Bonnéry, Daniel; Cheng, Yang; Ha, Neung Soo; Lahiri, Partha

Article

Triple-goal estimation of unemployment rates for U.S. states using the U.S. Current Population Survey data

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Bonnéry, Daniel; Cheng, Yang; Ha, Neung Soo; Lahiri, Partha (2015) : Triple-goal estimation of unemployment rates for U.S. states using the U.S. Current Population Survey data, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 16, Iss. 4, pp. 511-522,
https://doi.org/10.21307/stattrans-2015-030

STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY
Joint Special Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 511−522

TRIPLE-GOAL ESTIMATION OF UNEMPLOYMENT RATES FOR U.S. STATES USING THE U.S. CURRENT POPULATION SURVEY DATA

Daniel Bonnéry [1] Yang Cheng[2] Neung Soo Ha[3] Partha Lahiri[4]

ABSTRACT

In this paper, we first develop a triple-goal small area estimation methodology for simultaneous estimation of unemployment rates for U.S. states using the Current Population Survey (CPS) data and a two-level random sampling variance normal model. The main goal of this paper is to illustrate the utility of the triple-goal methodology in generating a single series of unemployment rate estimates for three separate purposes: developing estimates for individual small area means, producing empirical distribution function (EDF) of true small area means, and the ranking of the small areas by true small area means. We achieve our goal using a Monte Carlo simulation experiment and a real data analysis.

Key words: complex survey data; empirical distribution function; Monte Carlo Markov Chain; rank; risk; small area estimation,

1. Introduction

The national unemployment rate is one of the five key economic indicators published by the United States Bureau of Labor Statistics (BLS) and represents the number of unemployed as a percentage of the labor force. BLS publishes unemployment rate estimates for the nation and its different demographic and geographic subdomains. For example, unemployment rate estimates are produced for all states and the District of Columbia, all metropolitan statistical areas (MSA), all counties, cities and towns of New England, and all cities with population 25,000 or greater. The local unemployment rate estimates are used for regional planning and fund allocation under various federal assistance programs. The primary source of data for the unemployment rate statistics for both small and large domains is the Current Population Survey (CPS) conducted by the Census Bureau for BLS. The data are collected for about 729 MSAs consisting of more than 1,000 counties covering every state and the District of Columbia. More information about the CPS can be found at `http://www.bls.gov/cps/`

[1] Joint Program in Survey Methodology, University of Maryland. E-mail: dbonnery@umd.edu
[2] U.S. Census Bureau. E-mail: yang.cheng@census.gov
[3] Nielsen. E-mail: Neung.Ha@nielsen.com
[4] Joint Program in Survey Methodology, University of Maryland. E-mail: plahiri@umd.edu

The Census Bureau has been using the so-called AK composite estimation technique for generating national employment and unemployment levels and rates for the last several decades. The AK composite estimation technique, developed using the ideas of Gurney and Daly (1965), essentially improves on the standard survey-weighted estimates by borrowing strength over time. For more information on the AK estimation, see Lent et al. (1999). The estimation methodology for the BLS Local Area Unemployment Statistics (LAUS) can be found at `http://www.bls.gov/lau/laumthd.htm`. The state level unemployment statistics are based on a paper by Pfeffermann and Tiller (2006). For each month, model-based census division estimates are first benchmarked to the non-seasonally adjusted national A-K composite estimate and then similar model-based state estimates are benchmarked to the benchmarked estimate of the state's division. The unemployment estimates for the states or the census divisions can be viewed as benchmarked empirical best prediction (EBP) estimates, derived using a state-space model and implemented via an innovative Kalman Filter updating scheme that simplifies the computational burden in a complex production environment.

