Statistisk sentralbyrå
Statistics Norway

Flexible empirical Bayes estimation of local fertility schedules: reducing small area problems and preserving regional variation

Stefan Leknes and Sturla A. Løkken

TALL SOM FORTELLER

DISCUSSION PAPERS

953

Discussion Papers No. 953, April 2021
Statistics Norway, Research Department

Stefan Leknes and Sturla A. Løkken

Flexible empirical Bayes estimation of local fertility schedules: reducing small area problems and preserving regional variation

Abstract:

Reliable local demographic schedules are in high demand, but small area problems pose a challenge to estimation. The literature has directed little attention to the opportunities created by increased availability of high-quality geo-coded data. We propose the use of empirical Bayes methods based on a model with three hierarchical geographic levels to predict small area fertility schedules. The proposed model has a flexible specification with respect to age, which allows for detailed age heterogeneity in local fertility patterns. The model limits sampling variability in small areas, captures regional variations effectively, is robust to certain types of model misspecification, and outperforms alternative models in terms of prediction accuracy. The beneficial properties of the model are demonstrated through simulations and estimations on full-count Norwegian population data.

Keywords: small area estimation, hierarchical linear models, empirical Bayes method, shrinkage, age-specific fertility

JEL classification: J13, R58, C13, C18

Acknowledgements: We are grateful for helpful comments from Rolf Aaberge, Jacub Bijac, Nico Keilman, Marte Rønning, Terje Skjerpen, Astri Syse, Li-Chun Zhang and colleagues at the Research Department of Statistics Norway.

Address: Akersveien 26, Statistics Norway, Research Department. E-mail: sfl@ssb.no, sal@ssb.no

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no/en/forskning/discussion-papers
http://ideas.repec.org/s/ssb/dispap.html

ISSN 1892-753X (electronic)

Sammendrag

Det er stor etterspørsel etter pålitelige demografiske rater, også for mindre geografiske enheter som norske kommuner. Ratene brukes både av privat og offentlig sektor til planlegging, forskning og forretningsmessige formål. De er særlig etterspurte til beslutninger relatert til offentlig tjenestetilbud, som helse- og omsorgstjenester, skole og barnehage, samt til investeringer i infrastruktur og boligbygging.

Små geografiske områder har ofte liten befolkning som gjør det utfordrende å estimere lokale aldersspesifikke rater. Denne utfordringen faller inn under det som ofte er kalt «the small area problem» i statistiske termer. Forskningslitteraturen på dette feltet har i liten grad rettet oppmerksomheten mot mulighetene som har oppstått ved økt tilgang til rike administrative registre. Økt tilgjengelighet av rommelige data av høy kvalitet knyttet til befolkning og vitale hendelser skaper muligheter når det gjelder å fange opp lokale mønstre i demografisk atferd.

I denne artikkelen estimeres aldersspesifikke fruktbarhetsrater for små områder ved hjelp av empirisk Bayes-metode (EB). Vi finner at en modell med tre hierarkiske geografiske nivåer overgår alternative modellspesifikasjoner når det gjelder prediksjonenes treffsikkerhet. Metoden reduserer skjevheter i estimatene som stammer fra utilstrekkelig antall observasjoner i små områder, fanger opp regional variasjon på en effektiv måte og er robust overfor feilspesifikasjoner av modellen. Vi demonsterer de nyttige egenskapene til modellen gjennom Monte Carlo simulering og anvendelse på norske befolkningsdata for fertilitet.

EB-metoden er velkjent og har blitt brukt innenfor mange fagfelt. Slike modeller kan oppfattes som komplekse og virke tids- og ressurskrevende å anvende. Det kan ha forsinket mer utbredt bruk. Modellen som presenteres i denne artikkelen vil kunne hjelpe på dette ved at den er transparent, fleksibel og enkel å tallfeste. Prediksjonsresultatene er reproduserbare fra data og har en klassisk frekventistisk fortolkning. Disse egenskapene gjør at modellen er særskilt egnet for periodiske produksjonsprosesser, for eksempel beregninger av statistiske mål på dødelighet og fruktbarhet, samt befolkningsframskrivinger.

1 Introduction

Local demographic schedules are in great demand for planning, research, public policy and commercial purposes. However, obtaining reliable estimates of such schedules is often not straightforward. Even though the overall population may be large, the geographic subpopulations of interest are often small. Demarcation of data based on characteristics like sex and age curtails sample sizes further. To make things worse, demographic events are typically rare and concentrated in specific age intervals. As a consequence, random variation in demographic processes becomes prominent in small samples, which makes direct estimates noisy and unstable.[1] This is known as the small area problem and complicates identification of underlying demographic behavior.

Interest in small area estimation is one of the driving forces behind the recent upswing in statistical demography (Ahlo and Spencer, 2005). Multiple approaches have been proposed to handle small area problems, including optimized sampling design, aggregation of data over time and space, parametric modeling and indirect model-based methods. Reviews of the literature can be found in Pfeffermann (2013) and Rao and Molina (2015). Among the indirect methods, Bayesian approaches have gained in popularity, aided by increases in computing power (Bijak and Bryant, 2016).[2] Hierarchical Bayesian models have been employed with much success to deal with small area problems. They are especially advantageous for estimating many population parameters at the sub-national level (Alexander et al., 2017) and when units are similar but not identical, a trait commonly found in demography (Zhang and Bryant, 2019).

Empirical Bayes (EB) methods share these beneficial small sample properties, but differs from full Bayesian approaches in that they utilize priors that are generated directly from the data. For instance, a typical EB estimator (Gaussian-Gaussian) of local fertility rates will be a weighted mean of the local direct estimate and the global average. If the local estimate is unreliable the EB estimator will be weighted, or "shrunk", more heavily towards the global average. This curtails the over-dispersion that characterizes direct local estimates and limits the small area problems. According to Efron and Hastie (2016), the EB method exploits that a data set characterized by many parallel situations carry Bayesian information within itself.

Traditionally, indirect estimation methods have often been employed to counter the small area problems in situations where the data are partially unavailable or of low quality. Less attention has been directed at using such methods in situations where the practitioner

[1] Direct estimates refers to the traditional frequentist fraction of events relative to population.

[2] Some studies using Bayesian approaches in demography predominantly investigate patterns in data, whereas others aim at making projections. Examples are contributions on fertility, mortality and migration (Alkema and New, 2014; Alkema et al., 2012; Bijak, 2006), as well as the probabilistic population projections produced by the United Nations Population Division (Raftery et al., 2013, 2014).

possesses comprehensive high-quality data, although such data are becoming increasingly available (Poulain et al., 2013; Skinner, 2018). EB methods are well suited to exploit for instance rich administrative registers, where individuals and vital events are geo-coded, to uncover more of the geographic heterogeneity. There is convincing evidence of demographic processes displaying regional patterns (Matthews and Parker, 2013). A novel contribution is provided by Assunção et al. (2005) who use moving neighborhoods constructed from the closest geographic areas as shrinkage regions in a study of local fertility schedules in Brazil. In the study, the EB predictions borrow strength from observations that are geographically close, preserving the regional fertility patterns in the data.

The literature on area level models in demography has mostly focused on two-level hierarchies. We contribute to the literature by formulating a three-level hierarchical linear model from which we can make EB predictions of local fertility schedules. The model consists of global, intermediate, and local levels. The levels are nested such that the global mean (national) serves as a prior for intermediate level (regional) estimates, which again serve as priors for the local level (municipality) estimates. We argue that this specification is superior to alternative two-level models when there are systematic geographic differences in fertility patterns. As the global level functions as a fail-safe for small sample sizes, our proposed three-level model allows the practitioner to focus on specifying an intermediate level that captures the relevant geographic patterns. Specifically, the model allows for the extraction of both regional and local heterogeneity while avoiding unreliable estimates due to small area problems. We demonstrate that the proposed model also has other benefits over alternative two-level hierarchical models such as lower prediction bias and less overshrinkage.[3]

The performance of the model is evaluated in several ways. First, we formalize the model and discuss important statistical properties and their implications for choosing the intermediate regions. Next, we demonstrate model performance using simulated data where the true fertility rates are known. Applying an agnostic rule-based method of forming regions, we find that the three-level model consistently outperforms two-level models and traditional direct estimation methods in terms of lower mean square error. In fact, the three-level model displays a lower prediction bias in all simulations, not just on average. Finally, we provide an empirical application using data from a comprehensive administrative population register. The data quality ensures that the only non-negligible source of error in direct estimation of municipality means is sampling error originating from small population sizes. In Norway the smallest municipality has a population of about 200 persons and the median population size is just over 5 000 persons. Compared to direct estimates of the municipal fertility rates, the EB estimates are demographically plausible and reveal substantial variation in fertility level and timing of births across

[3] Overshrinkage refers to a phenomenon where between-area distribution of EB predictions is less dispersed than the true variation.

municipalities.

The rest of the paper is structured as follows. Section 2 describes the hierarchical model setup and the properties of the EB estimator. Section 3 presents a simulation exercise evaluating the performance of the model compared to alternative specifications. In Section 4, we apply our preferred model using Norwegian register data, and Section 5 provides discussion and concluding remarks.

2 Empirical Bayes strategy

The EB method was first described by Robbins (1964) and later extended to the parametric case by Morris (1983). One highly influential early application was provided by Fay and Herriot (1979), who exploited geographic hierarchies to estimate small area incomes. This type of area-level model has inspired many applications investigating a broad range of sociodemographic factors. The EB method has seen applications across many disciplines such as economics (Chetty et al., 2014; Angrist et al., 2017), epidemiology and public health (Manton et al., 1989; Marshall, 1991), and demography (Assunção et al., 2005; Schmertmann et al., 2013). In brief, the method borrows support from larger domains to produce estimates of small area statistics. Imprecise small area means will be weighted towards the larger domain mean. In a more abstract sense, EB method is useful when both the local parameters and in their distribution are of interest, e.g. the fertility rates of individual municipalities and the distribution of fertility rates across municipalities.

The connections between hierarchical linear models and EB estimators have been extensively documented, see for instance Robinson (1991). Hierarchical linear models consist of fixed and random effect components.[4] The random effect components are typically assumed to follow a Gaussian distribution.[5] The empirical estimates of the distributional moments from the hierarchical linear model, the fixed and random effects, are plugged into the EB estimator.[6]

The estimator is known to produce the empirical best linear unbiased predictors, which have favorable small area properties. Specifically, the EB method belongs to a class of shrinkage estimators that are known to outperform the maximum likelihood and ordinary least squares estimators under various mean squared error loss functions (Efron and Morris, 1973). The EB estimator shares methodology and terminology with the

[4] Hierarchical linear models are also known as mixed models, multilevel models or random effect models.

