Smoothing and Benchmarking for Small Area Estimation

Rebecca C. Steorts[*], Timo Schmid[†], and Nikos Tzavidis[‡]

February 2, 2020

Abstract

Small area estimation is concerned with methodology for estimating population parameters associated with a geographic area defined by a cross-classification that may also include non-geographic dimensions. In this paper, we develop constrained estimation methods for small area problems: those requiring smoothness with respect to similarity across areas, such as geographic proximity or clustering by covariates; and benchmarking constraints, requiring weighted means of estimates to agree across levels of aggregation. We develop methods for constrained estimation decision-theoretically and discuss their geometric interpretation. The constrained estimators are the solutions to tractable optimization problems and have closed-form solutions. Mean squared errors of the constrained estimators are calculated via bootstrapping. Our approach assumes the Bayes estimator exists and is applicable to any proposed model. In addition, we give special cases of our techniques under certain distributional assumptions. We illustrate the proposed methodology using web-scraped data on Berlin rents aggregated over areas to ensure privacy.

1 Introduction

Small area estimation (SAE) deals with estimating many parameters, each associated with an "area"—a geographic domain, a demographic group, an

[*]Department of Statistical Science and Computer Science, Duke University, Durham, North Carolina, Email: beka@stat.duke.edu

[†]Freie Universität Berlin Institute of Statistics and Econometrics, Berlin, Germany, Email: timo.schmid@fu-berlin.de

[‡]University of Southampton Southampton Statistical Sciences Research Institute, Southampton, United Kingdom, Email: n.tzavidis@soton.ac.uk

experimental condition, etc. Areas are "small" since there is little or no information about any one area. Estimates of a parameter based only on observations from the associated area, called direct estimates, can be imprecise. To increase precision, one tries to "borrow strength" from related areas, and hierarchical and empirical Bayesian models are one way to do so. Since the pioneering work of Fay and Herriot (1979) and Battese et al. (1988), such models have dominated SAE, with many successful applications in official statistics, sociology, epidemiology, political science, business, etc. (Rao and Molina 2015). Recently, SAE has been applied in other fields, such as neuroscience, and performs as well as common approaches such as smoothed ridge regression and elastic net (Wehbe et al. 2015).

We extend these classical approaches in two directions, both of which have been the subject of recent interest in the SAE literature. One direction is to take direct account of information about the proximity of areas in space or time. In many applications, it is reasonable to expect that the parameters will be smooth, so that nearby areas will have similar parameters, but this is not altogether standard within SAE (Rao and Molina 2015). Incorporating spatial dependence directly into Bayesian models leads to statistical and computational difficulties, yet it seems misguided to discard such information. The other direction is "benchmarking," the imposition of consistency constraints on (weighted) averages of the parameter estimates. A simple form of benchmarking is when the average of the parameter estimates must match a known global average. When there are multiple levels of aggregation for the estimates, there can be issues of internal consistency as well.

We provide a unified approach to smoothing and benchmarking by regarding them both as *constraints* on Bayes estimates. Benchmarking corresponds to equality constraints on global averages and variances. Similarly, smoothing corresponds to an inequality constraint on the "roughness" of estimates (how much the parameter estimates of nearby areas differ). The motivation of this smoothing is based upon manifold learning and frequentist non-parametrics, where loss functions are augmented by a penalty. Such a penalty term is in the spirit of ridge regression, where a transformation of the parameters is performed and additional shrinkage is carried out. Our penalty corresponds to how much estimates at nearby points in the domain should differ.

Decision-theoretically, we obtain smoothed, benchmarked estimates by minimizing the Bayes risk subject to these constraints, extending the approaches of Datta et al. (2011) and Ghosh and Steorts (2013) (themselves in the spirit of Louis (1984) and Ghosh (1992)). Geometrically, the constrained

Bayes estimates are found by projecting the unconstrained estimates into the feasible set. If the constraints are linear, then the resulting optimization can be solved in closed form, requiring nothing more than basic matrix operations on the unconstrained Bayes estimates. Another strong advantage of our decision-theoretic and geometric approach is its generality. We require no distributional assumptions on the data or on the unconstrained Bayes estimator. Our results apply whether the unconstrained estimator is linear or non-linear. The relevant notion of proximity between areas may be spatial or more abstract. It can also include clustering on covariates not directly included in the model. Finally, we are able to prove known cases under our proposed approach, where the Bayes and frequentist estimates are in fact the same. Finally, we illustrate our proposed methodology on rental prices in Berlin.

The rest of the paper proceeds as follows. Section 2 describes related work. Section 3 provides the proposed general framework for smoothing in small area estimation. Section 3.1, introduces notation used throughout the paper. Section 3.2 proposes a general result for SAE in the context of smoothing. Section 3.3 proposes special cases of our general framework under the area-level Fay-Herriot model. Section 4 extends our generalized result in Section 3.2 to benchmarking constraints. Section 4.2 derives special cases under the area-level Fay-Herriot model. Section 4.3 discusses choices of the smoothness penalties. Section 5 proposes a non-parametric bootstrap for mean squared error (MSE) estimation. Section 6 applies the proposed methodology to web-scraped data for estimating average rent prices in Berlin. Section 7 concludes with a discussion and future work.

2 Related Work

The proposed methodology for SAE with benchmarking and smoothing generalizes the work of Datta et al. (2011) and Ghosh and Steorts (2013), which take a decision theoretic approach to SAE. However, this literature does not allow for spatial smoothing. The approach proposed in this paper is also similar to that of Wehbe et al. (2015) in the sense that spatial smoothing is considered; however, these authors do not consider benchmarking. Moreover, the authors focus more on a neuroscience application and less on developing a general methodology for SAE methodology. Other relevant literature includes Pratesi and Salvati (2008), who proposed a spatial empirical best linear unbiased predictor under the Fay-Herriot model with a simultaneous autoregressive (SAR) structure for the random effects and an analytic based MSE. Souza et al. (2009) account for spatial relationships when fitting hierarchical Bayesian exponential growth models. Rao and Yu (1994) proposed a linking model that does not include area-specific random effects in essence to avoid over-smoothing, which is a valid concern when proposing any type of smoothing constraint.

Previous efforts at smoothing in SAE problems have smoothed either the raw data or direct estimates. In contrast, we smooth estimates based on models which do not themselves include spatial structure. Computationally, this is much easier than expanding the models. Our optimization problems can be solved in closed form and retain the advantages of model-based estimation. This approach to smoothing also combines naturally with the imposition of benchmarking constraints, which has never been handled to our knowledge in the literature before.

Our proposed methodology employs ideas about smoothing on graphs and manifolds from frequentist non-parametrics and machine learning. In particular, we take advantage of "Laplacian" regularization ideas (Belkin et al. 2006; Corona et al. 2008; Lee and Wasserman 2010), where the loss function is augmented by a penalty term which reflects how much estimates at nearby points in the domain differ. Such regularization is designed to ensure that estimates vary smoothly with respect to the intrinsic geometry of some underlying graph or manifold. (Smoothness on a domain is represented mathematically by the domain's Laplacian operator, which is the generator for diffusion processes.) This generalizes the roughness or curvature penalties from spline smoothing (Wahba 1990) to domains geometrically more complicated than $\mathbb{R}^d$. We are unaware of any previous application of Laplacian regularization to SAE problems, though spline smoothing is often used in spatial statistics, including traditional SAE applications to disease mapping (Kafadar 1996).

3 Smoothing for Small Area Estimation

In this section, we provide a generalized approach for SAE. First, we introduce notation used throughout the paper (Section 3.1). Second, we provide a general result for SAE in the context of smoothing (Section 3.2). Third, we provide special cases of this result under the area-level Fay-Herriot model (Section 3.3).

3.1 Notation and Terminology

In this section, we present general notation that is used throughout the remainder of the paper. We assume m areas, and for each area i, we estimate an associated scalar quantity θ_i, collectively $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. We denote the response (direct estimator) by $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$. "Areas" are often spatial regions, but they might be different demographic groups. Our goal is to estimate the unknown parameter $\boldsymbol{\theta}$ by some estimator $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$, where we denote the optimal estimator by $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_m)$.