In a statistical decision-theoretic framework, BLS addresses the problem of point estimation under a squared error loss function and the estimation of the corresponding risk measured by the mean squared prediction error (MSPE). These are indeed important statistical decision problems. It is expected that BLS will continue to focus on the point estimation and the corresponding MSPE because of a long history of such unemployment statistics series and official publication requirements. One can, however, envision a variety of statistical decision problems related to unemployment statistics. For example, different stakeholders may be interested in ranking different states in order of unemployment rates or identifying states with unemployment rates exceeding a certain specified threshold for regional planning and fund allocation problems. The need for answering research questions other than point estimation can be found in different contexts. For example, the goal can be estimating the performance evaluation, like the rank, among different companies; see Landrum et al. (2000). Reporting an ensemble of estimates can also provide useful interpretation in disease mapping to ascertain variation in disease rates for different geographical regions; see Conlon and Louis (1999) and Devine and Louis (1994).

Note that the research questions mentioned above correspond to different statistical decision-theoretic problems and thus, statistically speaking, a research question-specific unemployment series can be found, which is likely to be different from BLS published series. Of course, the published unemployment rates can be used to answer a variety of research questions, but they may not be well suited

for a wide range of problems. To elaborate this point, if the ranks of parameters are the target, under the Bayesian approach, the conditional expected ranks are optimal under squared error loss function, but ranking posterior means, which are optimal for point estimation under squared error loss, can perform poorly; see Goldstein and Spiegelhalter (1996). If the feature of interest is the histogram or the empirical distribution function (EDF) of the parameters, then the conditional expected EDF is optimal under integrated squared error loss function, and the histogram of the posterior means of the parameters is underdispersed; see Ghosh (1992). There are a number of papers on the estimation of parameters for an individual small area, e.g. Rao (2003), Jiang and Lahiri (2006), Pfeffermann (2013), a histogram of small area parameters, e.g. Louis (1984), Lahiri (1990), Ghosh (1992), and ranking small area parameters, e.g. Laird and Louis (1989).

Although different series can be produced to address different questions, reporting several ensembles for all different situations would be inefficient and may cause inconsistencies. While there does not exist a set of point estimates that simultaneously optimize all of these criteria (Gelman and Price, 1999), Shen and Louis (1998) developed an interesting method, called "triple-goal" estimation method, which produces estimates that perform reasonably well with respect to all three criteria.

In Section 2, we explore a triple-goal small area estimation methodology for simultaneous estimation of small area means using the CPS complex survey data. The main goal is to produce a set of small area estimates that are good for simultaneously meeting three different goals of developing estimates for individual small area means, producing histogram of true small area means, and ranking of the small areas by true small area means. We discuss evaluation of our methodology in Section 3.

2. Adaptation of the triple-goal estimation methodology to estimate unemployment rates for U.S. states

The main challenge for adapting the existing triple-goal methodology to estimate unemployment rates for U.S. states is to incorporate the complex survey features of the CPS. Let $\hat{\pi}_i$ be the survey-weighted direct estimate of the true unemployment rate π_i for the ith state $(i = 1, \cdots, m)$. We are interested in producing triple-goal estimates of $\pi = (\pi_1, \cdots, \pi_m)$. To obtain triple-goal estimates of π_i's and to compare with the corresponding Bayesian estimates (posterior means of π_i's), we consider the following hierarchical model.

For $i = 1, \ldots, m$,

$$\textit{Level 1} \text{ (sampling distribution)} : \hat{\pi}_i | \pi_i \overset{ind}{\sim} N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n_{i;\text{eff}}}\right);$$
$$\textit{Level 2} \text{ (prior distribution)} : \text{logit}(\pi_i) | \mu, A \overset{iid}{\sim} N(\mu, A),$$

where π_i and $n_{i;eff}$ are the "true" unemployment rate and the effective sample size for state i, respectively. The effective sample size for a state is the ratio of the CPS sample size for that state and the national estimate of design effect (deff). We assume flat priors on both μ and A.

We note that the BLS uses a two-level time series normality-based model to combine previous survey data. While the BLS model will be of interest to produce triple-goal unemployment rate estimates, in this paper we focus on the above relatively simple cross-sectional random sampling variance two-level normal model for demonstrating the utility of triple-goal estimation for multi-purpose estimation. Like the BLS model, we find it convenient to assume normality for the survey-weighted proportions, but use a random sampling model to incorporate uncertainty in estimating sampling variances of the survey weighted proportions. Such a model was considered earlier in different contexts by Liu et al. (2014) and Ha et al. (2014).