[5] Note that the random effects can be described by a range of distributions including non-parametric distributions.

[6] One important limitation of the empirical Bayes methodology is the need for a closed-form expression of the posterior distribution into which the empirical moments can be plugged. Hierarchical error structures that do not have such closed form posterior expressions can still be estimated using full Bayesian methods.

Bayesian statistics, but the predictions are completely data-driven and have frequentist interpretation (Carlin and Louis, 2008). Thus, the results are made reproducible by other practitioners by disclosing the model specification, the nesting of small areas within larger domains, and the data used.

2.1 A three-level hierarchical linear model

We propose an EB estimator based on a three-level hierarchical model. For simplicity, and for coherence with the empirical analysis later in the paper, we refer to the local "small area" geographic units as municipalities. The intermediate and global levels are denoted regions and country, respectively. Municipalities are nested within regions, which again are nested within the country. In such a setting, the EB estimator will borrow strength from both regional and national means, especially if the local estimates are unreliable. To fix ideas, we define the hierarchical linear model as follows:

$$Y_i = \boldsymbol{\theta} \boldsymbol{A}_i + \boldsymbol{\theta}_{r(i)} \boldsymbol{A}_i + \boldsymbol{\theta}_{j(i)} \boldsymbol{A}_i + \epsilon_i \tag{1}$$

$$\epsilon_i | \boldsymbol{\theta}_r, \boldsymbol{\theta}_j \sim N(0, \sigma^2_\epsilon) \tag{2}$$

where Y_i is a binary outcome describing whether woman i in municipality j and region r gives birth to a child or not. $\boldsymbol{A}_i$ is a vector of age indicators ranging over the fertile years, defined as ages 15-49.[7] The fixed part of the model, $\boldsymbol{\theta}$, is the national age-specific fertility rate. $\boldsymbol{\theta}_r$ is a vector of regional level random age effects, while $\boldsymbol{\theta}_j$ is a vector of municipality-level random age effects. The regional and municipal age-specific random effects ($\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}_j$) are both assumed to be normally distributed with no covariance across age groups:

$$\boldsymbol{\theta}_r \sim N(\boldsymbol{0}, \boldsymbol{\Omega}_r) \tag{3}$$

$$\boldsymbol{\theta}_j | \boldsymbol{\theta}_r \sim N(\boldsymbol{0}, \boldsymbol{\Omega}_j) \tag{4}$$

where $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_j$ are diagonal matrices representing the regional and municipal variance of the age-specific fertility rates, respectively.[8] Assumptions (3) and (4) characterize how regional age-specific fertility rates deviate from the national age-specific fertility rate and how municipal age-specific fertility rates deviate from regional age-specific fertility rates. This is a very flexible model specification in the sense that it decomposes the variation within each geographic level for each age group.

[7] Each vector have dimensions equal to the number of age groups between 15–35, 1×35.

[8] As we do not allow for covariance across age groups, $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_j$ will be diagonal matrices with dimensions equal to the number of age groups, 36×36.

2.2 Properties of the empirical Bayes estimator

The EB estimator can be expressed as the weighted sum of the means for each level of the hierarchical model.[9] For the sake of simplicity, we review the EB estimator defined for a single age group. Taking Equation (1) as our point of departure, the model can be rewritten as:

$$Y_i = \theta + \theta_{r(i)} + \theta_{j(i)} + \epsilon_i \tag{5}$$

where θ is the fixed effect or *grand mean* of the age fertility level. θ_r and θ_j are the regional and municipality-level random effects assumed to be independent and following Gaussian distributions with zero mean and the variances σ_r^2 and σ_j^2, respectively. The disturbance term, ϵ_i, is assumed to have the same properties with the variance σ_ϵ^2. The index $i = 1, ..., n$ denotes the individual women up to the population total n. The number of women within municipality j is denoted n_j, and within region r is denoted n_r. The index $j = 1, ..., J$ denotes the municipalities and the index $r = 1, ..., R$ denotes the regions.

The weights given to the mean of each geographic level in the EB estimator are determined by reliability factors. Following Raudenbush and Bryk (2002), we express these as:

$$\lambda_j = \frac{\sigma_j^2}{\sigma_j^2 + {}^{\sigma_\epsilon^2}/_{n_j}} \tag{6}$$

$$\lambda_r = \frac{\sigma_r^2}{\sigma_r^2 + \left\{ \sum\limits_{j \in r} \left[\sigma_j^2 + {}^{\sigma_\epsilon^2}/_{n_j} \right]^{-1} \right\}^{-1}} \tag{7}$$

The regional reliability factor λ_r measures the weight given to the regional mean relative to the national *grand mean* for the regional level EB estimator θ_r^{EB}, while the local reliability factor λ_j measures the weight given to the local mean relative to the regional EB estimator for the local EB estimator θ_j^{EB}. By plugging the empirical estimates of the hyperparameters, the estimated variances at each level $\hat{\sigma}_r^2$, $\hat{\sigma}_j^2$ and $\hat{\sigma}_\epsilon^2$, into Equations (6) and (7), we can express the EB estimators as:

$$\hat{\theta}_r^{EB} = \hat{\lambda}_r \hat{\theta}_r + (1 - \hat{\lambda}_r)\bar{y} \tag{8}$$

$$\hat{\theta}_j^{EB} = \hat{\lambda}_j \bar{y}_j + (1 - \hat{\lambda}_j)\hat{\theta}_r^{EB} \tag{9}$$

where the regional mean is a weighted combination of municipal means, $\hat{\theta}_r = (\sum\limits_j \hat{\omega}^{-1}\bar{y}_j)/(\sum\limits_j \hat{\omega}^{-1})$, with the estimated weights: $\hat{\omega} = \hat{\sigma}_j^2 + {}^{\hat{\sigma}_\epsilon^2}/_{n_j}$. The empirical estimate of the *grand mean*

[9] See Appendix A for a formal derivation of the general two-level case.

equals the overall sample mean of the outcome ($\hat{\theta} = \bar{y}$). Small (large) municipalities are generally weighted somewhat higher (lower) than when population weights are used. However, the regional mean will approach the population weighted mean if there is little variation at municipality level ($\hat{\sigma}_j^2$ is small) and there is much unexplained variation ($\hat{\sigma}_e^2$ is large).

By plugging Equation (8) into Equation (9), the EB estimator can be reformulated as a weighted sum of empirical estimates of the hierarchy means, weighted by functions of the estimated hierarchy variances:

$$\hat{\theta}_j^{EB} = \underbrace{\hat{\lambda}_j}_{w_j} \bar{y}_j + \underbrace{(1-\hat{\lambda}_j)\hat{\lambda}_r}_{w_r}\hat{\theta}_r + \underbrace{(1-\hat{\lambda}_j)(1-\hat{\lambda}_r)}_{w_c}\bar{y} \tag{10}$$

The local average weight w_j is equal to the local reliability factor $\hat{\lambda}_j$, the regional average weight, w_r, is given as the product of the local unreliability factor $(1-\hat{\lambda}_j)$ and the regional reliability factor $\hat{\lambda}_r$, and the residual (grand) mean weight $w_c = 1-w_j-w_r$ is the product of the local unreliability factor $(1-\hat{\lambda}_j)$ and the regional unreliability factor $(1-\hat{\lambda}_r)$. These sets of weights will vary depending on municipal and regional characteristics and will sum to unity for each municipality.

The mechanics of the framework are revealed by means of counterfactual manipulation of sizes of population and hyperparameters. Suppose we increase the population size of one municipality j' assuming the effect on the estimated hyperparameters ($\hat{\sigma}_r^2$, $\hat{\sigma}_j^2$, $\hat{\sigma}_\epsilon^2$) is second order and fixed. Then the local and regional reliability factors from Equations (6) and (7) would both increase. However, since the population of all other municipalities remains fixed, the regional reliability factor would increase less than the local reliability factor ($\frac{\partial \lambda_{j'}}{\partial n_{j'}} > \frac{\partial \lambda_r}{\partial n_{j'}} > 0$). The weight given to the local mean in Equation (10) will increase ($\frac{\partial w_1}{\partial n_{j'}} > 0$), the weight on the national mean will decrease ($\frac{\partial w_3}{\partial n_{j'}} < 0$) and the regional level weight will decrease in most cases but can theoretically go in either direction ($\frac{\partial w_2}{\partial n_{j'}} \lesseqgtr 0$).[10]

Next, we investigate what happens to the estimated model if the variation at one of the geographic levels is negligible (i.e. estimated hyperparameters are close to zero). Little variation in means across regions ($\hat{\sigma}_r^2$ close to zero) collapses the model to a two-level country-municipality hierarchical model, as the regional reliability factor ($\hat{\lambda}_r$) and the regional weight (w_2) approach zero. Correspondingly, if there is slight variation across municipalities conditional on the regional distribution ($\hat{\sigma}_j^2$ close to zero) the three-level model reverts to a two-level country-region hierarchical model, as the municipality reliability factor ($\hat{\lambda}_j$) and the municipal weight (w_1) approach zero. Furthermore, if there

[10] The derivative of the second weight $\frac{\partial w_2}{\partial n_{j'}}$ will almost always be negative. Only if municipality j' has a small reliability factor and a large population relative to the other municipalities in the region can the derivative be positive, which is highly unlikely.

is little residual variation left after taking out group means ($\hat{\sigma}_\epsilon^2$ close to zero), all variation is explained by the municipality level and the three-level model reverts to maximum likelihood estimation of the local means, as $\hat{\lambda}_j$ approaches one. The opposite case, where the variation that is unexplained by the model is substantial (large $\hat{\sigma}_\epsilon^2$) will reduce the regional and municipality-level reliability factors and increase the weight placed on the grand mean $\bar{y}$.

The formalized model can provide insights concerning the specification of the regional level. Thus, defining the regional level will entail a trade-off between number of regions R and region population size n_r. A favorable constellation has both a number of regions sufficient to provide a precise estimate of σ_r^2 and a population size within each region sufficient for precise estimation of the regional means, θ_r.

The optimal number of regions depends on the phenomenon under study and the available data. Kreft and de Leeuw (1998) argue that the number of regions should be at least 20, but having fewer groups typically leads to underestimation of the regional variation, σ_r^2. This will downplay the contribution of the regional level, as it reduces the weight placed on the regional means. Obviously there are no hard and fast rules, and specifying too few (R close to 0) or too many (R close to J) regions in our model will produce results close to those of a two-level model without the regional level. Also, the regional EB estimates will shrink towards the national *grand mean* if the regional group size, n_r, is small and the regional means are unreliable. Compared to the two-level model, these traits of the three-level model provide the practitioner with a relatively large degree of freedom for specifying the regional level. She can focus on specifying enough regions R to capture systematic regional heterogeneity and precisely estimate the hyperparameter σ_r^2 without worrying too much about sampling noise at the regional level.

