Universität Trier

Penalized Small Area Models for the Combination of Unit- and Area-level Data

Jan Pablo Burgard
Joscha Krause
Ralf Münnich

Research Papers in Economics
No. 5/19

Penalized small area models for the combination of unit- and area-level data

Jan Pablo Burgard, Joscha Krause, Ralf Münnich

Department of Economic and Social Statistics, Trier University

Abstract

The joint usage of unit- and area-level data for model-based small area estimation is investigated. The combination of levels within a single model encloses a variety of methodological problems. Firstly, it implies a critical decrease in degrees of freedom due to more model parameters that need to be estimated. This may destabilize model predictions in the presence of small samples. Secondly, unit- and area-level data has different distributional characteristics in terms of dispersion patterns and correlation structure. Thirdly, unit- and area-level data is usually subject to different kinds of measurement errors. We propose a multi-level model with level-specific penalization to overcome these issues and use unit- and area-level data jointly for model-based small area estimation. An application is provided on the example of regional health measurement in Germany. We combine health survey data on the unit-level and aggregated micro census records on the area-level to estimate hypertension prevalence.

Keywords: *disease mapping*, *multi-level model*, *multi-source estimation*, *penalization*, *stochastic gradient descent*

1 Introduction

Small area estimation (SAE) is frequently applied to obtain reliable estimates of aggregate-specific quantities (area-statistics) from small samples. A direct estimator that only uses data of one area at a time cannot produce area-statistic estimates with sufficient accuracy in that case. SAE was developed to solve this problem by combining data from multiple areas in suitable regression models. The objective is to improve estimation efficiency over a direct estimator by exploiting the functional relation between the area-statistic of interest and auxiliary data. Depending on data availability and privacy regulations in the field of application, unit- or area-level models are used for this purpose. The main difference between these model types is the aggregation level of the auxiliary data they consider for model parameter estimation. The original area-level model was proposed by Fay and Herriot (1979) and uses auxiliary data on the area-level. The original unit-level model was proposed by Battese et al. (1988) and considers auxiliary data below the area-level. For a comprehensive overview on SAE, we refer to Rao and Molina (2015). The efficiency gain of a small area estimator over a direct estimator is determined by the explanatory power of the underlying regression model. Accordingly, when auxiliary data for both unit- and area-level is available, both levels should be considered to maximize the explanatory power and thus to produce optimal area-statistic estimates.

The joint usage of unit- and area-level data for model-based SAE is not as well-established in the literature. A lot of multi-level SAE models use unit-level data while accounting for heterogeneity in fixed effects on the area-level, as for example proposed by Moura and Holt (1999). This marks an important generalization of the nested error structure in the original unit-level approach by Battese et al. (1988), which models differences between areas only via random intercept. Though, it does not allow for the direct combination of unit- and area-level data sets as area-level heterogeneity is assumed to be due to random deviations from unit-level fixed effects. Twigg et al. (2000) presented a multi-level approach to include both individual and ecological components to predict small area health-related behaviour as binary response. Still, the approach relies on a sequential procedure to calibrate a model on one data set first, and then use it in conjunction to another, which requires model specification to be very simple. Ghosh and Steorts (2013) developed a two-stage benchmarking approach that combines unit-and area-level in a single weighted squared loss function while benchmarking weighted means at both levels. However, the focus of our contribution is not on benchmarking, but on a direct and straightforward combination of unit- and area-level data.

Using unit- and area-level data in this manner encloses some methodological problems. Firstly, it requires model parameter estimation at both levels simultaneously. In the presence of small samples, the increased number of parameters may lead to considerably high model parameter estimate variances due to the lack in degrees of freedom. Model-based small area estimates then also suffer from high variance and are not reliable. Secondly, unit- and area-level data tend to have different distributional characteristics and correlation structures due to different degrees of aggregation (Clark and Avery, 1976). As a result, the levels should not be treated equally in terms variable selection

or model parameter estimation. Thirdly, unit- and area-level data is usually subject to different kinds of measurement errors. While unit-level data may suffer from false outliers (e.g. due to wrong coding), area-level data might be uncertain because its values are estimated. As ignoring measurement errors leads to suboptimal area statistic estimates, the researcher should account for this (Lohr and Ybarra, 2008). And finally, unit- and area-level data usually differs in terms of availability. Unit-level data is often rare due to privacy issues, whereas area-level data is less sensitive and easier to access, for example from registries. Accordingly, an approach must be able to deal with situations where there are a lot of variables on one level while there are only few on the other.

We propose to combine unit- and area-level data in a multi-level model under level-specific penalization. Level-specific penalization refers to penalized maximum likelihood estimation of the model parameters where the fixed effects on each level are penalized individually. For this purpose, ℓ_1-norm, squared ℓ_2-norm, elastic net, and mixed-norm penalties are considered. Using level-specific penalization in multi-level models for SAE solves the methodological problems mentioned before. Firstly, it allows for high-dimensional inference. Hence, even if the number of model parameters surpasses the number of observations, the underlying optimization problem for model parameter estimation is still well-posed. This is particularly attractive in the presence of small samples. Secondly, level-specific penalization marks an intuitive way to treat unit- and area-level data differently for model parameter estimation. The penalties can be defined dependent on the distributional characteristics of the corresponding auxiliary data. Further, if a sparsity-inducing penalty is chosen (e.g. ℓ_1-norm), an automatic level-specific variable selection is conducted. Thirdly, norm-based penalization implies a robustification against measurement errors in the auxiliary data (Bertsimas and Copenhaver, 2018; Burgard et al., 2019). Accordingly, level-specific penalization allows for different measurement errors on each level. And finally, the tuning parameter on each level required for penalization can be altered depending on the number of variables available for prediction.

Penalized maximum likelihood estimation of the model parameters is performed with a stochastic coordinate gradient descent algorithm using insights from Tseng and Yun (2009) as well as Schelldorfer et al. (2011). Random effect prediction is done in a Bayesian manner using a maximum a posteriori approach, as suggested by Schelldorfer et al. (2011). The methodology is tested under multiple scenarios in a simulation study. An empirical application is provided on the example of regional health measurement in Germany. We combine unit-level data from the German health survey *Gesundheit in Deutschland aktuell (GEDA)* with area-level data from aggregated micro census records to estimate regional hypertension prevalence. The remainder of the paper is organized as follows. In Chapter 2, the methodology is explained. This includes a description of the multi-level model, the level-specific penalties, and model parameter estimation. In Chapter 3, the simulation study is provided. Chapter 4 encloses the application to regional health measurement. Chapter 5 closes with an outlook on future research.