The triple-goal estimation method involves the following three steps (see Shen and Louis (1998) for further details):

Step 1: Produce element-specific point estimates with "optimality" qualities for the region of interest;

Step 2: Obtain an ensemble of point estimates that best approximate the histogram of the true parameter ensemble; see Louis (1984);

Step 3: Rank within a selected ensemble.

The procedure for obtaining triple-goal estimators follows along the line of Shen and Louis (1998), which is described below:

First, we need to obtain an estimate of the empirical distribution function (EDF) of π. The EDF of π is defined as:

$$F_m(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \mathscr{I}\{\pi_j \leq \alpha\},$$

where $\alpha \in \mathbb{R}$ and $\mathscr{I}$ is the indicator function. Under the following integrated squared error loss (ISEL) function for a given EDF estimator $\tilde{F}_m$:

$$\text{ISEL}(F_m, \tilde{F}_m) = \int \left[F_m(\alpha) - \tilde{F}_m(\alpha)\right]^2 d\alpha,$$

the Bayes estimator of EDF is given by

$$\hat{F}_m(\alpha) = E\left[F_m(\alpha)|\hat{\pi}\right] = \frac{1}{m}\sum_{j=1}^{m} P(\pi_j \leq \alpha|\hat{\pi}).$$

Secondly, we need to obtain the rank of the parameter ensemble π. The rank of π_i is defined as

$$R_i = \text{rank}(\pi_i) = \sum_{j=1}^{m} \mathscr{I}\{\pi_i \geq \pi_j\}.$$

Under the rank squared error loss (RSEL) function for a given rank estimator $\tilde{\mathbf{R}}$, defined as

$$\text{RSEL}(\mathbf{R}, \tilde{\mathbf{R}}) = \frac{1}{m}\sum_{j=1}^{m}(R_j - \tilde{R}_j)^2,$$

the Bayes estimator of R_i is given by

$$\bar{R}_i = E(R_i|\hat{\pi}) = \sum_{j=1}^{m} P(\pi_i \geq \pi_j|\hat{\pi}).$$

The $\bar{R}_i$'s are not integers in general; however, it is easy to transform them in order and denote it by:

$$\hat{R}_i = \text{rank}(\bar{R}_i|\mathbf{R}), i = \ldots, m.$$

Finally, we generate an ensemble of point estimates, conditional on the optimal estimate of the ensemble EDF, $\hat{F}_m$, and the optimal estimates of the ranks, $\hat{R}_i$. Furthermore, the added constraint that $\hat{F}_m$ is a discrete distribution with at most m mass points, the triple-goal estimator is defined as:

$$\hat{\pi}_i^{TG} = \hat{F}_m^{-1}\left(\frac{2\hat{R}_i - 1}{2m}\right), i = 1, \ldots, m.$$

We use MCMC to implement the triple-goal method. The simulated samples after deleting the first B "burn-in" samples, i.e.

$$\left\{\mu^{(B+\ell)}, A^{(B+\ell)}, \pi^{(B+\ell)},\ \ell = 1, \cdots, L\right\},$$

are considered as L simulated samples from the posterior distribution of β, A, π.

The posterior density of π is approximated by

$$\left\{\pi^{(B+\ell)},\ \ell = 1, \cdots, L\right\}.$$

In particular, we need the following approximations:

$$\hat{F}_m(\alpha) \approx \frac{1}{m}\sum_{j=1}^{m}\left\{\frac{1}{L}\sum_{\ell=1}^{L}\mathscr{I}\left[\pi_j^{(B+\ell)} \le \alpha\right]\right\},$$

$$\bar{R}_i \approx \sum_{j=1}^{m}\left\{\frac{1}{L}\sum_{\ell=1}^{L}\mathscr{I}\left[\pi_i^{(B+\ell)} \le \pi_j^{(B+\ell)}\right]\right\}.$$

