Flexible empirical Bayes estimation of local fertility schedules

Reducing small area problems and preserving regional variation

Stefan Leknes and Sturla A. Løkken

Research Department, Statistics Norway

IMA 2024 – Vienna

Statistisk sentralbyrå
Statistics Norway

Introduction

Local demographic schedules

- Local demographic schedules are in great demand for policy and planning, research and commercial purposes
- Important input in regional population projections
- Microsimulation models, in particular, allows for more heterogeneity and interactions between groups
- However, obtaining reliable estimates of such schedules is often not straightforward

Small area problem

- "... sample is not large enough to support direct estimates of adequate precision" (Rao and Molina, 2015)
- Random variation in demographic processes becomes prominent in small samples, which makes direct estimates noisy and unstable

Source: http://earthporm.com

Small area problems in the Norwegian setting

- Relevant local administrative units are municipalities (N=356)
- Large variation in population size – Oslo 680K / Utsira <0.2K
- Most have low population size – median 4.6K

Example: Number of estimates (cells) needed:

	Age range	Munici-palities	Sexes	Total number of cells
Mortality	0-100	365	2	73,730
Migration	0-69	365	2	51,100
Fertility	15-49	365	1	12,775

Solutions, drawbacks and motivation

- Common approaches are aggregation over time and space, parametric modeling and indirect model-based methods
- Pre-2020 regional projections for Norway relied heavily on aggregation of data
- These approaches reduce geographic variation in estimates
- We needed a method that:
  1. allows for local estimates
  2. preserve regional variation
  3. benefit from high-quality full-count population data
  4. is data-driven and transparent

Model and estimator

Empirical Bayes (EB) methods

Empirical Bayes estimator checks these boxes

- Excellent prediction properties (variance-bias trade-off)
- Shrinkage estimator borrows support from larger domains
    - Imprecise small area means → weighted towards larger domain mean
    - Precise local estimates less affected
    - If larger domain provides stat. support, cell sizes can be tiny (or zero)
- The standard is two-level models (Fay and Herriot, 1979)
- However, shrinking everything towards the population mean can wash away regional differences
- Instead, we use a three-level hierarchical linear model

Three-level hierarchical linear model

$$Y_{rji} = \theta + \theta_r + \theta_j + \epsilon_{rji} \tag{1}$$

$$\begin{aligned} \epsilon_{rji} &\sim N(0, \sigma_0^2), \\ \theta_j &\sim N(0, \sigma_1^2), \\ \theta_r &\sim N(0, \sigma_2^2) \end{aligned} \tag{2}$$

- Y_{rji}: binary for woman $\boldsymbol{i}$ in municipality $\boldsymbol{j}$ and region $\boldsymbol{r}$ gives birth
- θ: the fixed effect, population average fertility rate
- θ_r: the random effect at the regional level
- θ_j the random effect at the municipality level
- ϵ_{rji} is the individual error
- σ_0^2, σ_1^2, σ_2^2 are the random variances to be estimated

EB estimator

- The corresponding EB estimator, based on the three-level hierarchical model, can be formulated as a weighted sum of empirical estimates of the hierarchy means

$$\hat{\theta}_j^{EB} = w_j\bar{y}_j + w_r\bar{y}_r + w_c\bar{y} \tag{3}$$

- $\bar{y}_j$, $\bar{y}_r$, $\bar{y}_c$, is the municipality (j), regional (r), and population (c) means
- Weights sum to one, depend on pop. size and random effect variance estimates
- EB estimates bounded between 0 and 1

EB estimator

With a three-level hierarchical linear model, attention can be directed towards specifying a regional level to catch relevant heterogeneity

Regions can be based on:

- Administrative units (counties, hospital catchment areas)
- Official units (commuter zones, local labor markets, metropolitan areas, economic regions)
- or any other aggregated areas (arbitrary or optimal)

Simulation

Simulation design

- We simulate ASFRs for 400 municipalities
    - draw random population size and coordinates
    - systematic geographical component (nonlinear and continuous)
    - idiosyncratic local component (random)
- Agnostic region rule, subdivide municipalities into 64 equally sized regions (8×8 grid)
- Calculate our preferred EB estimator, along with 4 alternatives
- We run the procedure 1,000 times and collect the corresponding Bias and MSE
- Follow simulation design recommendations from Morris et al. (2019)

Simulation results

Table: Simulation bias and MSE of the estimators

	Linear models			Non-linear model	Direct estimates
	L3	L2C	L2R	NL3	DM
Bias (× 1000)					
Mean	0.00086	0.00090	-0.0014	-0.69	-0.011
SE	(0.45)	(0.49)	(0.44)	(0.50)	(0.63)
MCSE	[0.014]	[0.015]	[0.014]	[0.016]	[0.020]
MSE (× 1000)					
Mean	0.20	0.37	0.40	0.21	4.86
SE	(0.061)	(0.18)	(0.061)	(0.061)	(0.77)
MCSE	[0.0019]	[0.0056]	[0.0019]	[0.0019]	[0.024]

Models: 3-lvl linear (L3), 2-lvl linear (L2C), 2-lvl linear with region FE (L2R), 3-lvl logit (NL3), Frequentist mean (DM). Bias and MSE are calculated for each repetition, over 400 municipalities and 31 age groups.