2 Methods

2.1 Multi-level model

Let $\mathcal{U}$ be a finite population of size N containing m pairwise disjoint areas of size N_i with $i = 1, ..., m$ and $\sum_{i=1}^{m} N_i = N$. Let a random sample $\mathcal{S}$ of size n be drawn from $\mathcal{U}$ such that there are m area sub-samples of size $n_i > 1$ with $\sum_{i=1}^{m} n_i = n$. Let $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$ be a vector containing observations of some response variable y from which the area-statistic of interest in area i is calculated. Let $\mathbf{X}_i^u \in \mathbb{R}^{n_i \times p^u}$ be the fixed effect design matrix in area i containing unit-level auxiliary data for the description of $\mathbf{y}_i$. Let $\mathbf{X}_i^a = (\mathbf{x}_i^a, ..., \mathbf{x}_i^a)' \in \mathbb{R}^{n_i \times p^a}$ be the fixed effect design matrix resulting from an expansion of the vector $\mathbf{x}_i^a \in \mathbb{R}^{1 \times p^a}$ containing area-level auxiliary data. Note that $(p^u + p^a) > n$ is allowed. Let $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$ be the random effect design matrix in area i with $q \leq p$. In the majority of SAE models, the random effect structure is usually limited to an area-specific random intercept. However, the general formulation of the multi-level model allows for an area-specific random effect on potentially all covariates. The multi-level model combining unit- and area-level data is given by

$$\mathbf{y}_i = \mathbf{X}_i^u \boldsymbol{\beta}^u + \mathbf{X}_i^a \boldsymbol{\beta}^a + \mathbf{Z}_i \boldsymbol{b}_i + \mathbf{e}_i \quad \forall\ i = 1, ..., m, \tag{1}$$

where $\boldsymbol{\beta}^u \in \mathbb{R}^{p^u \times 1}$, $\boldsymbol{\beta}^a \in \mathbb{R}^{p^a \times 1}$ are the fixed effect coefficient vectors for each level and $\mathbf{b}_i \sim MVN(\mathbf{0}, \boldsymbol{\Psi})$ denotes the random effect coefficient vector under multivariate normality with some general positive-definite covariance matrix $\boldsymbol{\Psi}$. $\mathbf{e}_i \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ is a vector of i.i.d. random errors with model variance parameter σ^2. Note that $\boldsymbol{b}_1, ..., \boldsymbol{b}_m$, $\mathbf{e}_1, ..., \mathbf{e}_m$ are assumed to be stochastically independent. Thus, the response vector has the following multivariate normal distribution under the model:

$$\mathbf{y}_i \sim MVN\left(\mathbf{X}_i^u \boldsymbol{\beta}^u + \mathbf{X}_i^a \boldsymbol{\beta}^a, \mathbf{V}_i(\sigma^2, \boldsymbol{\psi})\right) \quad \forall\ i = 1, ..., m, \tag{2}$$

with $\mathbf{V}_i(\sigma^2, \boldsymbol{\psi}) = \sigma^2 \mathbf{I}_{n_i} + \mathbf{Z}_i \boldsymbol{\Psi} \mathbf{Z}_i'$, where the random effect covariance matrix $\boldsymbol{\Psi}$ is parametrised by a vector $\boldsymbol{\psi}$ of dimension $q^* < n$, for example resulting from a Cholesky decomposition. Restating the model over all areas obtains

$$\mathbf{y} = \mathbf{X}^u \boldsymbol{\beta}^u + \mathbf{X}^a \boldsymbol{\beta}^a + \mathbf{Z}\boldsymbol{b} + \mathbf{e}, \tag{3}$$

with $\mathbf{X}^u = ((\mathbf{X}_1^u)', ..., (\mathbf{X}_m^u)')'$, $\mathbf{X}^a = ((\mathbf{X}_1^a)', ..., (\mathbf{X}_m^a)')'$, $\mathbf{Z} = diag(\mathbf{Z}_1, ..., \mathbf{Z}_m)$ as stacked matrices and $\mathbf{y} = (\mathbf{y}_1', ..., \mathbf{y}_m')'$, $\mathbf{b} = (\mathbf{b}_1, ..., \mathbf{b}_m)'$, $\mathbf{e} = (\mathbf{e}_1', ..., \mathbf{e}_m')'$ as stacked vectors. Define $\boldsymbol{\theta} := (\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^k, \boldsymbol{\psi}, \sigma^2) \in \mathbb{R}^{p^u + p^a + q^* + 1}$ as the full parameter vector. The negative loglikelihood function is then:

$$-\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left[n \cdot log(2\pi) + log\left(|\mathbf{V}|\right) + \left(\mathbf{y} - \mathbf{X}^u \boldsymbol{\beta}^u - \mathbf{X}^a \boldsymbol{\beta}^a\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}^u \boldsymbol{\beta}^u - \mathbf{X}^a \boldsymbol{\beta}^a\right)\right], \tag{4}$$

with $\mathbf{V} = diag(\mathbf{V}_1, ..., \mathbf{V}_m)$ and $|\mathbf{V}|$ denoting the determinant of $\mathbf{V}$.

2.2 Penalizations

The penalties used for level-specific penalization in the multi-level model are described.

ℓ_1-norm penalty

The first penalty is the ℓ_1-norm, which is used in the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996) for linear models. It has been extended to linear mixed models (LMMLASSO) in contributions of for example Bondell et al. (2010), Ibrahim et al. (2011), or Schelldorfer et al. (2011). We use the LMMLASSO with level-specific penalization by defining individual penalization parameters $\lambda^u > 0$, $\lambda^a > 0$ for unit- and area-level. The resulting objective function is then given by

$$\mathcal{Q}_{\ell_1}(\boldsymbol{\theta}, \lambda^u, \lambda^a) = -\mathcal{L}(\boldsymbol{\theta}) + \lambda^u \sum_{j \in \mathcal{X}^u} |\beta_j^u| + \lambda^a \sum_{j \in \mathcal{X}^a} |\beta_j^a|, \tag{5}$$

with $\beta_j^u \in \boldsymbol{\beta}^u, \beta_j^a \in \boldsymbol{\beta}^a$ and $\mathcal{X}^u, \mathcal{X}^a$ denoting index sets that correspond to the auxiliary variables in $\mathbf{X}^u, \mathbf{X}^a$. The ℓ_1-norm marks a sparsity-inducing penalty in the sense that fixed effect coefficients irrelevant for the functional description of the response variable will not only be shrunken towards zero, but set exactly to zero in the estimation process. Thus, estimation and variable selection are performed simultaneously. Due to the level-specific penalization, the level of sparsity and shrinkage on each level can be controlled individually. If e.g. λ^u is raised, the sparsity in $\boldsymbol{\beta}^u$ is increased and more β_j^u are set to zero. However, note that the level of shrinkage has implications for bias in the fixed effect coefficient estimation. See Fan and Li (2001) as well as Zou (2006) for further details.