A criticism of EB methods is that the between-area dispersion of the predictions tend to be smaller than the real dispersion ($Var(\hat{\theta}_j^{EB}) < Var(\theta_j)$). Such underestimation of the variation is referred to as overshrinkage (Spjøtvoll and Thomsen, 1987; Zhang, 2003; Rao and Molina, 2015).[11] By utilizing the simulation framework described in Section 3, we can compare the distributional properties of different estimators. We demonstrate that the issue of overshrinkage is substantially reduced by using EB predictions from a three-level hierarchical linear model compared to a more traditional two-level model. For more details about overshrinkage properties see Appendix B.

[11] Analogously, direct estimates of small areas characteristics suffer from undershrinkage as sampling noise typically will increase the dispersion of the estimates.

2.3 Regional level specification

Selecting the appropriate model hierarchy is rarely trivial. By imposing a global prior we ensure that no local prediction lacks statistical support. However, many demographic outcomes have been found to have strong regional variations, for instance, studies from Norway show that hospital catchment areas and labor market conditions affect mortality and fertility decisions (Kravdal, 2002; Godøy and Huitfeldt, 2020). Hence, we aim to improve the local predictions by also including a regional level. This is supported by the hierarchical linear model literature, where several papers have found that ignoring a relevant level in the hierarchy can bias variance components and standard errors (Hutchison and Healy, 2001; Moerbeek, 2004; Opdenakker and van Damme, 2000; van Landeghem et al., 2005).[12]

In practice, we can take several different approaches to aggregating local units into a regional level. First, areas can be grouped using statistical criteria for clustering — minimizing variation within clusters and maximizing variation across clusters. Common implementations are iterative algorithmic methods from the machine learning toolkit, for instance tree-based methods and clustering algorithms (James et al., 2013). Second, areas can be grouped on the basis of commonality criteria, for instance related to adjacency, similar population size or sociodemographic characteristics like education level, income, and immigrant shares.

Third, regions can be based on groups of municipalities that belong to the same administrative, legal, or functional unit. Examples are counties, hospital catchment areas, local labor markets and areas with a common cultural history. Fourth, an arbitrary regional subdivision such as a grid can be used.[13] As long as there is systematic geographic variation in the outcome, a sufficient number of regions from clustered municipalities would capture a reasonable proportion of such variation and improve the accuracy of the predictions.

The gains from including a regional level depend on how well it explains the variation in the outcome, which can be tested. Consider regressing two alternative specifications on the outcome of interest using ordinary least squares. The first specification controls for the fixed effects part of the model (age-specific dummies) while the second specification controls for the fixed effects interacted with regions (regional age-specific dummies). If the regional interactions substantially increase the explained variance, R-squared, this indicates that the inclusion of an intermediate regional level will also improve the model fit and EB predictions. In the following, we will demonstrate how such an evaluation

[12] Obtaining correct standard errors is not a major concern in this paper as we are mainly interested in the predictions of local means.

[13] Such an approach is typically not considered because of the data requirements, but increased availability of detailed geo-coded data may increase usage of flexible spatial methods in the future.

method can contribute to the evaluation and specification of the regional level.

3 Simulation study

Using Monte Carlo simulations, we provide evidence of the benefits of our three-level model in estimating local age-specific fertility rates (ASFRs). We compare the predictions from the three-level model with those from more standard two-level models and direct estimation. We use the same nested geographic set-up as previously with "small area" municipalities, regions at intermediate level and the whole country at the global level.

First, we define a geographic plane with coordinates $(x, y) \in [0, 1]$ and, from uniform distributions, draw the positions of 400 municipalities. To construct intermediate regions, the plane is divided into 64 squares of equal size. The number of municipalities within each region depends on the draw of municipality coordinates. On average, each region will house 6.5 municipalities.

Second, we allocate unique fertility schedules to each municipality, determined by draws of three distributional characteristics of the ASFRs: η_j is the total fertility rate, *TFR* (the sum of the age specific fertility rates), μ_j is the age with the highest fertility rate denoted *peak fertility, and* ρ_j is the *fertility spread* given by the standard deviation of the fertility schedule. Each of these characteristics consists of a systematic component (s) that changes with geography and an idiosyncratic municipality-specific component (m):

$$\eta_j = \alpha_\eta + \eta_j^s + \eta_j^m, \quad \eta_j^m \sim N(0, \sigma_\eta) \tag{11}$$

$$\mu_j = \alpha_\mu + \mu_j^s + \mu_j^m, \quad \mu_j^m \sim N(0, \sigma_\mu) \tag{12}$$

$$\rho_j = \alpha_\rho + \rho_j^s + \rho_j^m, \quad \rho_j^m \sim N(0, \sigma_\rho) \tag{13}$$

The intercept parameters represent the national average of each fertility component and are set at realistic values, $(\alpha_\eta, \alpha_\mu, \alpha_\rho) = (2, 30, 3.5)$. The vectors of idiosyncratic municipality-specific components $(\eta_j^m, \mu_j^m, \rho_j^m)$ are randomly drawn from independent normal distributions with zero expectation and the following standard deviations: $(\sigma_\eta, \sigma_\mu, \sigma_\rho) = (0.1, 0.3, 0.1)$.[14]

Systematic geographic variation in fertility patterns is introduced by allowing *TFR, peak fertility* and *fertility spread* to vary non-linearly along the x and y coordinates. For each of the three characteristics, we draw five coefficients that determine how the fertility characteristics vary along coordinate polynomials. The coefficients are randomly drawn from uniform distributions with fixed intervals:

[14] While it may seem more realistic also to model the covariance between the fertility characteristics, it does not influence the results and will complicate the description of the data-generating process.

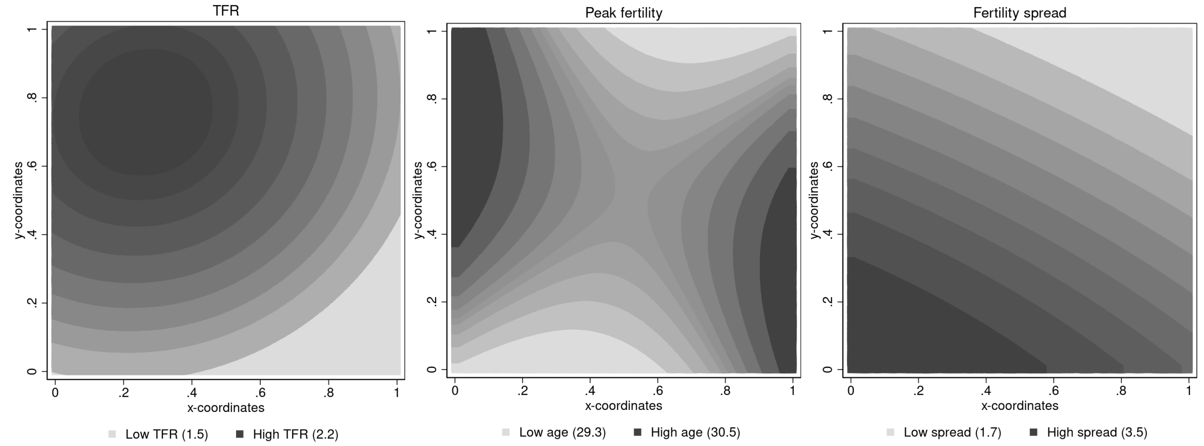

Figure 1: Simulated geographic distribution of fertility characteristics

Note: The figure shows the geographic variation of the three fertility characteristics from a simulated data set. The left-hand panel shows the geographic distribution of total fertility rate generated by Equation (14). The middle panel shows the geographic distribution of the peak fertility age as generated by Equation (15). The right-hand panel shows the geographic distribution of the fertility age spread as generated by Equation (16).

$$\eta_j^s = e_x^\eta x + e_y^\eta y + e_{xy}^\eta xy + e_{xx}^\eta x^2 + e_{yy}^\eta y^2, \quad e_k^\eta \sim U(-1,1) \tag{14}$$

$$\mu_j^s = e_x^\mu x + e_y^\mu y + e_{xy}^\mu xy + e_{xx}^\mu x^2 + e_{yy}^\mu y^2, \quad e_k^\mu \sim U(-3,3) \tag{15}$$

$$\rho_j^s = e_x^\rho x + e_y^\rho y + e_{xy}^\rho xy + e_{xx}^\rho x^2 + e_{yy}^\rho y^2, \quad e_k^\rho \sim U(-1,1) \tag{16}$$

where $k = (x, y, xx, yy, yx)$.

Note that the data-generating process does not impose the hierarchical structure of the model specification. Specifically, the choice of levels or regional subdivisions does not affect the simulated data and therefore should not influence the performance of the models we evaluate. Figure 1 illustrates the type of systematic geographic variation that is generated by Equations (14)–(16).

We generate age-specific fertility rates for each municipality by plugging the fertility characteristics generated by Equations (11)–(13) into a normal density function.[15] The normal distribution is centered around μ_j, has standard deviation ρ_j, and is scaled by the total fertility rate η_j. This is formalized in the following equation:

$$ASFR_j(age; \mu_j, \rho_j, \eta_j) = \eta_j \frac{1}{\rho_j\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{age-\mu_j}{\rho_j}\right)^2}, \quad age \in [15, 45] \tag{17}$$

Figure 2 shows the distribution of municipal fertility schedules produced by Equation (17) in a single simulation run.