Denote the ith area by a (column) vector of covariates

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix},$$

which may include spatial coordinates. We can represent the covariates as a design matrix in the following way:

$$X_{m \times p} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mp} \end{pmatrix}.$$

One Bayesian treatment of this problem of finding an optimal estimator is to define a loss function, and then minimize the posterior risk. That is, under a defined loss function $L(\boldsymbol{\theta}, \boldsymbol{\delta})$, one goal will be to minimize the posterior risk with respect to the estimator $\boldsymbol{\delta}$.

Turning to the loss function, we will assume for convenience and for the desirability of tractable solutions that our loss function is a weighted squared error, where the weight for area i is $\phi_i > 0$, which can be denoted by a matrix of weights, Φ. The total loss is denoted by

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{m} \phi_i (\theta_i - \delta_i)^2 = (\boldsymbol{\theta} - \boldsymbol{\delta})^T \Phi (\boldsymbol{\theta} - \boldsymbol{\delta}).$$

In many SAE applications, the weights Φ reflect variations in measurement precision and can be estimated from the survey design (Pfeffermann 2013; Rao and Molina 2015; Tzavidis et al. 2018). There exists a large literature regarding proposals for estimating loss function weights. Isaki et al.

(2004) proposed taking each weight as the reciprocal of the posterior variance of the Bayes estimator. For a full review of such choices for the loss function weights, we refer to Datta et al. (2011), and stress that the choice of the loss function weights is application specific.

Under our proposed methodology, we simply assume that a Bayes estimator exists, and under this framework, the modeling structure can be set by the user. Of course, in certain situations, the Bayes estimator is the same as other estimators in the frequentist literature, such as the Best Linear Unbiased Predictor (BLUP). A full review of such cases can be found in Molina and Rao (2010) and Ghosh et al. (1994).

3.2 General Result

In this section, we propose our general framework for smoothing in SAE. Before doing so, we introduce new terminology that is needed for the remainder of the paper. Consider two different areas i and i', where $i \neq i'$. We define a symmetric matrix, Q, with elements $q_{ii'} \geq 0$, to control how important it is that the estimate of θ_i is close to the estimate of $\theta_{i'}$. It may often be the case that $q_{ii'} = q(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$; i.e., the degree of smoothing of δ_i and $\delta_{i'}$ is a function of the covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$. Note also that the $q_{ii'}$ may be discrete-valued, corresponding to clustering of areas, or continuous-valued, corresponding to a metric space of areas. Writing Q in matrix form, we see that

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{pmatrix}.$$

A natural measure of the smoothness of δ_i is the Q-weighted sum of squared differences between elements, $\sum_{i=1}^m \sum_{i'=1}^m (\delta_i - \delta_{i'})^2 q_{ii'}$, where for the remainder of the paper we denote $\sum_{i=1}^m \sum_{i'=1}^m$ as $\sum_{i,i'}$. We add a penalty term

$$\gamma \sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'}$$

to our objective function, with the penalty factor $\gamma \geq 0$ chosen to specify the overall importance of smoothness. (We address the choice of Q below and of γ in Section 4.3.)

Therefore, we seek to minimize the posterior risk of the loss function

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_i \phi_i (\theta_i - \delta_i)^2 + \gamma \sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'}. \tag{3.1}$$

Minimizing the posterior expectation of equation 3.1 is equivalent to minimizing

$$\sum_i \phi_i E[(\theta_i - \delta_i)^2|\boldsymbol{y}] + \gamma \sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'}. \tag{3.2}$$

Finally, we define Ω to be a matrix such that $\sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'} = \boldsymbol{\delta}^T \Omega \boldsymbol{\delta}$, where Ω is a semi-positive definite matrix. See Lemma A in the Supplementary Material for the above equivalence.

Then minimizing equation 3.2 is equivalent to minimizing

$$(\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B)^T \Phi (\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B) + \gamma \boldsymbol{\delta}^T \Omega \boldsymbol{\delta}, \tag{3.3}$$

where $\hat{\boldsymbol{\theta}}^B = (\hat{\theta}_1^B, \ldots \hat{\theta}_m^B)$. See Datta et al. (2011) and Ghosh and Steorts (2013) for details on this equivalence. Then we have the following result.

Theorem 3.1. *The smoothed Bayes estimator is*

$$\tilde{\boldsymbol{\theta}}^S = (I_m + \gamma \Phi^{-1} \Omega)^{-1} \hat{\boldsymbol{\theta}}^B.$$

Proof. Differentiating equation 3.3 with respect to $\boldsymbol{\delta}$ and setting the gradient to zero at $\tilde{\boldsymbol{\theta}}^S$ yields $\Phi(\tilde{\boldsymbol{\theta}}^S - \hat{\boldsymbol{\theta}}^B) + \gamma \Omega \tilde{\boldsymbol{\theta}}^S = \mathbf{0}$. Then

$$(\Phi + \gamma \Omega) \tilde{\boldsymbol{\theta}}^S = \Phi \hat{\boldsymbol{\theta}}^B \implies \tilde{\boldsymbol{\theta}}^S = (I_m + \gamma \Phi^{-1} \Omega)^{-1} \hat{\boldsymbol{\theta}}^B.$$

Since equation 3.3 is a positive-definite quadratic form in $\boldsymbol{\delta}$, the solution is unique. □

3.3 Area-Level Fay-Herriot Model

In this section, we consider a special case of our general result in Section 3.2, where our only assumption was that the Bayes estimate exists. In this section, we consider a special case of Theorem 3.1 in Section 3.2, where we assume the standard Fay-Herriot model (Fay and Herriot 1979).

Before proceeding, we review the Fay-Herriot model and a few standard results that follow from assuming this model, which is a special case of the general framework proposed in Section 3.2. More specifically, we consider the area-level Fay-Herriot model

$$\begin{aligned} \boldsymbol{y}_{m \times 1} &= \boldsymbol{\theta}_{m \times 1} + \boldsymbol{e}_{m \times 1} \\ \boldsymbol{\theta}_{m \times 1} &= X_{m \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{u}_{m \times 1}, \end{aligned} \tag{3.4}$$

where

$$\boldsymbol{e}_{m \times 1} \overset{ind}{\sim} \text{MVN}(\mathbf{0}, D) \quad \text{and} \quad \boldsymbol{u}_{m \times 1} \overset{ind}{\sim} \text{MVN}(\mathbf{0}, \sigma_u^2 I_m), \tag{3.5}$$

and

$$D_{m\times m} = \text{Diag}(D_1, \ldots, D_m),$$

$$B_{m\times m} = \text{Diag}(D_1(\sigma_u^2 + D_1)^{-1}, \ldots, D_m(\sigma_u^2 + D_m)^{-1}).$$

Note that $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and the rank$(X) = p(< m)$. Equation 3.4 was first considered in the context of estimating income for small areas by Fay and Herriot (1979).

If both σ_u^2 or D are unknown, then the model is not identifiable. This is seen by writing the marginal distribution of $\boldsymbol{y}$, which can be shown to be $\boldsymbol{y} \sim \text{MVN}(X\boldsymbol{\beta}, V)$, where $V = \text{Diag}(\sigma_u^2 + D_1, \ldots, \sigma_u^2 + D_m)$. There is clearly an identifiability issue when both σ_u^2 or D are unknown. (In a Bayesian setting, this is the marginal distribution of $\boldsymbol{y}$ after integrating out $\boldsymbol{\theta}$).

When the variance component σ_u^2 is known and $\boldsymbol{\beta}$ has a uniform prior on $\mathbb{R}^p$, then the Bayes estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^B = \hat{\boldsymbol{\theta}}^{\text{BLUP}} = (I_m - B_{m\times m})\boldsymbol{y}_{m\times 1} + B_{m\times m}X_{m\times p}\tilde{\boldsymbol{\beta}}_{p\times 1}, \tag{3.6}$$

where $\tilde{\boldsymbol{\beta}} \equiv \tilde{\boldsymbol{\beta}}(\sigma_u^2) = (X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y}$. It is well known that under the conditions given above, the Bayes estimator is also the best unbiased predictor of $\boldsymbol{\theta}$ (Datta and Ghosh 1991; Datta et al. 2011; Rao and Molina 2015).