Squared ℓ_2-norm penalty

The next penalty is the squared ℓ_2-norm, which is used in ridge regression proposed by Hoerl and Kennard (1970) for linear models. It has has been extended to linear mixed models in contributions of for example Eliot et al. (2011), Li and Yang (2010), or Schulz-Streeck and Piepho (2010). Again, under level-specific penalization, the corresponding objective function is given by

$$\mathcal{Q}_{\ell_2}(\boldsymbol{\theta}, \lambda^u, \lambda^a) = -\mathcal{L}(\boldsymbol{\theta}) + \lambda^u \sum_{j \in \mathcal{X}^u} (\beta_j^u)^2 + \lambda^a \sum_{j \in \mathcal{X}^a} (\beta_j^a)^2. \tag{6}$$

Unlike the ℓ_1-norm, the squared ℓ_2-norm does not induce are sparse solution for the fixed effect coefficients, but a dense solution. The fixed effect coefficients are shrunken towards zero, but not set exactly to zero. Further, their estimated values are smoothed in the sense that the individual contributions of the corresponding predictors are equalized depending on the value of the penalization parameters. Due to the level-specific penalties, the level of smoothness and shrinkage on each level can be controlled individually. Note that because of the absence of sparsity, no automatic variable selection is conducted. The squared ℓ_2-norm has shown to be superior in predictive power relative to the ℓ_1-norm in the presence of highly correlated auxiliary data. Further details can be found in Zou and Hastie (2005).

Elastic net penalty

The third penalty is a linear combination of the ℓ_1-norm and ℓ_2-norm, which is used in the elastic net proposed by Zou and Hastie (2005) for linear models. The elastic net has been extended to linear mixed models in contributions of for example Ogutu et al. (2012) or Sidi (2017). Under level-specific penalization, the objective function can be stated as

$$\begin{aligned}\mathcal{Q}_{\ell_{1\&2}}(\boldsymbol{\theta}, \lambda^u, \lambda^a) = &- \mathcal{L}(\boldsymbol{\theta}) + \lambda^u \left[(1-\alpha^u) \sum_{j \in \mathcal{X}^u} \left(\beta_j^u\right)^2 + \alpha^u \sum_{j \in \mathcal{X}^u} |\beta_j^u| \right] \\ &+ \lambda^a \left[(1-\alpha^a) \sum_{j \in \mathcal{X}^a} \left(\beta_j^a\right)^2 + \alpha^a \sum_{j \in \mathcal{X}^a} |\beta_j^a| \right],\end{aligned} \tag{7}$$

where $\alpha^u \in [0,1], \alpha^a \in [0,1]$ are hyperparameters controlling the contribution of the ℓ_1-norm and ℓ_2-norm on each level. The elastic net can be viewed as a compromise of the LASSO and ridge regression, incorporating properties of both methods. It induces a sparse solution due to the ℓ_1-norm. However, the resulting variable selection and prediction performance has shown to be superior for highly correlated covariates relative to a ℓ_1-norm penalization only. The additional influence of the squared ℓ_2-norm allows for the anticipation of grouping structures in terms of correlation within the covariates while variable selection and stabilizes model predictions. See Zou and Hastie (2005) as well as Zou and Zhang (2009) for further details.

Mixed-norm penalty

The final penalty is a generic term for different ℓ_q-norm combinations. It basically refers to penalizations where each level is associated with an individual norm. To the best of our knowledge, mixed-norm penalties have mostly been studied in machine learning contexts and biomedical applications, e.g. by Flamary et al. (2014) or Nath et al. (2009). The objective function is given by

$$\mathcal{Q}_{mix}(\boldsymbol{\theta}, \lambda^u, \lambda^a) = -\mathcal{L}(\boldsymbol{\theta}) + \lambda^u \mathcal{P}^u(\boldsymbol{\beta}^u) + \lambda^a \mathcal{P}^a(\boldsymbol{\beta}^a), \tag{8}$$

where $\mathcal{P}(\cdot)$ denotes some level-specific penalty. Within this paper, mixed-norm penalization is achieved by associating one level with a ℓ_1-norm while associating the other with a ℓ_2-norm. This allows for simultaneous sparse and dense solutions for model parameter estimation, depending on the distributional characteristics of the auxiliary data.

2.3 Model parameter estimation

In order to estimate the model parameters under a given penalization, the following minimization problem has to be solved:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \boldsymbol{\psi}, \sigma^2 > 0, \boldsymbol{\Psi} > 0}{\arg\min} \left\{ \mathcal{Q}_{(\cdot)}(\boldsymbol{\theta}, \lambda^u, \lambda^a) = -\mathcal{L}(\boldsymbol{\theta}) + \lambda^u \mathcal{P}^u(\boldsymbol{\beta}^u) + \lambda^a \mathcal{P}^a(\boldsymbol{\beta}^a) \right\}. \tag{9}$$

For this purpose, a stochastic coordinate gradient descent algorithm is used. We primarily draw from Schelldorfer et al. (2011) as well as Tseng and Yun (2009) and modify their (block) coordinate gradient descent method by a randomized cycling order. This improves the convergence probability in the light of the non-convexity of (9), as an unfortunate series of coordinates less likely to occur (Bottou et al., 2018). Minimization via coordinate gradient descent implies that the value of the objective function is minimized gradually by updating a single element of the target parameter vector $\boldsymbol{\theta}$ at a time while keeping the others fixed. Then the remaining elements are updated accordingly in an iterative manner, such that a cyclic movement through all coordinates of $\boldsymbol{\theta}$ is achieved. This approach is particularly useful for the proposed multi-level model as it allows for an easy implementation of level-specific penalization in the estimation process.

Due to the unknown variance parameters σ^2 and $\boldsymbol{\psi}$ in the negative loglikelihood function (4), the minimization problem (9) is non-convex. This complicates model parameter estimation significantly, as the algorithm is not guaranteed to achieve the global minimum. The non-convexity of the objective functions $\mathcal{Q}_{(\cdot)}$ favours the existence of local minima, which implies that the resulting model parameter estimates may be sensitive to starting values. However, the minimization with respect to $\boldsymbol{\beta}^u, \boldsymbol{\beta}^a$ is convex under fixed variance parameters. Following the argumentation of Schelldorfer et al. (2011), this is exploited in the estimation process. For $s_t = 1, 2, ...,$ let $\mathcal{R}^{s_t}$ be the index cycling through the coordinates $\{1\}, \{2\}, ..., \{p + q^* + 1\}$ in the t-th iteration of the algorithm. Note that the order of the coordinates changes randomly after each iteration. Let $\theta^t_{\mathcal{R}^{s_t}}$ denote the s-th element of $\boldsymbol{\theta}^t$, where $\boldsymbol{\theta}^t$ is the full parameter vector in the t-th iteration. Let $d^t_{s_t}$ be the descent direction and let $h^t_{s_t} = \partial^2 \mathcal{Q}_{(\cdot)}(\boldsymbol{\theta}^t)/\partial(\theta^t_{\mathcal{R}^{s_t}})^2$ be the second partial derivative of the objective function with respect to the current element in the t-th iteration. Further, let $\mathcal{I}(\boldsymbol{\theta}^t)_{\mathcal{R}^{s_t}\mathcal{R}^{s_t}}$ denote the diagonal element of the Fisher Information matrix $\mathcal{I}(\boldsymbol{\theta}^t)$ corresponding to $\theta^t_{\mathcal{R}^{s_t}}$. Let $l \in \{u, a\}$. The stochastic coordinate gradient descent algorithm shown in Algorithm 1.