3. Evaluation

Our ultimate goal is to develop a triple-goal estimation system for the state unemployment rates using the CPS data. As in any real life data analysis, we encounter the challenging problem of evaluation of triple-goal estimates relative to the commonly used direct and posterior means since we do not have true unemployment values. We consider two options. First, we compare different estimates using simulated data generated using the model given in Section 2 and the CPS data. While such an evaluation is model-dependent, we argue that this is a reasonable approach since our main goal in this paper is to compare direct estimates, posterior means and triple-goal estimates for three separate purposes given a working model. In Subsection 3.1, we present results from such an evaluation study. The other option for evaluation is to use a real data that contain the truth or a gold standard. We do not have such data for unemployment rate estimation research. Since estimation of unemployment rates is essentially a problem of estimation of proportions, in Subsection 3.2 we use the well-known batting average data described in Efron and Morris (1975), which contain true batting averages (true proportions).

We now evaluate direct, posterior mean, triple-goal estimators of ranks, EDFs, and individual parameters. To be specific, we compare different estimators using the following four summary evaluation measures:

(i) Root Average Squared Deviation (RASD): $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\tilde{\pi}_i - \pi_i)^2}$

(ii) Root Integrated Squared Error Loss (RISEL): $\sqrt{\int\left[F_m(t) - \tilde{F}_m(t)\right]^2 dt}$

(iii) Variance Ratio (VR): $\frac{\sum_{i=1}^{m}(\tilde{\pi}_i - \bar{\tilde{\pi}})^2}{\sum_{i=1}^{m}(\pi_i - \bar{\pi})^2}$

(iv) Root Rank Average Squared Deviation (RRASD): $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\tilde{R}_i - R_i)^2}$,

where $\bar{\pi}_i$ ($\bar{\tilde{\pi}}$) is the average of the π_i's ($\tilde{\pi}_i$'s), average being taken over all m states.

3.1. Evaluation using simulated data

Using the two-level normal model described earlier with $\mu = \hat{\pi}$, the national unemployment rate estimate, and $A = \sum_{i=1}^{51}(\hat{\pi}_i - \bar{\hat{\pi}})^2/51$, where $\hat{\pi}_i$ is the survey-weighted CPS unemployment rate for state i $(i = 1, \cdots, m)$, we generate unemployment rate direct estimates and simulated true values for the states. We can then compare different methods using simulated values.

Table 1 displays values of the four evaluation measures for the three estimators. From the VR measure, it is clear that the variability of the direct estimates of the state unemployment rates overestimates the corresponding variability of the simulated unemployment rates across the states. On the other hand, the posterior means of the state unemployment rate estimates overshrink. The triple-goal estimates are almost perfect in terms of this criterion. Based on the RISEL criterion, the triple goal estimates are also the best among the three sets of estimates in terms of estimating the EDF of the simulated unemployment rates. The criterion RRASD suggests that in terms of the rank, triple-goal estimates are the best, but they are only marginally better than the posterior means. In terms of the RASD criterion, posterior means are the best as expected, but are only marginally better than the triple-goal estimates.

Figure 1 provides histograms of three sets of estimates for the states and the simulated values. From a visual inspection, it is clear that the histogram for the triple-goal estimates is the closest to that of the true values when compared to the histograms of the posterior means and the direct estimates.

	RASD	RISEL	VR	RRASD
direct	0.0097	0.0122	1.2861	8.2652
post. mean	0.0086	0.0121	0.8316	8.1889
triple-goal	0.0095	0.0091	1.0200	8.1746

Table 1: Summary statistics for the unemployment data

3.2. Evaluation using a real data with true values

As mentioned before, in this subsection we use the well-known baseball data, which were used earlier by researchers in evaluating different small area methodologies. The data contain batting averages of eighteen major league baseball players in the 1970 season. Each player had batted 45 times and their batting averages are recorded up to that point. Using this data alone, Efron and Morris (1975) wanted to predict each player's batting average for the remainder of the 1970 season. Here, a player corresponds to a small area like a state in the unemployment rate estimation.