[15] Other parametric functions may characterize age-specific fertility rates more precisely, but for our

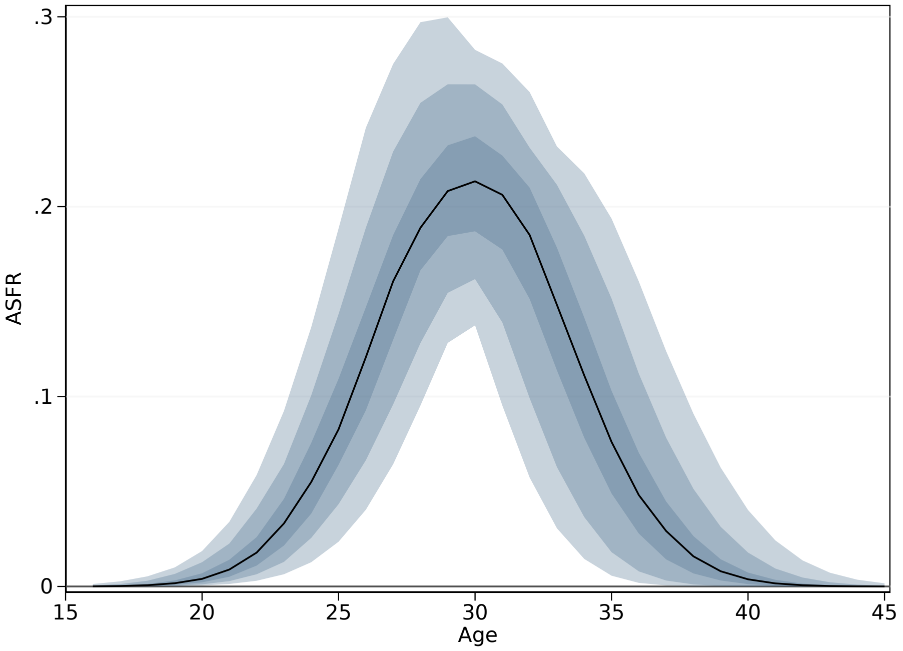

Figure 2: Simulated age-specific fertility rates

Note: This figure show the distributions of municipality-specific fertility rates by age from one draw of the simulation procedure. The shaded areas represent the 99/90/50-percent prediction interval at each age and the solid line represents the median age-specific fertility rate.

The next step is to populate the municipalities by drawing the number of fertile women in the age interval 15-45 in each municipality. For the sake of simplicity, we assume that within municipalities each one-year age group has the same number of women. To set the number, we draw uniformly an integer value in the range 1–50, leaving each municipality with between 31–1 550 women. For each individual (i), we use the municipality-level age-specific fertility rates to draw the binary random outcome of birthing a child ($child_i = \mathbb{1}[ASFR_j(age_i) > x_i], \quad x_i \sim U(0,1)$).

Finally, we fit three separate hierarchical models to the data. Our main model is the three-level model (L3) using the country-region-municipality hierarchy outlined in Section 2. We also fit two two-level models, L2C and L2R, where the top level of the hierarchy consists of the country and regions, respectively. We conduct 1 000 simulations. After each run of the simulation, we calculate the root mean squared error (RMSE) for the predicted values of each model. The RMSE measures the average difference between the predicted and the true age-specific fertility rates across all municipalities and age groups. In other words, it captures the average bias of the models.

simulation the normal density function will suffice.

Table 1: Prediction bias measured by root mean square error

	Model specifications			
RMSE (×100)	L3	L2R	L2C	Direct
Mean	1.63	2.05	2.21	5.77
Std. Dev.	0.22	0.16	0.50	0.41
Min	0.99	1.55	1.01	4.21
Max	2.38	2.60	4.21	7.18
Simulations:	1 000	1 000	1 000	1 000
Municipalities:	400	400	400	400

Note: Statistics for RMSE (×100) are based on 1 000 simulations with 400 municipalities. L3 is a three-level model with levels at municipality, regional and country level. The regional two-level model (L2R) and the country two-level model (L2C) have municipality as the local level and either region or country as the global level. The average total population across the simulations was 328 973 individuals.

3.1 Simulation results

Table 1 shows RMSE statistics for all models, based on 1 000 simulations. The predictions of the three-level model (L3) consistently outperform all the other models in terms of root mean square error. The average RMSE of the direct estimator is 354 percent higher than that of L3, illustrating the need to consider sampling variability due to small area problems.[16] The average RMSE of the regional two-level model (L2R) and the country two-level model (L2C) predictions are 26 and 36 percent, respectively, higher than those of L3. This means the EB predictions of the three-level model have the lowest average bias of all models. They also have the lowest minimum bias and the lowest maximum bias across all simulations.

However, these average comparisons obscure two important results. First, the three-level model predictions outperform those of the two-level models across all simulations. Second, under certain conditions the predictions of the two-level models can be severely biased relative to the three-level model. Figure 3 shows the distribution of the bias from both two-level models relative to the bias of the three-level model.[17] The relative bias is a ratio calculated as the RMSE of the two-level models divided by the RMSE of the three-level model. The distribution of the relative bias of the L2R-model is more left-skewed than the L2C-model, indicating that the L2R-model is typically less biased. Compared to the L2C-model, the L2R-model has a relative bias distribution with a fatter right tail, which means this model has a higher risk of severely biased results. Over 1 000 simulations, the average relative biases of the L2R-model and the L2C-model are 1.28 and 1.34, respectively.

The benefits of including a regional level will depend on the overall geographic variation. If

[16] The direct estimator is given by *number of births/number of females* for each age and municipality combination. This is an asymptotically unbiased estimator of fertility rates.

[17] We leave out any comparisons with the direct estimator as these have so large RMSE that the results obscures any nuances between the hierarchical linear models.

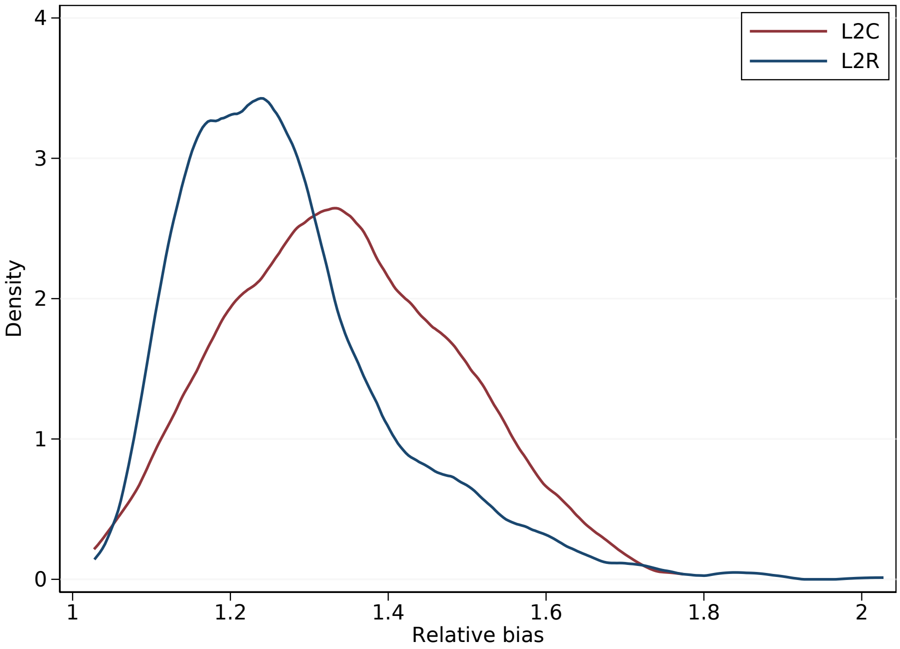

Figure 3: Distribution of relative bias

Note: The figure shows the distribution of relative biases for the two-level models compared to the three-level model. For each simulation, the relative bias is calculated as the ratio of the RMSE of each two-level model relative to the RMSE of the three-level model. Thus, values higher than 1 means the model results are more biased than the three-level model.

the regional variation is sizable, it may be optimal to increase the number of intermediate regions to capture this heterogeneity. Conversely, if the geographic variation is minor, there are concerns that a high number of regions may pick up mostly statistical noise, which may bias the model predictions. As a measure of the underlying regional variation, we propose to calculate an explanatory power ratio using R-squared from two separate regressions. We calculate R^2_C by regressing childbirth on age dummies at country level and R^2_R by regressing childbirth on interactions between age and regional dummies. We then calculate the ratio, $\varphi = {}^{R^2_R}/_{R^2_C}$, which may indicate the relative gain achieved by adding a regional level. A φ close to unity indicates that the potential gains from including a regional level are minor. A φ larger than unity indicates that the regional level might be beneficial in modeling local fertility schedules.

Figure 4 shows the relative bias of the two-level models compared to the three-level model and how the biases change with the explanatory power ratio (φ). As expected, we find that the relative bias of the L2C-model increases with the level of regional variation (φ), while the opposite is the case for the L2R-model. Most importantly, we find that the three-level hierarchical linear model has lower mean bias than both two-level models no matter the level of regional variance, suggesting that the three-level model is robust to misspecification of the regional level.[18]

[18] Each simulation run produces a different number of municipalities within each region and different population sizes in these municipalities. In Appendix B, we compare relative bias along these character-

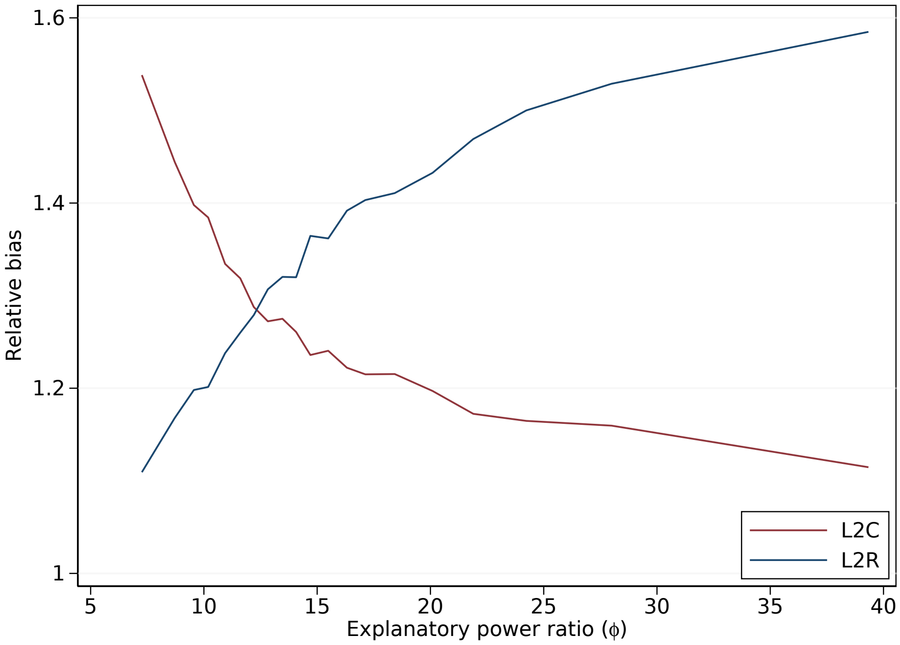

Figure 4: Relative bias and regional variation

Note: The figure shows how the relative bias of the two-level models is affected by overall regional variation as measured by the explanatory power ratio, φ. A relative bias of 1 means that the model prediction is the same as the bias of the three-level model and values higher than 1 mean that the predictions from the two-level model are more biased than those from the three-level model. The figure is produced by sorting simulations by relative bias into 20 equal-sized bins and plotting the average relative bias within each bin.

