantities, such as the finite-population smoking prevalence (ψ_l^{fin}) in the smoking simulation in Subsection 4.2, or life expectancy (ψ_{sr}) in the mortality application in Section 5.

With some applications, it is not necessary to obtain complete agreement between benchmarks and model-based estimates. Policy analysts, for instance, may have weaker requirements for agreement among estimates than administrators. In such cases, inexact benchmarking, using the methods described in this article, gives statisticians the ability to control the level of agreement with the benchmarks. As illustrated by the fertility and smoking simulations, using inexact benchmarking can achieve smaller mean squared errors than using exact benchmarking.

Finally, in some applications, the most natural benchmarks are quantities that have a non-linear relationship with the small area parameters, which can be accommodated under our approach. In this article, we consider the case of life expectancies, but other non-linear benchmarks such as growth rates and ratios can also be implemented using our methods.

7. References

Albert, I., S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, and J. Rousseau. 2012. “Combining expert opinions in prior elicitation.” *Bayesian Analysis* 7(3): 503–532. DOI: https://doi.org/10.1214/12-BA717.

Bell, W.R., G.S. Datta, and M. Ghosh. 2013. “Benchmarking small area estimators.” *Biometrika* 100(1): 189–202. DOI: https://doi.org/10.1093/biomet/ass063.

Berg, E. and W. Fuller. 2009. “A SPREE Small Area Procedure for Estimating Population Counts”. In *Proceedings of the Survey Methods Section, Statistical Society of Canada*. Section on Survey Methods, Statistical Society of Canada. Available at: http://www.ssc.ca/survey/documents/SSC2009_EBerg.pdf (accessed August 2019).

Berg, E.J., W.A. Fuller, and A.L. Erciulescu. 2012. “Benchmarked small area prediction.” In *Proceedings of the Section on Research Methods, Joint Statistical Meeting*. Section on Research Methods, Joint Statistical Meeting. Available at: http://www.asasrms.org/Proceedings/y2012/Files/305110_74288.pdf (accessed August 2019).

Datta, G.S., M. Ghosh, R. Steorts, and J. Maples. 2011. “Bayesian benchmarking with applications to small area estimation.” *TEST* 20(3): 574–588. DOI: https://doi.org/10.1007/s11749-010-0218-y.

De Waal, T. 2016. “Obtaining numerically consistent estimates from a mix of administrative data and surveys.” *Statistical Journal of the IAOS* 32(2): 231–243. DOI: https://doi.org/10.3233/SJI-150950.

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. “Micro-level estimation of poverty and inequality.” *Econometrica* 71(1): 355–364. DOI: https://doi.org/10.1111/1468-0262.00399.

Fabrizi, E., C. Giusti, N. Salvati, and N. Tzavidis. 2014. “Mapping average equivalized income using robust small area methods.” *Papers in Regional Science* 93: 685–701. DOI: https://doi.org/10.1111/pirs.12015.

Fabrizi, E., N. Salvati, and M. Pratesi. 2012. “Constrained small area estimators based on M-quantile methods.” *Journal of Official Statistics* 28(1): 89–106. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/constrained-small-area-estimators-based-on-m-quantile-methods.pdf (accessed August 2019).

Fay, R.E. and R.A. Herriot. 1979. “Estimates of income from small places: an application of James-Stein procedures to census data.” *Journal of the American Statistical Association* 74: 269–277. DOI: https://doi.org/10.1080/01621459.1979.10482505.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. 2014. *Bayesian Data Analysis*, Third Edition. New York: Chapman and Hall.

Gelman, A., A. Jakulin, M.G. Pittau, and Y.-S. Su. 2008. “A weakly informative default prior distribution for logistic and other regression models.” *The Annals of Applied Statistics* 2: 1360–1383. DOI: https://doi.org/10.1214/08-AOAS191.

Ghosh, M., T. Kubokawa, and Y. Kawakubo. 2015. “Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation.” *Biometrika* 102(3): 647–659. DOI: https://doi.org/10.1093/biomet/asv010.

Lindley, D.V. 1983. “Reconciliation of Probability Distributions.” *Operations Research* 31: 866–880. DOI: https://doi.org/10.1287/opre.31.5.866.

Lindley, D.V., A. Tversky, and R.V. Brown. 1979. “On the Reconciliation of Probability Assessments (with discussion).” *Journal of the Royal Statistical Society, Series A* 142: 146–180. DOI: https://doi.org/10.2307/2345078.

Little, R.J. 2012. “Calibrated Bayes, an alternative inferential paradigm for official statistics.” *Journal of Official Statistics* 28(3): 309. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/calibrated-bayes-an-alternative-inferential-paradigm-for-official-statistics.pdf (accessed August 2019).

Lumley, T. 2004. “Analysis of Complex Survey Samples.” *Journal of Statistical Software* 9(1): 1–19. DOI: https://doi.org/10.18637/jss.v009.i08.

Lumley, T. 2011. *Complex surveys: A guide to analysis using R,* Volume 565. John Wiley & Sons.