Note that if $\theta^t_{\mathcal{R}_{s_t}}$ is subject to a penalization including the ℓ_1-norm, the first and second partial derivatives don't exist because the corresponding objective function is not continuously differentiable at the origin. In that case, $h^t_{s_t}$ and $d^t_{s_t}$ are determined according to Tseng and Yun (2009). If $h^t_{s_t}$ is not truncated, Schelldorfer et al. (2011) propose an analytical update of the element that is obtained from setting $a^t_{s_t}$ and the fact that $\mathcal{L}(\boldsymbol{\theta})$ is quadratic with respect to the fixed effect coefficients. For the elastic net penalty, additional shrinkage is achieved by dividing the ℓ_1-norm solution by $1 + \lambda^j(1 - \alpha)$, as suggested by Friedman et al. (2010). We further use an active-set algorithm in the case of a sparsity-inducing penalty for a level according to Friedman et al. (2010) as well as Schelldorfer et al. (2011). This implies that in each iteration of the stochastic coordinate gradient descent algorithm elementwise minimization is only performed with respect to non-zero parameters. Once a parameter is set to zero, it remains zero throught the estimation process. This of course does not apply to the squared ℓ_2-norm, which does not induce a sparse solution.

Algorithm 1 Stochastic coordinate gradient descent

choose a starting value $\boldsymbol{\theta}^0$

while not converged **do**

  define a random cycling order s_t

  for $s_t = 1, 2, ...$:

    if $\theta^t_{\mathcal{R}^{s_t}}$ subject to ℓ_1-norm or elastic net penalization

      then $h^t_{s_t} \approx \mathcal{I}(\boldsymbol{\theta})_{\mathcal{R}^{s_t}\mathcal{R}^{s_t}}$

      if $h^t_{s_t} \neq min(max(\mathcal{I}(\boldsymbol{\theta}^t)_{\mathcal{R}^{s_t}\mathcal{R}^{s_t}}, c_{min}), c_{max})$

        then $d^t_{s_t} = median\left(\frac{\lambda^l - \frac{\partial g(\boldsymbol{\theta}^{s_t})}{\partial \theta^t_{\mathcal{R}^{s_t}}}}{h^{s_t}}, -\beta^j_{\mathcal{R}^{s_t}}, \frac{-\lambda^l - \frac{\partial g(\boldsymbol{\theta}^{s_t})}{\partial \theta^t_{\mathcal{R}^{s_t}}}}{h^{s_t}}\right)$

        Choose $a^t_{s_t}$ according to the Armijo Rule $\theta^{t+1}_{\mathcal{R}^{s_t}} = \theta^t_{\mathcal{R}_{s_t}} - a^t_{s_t} d^t_{s_t}$

      else if $\theta^t_{\mathcal{R}^{s_t}}$ subject to ℓ_1-norm penalization

        then $\theta^{t+1}_{\mathcal{R}^{s_t}} = \left[sign\left((\mathbf{y} - \tilde{\mathbf{y}})'\mathbf{V}^{-1}\mathbf{X}_{\mathcal{R}^{s_t}}\right) \cdot max\left(|(\mathbf{y} - \tilde{\mathbf{y}})'\mathbf{V}^{-1}\mathbf{X}_{\mathcal{R}^{s_t}}| - \lambda^l, 0\right)\right] / h^t_{s_t}$

      else $\theta^{t+1}_{\mathcal{R}^{s_t}} = \frac{\left[sign\left((\mathbf{y} - \tilde{\mathbf{y}})'\mathbf{V}^{-1}\mathbf{X}_{\mathcal{R}^{s_t}}\right) \cdot max\left(|(\mathbf{y} - \tilde{\mathbf{y}})'\mathbf{V}^{-1}\mathbf{X}_{\mathcal{R}^{s_t}}| - \alpha\lambda^j, 0\right)\right]}{1 + \lambda^l(1-\alpha)]} / h^t_{s_t}$

    else $h^t_{s_t} = \frac{\partial^2 \mathcal{Q}_{(\cdot)}}{\partial(\theta^t_{\mathcal{R}^{s_t}})^2}$ $\theta^{t+1}_{\mathcal{R}^{s_t}} = \theta^t_{\mathcal{R}_{s_t}} - \frac{\partial \mathcal{Q}_{(\cdot)}}{\partial \theta^t_{\mathcal{R}^{s_t}}} / h^t_{s_t}$

end while

return $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^t$

2.4 Prediction

Beside model parameter estimation, the random effects must be predicted. For this, we use maximum a posteriori estimation, as suggested by Schelldorfer et al. (2011). This is a Bayesian approach where the quantity of interest is estimated from the mode of the posterior distribution. Let f denote a normal probability density. We have

$$\begin{aligned}
\widetilde{\mathbf{b}}_i &= \arg\max_{\mathbf{b}_i} \left\{ f(\mathbf{b}_i | \mathbf{y}_1, ..., \mathbf{y}_m, \boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \boldsymbol{\psi}, \sigma^2) \right\} \\
&= \arg\max_{\mathbf{b}_i} \left\{ f(\mathbf{b}_i | \mathbf{y}_i, \boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \boldsymbol{\psi}, \sigma^2) \right\} \\
&= \arg\max_{\mathbf{b}_i} \left\{ \frac{f(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \boldsymbol{\psi}, \sigma^2) \cdot f(\mathbf{b}_i | \boldsymbol{\psi})}{f(\mathbf{y}_i | \boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \boldsymbol{\psi}, \sigma^2)} \right\} \\
&= \arg\min_{\mathbf{b}_i} \left\{ \frac{1}{\sigma^2} || \mathbf{y}_i - \mathbf{X}_i^u \boldsymbol{\beta}^u - \mathbf{X}_i^a \boldsymbol{\beta}^a - \mathbf{Z}_i \boldsymbol{b}_i ||^2 + \mathbf{b}_i' \boldsymbol{\Psi}^{-1} \mathbf{b}_i \right\}.
\end{aligned} \tag{10}$$

Solving (10) under the model assumption delivers

$$\widetilde{\mathbf{b}}_i = \left(\mathbf{Z}_i' \mathbf{Z}_i + \sigma^2 \boldsymbol{\Psi}^{-1} \right)^{-1} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i^u \boldsymbol{\beta}^u - \mathbf{X}_i^a \boldsymbol{\beta}^a \right), \tag{11}$$

which is then estimated by

$$\widehat{\mathbf{b}}_i = \left(\mathbf{Z}_i' \mathbf{Z}_i + \widehat{\sigma}^2 \widehat{\boldsymbol{\Psi}}^{-1} \right)^{-1} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i^u \widehat{\boldsymbol{\beta}}^u - \mathbf{X}_i^a \widehat{\boldsymbol{\beta}}^a \right), \tag{12}$$

using the model parameter estimates obtained from the minimization of $\mathcal{Q}_{(\cdot)}$.