4 Application to Norwegian municipalities

In the following, we apply our model to individual-level Norwegian administrative data to estimate age-specific fertility rates for municipalities. The municipalities are administrative units responsible for a range of public services, like nurseries and kindergartens, primary and lower secondary school, primary healthcare and social services, and local area planning and roads. The provision needs to be planned years in advance and scaled to meet future demand. As such, reliable demographic schedules that can help inform such decisions are valued and highly demanded by local governments and policy-makers.

Norway comprises 356 municipalities, which vary widely in population size. The first panel of Table 2 shows that while the mean municipality has almost 15 000 inhabitants, the largest municipality, Oslo, has about 680 000 inhabitants and the smallest, the island community of Utsira, has less than 200 inhabitants. The median municipal population is about 4 600 inhabitants. Sample sizes tend to be small for Norwegian municipalities; for instance, the municipality at the 50th percentile has just over 27 females aged 30, which renders estimation of age-specific demographic rates fraught with small area problems. As a hypothetical experiment, let us assume that the sample size is fixed at 27 and the

istics, and demonstrate that EB estimates based on the three-level model suffer substantially less from over-shrinkage issues than the EB estimates from the two-level models.

women have a true fertility rate of 0.11, i.e. they are expected to give birth to a total of three children a year in this particular municipality. In a random draw, these women will only give birth to 3 children in 24 percent of the cases. In 4.5 percent of the cases they will give birth to no children, in 14 percent of the cases to one child, and in 23.5 percent of the cases to two children. The small sample size means that the estimated fertility rate will fluctuate wildly; and in this case, the sample estimate will be either 50 percent larger or smaller than the underlying rate in more than 35 percent of the cases.

As in most developed countries, Norway has experienced a fall in fertility over time. This has been especially pronounced after paid work for women and contraception became more common in the 1970s. Since the mid 70s, TFR has fluctuated appreciably but has remained below 2 children per woman. In recent years, it has been falling continuously from its high point of 1.98 in 2009 and is now at the lowest level ever measured for Norway, 1.53 in 2019. In the same period, the average age of giving birth has increased steadily.[19] There is substantial geographic variation in fertility in Norway. Typically, fertility has been high in the south-west of the country, whereas the south-eastern part of the country had low fertility. In 2019, the maximum difference in TFR across the eleven Norwegian counties was 0.25. Substantial differences across smaller geographic units have also been documented by Leknes and Løkken (2020).

With direct estimation approaches, aggregation of data across age groups and/or time is necessary to obtain stable small area estimates of fertility. In comparison, the EB method relies on parallel sets of similar observations which reduce reliance on longer data panels and preserve age-specific heterogeneity.[20]This is particularly useful in a setting where fertility levels and birth age of mothers are changing rapidly, as is currently the case in Norway and many other Western countries.

4.1 Data and regions

Norwegian full-count population data are available from an administrative register (Folkeregisteret). The data represent the de jure population in each municipality. The administrative register is comprehensive and missing observations and measurement error are minimal. We can therefore focus on extracting local heterogeneity in demographic arrays in a setting where the lack of statistical support is attributable solely to insufficient population scale. Our analysis was conducted on a 2019 sample of women aged 15 to 49 with information on whether they gave birth or not.

[19] These two processes are connected, as the total fertility rate is sensitive to changes in the timing of births.

[20] Administrative borders are frequently changed or adjusted, for instance because of municipal amalgamation or regional policy reforms. This further will limit the availability and quality of population panel data sets.

Table 2: Summary statistics of population and births in Norway, 2019

		Municipality	Region	Country
Population:				
	Mean	14 967	57 293	5 328 212
	Min	196	7 878	-
	Max	681 071	681 071	-
Women (15-49):				
	Mean	4 685	17 936	1 668 024
	Min	53	2 284	-
	Max	236 108	236 108	-
Births:				
	Mean	153	586	54 495
	Min	2	73	-
	Max	9 343	9 343	-
	N	356	93	1

Summary statistics are based on the Norwegian population register for the year 2019 and all statistics are rounded to the closest integer value.

Official economic regions form the basis for the intermediate regional level. These 89 economic regions consist of travel-to-work areas derived from commuting intensities across municipalities and correspond to the EU NUTS-4 level (Hustoft et al., 1999). To take into account the fertility differences between urban and sub-urban areas, the largest urban municipalities are specified as separate regions, leaving us with 93 distinct geographic subdivisions to be used in the analysis. As shown in Table 2, the regions vary in population size from about 7800 to 681 000 inhabitants, while the number of women of fertile age varies from about 2300 to 236 100. The regions with the fewest number of females are quite small. However, the three-level model takes account of this margin of freedom since a noisy estimate of the regional average will shrink towards the national mean. The intermediate level can therefore be specified from objective commonality criteria, i.e. groupings that for instance make sense from a geographic or administrative perspective.

To evaluate the gain from adding a regional level, we estimate the explanatory variation ratio φ - the relative increase in R-squared due to going from a regression with age dummies to an extended model in which age is interacted with regions. In the 2019 data we find a φ of 1.18 which means R-squared increased by more than 18 percent as a result of including regional information. Drawing on the lessons from the simulation exercise, the results in Figure 4 indicate a scenario where a three-level model setup substantially outperforms both types of two-level models (and direct estimates) in terms of low prediction bias.

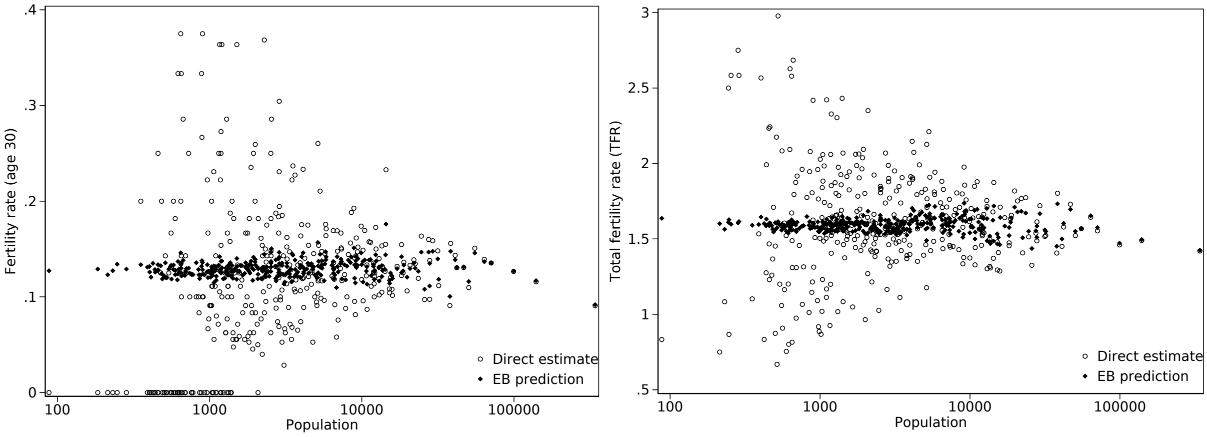

Figure 5: Fertility rates for municipalities of different sizes. Comparison of direct estimates and empirical Bayes predictions

Note: The figure shows differences in fertility rates at age 30 (left-hand panel) and TFR (right-hand panel) across municipalities of different population sizes. Using administrative registry data from 2019, age-specific fertility rates are derived using two different methods: direct estimates, calculated as the number of births relative to the female population, and EB predictions. Five municipalities with direct estimates of fertility rates at age 30 higher than 0.4 are excluded from the right-hand panel. Three of these municipalities have fertility rates equal to one. In 53 municipalities the direct fertility rate estimates are equal to zero. Two municipalities with TFR below 0.5 and two municipalities with a TFR above three are excluded from the left-hand panel.

4.2 Empirical results

In Figure 5, we compare the EB predictions of municipal fertility schedules with direct estimates, calculated as the number of births divided by the number of women in each age category, across municipalities of different population sizes. The left-hand panel shows the distribution of age-specific fertility rates at age 30. For small municipalities, the rates derived from direct estimation are often extreme and demographically implausible and range from zero to 100 percent. The dispersion of these rates decreases with population size, as statistical support increases and sampling variability becomes less prominent. The EB predictions display less dispersion, with fertility rates ranging from 9.2 to 17.6 percent, and do not exhibit the same funnel shape with respect to population size as the direct estimates.

The right-hand panel shows the corresponding TFRs with both methods. TFR is a more robust measure than ASFR, as sampling errors in opposite directions offset one another when the ASFRs are totaled. Nevertheless, the dispersion is much larger for the TFRs calculated from direct estimates than for those based on EB predictions, and again small municipalities are more strongly affected. TFRs based on direct estimation range from a little under 0.4 up till 4 children per woman, while the TFRs based on the EB predictions are distributed between 1.4 and 1.75 children per woman. The variation is about ten times lower with the EB method.

Figure 6 shows the distribution of all EB predictions of ASFRs across municipalities. The

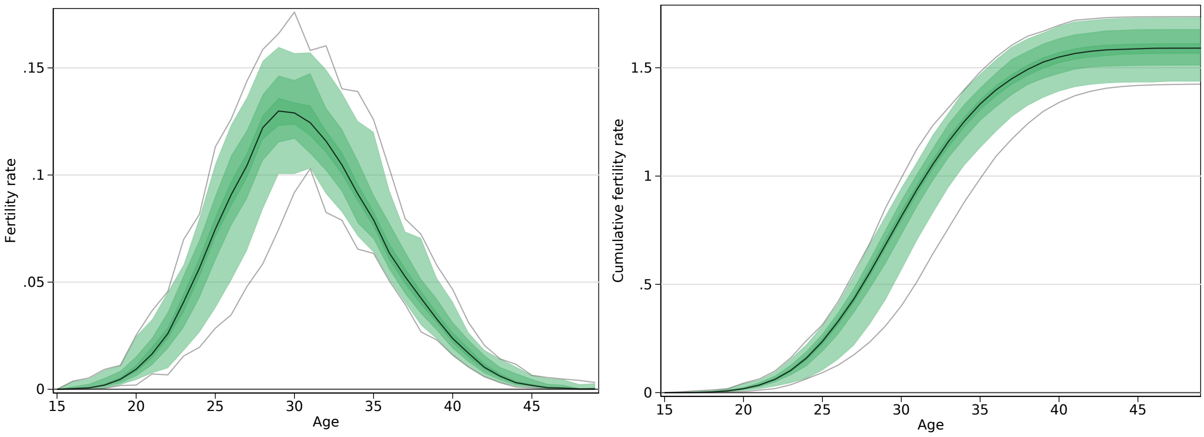

Figure 6: Distribution of empirical Bayes predictions of age-specific fertility rates across municipalities

Note: The figure shows ASFR (left) and cumulative fertility (right) across municipalities by age. Fertility rates are EB predictions from a three-level hierarchical linear model estimated on data for 2019. The shaded areas (from light to dark green) cover 99, 90 and 50 percent of the municipal fertility rates, while the black line in the center represents the median. The upper/lower gray lines represent the maximum/minimum fertility rate at each age.

left panel displays substantial variation between municipalities at almost all ages. The right panel shows the cumulative distribution, which converges to the total fertility rate as the age approaches 49 years. Figure 7 shows the geographic distribution of the corresponding TFRs. A well-known overall pattern is reproduced, with the TFR highest in the south-western part of Norway and around the capital Oslo, while the south-eastern and northern part of the country have relatively low fertility. The EB method produces demographically plausible results by limiting small sample errors and reducing the occurrences of rates with extreme values.[21] In that sense, it provides conservative rates, which may be especially suitable for local planning or projection purposes.

