Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Combining Unit- and Area-Level Data for Small Area Estimation via Penalized Multi-Level Models

Jan Pablo Burgard, Joscha Krause, and Ralf Münnich[1]

Abstract

In small area applications, often unit- and area-level information is available. The joint usage of unit- and area-level data for small area estimation within one single model encloses a variety of methodological problems. Firstly, it implies an increasing number of model parameters that need to be estimated. A careful selection of variables needs to be considered in order to avoid destabilized model predictions in the presence of small samples. Secondly, unit- and area-level data may have different distributional characteristics in terms of dispersion patterns and correlation structures. Thirdly, unit- and area-level data are 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 small area estimation. In an example, we combine health survey data on the unit-level and aggregated micro census records on the area-level to estimate regional hypertension prevalence in Germany.

Key Words: Coordinate Gradient Descent; Multi-Level Model; Penalized Maximum Likelihood.

1. Introduction

Small area estimation is frequently applied to obtain reliable estimates of aggregate-specific quantities (area-statistics) from small samples (Rao and Molina, 2015; Münnich et al., 2016;). A direct estimator that only uses data of one area at a time cannot produce area-statistic estimates with sufficient accuracy in that case. Small area estimation was developed to solve this problem by combining data from multiple areas in suitable regression models. The general idea is to improve estimation efficiency over a direct estimator by exploiting the functional relation between the area-statistic of interest and suitable auxiliary data. Depending on the field of application, either area-level models (Fay and Herriot, 1979) or unit-level models (Battese et al., 1988) are proposed. The efficiency gain of a small area estimator over a direct estimator is determined by the explanatory power of the underlying regression model. Accordingly, in a situation where auxiliary data for both unit- and area-level is available, both levels should be considered in order to maximize the explanatory power and thus to produce optimal area-statistic estimates.

However, using unit- and area-level data jointly for model-based small area estimation encloses some methodological problems. Firstly, it requires model parameter estimation on both levels simultaneously. In the presence of small samples, the increased number of parameters may lead to considerably high variance in model parameter estimates due to the lack in degrees of freedom. Model-based area-statistic 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). Subsequently, the levels should not be treated equally in terms variable selection and model parameter estimation. Thirdly, unit- and area-level data is usually subject of different kinds of measurement errors. While unit-level data may suffer from imprecise responses, 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 availability. Unit-level data is often not available due to confidentiality, whereas area-level data is less sensitive and easier to access, for example from registries. Accordingly, a combined model must be able to handle situations where there are a large number of variables on area-level while only few are available on unit-level.

[1]Jan Pablo Burgard, Trier University, Universitätsring 15, Trier, Germany, D-54296; Joscha Krause, Trier University, Universitätsring 15, Trier, Germany, D-54296; Ralf Münnich, Trier University, Universitätsring 15, Trier, Germany, D-54296 (muennich@uni-trier.de)

We propose to combine unit- and area-level data for small area estimation 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 effect coefficients on each level are shrunk by an individual penalty. For this purpose, the least absolute shrinkage and selection operator (LASSO), the ridge penalty and the elastic net are considered. Using level-specific penalization in multi-level models 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 a simple 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. the LASSO), an automatic level-specific variable selection is conducted. Thirdly, norm-type regularization implies a robustification against measurement errors in the auxiliary data (Bertsimas and Copenhaver, 2018; Burgard et al., 2019a). Accordingly, level-specific penalization allows for robust estimation despite different measurement errors on each level. And finally, the degree of penalization on each level 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 from a Bayesian approach using maximum a posteriori, as suggested by Schelldorfer et al. (2011). An application is provided on the example of 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 multi-level model and model parameter estimation are described. In Chapter 3, the application to health measurement is presented. Chapter 4 closes with some conclusive remarks. This contribution is a short version of the working paper *Penalized Small Area Models for the Combination of Unit- and Area-Level Data* by the same authors. It gives a general overview on the methodology. For deeper insights and computational details of our approach, we refer to Burgard et al. (2019b).