In more realistic settings, σ_u^2 is unknown and must be estimated. Thus, the empirical Bayes estimator becomes

$$\hat{\boldsymbol{\theta}}^{EB} = (I_m - \hat{B}_{m\times m})\boldsymbol{y}_{m\times 1} + \hat{B}_{m\times m}X_{m\times p}\tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2)_{p\times 1}. \tag{3.7}$$

Note that $\hat{\sigma}_u^2$ in equation 3.7 can be estimated many different ways. For example, many common estimators are moment estimators or maximum likelihood estimation. Moment estimation is quite convenient in particular situations, as shown in Prasad and Rao (1990) and Steorts and Ghosh (2013). However, more general techniques can be found in Rao and Molina (2015). Thus, in the standard Fay-Herriot situation, where one considers this very specialized situation, the Bayes and frequentist estimates are the same, and one may find these estimates without resorting to Markov chain Monte Carlo (MCMC).

Assuming the standard area-level Fay-Herriot model in equation 3.4, we prove a special case of Theorem 3.1 in Lemma 3.1.

Lemma 3.1. *Assume the loss function in equation 3.3 and assume the area-level Fay-Herriot model in equation 3.4. Consider two choices of the Bayes estimate, which are* $\hat{\boldsymbol{\theta}}^{BLUP}$ *and* $\hat{\boldsymbol{\theta}}^{EBLUP}$ *and are given in equations 3.6 and 3.7. The smoothed best linear unbiased Bayes (SBLUP) estimator is*

$$\tilde{\boldsymbol{\theta}}^{SBLUP} = (I_m + \gamma\Phi^{-1}\Omega)^{-1}\hat{\boldsymbol{\theta}}^{BLUP}.$$

and the smoothed empirical best linear unbiased Bayesian (SEBLUP) estimator is

$$\tilde{\boldsymbol{\theta}}^{SEBLUP} = (I_m + \gamma\Phi^{-1}\Omega)^{-1}\hat{\boldsymbol{\theta}}^{EBLUP}.$$

Proof. Under the assumption of the Fay-Herriot model and by equations 3.6 and 3.7, the results follows by direction substitution into Theorem 3.1. □

As already mentioned in Section 3.1, there are many ways to choose the loss function weights, and this is typically application specific. We define the loss function weights used in our application in Section 6.

4 Benchmarking and Smoothing

We now turn to situations where our estimates should not just be smooth, minimizing equation 3.3, but also obey benchmarking constraints. As the benchmarking constraints are relaxed, we should recover the results of Section 3.

Definition 4.1 (Benchmarking constraints, benchmarked Bayes estimator)**.** Benchmarking constraints *are equality constraints on the weighted means or weighted variances of subsets (possibly all) of the estimates. The* benchmarked Bayes estimator *is the minimizer of the posterior risk subject to the benchmarking constraints.*

The levels to which we benchmark, i.e., the values of the equality constraints, are assumed to be given *externally* from some other data source. For internal benchmarking, we refer to Bell et al. (2013). Our methods address linear, weighted mean constraints, in a similar manner to that of Datta et al. (2011) and Ghosh and Steorts (2013); however, our results are more general.

4.1 General Linear Benchmarking Constraints

We now return to our general assumptions in Section 3.2. We first consider benchmarking constraints which are linear in the estimate $\boldsymbol{\delta}$, such as means or totals. The general problem is now to minimize the posterior risk in equation 3.3 subject to the constraints

$$M\boldsymbol{\delta} = \boldsymbol{t}, \tag{4.1}$$

where $\boldsymbol{t}$ is a given k-dimensional vector and M is a $k \times m$ matrix. This is equivalent to introducing a k-dimensional vector of Lagrange multipliers $\boldsymbol{\lambda}$ and minimizing $(\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B)^T \Phi (\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B) + \gamma \boldsymbol{\delta}^T \Omega \boldsymbol{\delta} - 2\boldsymbol{\lambda}^T (M\boldsymbol{\delta} - \boldsymbol{t})$. The full details on this equivalence can be found in Datta et al. (2011).

Theorem 4.1. *Suppose that equation 4.1 has solutions. Then the constrained Bayes estimator under the constraint in equation 4.1 is*

$$\tilde{\boldsymbol{\theta}}^{BM} = \Sigma^{-1}\Big[\Phi\hat{\boldsymbol{\theta}}^B + M^T(M\Sigma^{-1}M^T)^{-1}\Big(\boldsymbol{t} - M\Sigma^{-1}\Phi\hat{\boldsymbol{\theta}}^B\Big)\Big],$$

where $\Sigma = \Phi + \gamma\Omega$.

Remark 4.1. *Note that the Theorem 4.1 estimator* $\tilde{\boldsymbol{\theta}}^{BM}$ *can be expressed in terms of the Theorem 3.1 estimator* $\tilde{\boldsymbol{\theta}}^S$ *as*

$$\tilde{\boldsymbol{\theta}}^{BM} = \tilde{\boldsymbol{\theta}}^S + \Sigma^{-1}M^T(M\Sigma^{-1}M^T)^{-1}(\boldsymbol{t} - M\tilde{\boldsymbol{\theta}}^S).$$

Thus, it can be seen that the benchmarking essentially "adjusts" the estimator $\tilde{\boldsymbol{\theta}}^S$ *based on the discrepancy between* $M\tilde{\boldsymbol{\theta}}^S$ *and the target* $\boldsymbol{t}$.

Proof of Theorem 4.1. Differentiating with respect to $\boldsymbol{\delta}$ and setting the result equal to zero at $\tilde{\boldsymbol{\theta}}^{BM}$ yields

$$\begin{aligned} M^T\boldsymbol{\lambda} &= \Phi(\tilde{\boldsymbol{\theta}}^{BM} - \hat{\boldsymbol{\theta}}^B) + \gamma\Omega\tilde{\boldsymbol{\theta}}^{BM} \\ &\implies \tilde{\boldsymbol{\theta}}^{BM} = \Sigma^{-1}(\Phi\hat{\boldsymbol{\theta}}^B + M^T\boldsymbol{\lambda}). \end{aligned}$$

Then by the constraint,

$$\begin{aligned} \boldsymbol{t} &= M\Sigma^{-1}(\Phi\hat{\boldsymbol{\theta}}^B + M^T\boldsymbol{\lambda}) \\ &= M\Sigma^{-1}\Phi\hat{\boldsymbol{\theta}}^B + M\Sigma^{-1}M^T\boldsymbol{\lambda}, \end{aligned} \tag{4.2}$$

so $\boldsymbol{\lambda} = (M\Sigma^{-1}M^T)^{-1}(\boldsymbol{t} - M\Sigma^{-1}\Phi\hat{\boldsymbol{\theta}}^B)$. The result follows immediately. □

Often there is only one linear constraint of the form $\sum_i w_i\delta_i = t$, or equivalently $\boldsymbol{w}^T\boldsymbol{\delta} = t$, for some nonnegative weights w_i and some $t \in \mathbb{R}$. This is simply a special case of Theorem 4.1 with $k = 1$ and $M = \boldsymbol{w}^T$, in which case the result simplifies to

$$\tilde{\boldsymbol{\theta}}^{BM} = \tilde{\boldsymbol{\theta}}^S + (t - \boldsymbol{w}^T\tilde{\boldsymbol{\theta}}^S)(\boldsymbol{w}^T\Sigma^{-1}\boldsymbol{w})^{-1}\Sigma^{-1}\boldsymbol{w}.$$

Geometric Interpretation: Our formulation of benchmarking and smoothing as constrained optimization problems has a geometric interpretation. It is well known that the Bayes estimate is the minimizer of the conditional expectation of the MSE. Since the minimization is taken over *all* possible values of $\boldsymbol{\theta}$, the Bayes estimate will not respect any constraints we might wish to impose (except by chance) or unless these constraints are included in the specification of the prior. Instead, we minimize the MSE within the feasible set of the constraints. We find the point in the feasible set which is as close (in the sense of expected weighted squared error) to the Bayes estimate as possible. That is, we *project* the Bayes estimate into the feasible set.

The geometry of the feasible set is itself slightly complicated, because of the constraints imposed. Note that the smoothness penalty in the loss function may be reformulated as a smoothness constraint of the form $\boldsymbol{\delta}^T \Omega \boldsymbol{\delta} \leq s$ for some $s > 0$. This constraint defines an ellipsoid centered at the origin. Constraints on weighted means define linear sub-spaces, e.g., planes, depending on the number of constraints and the number of variables.