[21] Although EB estimates typically are relatively smooth, practitioners may want to smooth the local ASFRs further. In Appendix C, we outline a local polynomial regression smoothing procedure that conserves local heterogeneity.

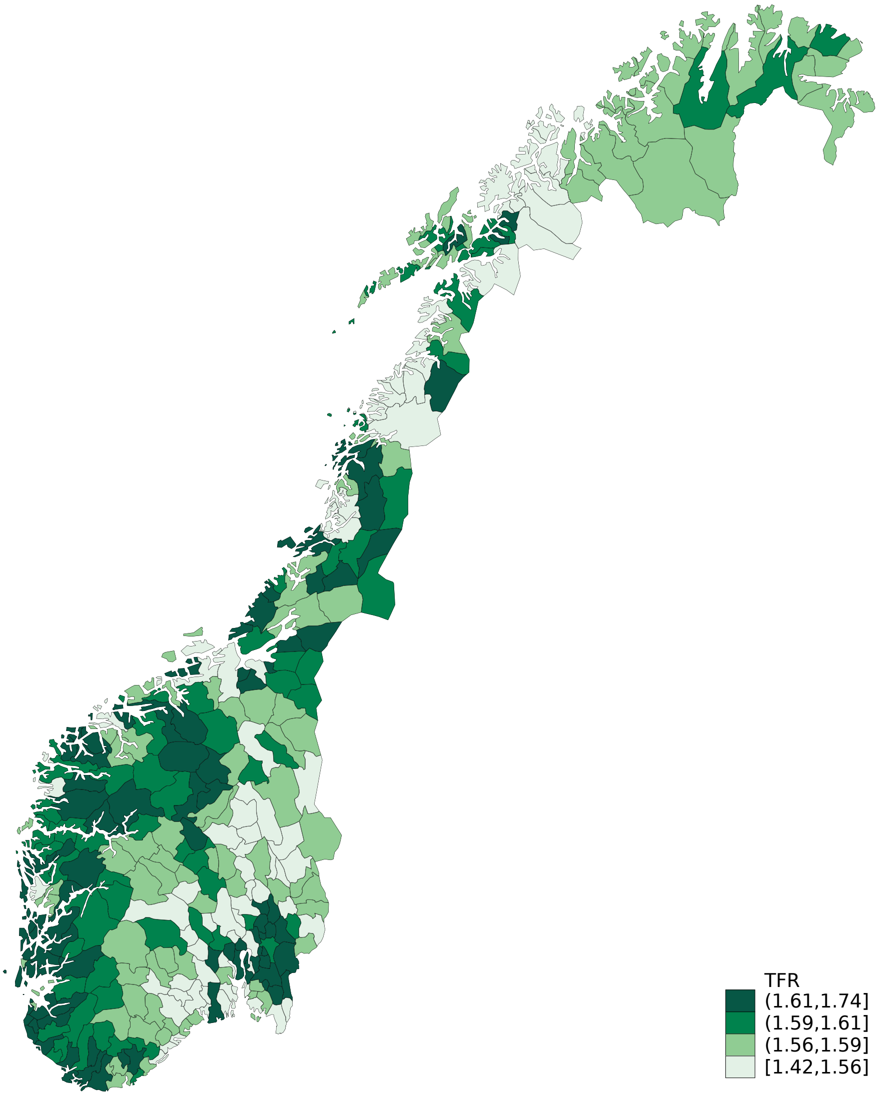

Figure 7: Geographic distribution of total fertility rates estimated by the empirical Bayes method

Note: The figure shows geographic distribution of TFRs across Norwegian municipalities based on empirical Bayes predictions of age-specific fertility rates. Light green colors indicate relative low total fertility while darker greens indicate relatively higher total fertility rates.

5 Discussion and concluding remarks

Demographic estimation of local schedules becomes a problem of small area estimation when disaggregation leads to sample sizes that are insufficient for direct estimates. The empirical Bayes method handles such small area problems by borrowing statistical strength from plentiful observations at higher-level geographic areas. Inspired by work on the estimation of local fertility rates using such methods (Assunção et al., 2005; Schmertmann et al., 2013) and lessons from the literature on hierarchical linear modeling (Hutchison and Healy, 2001; Moerbeek, 2004; Opdenakker and van Damme, 2000; van Landeghem et al., 2005), we propose amendments to the standard hierarchical linear model for computing small area demographic schedules. Our main innovation is to expand the hierarchy by including an intermediate regional shrinkage level. Using Monte Carlo simulations and applying the method to full-count Norwegian register data, we substantiate the claim that a three-level hierarchy with an aggregate global level and intermediate regional level displays many positive properties.

Including an intermediate regional level will have consequences for the performance of the model. In general, the researcher faces a trade-off between specifying regions large enough to curb sampling variability, the small area problem, but small enough as to capture the relevant geographic variation. The challenge of balancing these two sources of bias is especially pronounced in a two-level model setup, where the regional level must contain sufficient observations to function as an unbiased *grand mean* for the local demographic rates. The challenge is exacerbated by the complex nature of demographic behavior, where important driving factors can have different spatial patterns. This makes it demanding to allocate individuals to the (most) appropriate geographic units. We show that having both a global and a regional level in a three-level model eases these concerns. The practitioner is then at liberty to reduce the size of the intermediate regions, and instead to prioritize capturing relevant regional heterogeneity. Age-specific estimates that lack statistical support at the intermediate level will lean more heavily on the global level. Through Monte Carlo simulations, we show that the three-level model performs substantially better than the two-level models, even with arbitrarily selected regions.

The process of computing demographic schedules for municipalities in Norway, which provide many important public services, is riddled with small area estimation problems. In most municipalities only a few demographic events happen within each sex and age group, causing the corresponding direct estimate rates to become unstable and demographically implausible. We estimate age-specific fertility rates for each municipality in Norway using our preferred model. We demonstrate that the extreme variability of the estimates is dramatically reduced for smaller municipalities. However, the estimates still reveal substantial local variations in fertility level and timing of births. The described method is not limited to the Norwegian context or to fertility rates, but can be readily used for many other types of behavior, demographic or otherwise.

The model setup of this paper relies on several simplifying assumptions that may provide fruitful avenues for future research. First, the model imposes a diagonal covariance structure on the hyperparameters, restricting influence from other age groups. Relaxing this restriction will allow the model to exploit information from adjacent age groups when estimating ASFRs (Assunção et al., 2005). Second, exploring modeling choices for handling time trends at the various level of the hierarchy could potentially improve model performance when estimation samples that span several years are used. Finally, there are potential gains to be realized by investigating data-driven approaches for the specification of the intermediate level regions.

The EB method is well-known and has seen applications across many fields of study. Nonetheless, hierarchical linear models may be perceived as complex (Moerbeek et al., 2003) and the Bayes approaches may seem potentially time- and resource-demanding (Wilson, 2015), which may have delayed even more wide-spread use among practitioners. The estimation framework presented in this article is arguably transparent, flexible, and computationally simple. The hierarchical nested model with detailed age effects at all levels ensures that the EB predictions will, if applied to the estimation population, always reproduce the overall fertility numbers of the estimation sample. The estimates are easily reproducible and have classical frequentist interpretations. These properties translate into model predictions highly suitable as inputs into established production frameworks, for instance related to publication of statistical measures of mortality and fertility and population projections.

References

Ahlo, J. and Spencer, B. (2005). *Statistical demography and forecasting.* Springer Series in Statistics. Berlin-Heidelberg: Springer.

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

Alkema, L. and New, J. (2014). Global estimation of child mortality using a bayesian b-spline bias-reduction method. *Annals of Applied Statistics*, 8:2122–2149.

Alkema, L., Raftery, A., Gerland, P., Clark, S., Pelletier, F., Buettner, T., and Heilig, G. (2012). Probabilistic projections of the total fertility rate for all countries. *Demography*, 48:815–839.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging Lotteries for School Value-Added: Testing and Estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Assunção, R. M., Schmertmann, C. P., Potter, J. E., and Cavenaghi, S. M. (2005). Empirical bayes estimation of demographic schedules for small areas. *Demography*, 42(3):537–558.

Bijak, J. (2006). Bayesian methods in international migration forecasting. In Raymer, J. and Willekens, F., editors, *International migration in Europe: data, models and estimates*, pages 253–281. John Wiley and Sons, Chichester, UK.

Bijak, J. and Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, 70(1):1–19.

Calonico, S., Cattaneo, M., and Farrell, M. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software*, 91(8):1–33.

Carlin, B. and Louis, T. (2008). *Bayesian Methods for Data Analysis.* Boca Raton: CRC Press.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science.* New York: Cambridge University Press.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors: an empirical bayes approach. *Journal of the American Statistical Society*, 68(341):117–130.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

Godøy, A. and Huitfeldt, I. (2020). Regional variation in health care utilization and mortality. *Journal of Health Economics*, 71:102254.

Hustoft, A., Hartvedt, H., Nymoen, E., Stålnacke, M., and Utne, H. (1999). Standard for økonomiske regioner. Etablering av et publiseringsnivå mellom fylke og kommune (Standard for economic regions. Establishing a new level between county and municipality for the purpose of publishing statistics). Reports 1999/6, Statistics Norway.

Hutchison, D. and Healy, M. (2001). The effect of variance component estimates of ignoring a level in a multilevel model. *Multilevel Modeling Newsletter*, (13):4–5.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kravdal, Ø. (2002). The impact of individual and aggregate unemployment on fertility in Norway. *Demographic Research*, 48(6):185–262.