2. Methods

2.1 Multi-Level Model

Let U be a finite population of size N consisting of m pairwise disjoint areas of size N_i with $i = 1, \dots, m$ and $\sum_{i=1}^{m} N_i = N$. Let a random sample S of size n be drawn from U such that there are m area sub-samples S_i of size $n_i > 1$ with $\sum_{i=1}^{m} n_i = n$. Let $\boldsymbol{y}_i \in \mathbb{R}^{n_i \times 1}$ be a vector containing observations of some response variable y from which the area-statistic of interest in area i is calculated. Let $\boldsymbol{X}_i^u \in \mathbb{R}^{n_i \times p^u}$ be the fixed effects design matrix in area i containing unit-level covariates. Let $\boldsymbol{X}_i^a = (\boldsymbol{x}_i^a, \dots, \boldsymbol{x}_i^a)' \in \mathbb{R}^{n_i \times p^a}$ be the fixed effect design matrix resulting from an expansion of the vector $\boldsymbol{x}_i^a \in \mathbb{R}^{1 \times p^a}$ containing area-level covariates. Note that $(p^u + p^a) > n$ is allowed. Let $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times q}$ be the random effect design matrix in area i with $q \leq (p^u + p^a)$. In the majority of small area 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

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

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 $\boldsymbol{b}_i \sim MVN(\boldsymbol{0}, \boldsymbol{\Psi})$ denotes the random effect coefficient vector under multivariate normality with some general positive-definite covariance matrix $\boldsymbol{\Psi}$. $\boldsymbol{e}_i \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_i})$ is a vector of i.i.d. random errors with model variance parameter σ^2. Note that $\boldsymbol{b}_1, \dots, \boldsymbol{b}_m, \boldsymbol{e}_1, \dots, \boldsymbol{e}_m$ are assumed to be stochastically independent. Thus, the response vector has the following distribution under the model

$$\boldsymbol{y}_i \sim MVN\big(\boldsymbol{X}_i^u \boldsymbol{\beta}^u + \boldsymbol{X}_i^a \boldsymbol{\beta}^a, \boldsymbol{V}_i(\sigma^2, \boldsymbol{\psi})\big),$$

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

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

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

$$-l(\boldsymbol{\theta}) = \frac{1}{2}\big(n \cdot \log(2\pi) + \log(|\boldsymbol{V}|) + (\boldsymbol{y} - \boldsymbol{X}^u\boldsymbol{\beta}^u - \boldsymbol{X}^a\boldsymbol{\beta}^a)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}^u\boldsymbol{\beta}^u - \boldsymbol{X}^a\boldsymbol{\beta}^a)\big),$$

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

2.2 Penalized Model Parameter Estimation

Penalized model parameter estimation in the model is characterized by the optimization problem

$$\widehat{\boldsymbol{\theta}} = \ argmin_{\boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \sigma^2 > 0, \boldsymbol{\Psi} > 0}\{-l(\boldsymbol{\theta}) + \lambda^u P^u(\boldsymbol{\beta}^u) + \lambda^a P^a(\boldsymbol{\beta}^a)\},$$

where $P^u(\boldsymbol{\beta}^u): \mathbb{R}^{p^u} \to \mathbb{R}$, $P^a(\boldsymbol{\beta}^a): \mathbb{R}^{p^a} \to \mathbb{R}$ are level-specific penalties on the fixed effect coefficients and $\lambda^u, \lambda^a > 0$ are level-specific tuning parameters that are determined from k-fold cross validation. The following penalties are considered ($l \in \{u, a\}$):

1. LASSO (Tibshirani, 1996): $P^l(\boldsymbol{\beta}^l) = \|\boldsymbol{\beta}^l\|_1$
2. Ridge (Hoerl and Kennard, 1970): $P^l(\boldsymbol{\beta}^l) = \|\boldsymbol{\beta}^l\|_2^2$
3. Elastic Net (Zou and Hastie, 2005): $P^l(\boldsymbol{\beta}^l) = \alpha^l\|\boldsymbol{\beta}^l\|_1 + (1-\alpha^l)\|\boldsymbol{\beta}^l\|_2^2$, where $\alpha^l \in [0,1]$ is a level-specific hyper parameter.