Remark 4.2. *We do not consider benchmarked constraints of weighted variabilities in this paper as the problem is non-convex. Geometrically, constraints of weighted variabilities define the surfaces of cones. The constrained Bayes estimator is the projection of the unconstrained Bayes estimator onto the intersection of the ellipsoid, the linear sub-space, and the cones. We return to this in our discussion of future work.*

4.2 Area-Level Fay-Herriot Model with Benchmarking and Smoothing

In this section, we return to the assumptions of the area-level Fay-Herriot model in Section 3.3 and equation 3.4, which allows us to derive a special case of our generalized approach for smoothing and benchmarking in Lemma 4.1.

Lemma 4.1. *Let us assume the conditions of Theorem 4.1. In addition, assume the area-level Fay-Herriot model in equation 3.4. Finally, let us assume that the Bayes estimator is either $\hat{\boldsymbol{\theta}}^{BLUP}$ or $\hat{\boldsymbol{\theta}}^{EBLUP}$. It immediately follows that the smoothed, benchmarked BLUP and EBLUP are*

$$\tilde{\boldsymbol{\theta}}^{SB\text{-}BLUP} = \tilde{\boldsymbol{\theta}}^{SBLUP} + (t - \boldsymbol{w}^T \tilde{\boldsymbol{\theta}}^{SBLUP})(\boldsymbol{w}^T \Sigma^{-1} \boldsymbol{w})^{-1} \Sigma^{-1} \boldsymbol{w} \tag{4.3}$$

and

$$\tilde{\boldsymbol{\theta}}^{SB\text{-}EBLUP} = \tilde{\boldsymbol{\theta}}^{SEBLUP} + (t - \boldsymbol{w}^T \tilde{\boldsymbol{\theta}}^{SEBLUP})(\boldsymbol{w}^T \Sigma^{-1} \boldsymbol{w})^{-1} \Sigma^{-1} \boldsymbol{w}, \tag{4.4}$$

where

$$\tilde{\boldsymbol{\theta}}^{SBLUP} = (I_m + \gamma \Phi^{-1} \Omega)^{-1} \hat{\boldsymbol{\theta}}^{BLUP}.$$

and the smoothed empirical best linear unbiased Bayesian (SEBLUP) estimator is

$$\tilde{\boldsymbol{\theta}}^{SEBLUP} = (I_m + \gamma \Phi^{-1} \Omega)^{-1} \hat{\boldsymbol{\theta}}^{EBLUP}.$$

Proof. In this situation, the result follows directly from Lemma 3.1. □

4.3 Choice of Smoothing Penalties

The choice of γ is assumed to be fixed *a priori*. But knowing γ is equivalent to knowing how smooth the estimate *ought* to be, and this knowledge is lacking in most applications. In such situations, we suggest obtaining γ by leave-one-out cross-validation (Corona et al. 2008; Stone 1974; Wahba 1990).

For each value of γ and each area i, define $\boldsymbol{\delta}^{(-i)}(\gamma)$ as the solution of the corresponding optimization problem with the loss-function term for area i dropped.[1] The smoothness penalty and any applicable benchmarking constraints are calculated over the *whole* of the vector $\boldsymbol{\delta}$, not just the non-i entries. (This ensures that $\boldsymbol{\delta}^{(-i)}(\gamma)$ does meet all the constraints, while still making a *prediction* about θ_i.)

The cross-validation score of γ is

$$V(\gamma) = \frac{1}{m} \sum_{i=1}^{m} \left[\delta_i^{(-i)}(\gamma) - \hat{\boldsymbol{\theta}}_i^B \right]^2 \phi_i,$$

where $\delta_i^{(-i)}(\gamma)$ denotes the ith component of $\boldsymbol{\delta}^{(-i)}(\gamma)$, and the minimizer of the cross-validation scores is $\hat{\gamma} = \operatorname{argmin}_{\gamma \geq 0} V(\gamma)$. Direct evaluation of $V(\gamma)$ can be computationally costly. See Wahba (1990) for faster approximations, such as "generalized cross-validation."

5 Mean Squared Error Estimation

It is traditional in SAE to report approximations to the overall prediction error. This is generally a challenging undertaking, since methods like cross-validation can be used to evaluate *prediction* error in a way which is comparable across models, but they do not work for *estimation* error. Thus, one needs to use more strictly model-based approaches, either analytic or based on the bootstrap.

[1] Instead of the sum of squared errors $\sum_{i'=1}^{m} \phi_{i'}(\delta_{i'} - \theta_{i'})^2$, we use $\sum_{i' \neq i} \phi_{i'}(\delta_{i'} - \theta_{i'})^2$. This amounts to replacing Φ with a matrix whose ith row and column are both 0.

Evaluating the MSE of our estimates is especially difficult, since we combine a model-based estimate with a non-parametric smoothing term. A straightforward model-based bootstrap would sample from the posterior distribution of equation 3.4 to generate a new set of true estimates $\boldsymbol{\theta}^*$ and observations $\boldsymbol{y}^*$, re-run the estimation on $\boldsymbol{y}^*$, and see how close the resulting estimates $\boldsymbol{\delta}^*$ came to $\boldsymbol{\theta}^*$. However, this presumes the correctness of the Fay-Herriot model in equation 3.4, which is precisely what we have chosen *not* to assume through our imposition of the benchmarking/smoothing constraints[2]. Note that such constraints do not fit naturally into the generative model.

We evade this dilemma by using a non-parametric bootstrap, a common approach when the functional form of a regression is known fairly securely but the distribution of the error terms is not. The bootstrap assumes that smoothing is appropriate and that we have chosen the right Ω matrix. We discuss the choice of Ω and the smoothing penalty in Section 6.3 for the application on the Empirica database. We assume that the loss function weight for the ith area (ϕ_i) is the inverse of the estimated MSE.

5.1 Non-parametric Bootstrap

In this section, we describe the use of a non-parametric bootstrap in order to estimate the MSE. Assume the area-level Fay-Herriot model in equation 3.4; however, the assumed model here can be a parametric, non-parametric, or semi-parametric. Assume an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. For example, one could consider $\hat{\boldsymbol{\theta}}^{EB}$ in equation 3.7, such as the Bayes estimator or the empirical Bayes estimator. In addition, assume a constrained Bayes estimator $\hat{\boldsymbol{\theta}}^{BM}$, such as a benchmarked estimator or a smoothed benchmarked estimator.

For convenience, we can re-write the Fay-Herriot model in equation 3.4 as the following:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{e}. \tag{5.1}$$

Now, define the residuals as $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\theta}}$ as in Carpenter et al. (2003). There are other non-parametric bootstraps that have been utilized in the small area literature that can be found in the work of Opsomer et al. (2008) and Molina et al. (2009). Next, center and scale the residuals $\boldsymbol{r}$ and the estimated random effects $\hat{\boldsymbol{u}}$, where we denote these by $\boldsymbol{r}^c_{\boldsymbol{e}}$ and $\boldsymbol{r}^c_{\boldsymbol{u}}$, respectively. These are centered at $\mathbf{0}$ and scaled to ensure that they have empirical covariances

[2] If we follow this procedure nonetheless, we always conclude that benchmarking and especially smoothing radically increase the MSE by introducing large biases.

equal to D and $\hat{\sigma}_u I_m$, respectively. Next, we bootstrap and re-sample the centered and scaled random effects $\boldsymbol{r}^c_{\boldsymbol{u}}$, and the residuals $\boldsymbol{r}^c_{\boldsymbol{e}}$, in the following way:

$$\boldsymbol{u}^* \overset{\text{iid}}{\sim} \{\boldsymbol{r}^c_{\boldsymbol{u}}\} \tag{5.2}$$

$$\boldsymbol{e}^* \overset{\text{iid}}{\sim} \{\boldsymbol{r}^c_{\boldsymbol{e}}\} \tag{5.3}$$

$$\boldsymbol{y}^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\hat{\sigma}^2_u) + \boldsymbol{u}^* + \boldsymbol{e}^*. \tag{5.4}$$

All the residuals are re-sampled independently.

In equations 5.2 and 5.3, we draw (with replacement) independent and identically distributed (iid) random variables $u^*_1, \ldots, u^*_m$ and $e^*_1, \ldots, e^*_m$ where each u^*_i is equal to each of $r^c_1, \ldots, r^c_m$ and each e^*_i is equal to each of $r^c_1, \ldots, r^c_m$ with probability $1/m$ respectively.