Morris, P.A. 1974. “Decision Analysis Expert Use.” *Management Science* 20: 1233–1241. DOI: https://doi.org/10.1287/mnsc.20.9.1233.

Morris, P.A. 1977. “Combining Expert Judgements: A Bayesian Approach.” *Management Science* 23: 679–693. DOI: https://doi.org/10.1287/mnsc.23.7.679.

Nandram, B. and H. Sayit. 2011. “A Bayesian analysis of small area probabilities under a constraint.” *Survey Methodology* 37: 137–152. Available at: www.150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11603-eng.pdf (accessed August 2019).

Nandram, B., M.C.S. Toto, and J.W. Choi. 2011. “A Bayesian benchmarking of the Scott-Smith model for small areas.” *Journal of Statistical Computation and Simulation* 81(11): 1593–1608. DOI: https://doi.org/10.1080/00949655.2010.496726.

O’Hagan, A., C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkenson, J.E. Oakley, and T. Rakow. 2006. *Eliciting Experts’ Probabilities*. John Wiley and Sons, Ltd.

Pfeffermann, D. 2013. “New important developments in small area estimation.” *Statistical Science* 28(1): 40–68. DOI: https://doi.org/10.1214/12-STS395.

Pfeffermann, D. and C.H. Barnard. 1991. “Some new estimators for small-area means with application to the assessment of farmland values.” *Journal of Business & Economic Statistics* 9(1): 73–84. DOI: https://doi.org/10.1080/07350015.1991.10509828.

Pfeffermann, D., A. Sikov, and R. Tiller. 2014. “Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation.” *TEST* 23(4): 631–666. DOI: https://doi.org/10.1007/s11749-014-0400-8.

Pfeffermann, D. and R. Tiller. 2006. “Small-area estimation with state-space models subject to benchmark constraints.” *Journal of the American Statistical Association* 101(476): 1387–1397. DOI: https://doi.org/10.1198/016214506000000591.

Poole, D. and A.E. Raftery. 2000. “Inference for deterministic simulation models: The Bayesian melding approach.” *Journal of the American Statistical Association* 95: 1244–1255. DOI: https://doi.org/10.1080/01621459.2000.10474324.

Prado, R. and M. West. 2010. *Time series: modeling, computation, and inference*. CRC Press.

Preston, S., P. Heuveline, and M. Guillot. 2001. *Demography: Modelling and Measuring Population Processes*. Oxford: Blackwell.

Ranalli, M.G., G.E. Montanari, and C. Vicarelli. 2018. “Estimation of small area counts with the benchmarking property.” *Metron* 76(3): 349–378. DOI: https://doi.org/10.1007/s40300-018-0146-2.

Rao, J.N.K. and I. Molina. 2015. *Small area estimation*, Second edition. John Wiley & Sons.

Roback, P.J. and G.H. Givens. 2001. “Supra-Bayesian pooling of priors linked by a deterministic simulation model.” *Communications in Statistics-Simulation and Computation* 30(3): 447–476. DOI: https://doi.org/10.1081/SAC-100105073.

Steorts, R.C. and M. Ghosh. 2013. "On estimation of mean squared errors of benchmarked empirical Bayes estimators." *Statistica Sinica* 23(2): 749–767. DOI: https://doi.org/10.5705/ss.2012.053.

Toto, M.C.S. and B. Nandram. 2010. "A Bayesian predictive inference for small area means incorporating covariates and sampling weights." *Journal of Statistical Planning and Inference* 140(11): 2963–2979. DOI: https://doi.org/10.1016/j.jspi.2010.03.043.

U.S. Census Bureau. 2014. "Model-based Small Area Income and Poverty Estimates (SAIPE) for School Districts, Counties, and States" (accessed July 2014).

Vesper, A.J. 2013. *Three Essays of Applied Bayesian Modeling: Financial Return Contagion, Benchmarking Small Area Estimates, and Time-Varying Dependence*. PhD thesis, Harvard University. Available at: https://dash.harvard.edu/handle/1/11124829 (accessed August 2019).

Wang, J., W.A. Fuller, and Y. Qu. 2008. "Small area estimation under a restriction." *Survey Methodology* 34(1): 29. Available at: www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10619-eng.pdf (accessed August 2019).

You, Y. and J. Rao. 2002. "A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights." *Canadian Journal of Statistics* 30(3): 431–439. DOI: https://doi.org/10.2307/3316146.

You, Y. and J. Rao. 2003. "Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights." *Journal of Statistical Planning and Inference* 111: 197–208. DOI: https://doi.org/10.1016/S0378-3758(02)00301-4.

You, Y., J. Rao, and P. Dick. 2004. "Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation." *Statistics in Transition* 6: 631–640. Available at: https://pts.stat.gov.pl/en/journals/statistics-in-transition/ (accessed February 2020).

You, Y., J. Rao, and M. Hidiroglou. 2013. "On the performance of self benchmarked small area estimators under the Fay-Herriot area level model." *Survey Methodology* 39: 217–230. Available at: www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-eng.htm (accessed August 2019).

Received November 2018
Revised May 2019
Accepted September 2019