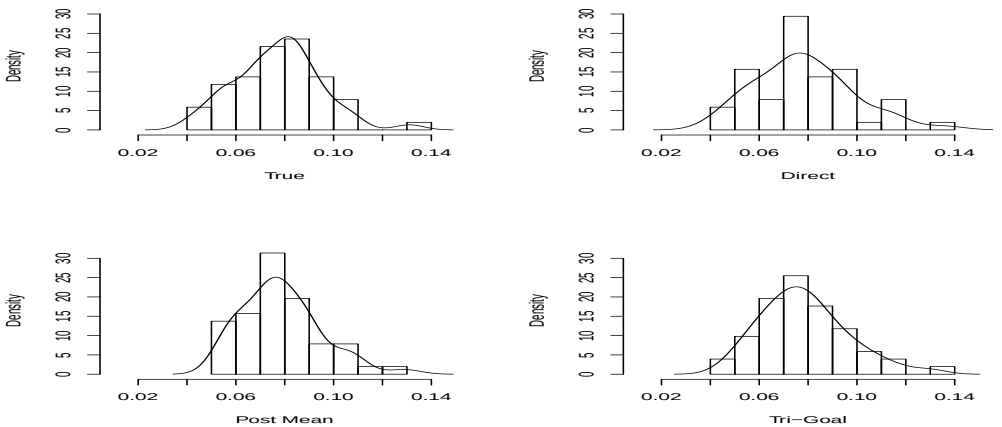

Figure 1: Histograms for the unemployment data

We report the four summary evaluation measures for the three sets of estimates in Table 2. In Figure 2, we plot the histograms for the three sets of estimates of batting averages and the true batting averages. The conclusion is similar to the one in Subsection 3.1.

	RASD	RISEL	VR	RRASD
direct	0.0572	0.0486	3.3920	5.8878
post.mean	0.0311	0.0311	0.1899	5.8214
triple-goal	0.0334	0.0094	1.0328	5.8022

Table 2: Summary statistics for the baseball data

4. Concluding Remarks

In this paper, we extend the triple-goal methodology, originally proposed by Shen and Louis (1998), to a hierarchical model not considered earlier in modeling unemployment rates for small areas. First, instead of using fixed and known sampling variance of a survey-weighted unemployment rate for a small area, we have used the true variance formula of a sample proportion with sample size replaced by the effective sample size in order to incorporate the complex survey design. Secondly, to borrow strength from small areas, we use normality on the logistic function of the unknown true unemployment rates, which appear in both the means and variances in the sampling distribution.

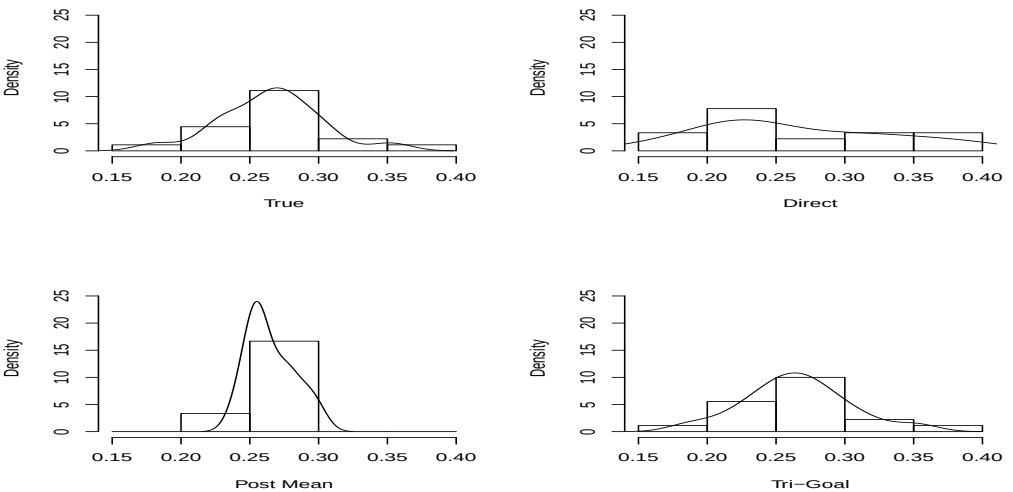

Figure 2: Histograms for the baseball data

We reiterate that the triple-goal method is for multi-purpose inferences. In theory, this approach should reduce the overshrinking problem associated with the standard Bayesian estimates (posterior means) targeted for point estimation and should do better than rival methods in estimating ranks and empirical distribution function of the true values. While our evaluation studies demonstrate a clear superiority of the triple-goal method in reducing the overshrinking problem and estimating the empirical distribution function of the true values, it is only marginally better than the posterior means and direct estimates in estimating ranks. This could be due to certain approximations applied to the optimal rank estimates in order to produce integer valued ranks of the small areas. Under the theoretical setting, posterior means should perform better than the triple-goal estimates in terms of point estimation of the small area proportions. Our evaluation studies, however, show that they are only marginally better.