Re-sampling–based bootstraps are commonly used in assessing uncertainty for regression models. They presume the correctness of the functional form of the regression, but not of distributional assumptions about the noise. To summarize, the resampling procedure proceeds in the following way:

1. From data $(\boldsymbol{x}, \boldsymbol{y})$, obtain estimates $\hat{\boldsymbol{\theta}}$ and centered and scaled residuals $\boldsymbol{r}^c_{\boldsymbol{u}}$ and $\boldsymbol{r}^c_{\boldsymbol{e}}$.
2. Repeat B times:
   (a) Draw $\boldsymbol{u}^*$ and $\boldsymbol{e}^*$ by resampling with replacement from $\boldsymbol{r}^c_{\boldsymbol{u}}$ and $\boldsymbol{r}^c_{\boldsymbol{e}}$ respectively.
   (b) Set $\boldsymbol{y}^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\hat{\sigma}^2_u) + \boldsymbol{u}^* + \boldsymbol{e}^*$.
   (c) Refit the model on $(\boldsymbol{x}, \boldsymbol{y}^*)$ to obtain $\hat{\boldsymbol{\theta}}^*$.
   (d) Calculate the constrained Bayes estimate $\hat{\boldsymbol{\theta}}^{BM*}$.
3. Use the distribution of $\hat{\boldsymbol{\theta}}^*$ and $\hat{\boldsymbol{\theta}}^{BM*}$ in bootstrap calculations to obtain the estimated MSE.

Thus, we have proposed a non-parametric bootstrap, where we define this using the unconstrained estimates. This is important to ensure that the bootstrap produces synthetic data closer to the observed data. This can of course be checked for a few situations under the Fay-Herriot model, and we do so in our application given previous work done by Prasad and Rao (1990).

6 Application: Estimating Rental Prices in Berlin

In this section, our goal is to estimate the average rent in each of the 447 low geographical areas called *Lebensweltlich orientierte Räume* (LORs) in Berlin in 2015. The Berlin Senate Department for Urban Development and Housing is officially responsible for providing official comparative rents for the consumer price index in Berlin. This official data set is comprised of roughly 2,000 apartments. Furthermore, it is collected by the Federal Statistical Office in Berlin-Brandenburg using a stratified sampling design. The strata are defined by type of the apartment, districts, and type of the landlord.[3] Unfortunately, this survey is confidential due to privacy constraints, and it is not possible to access this database.

In order to mimic the official data set collected by the Federal Statistical Office in Berlin-Brandenburg, Empirica provided us with a similar data set for 2,000 apartments available for rent in Berlin in 2015. The Empirica data set was obtained via web-scraping and print media. The Empirica data set creates the following two new challenges: (1) reliable estimates at the LOR level are not available due to the very small or zero sample sizes in some LORs and (2) the sample from the Empirica data set may fail to capture important parts of the Berlin rental market. Therefore, we combine direct estimates from the sample of the Empirica data set with small area models that use area level predictors. The small area model is described in detail in Section 6.2. To match the official rent per square meter in Berlin we incorporate a benchmarking constraint (i.e. the fixed amount of rent set by law for the city of Berlin) and investigate the effect of smoothing in Section 6.3. The spatial distribution of rent prices in Berlin is also discussed. Before proceeding, we first further describe the Empirica data set in Section 6.1.

6.1 The Empirica Data Set

The Empirica data set is chosen (in order to mimic the data collected by the Federal Statistical Office in Berlin-Brandenburg) according to a stratified sampling design (strata are defined using the region and the size of the apartment) with a sample size of around 2,000. There are a total of 100 covariates such as rental price per square meter (excluding costs for water, sewage, trash collection, etc.), number of bedrooms and bathrooms, year of construction, balcony, and the address (including longitude and latitude coordinates).

[3] We refer to https://www.statistik-berlin-brandenburg.de/publikationen/aufsaetze/2016/HZ_201602-04.pdf for further information.

The 48 strata are defined by the cross-classification of the 12 districts with the living space categories (four categories: $< 40m^2$, $40-60m^2$, $60-90m^2$, $> 90m^2$). This leads to a sample size of 2,083 apartments with 302 in-sample LORs and 142 out-of-sample LORs. The summary statistics of the sample sizes by LORs are presented in Table 1.

Table 1: Summary statistics over LOR (Empirica Database)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample size	3	4	5	6.90	9	24

Direct estimation—using only the sample data—of the average rental price per square meter is not an option because direct estimates are only available for 302 out of the 447 LORs. In addition, for some areas where sample data is available, the small samples sizes lead to direct estimates with a low precision. As such, we attempt to improve the precision of small area direct estimates by combining the direct estimates from the sample of the database with small area models. The small area model from which we derive our initial estimates is described in Section 6.2. It provides an estimate of the average rental price per square meter at LOR level in Berlin based on the Empirica database.

In addition, because rentals often directly change hands from outgoing tenants to incoming tenants in Berlin, a market is not covered by online and print media sources. More specifically, in this secondary market, it is likely that the rental price remains constant. For this reason, we may expect some overestimation of the average rental price per square meter by using the Empirica database. To adjust for the potential lack of representativeness of the sample data, we incorporate a benchmarking constraint such that the weighted mean of the average LOR rental price estimates matches the official rent per square meter of €8.02 in Berlin as published by the Berlin Senate Department for Urban Development and Housing. In addition to benchmarking, we add a spatial smoothness constraint across LORs since we may expect rental prices to be spatially related. Our choice of the Laplacian and smoothing penalty is discussed in Section 6.3.

6.2 The Fay-Herriot Model applied to Empirica

In this section, we describe our methodology that we use for analysis and estimation. In the context of the Empirica data set, $\boldsymbol{\theta}$ denotes the true rental price per square meter for all the LORs, $\boldsymbol{y}$ denotes the direct estimates based on the survey from the Empirica data set, $D = \mathrm{Diag}(D_1, \ldots, D_m)$ denotes the sampling covariance matrix of the direct estimates $\boldsymbol{y}$, σ_u^2 denotes the area-level variance parameter, $\boldsymbol{x}_i$ denotes the vector of covariates, and $\boldsymbol{\beta}$ denotes the unknown regression coefficients. Our initial unconstrained estimates for the average rental price per square meter are derived from the area-level Fay-Herriot model in equation 3.4 (Fay and Herriot 1979).

The final model is selected following the ideas by Marhuenda et al. (2014). In particular, we used a Kullback symmetric divergence criterion with a bootstrap adjustment, KICb2. The final model includes seven aggregated (LOR-level) predictors obtained from the Empirica data set: 1) the average year of construction, 2) the average floor of the apartment in the building, and 3-7) share of apartments with an energy performance certificate available (EPC)/ balcony/ elevator/ fitted kitchen/ open fireplace, respectively. The distribution of the predictors over LORs is presented in Table 2. The inclusion of these covariates led to an $R^2 = 52\%$ for the linear model at the aggregated level.

Table 2: Summary statistics over LOR

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Year of construction	1909	1936	1955	1956	1973	2010
Floor	0.44	1.77	2.27	2.50	2.83	7.64
Share of EPC	0.00	0.55	0.66	0.64	0.75	1.00
Share of balcony	0.36	0.62	0.72	0.72	0.83	1.00
Share of elevator	0.00	0.16	0.31	0.36	0.54	0.98
Share of fitted kitchen	0.000	0.33	0.47	0.47	0.60	0.94
Share of open fireplace	0.00	0.00	0.01	0.02	0.02	0.28

The Fay-Herriot model is estimated by using the *emdi* package in R (Kreutzmann et al. 2019). Figure 1 presents the average rental price per square meter based on the Fay-Herriot estimator. We observe that the most expensive parts of Berlin are around the city centre and the area in the south-west (Zehlendorf and Grunewald) of Berlin, which is consistent with official results by the Berlin Senate.

6.3 Benchmarking/Smoothing applied to Empirica

Figure 1 offers a first picture about the rent per square meter at the LOR-level in Berlin using a sample from the Empirica data set. As already mentioned, the Empirica data set may exclude certain parts of the rental market.