Now, in order to obtain estimates for the area statistic of interest, predictions from the model must be produced. However, note that the exact procedure depends on the nature of the area statistic. If it is linear, e.g. the area-specific mean of y, then it is sufficient to use the area-specific means of unit level covariates $\mathbf{X}^u$ and the area-specific covariate values of $\mathbf{X}^a$. However, if the area statistic non-linear, then unit level predictions are required. In the following, we consider the latter case. Let $\mathbf{x}_{i\iota}^u, \mathbf{x}_{i\iota}^a$ and $\mathbf{z}_{i\iota}$ denote the fixed and random effect vector corresponding to some new individual ι in area i. The response prediction of ι under model (3) is obtained from:

$$\widehat{y}_{i\iota} = \mathbf{x}_{i\iota}^u \widehat{\boldsymbol{\beta}}^u + \mathbf{x}_{i\iota}^a \widehat{\boldsymbol{\beta}}^a + \mathbf{z}_{i\iota} \widehat{\mathbf{b}}_i. \tag{13}$$

Note that $\widehat{y}_{i\iota}$ is the empirical best predictor for $y_{i\iota}$ under the model if $\widehat{\boldsymbol{\theta}}$ is a consistent estimator for $\boldsymbol{\theta}$ (Boubeta et al., 2016). Schelldorfer et al. (2011) provided a consistency proof for high-dimensional ℓ_1-penalized LMM estimation with asymptotics $m \to \infty$ assuming bounded eigenvalues of $\mathbf{Z}_i' \mathbf{Z}_i$ and some sparsity condition for $\boldsymbol{\beta}$. However, we are not aware of corresponding proofs for high-dimensional LMM estimation with other penalizations.

3 Simulation study

3.1 Set up

A monte carlo simulation with $R = 500$ iterations ($r = 1, ..., R$) is conducted in order to test the methodology against established small area estimators in a controlled environment. For this, a synthetic population of $N = 20\,000$ individuals in $m = 100$ areas of size $N_i = 200$. In each iteration, a stratified random sample of size $n = 200$ with strata sample size $n_i = 2$ is drawn. For unit level auxiliary data, a total of 40 variables with a weak internal correlation structure is drawn from a multivariate normal with area-specific means. For area level auxiliary data, a total of 100 variables with a strong internal correlation structure is drawn from a multivariate normal. The response variable is created on the unit level according to

$$y_{i\iota} = 100 + \beta^{u1} x_{i\iota}^{u1} + \beta^{u2} x_{i\iota}^{u2} + \beta^{a1} x_i^{a1} + \beta^{a2} x_i^{a2} + v_i + e_{i\iota}, \tag{14}$$

where $v_i \sim N(0, 200^2)$ is a random area intercept and $e_{i\iota} \sim N(0, 100^2)$ is a unit level error term. For simplicity, all fixed effect regression coefficients are set to 3. Note that from the unit level and area level auxiliary data sets described above, only 2 variables per set is relevant for the functional description of y. This is done in order to include variable selection aspects in the simulation study. The area statistic of interest is the area-specific mean of the response variable:

$$\bar{y}_i = \frac{1}{N_i} \sum_{\iota=1}^{N_i} y_{i\iota}. \tag{15}$$

In order to estimate $\bar{y}_i$, the following estimators are considered:

- *LMM.Oracle*: EBLUP under the true multi-level model (14) with known covariates for all units of the population
- *FH.Oracle*: Fay-Herriot EBLUP considering the true auxiliary variables, but using the true area-specific means of $x_{i\iota}^{u1}, x_{i\iota}^{u2}$ as substitute for unit level data
- *LMM.Select*: EBLUP under a multi-level model where variable selection is performed via the corrected conditional AIC proposed by Greven and Kneib (2010)
- *LMMLASSO*: Prediction from the multi-level model (3) with uniform ℓ_1-norm penalization
- *Mixed.Ridge*: Prediction from the multi-level model (3) with uniform ℓ_2-norm penalization
- *LMMEN*: Prediction from multi-level model (3) with uniform $\ell_{1\&2}$-norm penalization
- *Multi.L1*: Prediction from the multi-level model (3) with level-specific ℓ_1-norm penalization

- *Multi.L2*: Prediction from the multi-level model (3) with level-specific ℓ_2-norm penalization
- *Multi.EN*: Prediction from the multi-level model (3) with level-specific $\ell_{1\&2}$-norm penalization
- *Multi.MX*: Prediction from the multi-level model (3) with mixed-norm penalization

Note that the EBLUP under original unit level model proposed by Battese et al. (1988) is not included since exclusively considering the unit level variables cannot provide any reasonable results given the way the response variable is generated. However, the EBLUP under the Fay-Herriot model is included, as the unit level variables can be aggregated in order to use them on the area level. The efficiency of the area-statistic estimation is evaluated in terms of the relative root mean squared error, which is given by

$$\text{Relative RMSE}(\widehat{\bar{y}}_i) = \frac{1}{R \cdot m} \sum_{r=1}^{R} \sum_{i=1}^{m} \frac{\sqrt{\left(\widehat{\bar{y}}_i^r - \bar{y}_i^r\right)^2}}{\bar{y}_i^r}. \tag{16}$$

In order to obtain more insights on the estimators' performances, we further consider the relative bias

$$\text{Relative Bias}(\widehat{\bar{y}}_i) = \frac{1}{R \cdot m} \sum_{r=1}^{R} \sum_{i=1}^{m} \frac{\left(\widehat{\bar{y}}_i^r - \bar{y}_i^r\right)}{\bar{y}_i^r} \tag{17}$$

as well as the coefficient of variation

$$\text{Coeff.Variation}(\widehat{\bar{y}}_i) = \frac{1}{R \cdot m} \sum_{r=1}^{R} \sum_{i=1}^{m} \frac{\sqrt{\left(\widehat{\bar{y}}_i^r - E(\widehat{\bar{y}}_i^r)\right)^2}}{E(\widehat{\bar{y}}_i^r)}. \tag{18}$$

3.2 Results

Table 1 shows the overall performance of the estimators in terms of the point estimation over all areas and iterations of the monte carlo simulation. As can be seen, all penalized multi-level estimators except for the Mixed.Ridge outperform the unpenalized single-level estimator FH.Oracle as well as the unpenalized multi-level estimator LMM.Select. Their relative RMSEs are considerably smaller by a range of 12.9% to 23.8%. The LMM.Select is slightly less efficient than the FH.Oracle despite using information from both the unit- and the area-level. This is most likely due to additional uncertainty resulting from multi-level covariate selection with a single-level information criterion, as the LMM.Select does not know the true model. The most efficient estimator is the LMM.Oracle, which has perfect information by knowing the true model and the covariate values for all individuals in each area. It has the smallest relative bias and the smallest relative RMSE. However, this was expected since it serves as reference estimator within the simulation due to its unrealistic information advantage. An interesting observation is that despite having perfect information, the LMM.Oracle does not have the smallest

Predictor	Relative Bias	Coeff.Variation	Relative RMSE
LMM.Oracle	0.00063	0.01977	0.01667
FH.Oracle	0.00094	0.02606	0.02313
LMM.Select	0.00026	0.02938	0.02368
LMMLASSO	0.00080	0.02157	0.01981
Mixed.Ridge	0.00343	0.01762	0.02486
LMMEN	0.00187	0.01863	0.02014
Multi.L1	0.00079	0.02148	0.01975
Multi.L2	0.00142	0.01846	0.01995
Multi.EN	0.00128	0.01951	0.01960
Multi.MX	0.00073	0.02084	0.01762

Table 1: Point estimation results

coefficient of variation. Four of the seven penalized estimators show less relative standard deviation. This is due to additional variables they can consider since they don't use the true model, but more complex model fits that further decrease variance as a result from sampling-induced dependencies. Though, the penalized estimators have a larger relative bias that makes them ultimately less efficient than the unpenalized LMM.Oracle.