Simulation results - MSE heterogeneity

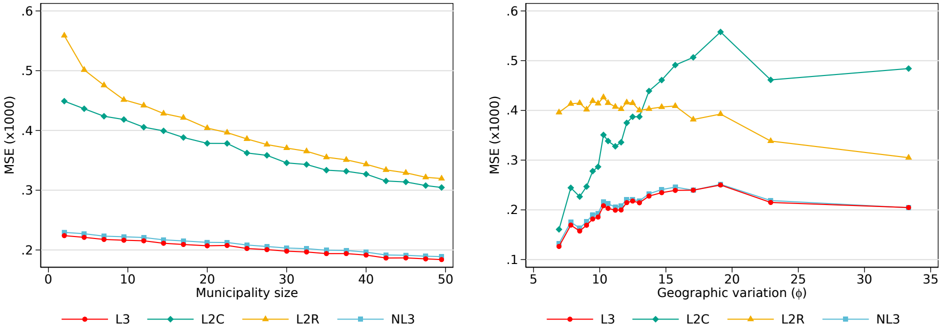


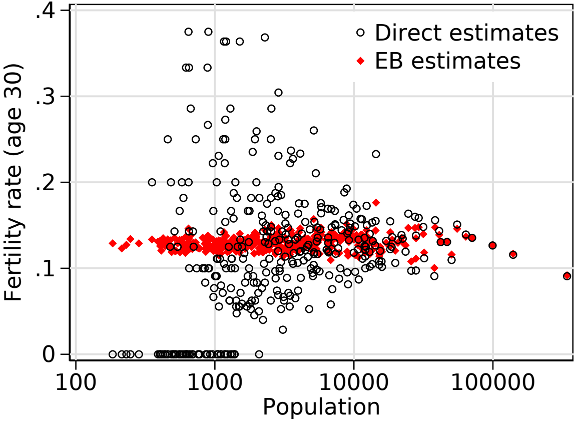
Figure: MSE by simulated characteristics

Real world application

Application with Norwegian data on fertility

- Context
  - Falling fertility: 1.98 in 2009 to 1.53 in 2019
  - The average age of giving birth has increased steadily
  - Substantial geographic variation in fertility: In 2019, the maximum difference in TFR across counties was 0.25
- EB method useful when ASFRs change rapidly, not relying on long panels of data
- Population and fertility data from 2019, for women aged 15-45
- Official economic regions, corresponding to the EU NUTS-4 level, form the basis for the intermediate regional level (N=89)

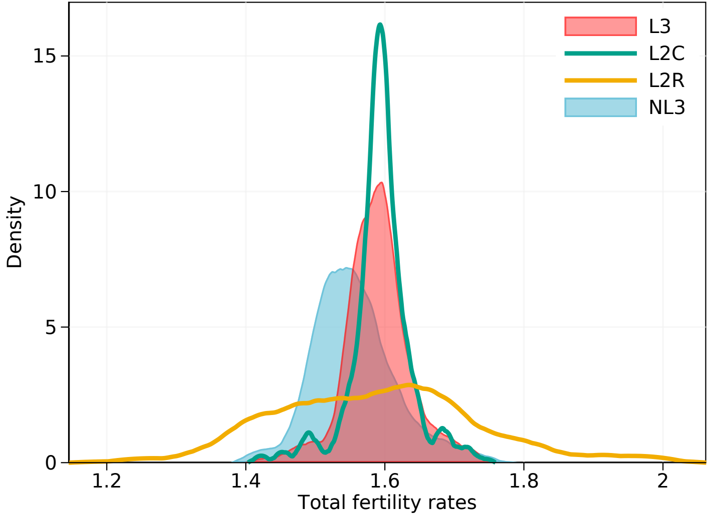
Fertility rates for municipalities by sizes



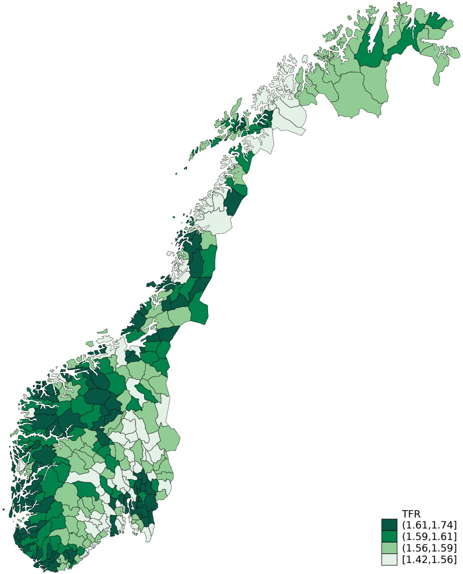
Note: Five municipalities with direct estimates higher than 0.4 are excluded. Three of these have fertility rates equal to one. In 53 municipalities the direct fertility rate estimates are equal to zero.

Municipal distribution of local TFR

Statistisk sentralbyrå
Statistics Norway



Geographic variation of local TFR



TFR based on our preferred EB estimator (L3)

- Fits well with the experiences
- High fertility in western Norway
- Low fertility in North and central east
- ...except around large cities

Conclusion

Concluding remarks

- The three-level model estimates outperforms alternative models and provides demographically plausible results
- An advantage is the preservation of regional heterogeneity, while still limiting sampling variability
- The method presented is arguably transparent, flexible, and computationally simple - making it suitable for established production frameworks
- Future research:
  - Introduce age group dependency
  - Incorporate time trends with panel data
  - Investigate data-driven approaches to the specification of the regions

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

Morris, T. P., I. R. White, and M. J. Crowther (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 38(11), 2074–2102.

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