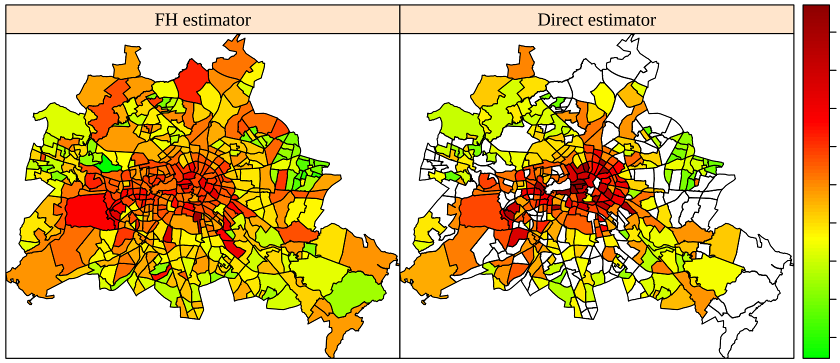

Figure 1: Average rent per square meter in € based on the unconstrained Fay-Herriot estimator (left map) and the direct estimator (right map).

To correct for this, we incorporate a benchmarking constraint requiring that the weighted mean of the average rental price estimates matches the official rent per square meter of €8.02 in Berlin and in addition consider smoothing the estimates over space.

In particular, we consider the following two options for benchmarking and/or smoothing: (i) benchmarking the mean without spatial smoothing, and (ii) benchmarking the mean with spatial smoothing. We expect that smoothing will reduce the variability in the resulting benchmarked estimates. In addition, in each case (i) and (ii), we choose the benchmarked weights w_i to be proportional to the number of apartments in each LOR. There is a rich literature on the choice of such weights. We refer to Ghosh and Steorts (2013) and Datta et al. (2011) for further details.

Figure 2 compares the unconstrained Fay-Herriot estimator to the benchmarked Fay-Herriot estimator. As expected, the Fay-Herriot estimates of the average rental price per square meter are higher than the benchmarked estimates for each of the 447 LOR. Observe that the benchmarked Fay-Herriot estimates are on average around €0.348 lower than the unconstrained Fay-Herriot estimates. This is expected as the Empirica data set excludes the secondary rental market, which means that rental prices in the Empirica data set tend to be lower than those advertised online or in print media. Intuitively, properties that are advertised in the open market may ignore the law on rental prices, altogether.

Turning now to the use of spatial smoothing, the most important part of the smoothing procedure is selecting the matrices Ω and Q. Recall that

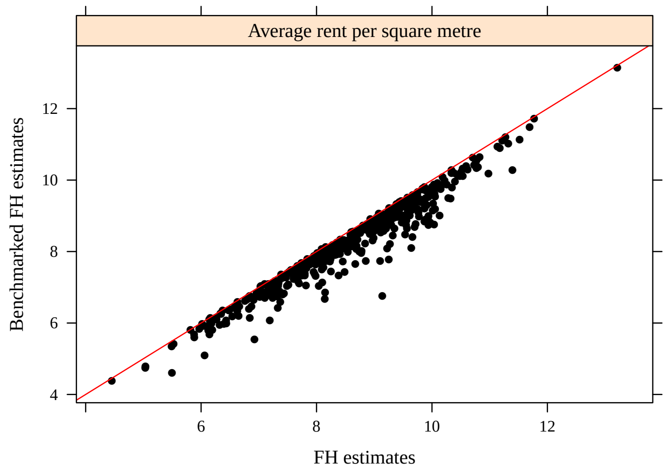

Figure 2: Average rent per square meter in € using (a) the benchmarked Fay-Herriot estimates and (b) the unconstrained Fay-Herriot estimates.

Ω is used to measure the smoothness of estimates; and Q shows how similar the estimates for any two domains should be. This is inevitably application-specific. In our application, we utilize a simple choice, where where $q_{ii'} = 1$ if the LORs i and i' shared a border, and 0 otherwise. This treats the LOR as nodes in an unweighted graph, with Q being its adjacency matrix and Ω its Laplacian. In addition, we considered several alternative ways of smoothing the Fay-Herriot estimates. One can choose $q_{ii'}$ such that it decreases with the geographic distance between LORs, regarding the points at their respective centers. A second approach was to treat the 12 districts in Berlin as clusters, setting $q_{ii'} = 1$ for LORs within a cluster and $q_{ii'} = 0$ for LORs between them, but neither of these two approaches worked well under cross-validation. Note that choosing the spatial smoothing parameter is not an issue in our application as we do not encounter spatial islands, however, if one does encounter such issues, we would recommend modifying the definition of a neighbor to be the minimum geographic distance criterion.

As described in Section 4.3, the smoothing factor γ was picked by leave-one-out cross-validation and the final value was $\gamma \approx 0.146$. Figure 3 shows the smoothed and benchmarked Fay-Herriot estimates versus the unconstrained Fay-Herriot estimates. In general, the effect of spatial smoothing causes an upward adjustment of low values of the unconstrained Fay-Herriot estimates, and causes a downwards adjustment of higher unconstrained Fay-

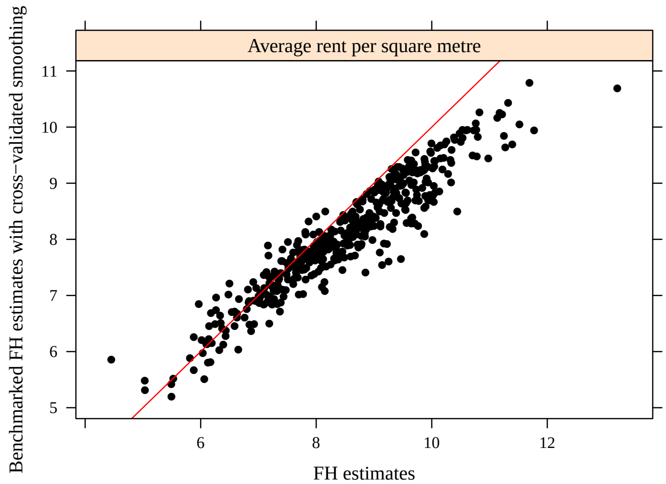

Figure 3: Average rent per square meter in € using (a) the benchmarked Fay-Herriot estimates with cross-validated smoothing and (b) the unconstrained Fay-Herriot estimates.

Herriot estimates. However, the majority of the unconstrained Fay-Herriot estimates are pulled down as a result of both smoothing and benchmarking. This observation is further confirmed by Figure 4 where the effect of combining smoothing with benchmarking is illustrated this time for the twelve districts in Berlin. Observe that in each of the twelve districts the smooth benchmarked Fay-Herriot estimates fall on the line with a slope of less than 1.

Table 3 reports the MSEs under the non-parametric bootstrap of Section 5 for different combinations of benchmarking and smoothing. In particular, *FH* denotes the unconstrained Fay-Herriot estimates, *FH Bench* the benchmarked estimates and *FH Bench/Smooth* the corresponding benchmarked estimates with cross-validated smoothing. The results are based on $B = 1000$ bootstrap replications. In addition, we ran a Fay-Herriot model with spatially correlated random effects, *FH SAR*, using the same adjacency matrix Q used for the *FH Bench/Smooth*. We followed Pratesi and Salvati (2009) and used a SAR specifications for the random effects. The MSE of the *FH SAR* is estimated by a non-parametric bootstrap ($B = 1000$) as proposed by Molina et al. (2009). Please note that in the case of MSE estimation under the benchmarked approach with spatially correlated random effects, *FH SAR Bench*, benchmarking is being implemented with each bootstrap