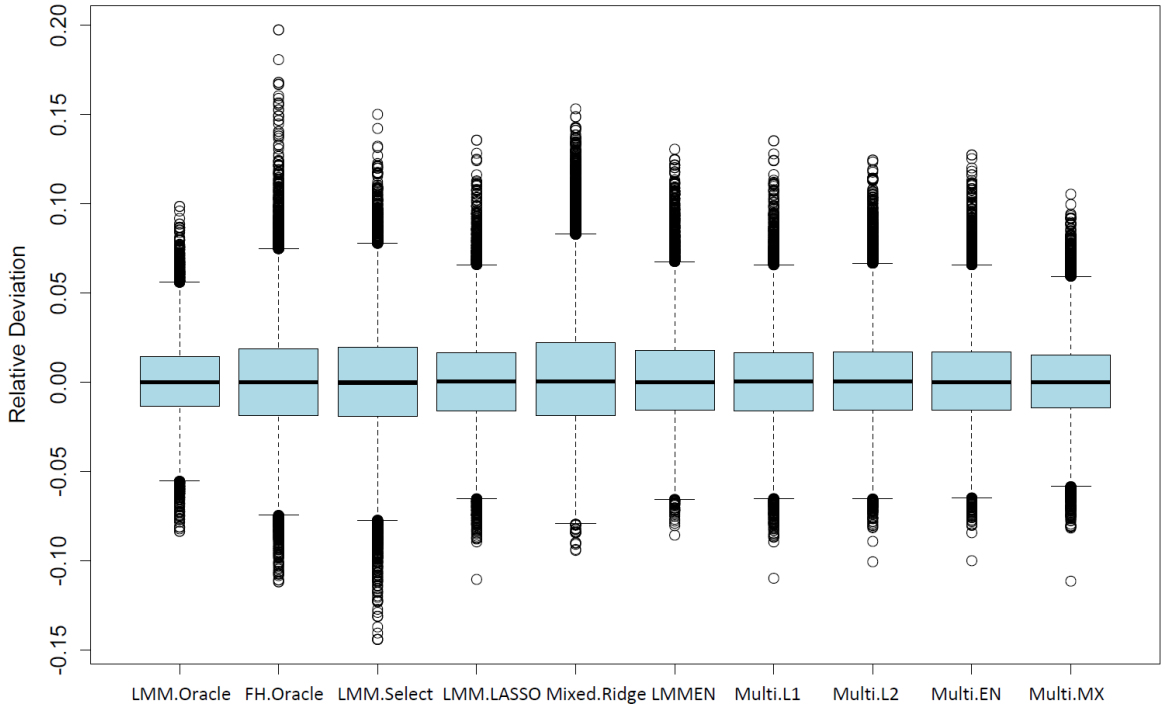

Figure 1: Boxplot of relative deviations

Figure 1 shows boxplots of the relative deviations of the estimates from the true values of the area-specific means $(\widehat{\bar{y}}_i^r - \bar{y}_i^r)/\bar{y}_i^r$. As can be seen, within the group of penalized estimators, the approaches with level-dependent penalization (Multi.L1, Multi.L2,

Multi.EN) perform better than their counterparts with uniform penalization (LMM-LASSO, Mixed.Ridge, LMMEN). The efficiency gain ranges from 0.3% to 19.8%. The overall best estimator without perfect information is the Multi.MX, which uses the ℓ_1-norm penalty on the unit-level and the ℓ_2-norm penalty on the area-level. Its efficiency is close to the reference estimator LMM.Oracle.

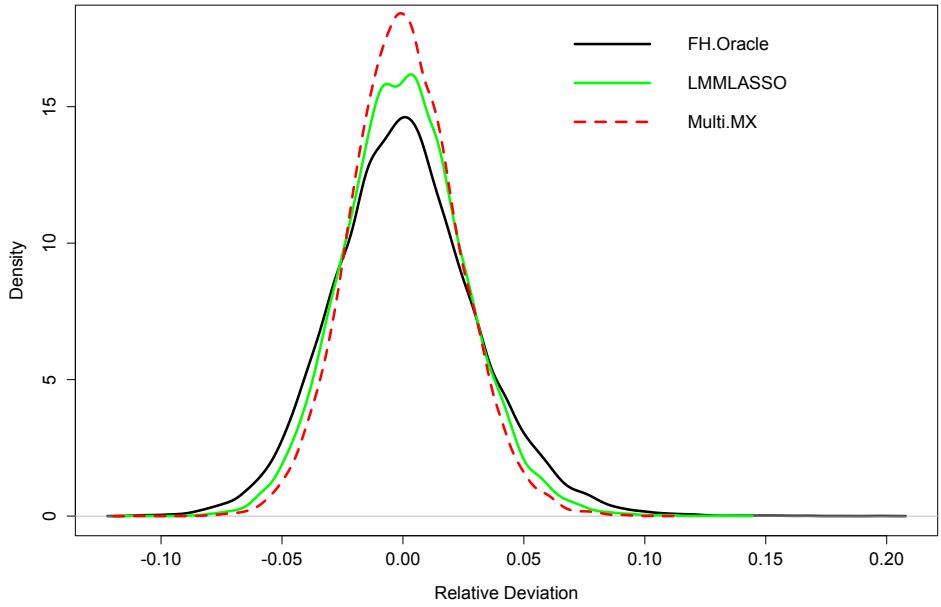

Figure 2: Density of relative deviations

The efficiency gain resulting from level-specific penalization is further visualized in Figure 2. The graph shows the density of relative deviations of three estimators with different information usage. The results from single-level estimator FH.Oracle are depicted in black, the multi-level estimator with uniform penalization LMMLASSO is displayed in green, and the multi-level estimator with level-specific penalization Multi.MX is ablined in red. As can be seen, the additional (implicit) information used by the three estimators leads to a stronger concentration of the density mass around 0. The FH.Oracle is the baseline estimator in this graph, as it only uses information on the area-level. The LMMLASSO considers information from the unit- and the area-level and is thus more efficient. However, it uses a uniform penalization that does not account for the different covariate properties on the levels. Subsequently, the Multi.MX has a visible advantage of the LMMLASSO by not only using level-specific tuning parameters, but even level-specific penalties.