Kreft, I. G. and de Leeuw, J. (1998). *Introducing multilevel modeling.* Sage.

Leknes, S. and Løkken, S. (2020). Befolkningsframskrivinger for kommunene, 2020-2050 (Municipal population projections, 2020-2050). Reports 2020/27, Statistics Norway.

Manton, K., Woodbury, M., Stallard, E., Riggan, W., Creason, J., and Pellom, A. (1989). Empirical bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association*, (84):637–650.

Marshall, R. (1991). Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, (40):283–294.

Matthews, S. and Parker, D. (2013). Progress in spatial demography. *Demographic Research*, 28:271.

Moerbeek, M. (2004). The consequences of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, 39(1):129–149.

Moerbeek, M., van Breukelen, G., and Berger, M. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, 56:341–350.

Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Opdenakker, M. and van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, 11:103–130.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, (28):40–68.

Poulain, M., Herm, A., and Depledge, R. (2013). Central population registers as a source of demographic statistics in europe. *Population*, (68):183–212.

Raftery, A., Chunn, J., Gerland, P., and Sevcikova, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50:777–801.

Raftery, A., Sevcikova, H., Gerland, P., and Heilig, G. (2014). Bayesian probabilistic projections for all countries. *Proceedings of the National Academy of Sciences of the USA*, 109:13915–13921.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation.* New Jersey: John Wiley and Sons Inc.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage.

Robbins, H. (1964). The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1–20.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32.

Schmertmann, C. P., Cavenaghi, S. M., Assunção, R. M., and Potter, J. E. (2013). Bayes plus brass: Estimating total fertility for many small areas from sparse census data. *Population Studies*, 67(3):255–273. PMID: 24143946.

Skinner, C. (2018). Issues and challenges in census taking. *Annual Review of Statistics and Its Applications*, (5):49–63.

Spjøtvoll, E. and Thomsen, I. (1987). Application of some empirical bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, 2:435–449.

van Landeghem, G., de Fraine, B., and van Damme, J. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research*, 40:423–434.

Wilson, T. (2015). New evaluations of simple models for small area population forecasts. *Population, Space and Place*, (21):335–353.

Zhang, J. and Bryant, J. (2019). Combining multiple imperfect data sources for small area estimation: a Bayesian model of provincial fertility rates in Cambodia. *Statistical Theory and Related Fields*, (3):178–185.

Zhang, L.-C. (2003). Simultaneous estimation of the mean of a binary variable from a large number of small areas. *Journal of Official Statistics*, 19(3):253.

A Empirical Bayes approach

In the following, we will provide a formal description of the empirical Bayes model with two hierarchical levels and how it may be operationalized. Let $j \in \{1, ..., J\}$ denote index groups (e.g. municipalities), and let $i \in \{1, ..., N\}$ index individuals within groups. Let θ_j be an unknown parameter for the age- and municipality-specific group j (e.g. the fertility rate for 30-year-old women in municipality j) and Y_{ij} be an observed outcome (e.g. childbirth or not) for individual i in group j, assumed to follow the distribution:

$$Y_{ij}|\theta_j \sim f(y; \theta_j) \tag{A1}$$

In the next level of the hierarchy, we assume a distribution of the group level parameters:

$$\theta_j \sim g(\theta; \Omega) \tag{A2}$$

In the Bayesian framework, $g(\cdot\,)$ is a prior distribution, and Ω is a hyperparameter describing the prior. In the case of fertility, this distribution would characterize the spread of municipality-specific fertility rates. Alternatively, we can think of this as a random coefficient model where $g(\cdot\,)$ is the distribution of the random coefficients. It may be worth emphasizing that this is not the distribution of the measured outcomes, but rather the distribution of the unobserved group parameters.

We want to predict the individual θ_j, which tells us about each group parameter (e.g. municipality fertility rates). But to estimate the group parameters, we first need to estimate the hyperparameter Ω which informs us about the inter-group heterogeneity (the distribution of rates across municipalities).

To estimate Ω, we construct an integrated likelihood function from Equations (A1) and (A2) that expresses the distribution of the data for group j, $Y_j = (Y_{1j}, ..., Y_{Nj})$, as a function of the hyperparameter:

$$\mathcal{L}(Y_j|\Omega) = \int \prod_i f(Y_{ij}; \theta) g(\theta; \Omega) d\theta \tag{A3}$$

From this function we can write the EB maximum likelihood estimator as:

$$\hat{\Omega}_{EB} = \arg\max_{\Omega} \sum_j log \mathcal{L}(Y_j|\Omega) \tag{A4}$$

Using Bayes' rule, the posterior density for the group-specific parameter θ_j conditional on the observed data is given by:

$$h(\theta_j|Y_j;\Omega) = \frac{\prod_i f(Y_{ij};\theta_j)g(\theta_j;\Omega)}{\mathcal{L}(Y_j|\Omega)} \tag{A5}$$

$$\theta_j^* = \int \theta h(\theta|Y_j;\Omega)d\theta \tag{A6}$$

The empirical part of EB estimator comes from plugging the $\hat{\Omega}_{EB}$ estimate into Equations (A5) and (A6).

In many respects, this approach is more frequentist than Bayesian. The prior does not contribute any new information to the likelihood function other than the structure of the data, which is why statisticians sometimes criticize this approach for using the same data twice.

Consider a Gaussian model where $Y_{ij}|\theta_j \sim N(\theta_j, \sigma_\theta^2)$ and $\theta_j \sim N(0, \sigma_{\theta_j}^2)$. In this case the posterior distribution has a closed form solution and the EB estimator can be written as a weighted sum of the local mean $\bar{Y}_j$ and the *grand mean* $\bar{Y}$ which takes the form:[22]

$$\theta_j^* = \tau_j \bar{Y}_j + (1 - \tau_j)\bar{Y} \tag{A7}$$

$$\tau_j = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\theta_j}^2 / N_j} \tag{A8}$$

The weight τ_j is typically referred to as the *shrinkage factor* and is a function of the overall variation in the *grand mean* (σ_θ^2), the variation of the local mean ($\sigma_{\theta_j}^2$) and the municipality sample size (N_j).

Plugging the corresponding sample moments (estimated from the data) into Equations (A7) and (A8) returns the EB estimator. From Equation (A8) we see that the EB estimator is weighted closer to the local mean if the local mean is either precisely estimated or the local population size is large. Also, it is apparent that the EB estimates are unbiased, as τ_j will approach 1 as $N_j \to \infty$, which again means EB estimates will approach the unbiased sample means. This is exactly why the EB estimates are considered to be the best linear unbiased predictors (BLUP).

[22] Grand mean (or pooled mean) is the mean across all subsamples. In hierarchical models it refers to the mean of the top hierarchical level.

B Regional level bias and overshrinking

Our simulation exercise produces a different number of municipalities within each region and different population sizes in these municipalities for each run. Figure B1 displays the relative biases when we distinguish between regions that vary along these characteristics. The upper left panel shows the relative bias for regions that differ in the number of municipalities. For the L2R model, the bias is highest when the number of municipalities is low, but outperforms the L2C model when the number of municipalities increases. The result is related to regional population size, and we investigate this further in the upper right panel. For the L2C model, the relative bias is smallest when the regional population size is small, but this model is outperformed by the L2R model when the population size increases. The lower left panel shows the relative bias for the two models with respect to average municipality size in the region. The pattern resembles what we see in the two upper panels. The lower right panel shows the relative bias of the two models with respect to the standard deviation of the municipality population size within the regions. Here, the L2R model generally has a higher bias when the standard deviation is either very high or very low, but a lower bias when the standard deviation is average (which is where most of the observations tend to be located).

As mentioned, the regional characteristics we compare will typically be correlated, which may produce similar patterns across the graphs of Figure B1. The results indicate that the L2R model has the lowest relative bias as long as the number of individuals in the region is large. Since the systematic regional variation is orthogonal to population size in the simulation, this suggests that the increase in the relative bias of the L2R model in small samples is caused by increased variation in the regional level estimates. However, neither of the two-level models ever performs better than the three-level model.

A known issue with EB method is that the distribution of the predictions tend to be overshrunk relative to the real distribution. This problem has been highlighted in the statistical literature but rarely discussed in the three-level model case. See for instance Spjøtvoll and Thomsen (1987), Zhang (2003) and (Rao and Molina, 2015). Intuitively, it makes sense that EB estimators based on three-level hierarchical linear models should suffer less from overshrinkage. Since the local estimates are weighted towards the regional EB estimates (see Eq. 9) they are in a sense shrunk towards a more representative prior than in the two-level case. By comparing the variance of the municipal fertility rate EB predictions of the hierarchical models in the simulations with the variance of the true rates, we obtain a measure of the overshrinkage. Figure B2 shows that the three-level model suffers much less from overshrinkage than the two-level model (L2C). While the three-level EB predictions on average have a variance of 0.65 of the true variance, the EB predictions from the two-level model have a variance of 0.31 of the true variance.