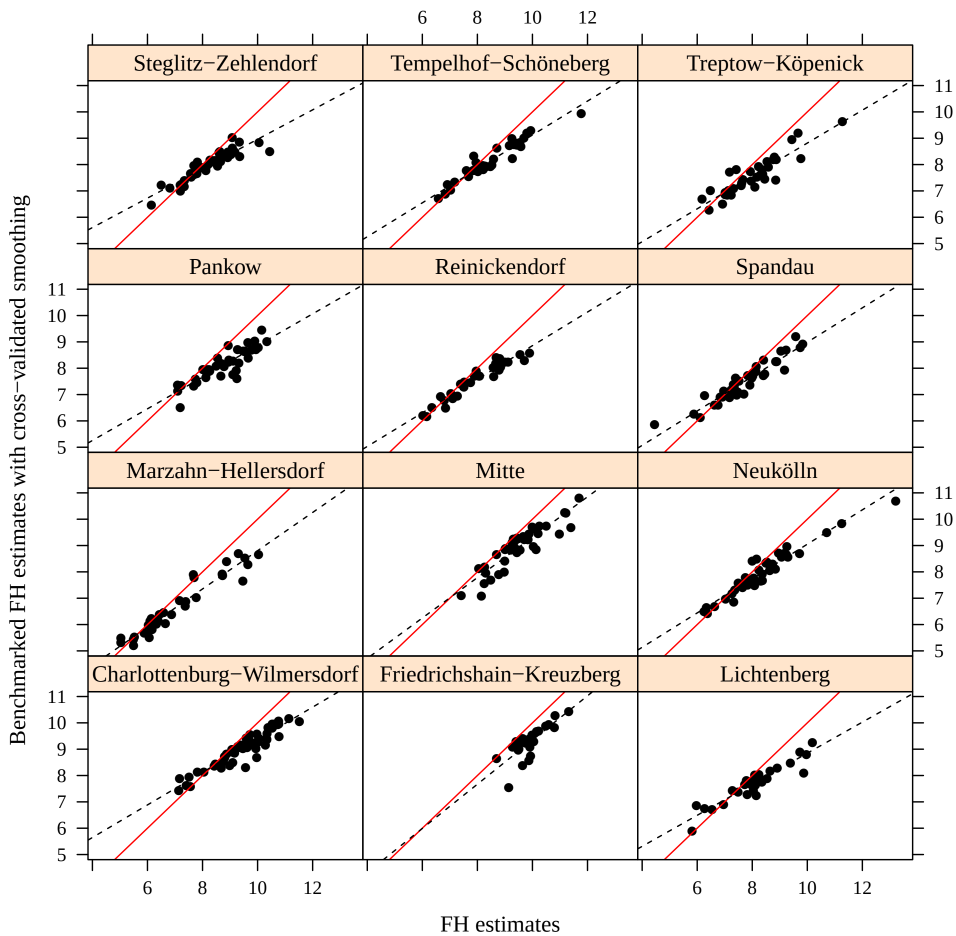

Figure 4: Benchmarked and Spatially Smoothed Fay-Herriot estimates versus unconstrained Fay-Herriot estimates, by region. Large unconstrained Fay-Herriot estimates are adjusted downwards by the benchmarked and spatially smoothed Fay-Herriot estimators, while small unconstrained Fay-Herriot estimators are adjusted upwards. This effect can be seen by the dotted line, denoting the regression line, and the red line, denoting the intersection line.

sample. First, we observe that the unconstrained Fay-Herriot estimates (*FH* and *FH SAR*) have a smaller MSE compared to the benchmarked estimates. This is expected as the constraint in the estimation introduces additional variability. However, as benchmarking is required in the application, we focus on the three constrained estimates (*FH Bench*, *FH SAR Bench* and *FH Bench/Smooth*). Incorporating the spatial effect via smoothing or a SAR structure reduces the variability for most LORs compared to the benchmarked estimates (*Bench*). In addition, the estimated MSEs under the *FH SAR Bench* approach and the *FH Bench/Smooth* are comparable. In particular, on average both methods provide similar estimated MSEs with the Fay-Herriot benchmarked and smooth approach also offering less extreme estimated MSEs.

Table 3: Summary statistics of RMSE estimates over LOR

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FH	0.07	0.60	0.67	0.63	0.71	0.89
FH SAR	0.10	0.46	0.64	0.61	0.74	2.17
FH SAR Bench	0.10	0.46	0.67	0.71	0.86	2.91
FH Bench/Smooth	0.07	0.67	0.74	0.75	0.81	1.86
FH Bench	0.07	0.68	0.75	0.77	0.85	1.80

Having assessed the variability of the three constrained estimates, we have a closer look to the point estimates of the actual rent per square meter in Berlin. Figure 5 presents the benchmarked estimates with and without cross-validated smoothing and the benchmarked Fay-Herriot with spatially correlated random effects. Overall, all maps reflect the current situation of the rental market for apartments in Berlin with higher rents in the city center and in the district Steglitz-Zehlendorf (in the south-west), whereas the districts of Spandau (in the west) and Marzahn-Hellersdorf (in the east) have lower rents compared to other parts in Berlin. For instance, vast housing estates (plattenbau style - large panel system building) were built in the 1980s in Marzahn-Hellersdorf. Most of the plattenbau apartments were built in large settlements on the edge of Berlin making them inconveniently located leading to high vacancy rates and low rent prices. In contrast, the district of Steglitz-Zehlendorf consists of very affluent localities like Dahlem or Zehlendorf. The localities Nikolasee and Wannsee of Steglitz-Zehlendorf are located around the forest of Grunewald and two lakes (Greater and Little Wannsee). These localities are some of the most expensive areas in Berlin for housing.

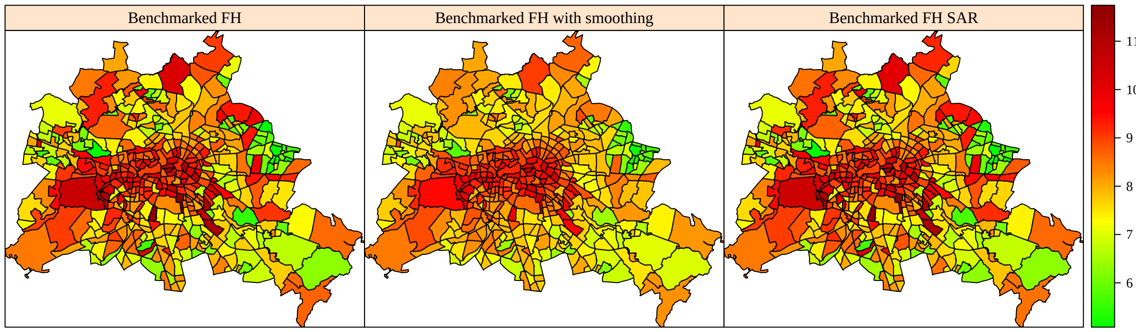

Figure 5: Average rent per square meter in € based on the benchmarked Fay-Herriot estimator with (middle) and without (left map) cross-validated smoothing and the benchmarked Fay-Herriot with spatially correlated random effects (right map).

However, we also observe some differences between the three maps. First, some LORs in the city center around the governmental quarter (with the Reichstag building, German Chancellery, and Bellevue Palace) have lower than expected rental prices based on the benchmarked estimates (*FH Bench* and *FH SAR Bench*) in Figure 5 and on the unconstrained Fay-Herriot estimates in Figure 1. Additional discrepancies occur for LORs in the suburbs in the north (for instance the locality Blankenfelde in the district Pankow) and south-east (for instance the locality Karolinenhof in the district Treptow-Köpenick). In the latter case LORs have higher than expected rental prices based on the benchmarked estimates (*FH Bench* and *FH SAR Bench*) in Figure 5 and on the unconstrained Fay-Herriot estimates in Figure 1. These localities are bordering the federal state of Brandenburg, are located in rural parts of Berlin and they are the least densely populated areas in Berlin. In addition, Figure 1 reveals that sample information is missing from these LORs, and thus, we heavily rely on the Fay-Herriot equation 3.4. It is possible that especially for LORs with somehow different characteristics (infrastructure and environment) our estimates based on equation 3.4 suffer from some misspecification. Nevertheless, it appears that our benchmarked estimates with cross-validated smoothing are able to adjust the estimates and protect against potential model misspecification.

7 Discussion

We have provided a general approach to area-level SAE, where we smooth and benchmark model-based estimates. Our approach yields closed-form solutions without requiring any distributional assumptions. Furthermore, our results apply for linear and non-linear estimators. Finally, we show in the application that smoothing has the potential to improve estimation of rental prices on LOR level in Berlin for most LORs.

We now outline some possible extensions, namely extensions to weighted variability constraints, moving beyond squared error loss, and moving from point estimation to full posterior estimates for maximal flexibility. As mentioned earlier, working beyond a weighted mean constraint and with both a weighted mean and weighted variability would be a more general benchmarking framework. The question of how to incorporate variability constraints while maintaining tractability of the model is a potential direction of future research and is beyond the scope of this paper, as the problem may not always be a convex optimization problem. In addition, throughout our paper, we have worked with the squared error loss function. However, it should be possible to replace this with any other loss function. Once the Bayes estimate is obtained, the constrained Bayes estimate would be found by a projection onto the corresponding feasible set.