Figure 3 elaborates further on the comparison between uniform and level-specific penalization. It shows point estimates of the Mixed.Ridge (black) and its counterpart, the Multi.L2 (green). As can be seen, the Mixed.Ridge has problems with capturing the tails of the distribution. The uniform shrinkage behaviour leads to solid estimates on

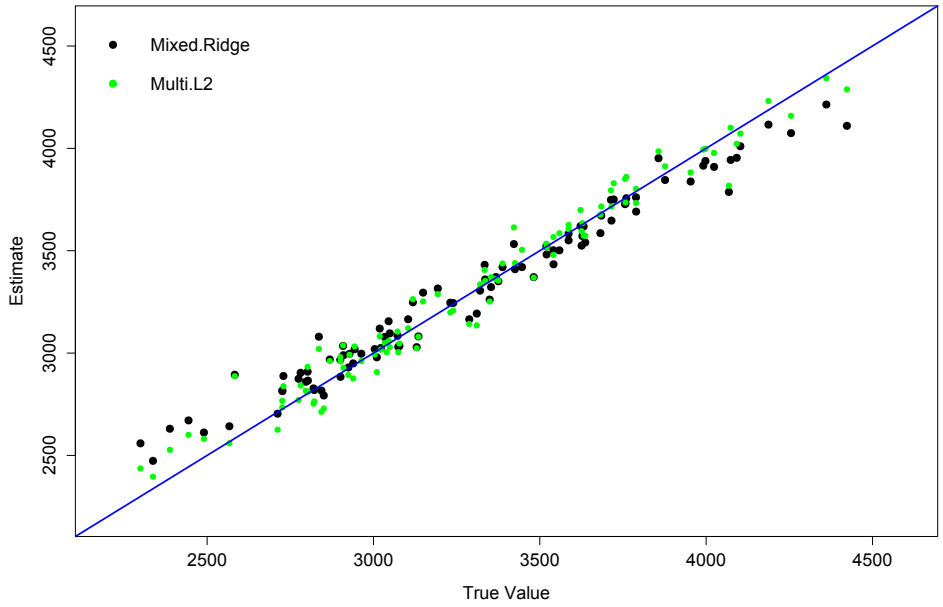

Figure 3: Scatterplot of point estimates

average, but for more extreme values, there are larger deviations. With the level-specific shrinkage behaviour of the Multi.L2, also the tails of the distribution can be captured with decent accuracy. Again, the information advantage by anticipating level-specific data properties leads to an increase in estimation efficiency. Further, by looking at Table 1, the Mixed.Ridge performs slightly worse than the unpenalized approaches FH.Oracle and LMM.Select. This is likely due to suboptimal shrinkage behaviour in the light of the strongly different level-specific correlation structures in combination with uniform penalization and dense model parameter estimates.

4 Application

The methodology is applied to health measurement in Germany. The objective is to estimate the hypertension prevalence for the population of age 18+ on regional levels. The definition of the disease profile is adapted from the Robert Koch Institute (2012). We combine two different data sources for this purpose. The first data source is the German health survey Gesundheit in Deutschland aktuell (GEDA) from 2010. It contains detailed medical and health-related information on roughly 20.000 participants of age 18+. The observations of this survey are used as unit-level data source. The second data source is aggregated records of the German micro census from 2010. The micro census is a large-scale survey that covers 1%-sample of the German population. It contains (among others) socio-demographic and economic information that we use in aggregated form on regional levels to maximize the explanatory power for hypertension

prevalence estimation. The elastic net penalty with hyperparameters $\alpha^u = \alpha^a = 0.5$ is used for penalized maximum likelihood estimation of the model parameters. The tuning parameters λ^u, λ^a are determined by k-fold cross validation with a bivariate grid search.

As the elastic net is a sparsity-inducing penalty, an automatic variable selection is conducted in the estimation process. In the following, we provide a brief overview of covariates selected for hypertension prevalence estimation. Further, we gradually increased the optimal level-specific tuning parameter values obtained from k-fold cross validation in order to assess the relevance of the corresponding fixed effect for the regional hypertension prevalence within the underlying regression model. With this, we obtain a rough measure of (pseudo-) significance for the selected covariates. We distinguish three levels of significance: strong (+++), medium (++), weak (+). From GEDA, variables were selected on the unit-level.

- Demographic variables, e.g. sex (+++), age group affiliation (+++)
- Comorbidity variables, e.g. having other cardiovascular diseases (+++)
- Lifestyle variables, e.g. smoking, drinking (+++), sport activities (+++)
- Medical care variables, e.g. visits to the doctor (++), health insurance membership (+)
- Living condition variables, e.g. degree of urbanisation (++)

From the micro census, variables were selected on the area-level. Examples are

- Socioeconomic variables, e.g. income distribution (+++), education structure (+++)
- Labour market variables, e.g. working time, industrial sectors (++), unemployment (++)
- Ethnical variables, e.g. population structure in terms of nationalities (+)

Using the mentioned variables for regional hypertension prevalence estimation obtains the following results. Figure 4 is a heat map of Germany in which the estimated hypertension prevalence per federal state are displayed. The nationwide hypertension prevalence is at 26.8%. This is consistent with the results of the Robert Koch Institute (2012), which calculated a survey-based 95%-confidence interval of [25.9%; 27.6%]. By looking at the federal state estimates, one can see that the lowest prevalence is located in the south of the country, which consists of the federal states Baden-Württemberg and Bavaria. The highest prevalence can be found in the east of the country, which is the former territory of the German Democratic Republic. The estimated regional distribution is plausible, as in past studies similar distributions of closely related diseases, like diabetes mellitus type 2, have been found (see e.g. Schipf et al., 2014).