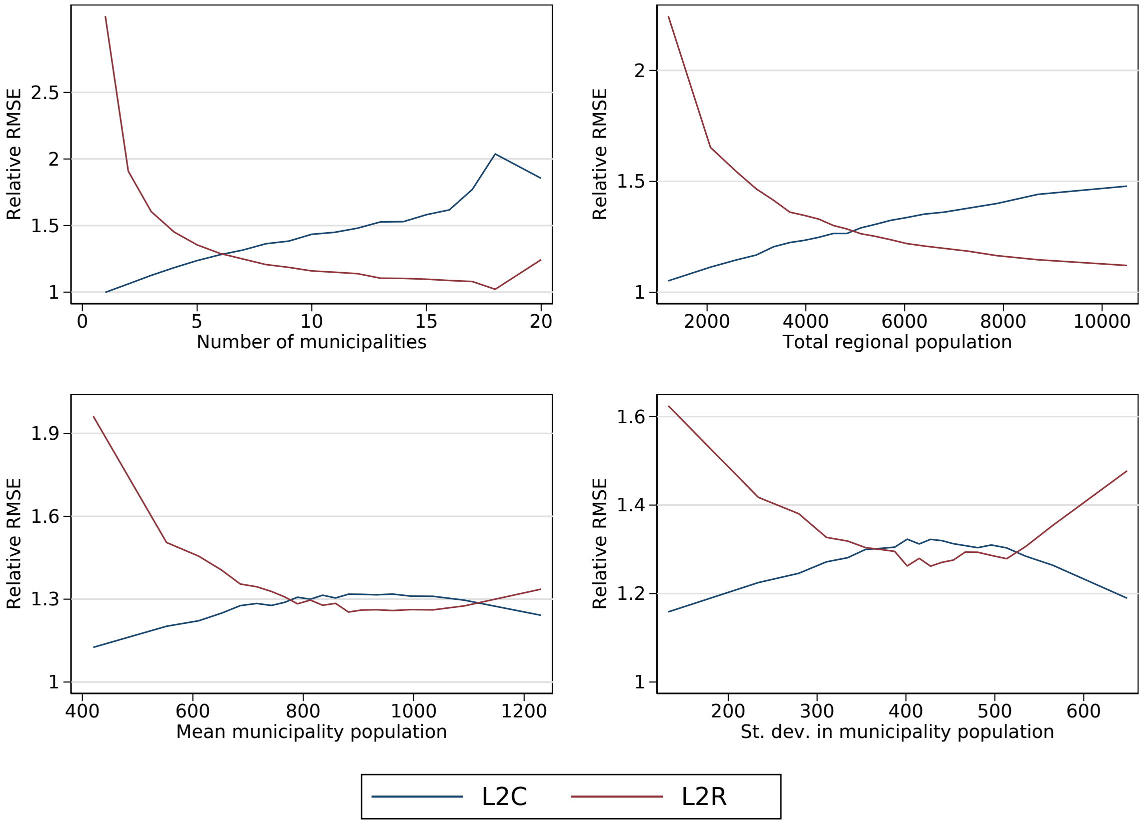

Figure B1: Relative bias and regional variation

Note: The figure displays how the relative bias of the two-level models is affected by regional characteristics based on data from 64 000 "regions" (64 regions×1000 simulations). The upper left panel shows the relative bias for regions with different number of municipalities, while the upper right panel shows the relative bias for regions with respect to total population size in the regions. The lower-left panel shows the regional relative bias with respect to average municipality size in the region, while the lower right panel shows the relative bias with respect to the standard deviation of the municipality population size within the region. Each sub-figure is produced by splitting the different regional characteristics into 20 equal-sized bins and plotting the average relative bias within each bin.

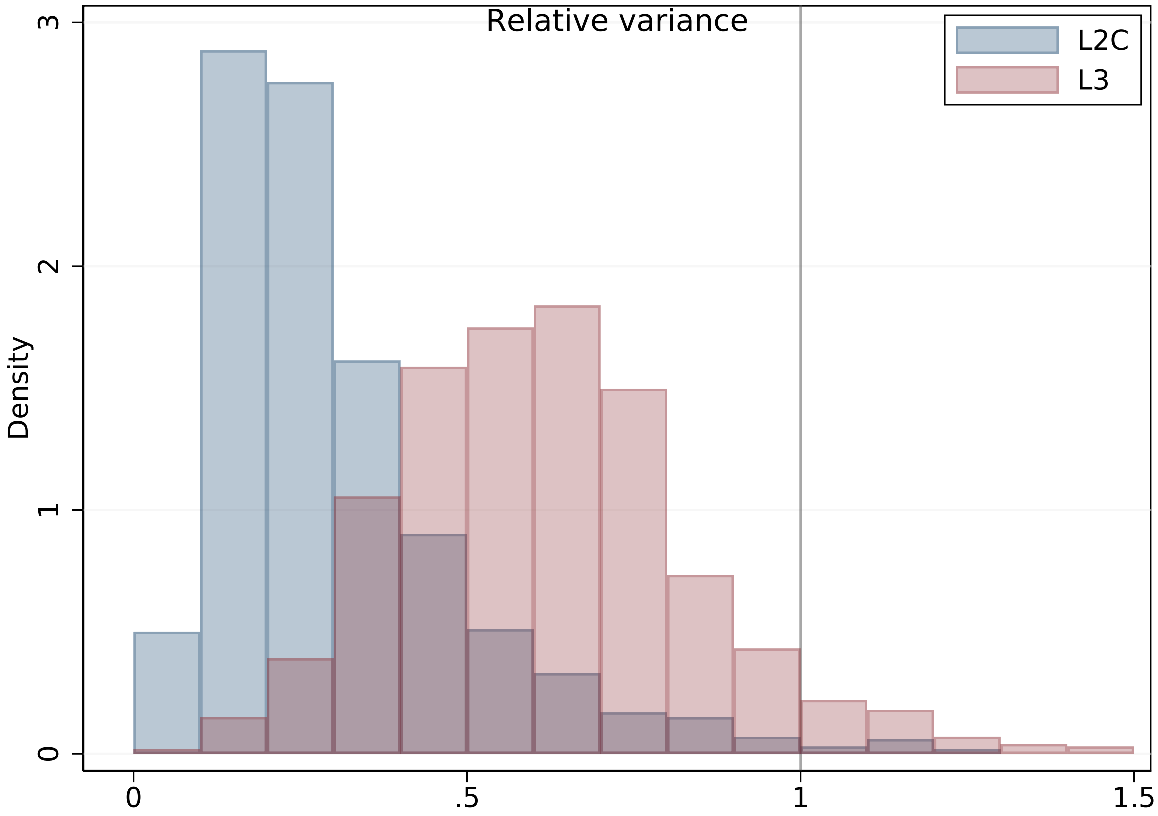

Figure B2: Overshrinkage of the empirical Bayes predictions

Note: The figure shows the distribution of overshrinkage from the three-level (L3) model and the two-level (L2C) model. The overshrinkage is measured by comparing the variance of the EB predictions of the municipality fertility rate for women of age 30 from both models with the variance of the true municipal fertility rate. If this measure is below 1 the estimation is overshrunk, whereas if the measure is above 1 the predictions are undershrunk. Results are based on data from 1000 simulations.

C Smoothing procedure

The demographic rates, generated using the EB method, may be used directly. However, smoothing demographic rates is not unusual over and preferred by many users. It is also in some sense more plausible in that the smoothed rates are "well-behaved" and do not jump and dive from one age group to the next. Therefore, we smooth the rates for each municipality over age.

We want to use a smoothing procedure that does not systematically bias the results. For this reason, we implement a bias-corrected smoothing procedure based on local polynomial regressions. The bias-correction ensures that the smoothed rates do not deviate unduly from the EB estimates. The user-written Stata package *nprobust* is used for this purpose and a description of the method can be found in Calonico et al. (2019).

The package offers several kernel functions for constructing local polynomial estimators. We use the default kernel function, Epanechnikov. The package also provides procedures for estimating optimal bandwidth size. For communication reasons we set the bandwidth at fixed values for each one-year age group in the smoothing procedure.The bandwidth is set at 3 for all age groups.

Figure C1 illustrates the difference between the smoothed and unsmoothed EB estimates of age-specific fertility rates. The local polynomial-based procedure preserves the overall shape while still smoothing out the jaggedness of the EB estimates from age group to age group. The bias correction ensures that the differences between overall fertility (sum of the AFSRs) is minimized, as well as the difference between the smoothed and unsmoothed rates.

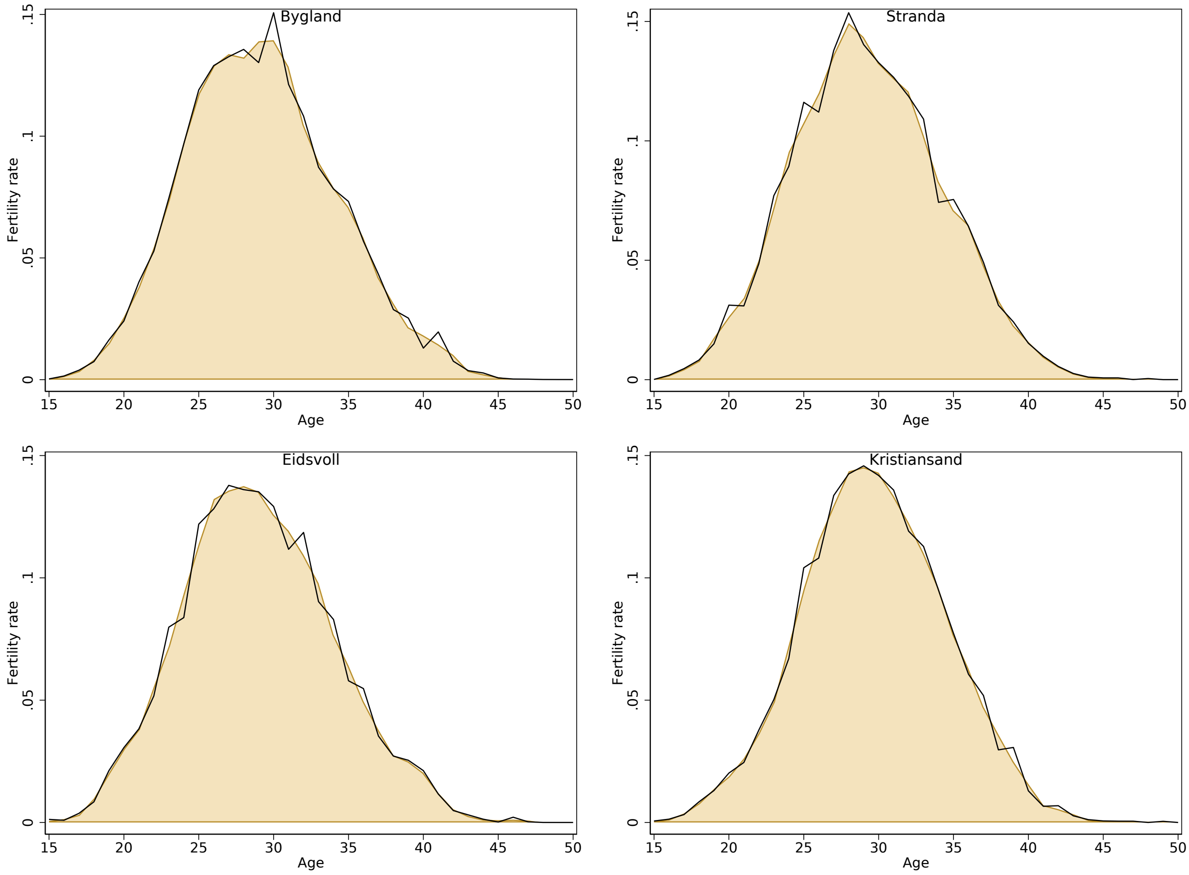

Figure C1: Comparison of smoothed and unsmoothed EB estimates of age-specific fertility rates in selected municipalities of different sizes

Note: The top left panel shows the smoothed (yellow area) and unsmoothed (black line) empirical Bayes estimates of age-specific fertility rates for a municipality with a population at the 10th percentile. The top right, bottom left and bottom right panels show the corresponding rates for municipalities with populations at the 50th, 90th and 99th percentile, respectively.