This would presumably mean a need for using numerical optimization when the optimization problem in not tractable. Finally, it may be possible to go beyond point estimates to distributional estimates. Given a sample from the posterior distribution (e.g., from MCMC), it is possible to project each sample point into the feasible set, giving rise to a posterior distribution whose support respects the constraints. This idea is related to that of Dunson and Neelon (2003), however, cannot be directly adapted to our setting. Dunson and Neelon (2003) have proposed constrained Bayes estimation through a posterior projection approach, which is appealing in the sense that one fully achieves a Bayesian posterior distribution to the constrained optimization problem. The constraints considered by the authors are ordered parameters, and do not easily generalize to both weighted means and weighted variabilities in our general framework.

Acknowledgements

Schmid and Tzavidis are supported by ES/N011619/1 — *Innovations in Small Area Estimation Methodologies* from the UK Economic and Social

Research Council. Tzavidis is also supported by the InGRID 2 EU-Horizon 2020 infrastructure grant (http://www.inclusivegrowth.eu). The authors thank Empirica-Systeme GmbH (www.empirica-systeme.de) for providing the data set used in the application. The ideas of this paper are of the authors and not of the funding organizations or the data providers. Finally, the authors thank the Editor and the reviewers for comments that significantly improved the paper. The authors also thank David Banks for providing minor comments regarding manuscript.

References

Battese, G., Harter, R., and Fuller, W. (1988), "An Error-Components Model for Prediction of County Crop Area Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006), "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, 7, 239–2434.

Bell, W., Datta, G., and Ghosh, M. (2013), "Benchmarked Small Area Estimators," *Biometrika*, 100, 189–202.

Carpenter, J. R., Goldstein, H., and Rasbash, J. (2003), "A novel bootstrap procedure for assessing the relationship between class size and achievement," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 431–443.

Corona, E., Lane, T., Storlie, C., and Neil, J. (2008), "Using Laplacian Methods, RKHS Smoothing Splines and Bayesian Estimation as a framework for Regression on Graph and Graph Related Domains," Tech. Rep. TR-CS-2008-06, Department of Computer Science, University of New Mexico.

Datta, G. and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *The Annal of Statistics*, 19, 1748–1770.

Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2011), "Bayesian benchmarking with applications to small area estimation," *TEST*, 20, 574–588.

Dunson, D. B. and Neelon, B. (2003), "Bayesian inference on order-constrained parameters in generalized linear models," *Biometrics*, 59, 286–295.

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

Ghosh, M. (1992), "Constrained Bayes estimation with applications," *Journal of the American Statistical Association*, 87, 533–540.

Ghosh, M., Rao, J., et al. (1994), "Small area estimation: an appraisal," *Statistical science*, 9, 55–76.

Ghosh, M. and Steorts, R. C. (2013), "Two-Stage Bayesian Benchmarking as Applied to Small Area Estimation," *TEST*, 22, 670–687.

Isaki, C., Tsay, J., and Fuller, W. (2004), "Weighting sample data subject to independent controls," *Survey Methodology*, 20, 35–44.

Kafadar, K. (1996), "Smoothing Geographical Data, Particularly Rates of Disease," *Statistics in Medicine*, 15, 2539–2560.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019), "emdi: estimating and mapping disaggregated Indicators," *Journal of Statistical Software*, 91, 1–33.

Lee, A. B. and Wasserman, L. (2010), "Spectral Connectivity Analysis," *Journal of the American Statistical Association*, 105, 1241–1255.

Louis, T. (1984), "Estimating a population of parameter values using Bayes and empirical Bayes methods," *Journal of the American Stastical Association*, 79.

Marhuenda, Y., Morales, D., and del Carmen Pardo, M. (2014), "Information criteria for Fay–Herriot model selection," *Computational Statistics and Data Analysis*, 70, 268 – 280.

Molina, I. and Rao, J. (2010), "Small area estimation of poverty indicators," *Canadian Journal of Statistics*, 38, 369–385.

Molina, I., Salvati, N., and Pratesi, M. (2009), "Bootstrap for estimating the MSE of the Spatial EBLUP," *Computational Statistics*, 24, 441–458.

Newman, M. E. J. (2010), *Networks: An Introduction*, Oxford, England: Oxford University Press.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. (2008), "Non-parametric small area estimation using penalized spline regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 265–286.

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

Prasad, N. and Rao, J. (1990), "The estimation of the mean squared error of small-area estimators," *Journal of the American Stastical Association*, 85, 163–171.

Pratesi, M. and Salvati, N. (2008), "Small area estimation: the EBLUP estimator based on spatially correlated random area effects," *Statistical methods and applications*, 17, 113–141.

— (2009), "Small Area Estimation in the Presence of Correlated Random Area Effects," *Journal of Official Statistics*, 25 (1), 37–53.

Rao, J. and Molina, I. (2015), *Small Area Estimation*, John Wiley & Sons, New York.

Rao, J. N. K. and Yu, M. (1994), "Small-area estimation by combining time-series and cross-sectional data," *Canadian Journal of Statistics*, 22, 511–528.

Souza, D. F., Moura, F., and Migon, H. (2009), "Small area population prediction via hierarchical models," *Catalogue no. 12-001-X*, 203.

Steorts, R. C. and Ghosh, M. (2013), "On estimation of mean squared errors of benchmarked empirical Bayes estimators," *Statistica Sinica*, 749–767.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society B*, 36, 111–147.

Tzavidis, N., Luna, A., Zhang, L. C., Schmid, T., and Rojas-Perilla, N. (2018), "From start to finish: A framework for the production of small area official statistics," *Journal of the Royal Statistical Society: Series A*, 181, 927–979.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

Wehbe, L., Ramdas, A., Steorts, R. C., and Shalizi, C. R. (2015), "Regularized Brain Reading with Shrinkage and Smoothing," *Annals of Applied Statistics*, 9, 1997–2022.

Supplementary Material

A Lemma on Squared Differences

Lemma A.1. *For a suitable matrix* Ω*,*

$$\sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'} = \boldsymbol{\delta}^T \Omega \boldsymbol{\delta}.$$

Proof. Begin by expanding the square and collecting terms:

$$\begin{aligned}
&\sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{ii'} \\
&= \sum_{i,i'} \delta_i^2 q_{ii'} + \sum_{i,i'} \delta_{i'}^2 q_{ii'} - 2 \sum_{i,i'} \delta_i \delta_{i'} q_{ii'} \\
&= \sum_{i} \delta_i^2 \sum_{i'} q_{ii'} + \sum_{i'} \delta_{i'}^2 \sum_{i} q_{ii'} - 2 \sum_{i,i'} \delta_i \delta_{i'} q_{ii'}.
\end{aligned}$$

Now define the diagonal matrix $Q^{(r)}$ with elements $q_{ii}^{(r)} = \sum_{i'} q_{ii'}$, and define the diagonal matrix $Q^{(c)}$ with elements $q_{jj}^{(c)} = \sum_i q_{ij}$. Substituting,

$$\begin{aligned}
\sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{i,i'} &= \boldsymbol{\delta}^T Q^{(r)} \boldsymbol{\delta} + \boldsymbol{\delta}^T Q^{(c)} \boldsymbol{\delta} - 2 \boldsymbol{\delta}^T Q \boldsymbol{\delta} \\
&= \boldsymbol{\delta}^T \left(Q^{(r)} + Q^{(c)} - 2Q \right) \boldsymbol{\delta},
\end{aligned}$$

which defines Ω. □

Remark A.1. *In an unweighted, undirected graph with adjacency matrix* A*, the degree matrix* D *is defined by* $D_{ii} = \sum_j A_{ij}$*,* $D_{ij} = 0$*; the graph Laplacian in turn is* $L = D - A$ *(Newman 2010). If* Q *is an adjacency matrix, then* $Q^{(r)} = Q^{(c)} = D$*, and* $\Omega = 2L$*.*

Remark A.2. *By construction,* Ω *is clearly positive semi-definite. It is not positive definite, because* $(1\ 1\ \cdots\ 1)$ *is always an eigenvector, of eigenvalue zero. This corresponds to the fact that adding the same constant to each* δ_i *does not change* $\sum_{i,i'} (\delta_i - \delta_{i'})^2 q_{i,i'}$*. (These are of course basic properties of graph Laplacians.)*