While the goal of this paper is not to find the posterior means and triple-goal estimates under the best possible working model, a good working model is expected to improve on both the standard Bayesian and triple-goal methods. Thus model selection will be a problem of great interest before implementing the triple-goal

method for producing a new unemployment rate series for multi-purpose uses. In the future, we plan to develop a benchmarked triple-goal estimation system using the multi-level time series model used by BLS for its production of official statistics for the states. Neither of the two methods of evaluation considered in the paper should be considered an ideal method, which does not seem to exist in small area estimation evaluation. But nonetheless our evaluation study should shed some light on the merit of triple-goal for multi-purpose inferences and should encourage researchers to think of new ideas for evaluating small area methods.

Acknowledgements

Authors are listed in alphabetical order. The research of the first and last authors has been supported by the U.S. Census Bureau Prime Contract No: YA1323-09-CQ-0054 (Subcontract No: 41-1016588). The research of the third author has been partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Authors would like to thank Professor Thomas Louis for some useful discussions on the triple-goal method. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the U.S. Census Bureau.

REFERENCES

CONLON, E. M. and LOUIS, T. A. (1999). Addressing Multiple Goals in Evaluating Region-specific Risk Using Bayesian Methods. In *Disease Mapping and Risk Assessment for Public Health*, Chichester: Wiley, 31–47.

DEVINE, O. J. and LOUIS, T. A. (1994). A Constrained Empirical Bayes Estimator for Incidence Rates in Areas with Small Populations. *Statistics in Medicine*, 13, 1119–1133.

EFRON, B. and MORRIS, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, 70, 311–319.

GELMAN, A. and PRICE, P. N. (1999). All Maps of Parameter Estimates are Misleading. *Statistics in Medicine*, 18, 3221–3234.

GHOSH, M. (1992). Constrained Bayes Estimation with Applications. *Journal of the American Statistical Association*, 87, 533–540.

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of Royal Statistical Society, Series A.*, 159, 385–409.

GURNEY, M. and DALY, J. F. (1965). A Multivariate Approach to the Estimation in Periodic Sample Surveys. In *Proceedings of the Social Statistics Section, ASA.*, 242–257.

HA, N. S., LAHIRI, P., and PARSONS, P. (2014). Methods and Results for Small Area Estimation Using Smoking Data from The 2008 National Health Interview Survey. *Statistics in Medicine*, 33, 3932–3945.

JIANG, J. and LAHIRI, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, 15, 111–999.

LAHIRI, P. (1990). Adjusted Bayes and Empirical Bayes Estimation in Population Sampling. *Sankhya*, 52, 50–66.

LAIRD, N. M. and LOUIS, T. A. (1989). Empirical Bayes Ranking Methods. *Journal of Educational Statistics*, 14, 29–46.

LANDRUM, M. B., BRONSKILL, S. E., and NORMAND, S. (2000). Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers. *Health Services Outcomes Research Methodology*, 1, 23–47.

LENT, J., MILLER, S., CANTWELL, P., and DUFF, M. (1999). Effects of composite weights Current Population Survey. *Journal of Official Statistics*, 15, 431–448.

LIU, B., LAHIRI, P., and KALTON, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology*, 40, 1–13.

LOUIS, T. (1984). Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods. *Journal of the American Statistical Association*, 79, 393–398.

PFEFFERMANN, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, 28, 1, 40–68

PFEFFERMANN, D. and TILLER, R. B. (2006). Small Area Estimation With State-Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, 1387–1397.

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

SHEN, W. and LOUIS, T. (1998). Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of Royal Statistical Society, Series B.*, 60, 455–471.