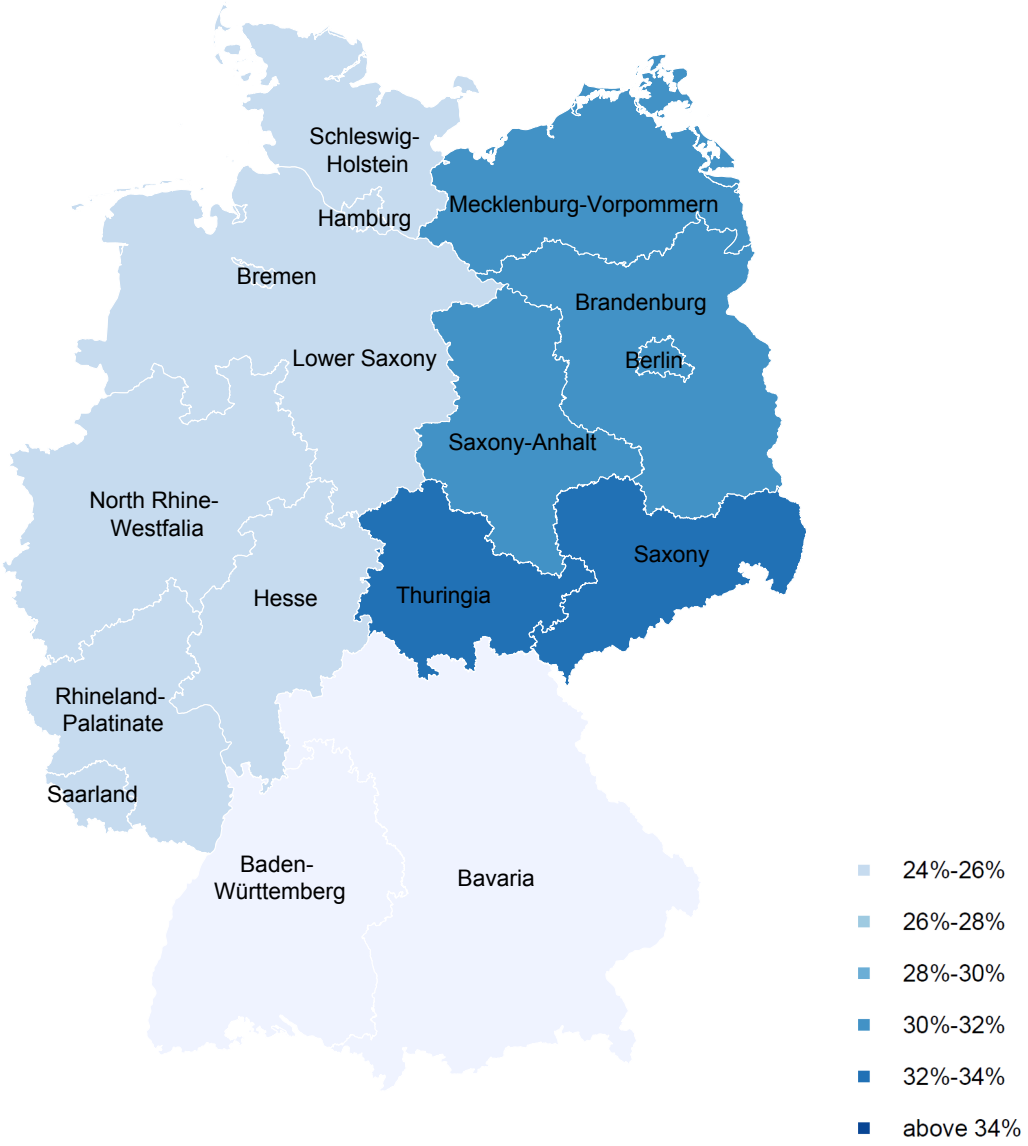

Figure 4: Estimated hypertension prevalence

5 Outlook

A multi-level model for the joint usage of unit- and area-level data was proposed. The model allows to combine multi-level data from different data sources to optimize model-based small area estimation by maximizing the explanatory power of the underlying regression model. The methodological problems associated with the level combination are solved by level-specific penalization using the LASSO, the ridge penalty, and the elastic net. Regularization parameter tuning is done via k-fold cross validation. Model parameter estimation is performed by a stochastic gradient descent algorithm. For random effect prediction, a maximum a posteriori approach is used.

Future research may focus on estimating the mean squared error of the area-statistic estimates under level-specific penalization. On the one hand, the penalized model parameter estimates don't have a closed-form solution. On the other hand, the penalized maximum likelihood approach introduces some bias to the model parameter estimates that is hard to quantify. Burgard et al. (2019) propose a modified Jackknife approach to estimate the MSE of a penalized Fay-Herriot model. While the general procedure is applicable to our approach, some further modifications may be required in order to include the level-specific penalization in the estimation process.

References

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*(401), 28–36.

Bertsimas, D. and M. S. Copenhaver (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research 270*, 931–942.

Bondell, H. D., A. Krishna, and S. K. Ghosh (2010). Joint variable selection of fixed an random effects in linear mixed-effects models. *Biometrics 66*(4), 1069–1077.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). Optimization methods for large-scale machine learning. URL: https://arxiv.org/abs/1606.04838v3.

Boubeta, M., M. J. Lombardía, and D. Morales (2016). Empirical best prediction under area-level poission mixed models. *TEST 25*, 548–569.

Burgard, J. P., J. Krause, and D. Kreber (2019). Regularized area-level modelling for robust small area estimation in the presence of unknown covariate measurement errors. *Trier University Working Paper Series*.

Clark, W. A. V. and K. L. Avery (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis 8*(4), 428–438.

Eliot, M., J. Ferguson, M. P. Reilly, and A. S. Foulkes (2011). Ridge regression for longitudinal biomarker data. *The International Journal of Biostatistics 7*(1), Article 37.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

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.

Flamary, R., N. Jrad, M. Congedo, and A. Rakotomamonjy (2014). Mixed-norm regularization for brain decoding. *Computational and Mathematical Methods in Medicine 2014*(ID 317056), 13 pages.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate gradient descent. *Journal of Statistical Software 33*(1), 1–22.

Ghosh, M. and R. C. Steorts (2013). Two-stage benchmarking as applied to small area estimation. *Test 22*, 670–687.

Greven, S. and T. Kneib (2010). On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika 97*(4), 773–789.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Techometrics 12*(1), 55–67.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011). Fixed and random effects selection in mixed effects models. *Biometrics 67*(2), 495–503.

Li, Y. and H. Yang (2010). A new stochastic mixed ridge estimator in linear mixed model. *Statistical Papers 51*, 315–323.

Lohr, S. and L. Ybarra (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*, 919–931.

Moura, F. A. S. and D. Holt (1999). Small area estimation using multilevel models. *Survey Methodology 25*(1), 73–80.

Nath, J. S., S. Raman, A. Ben-tal, G. Dinesh, C. Bhattacharyya, and K. R. Ramakrishnan (2009). On the algorithmics and applications of a mixed-norm based kernel learning formulation. In *In Advances in Neural Information Processing Systems*, pp. 844–852.

Ogutu, J. O., T. Schulz-Streeck, and H.-P. Piepho (2012). Genomic selection using regularized regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings 2012 6*(Suppl 2), S10.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation* (2 ed.). Wiley Series in Survey Methodology. Jon Wiley & Sons, Inc., Hoboken, New Jersey.

Robert Koch Institute (2012). Daten und Fakten: Ergebnisse der Studie "Gesundheit in Deutschland aktuell 2010". *Beiträge zur Gesundheitsberichterstattung des Bundes*. RKI, Berlin.

Schelldorfer, J., P. Bühlmann, and S. van de Geer (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics 38*, 197–214.

Schipf, S., T. Ittermann, T. Tamayo, R. Holle, M. Schunk, W. Maier, C. Meisinger, B. Thorand, A. Kluttig, K. H. Greiser, K. Berger, G. Müller, S. Moebus, U. Slomiany, W. Rathmann, and H. Völzke (2014). Regional differences in the incidence of self-reported type 2 diabetes in germany: Results from five population-based studies in germany (diab-core consortium). *Journal of Epidemiology and Community Health 68*, 1088–1095.

Schulz-Streeck, T. and H.-P. Piepho (2010). Genome-wide selection by mixed ridge regression and extensions based on geostatistical models. *BMC Proceedings 2010 4*(Suppl 1), S8.

Sidi, J. (2017). *lmmen: Linear Mixed Model Elastic Net* (1.0 ed.). URL: https://CRAN.R-project.org/package=lmmen.

Tibshirani, R. (1996). Regressoin shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tseng, P. and S. Yun (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming 117*(1-2), 387–402.

Twigg, L., G. Moon, and K. Jones (2000). Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators. *Social Science & Medicine 50*, 1109–1120.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological) 67*(2), 301–320.

Zou, H. and H. H. Zhang (2009). On the adapative elastic-net with a diverging number of parameters. *The Annals of Statistics 37*(4), 1733–1751.