The penalties have different effects on the optimal solutions for the fixed effect coefficients. The LASSO induces are sparse solution for $\widehat{\boldsymbol{\beta}}^u$ and $\widehat{\boldsymbol{\beta}}^a$. The level-specific sparsity is controlled by λ^u, λ^a. Including this penalty leads to an automatic variable selection in the estimation process, as coefficients that are irrelevant for the description of the response variable are set to zero. On the contrary, the ridge penalty induces a dense solution for $\widehat{\boldsymbol{\beta}}^u$ and $\widehat{\boldsymbol{\beta}}^a$. It smooths the individual contributions of the coefficients in the estimation process. The level-specific smoothness is controlled by λ^u, λ^a. The ridge penalty does not perform variable selection, but is known to deliver more stable results in the presence of multicollinearity and grouping structures in the covariates. The elastic net is a linear combination of the LASSO and the ridge penalty. It induces a sparse solution while allowing for grouping structures in the covariates. The level-specific weight of each penalty is controlled by α^u, α^a. The idea of level-specific penalization is to choose $P^u(\boldsymbol{\beta}^u), P^a(\boldsymbol{\beta}^a), \lambda^u, \lambda^a$ dependent on the distributional characteristics of the covariates on each level to obtain optimal model predictions.

In order to solve the minimization problem under a given penalization, a stochastic modification of the coordinate gradient descent algorithm proposed by Tseng and Yun (2009) as well as Schelldorfer et al. (2011) is used. Minimization via coordinate descent implies that the value of the objective function is minimized gradually by updating a single element $\theta_r \in \boldsymbol{\theta}$ at a time while keeping $\boldsymbol{\theta}_{-r}$ fixed. The remaining elements $\boldsymbol{\theta}_{-r}$ are updated accordingly in an iterative manner such that there is a cyclic movement through all coordinates of $\boldsymbol{\theta}$. This cyclic approach is particularly useful for the proposed multi-level model, as it allows for an easy implementation of the level-specific penalization in the estimation process. The order of the coordinates that correspond to fixed effect coefficients is changed randomly in each iteration to improve the convergence probability of the algorithm in the light of the nonconvex minimization problem. The general estimation order that is common in small area models (fixed effects conditionally on the variance parameters and vice versa) is not varied. Note that if $\theta_r \in (\boldsymbol{\beta}^u, \boldsymbol{\beta}^a)$, then the required update for minimization is additionally dependent on the penalty chosen for its respective level. This has further implications on how to calculate the descent direction and how to determine a suitable step length in each iteration. However, these details are skipped here. For more computational insights on the algorithm used for model parameter estimation, we refer to Burgard et al. (2019b).

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

$$\begin{aligned}\widetilde{\boldsymbol{b}}_i &= argmax_{\boldsymbol{b}_i}\left\{\frac{f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \sigma^2, \boldsymbol{\psi}) \cdot f(\boldsymbol{b}_i|\boldsymbol{\psi})}{f(\boldsymbol{y}_i|\boldsymbol{\beta}^u, \boldsymbol{\beta}^a, \sigma^2, \boldsymbol{\psi})}\right\} \\ &= argmin_{\boldsymbol{b}_i}\left\{\frac{1}{\sigma^2}\|\boldsymbol{y}_i - \boldsymbol{X}_i^u\boldsymbol{\beta}^u - \boldsymbol{X}_i^a\boldsymbol{\beta}^a - \boldsymbol{Z}_i\boldsymbol{b}_i\|_2^2 + \boldsymbol{b}_i'\boldsymbol{\Psi}^{-1}\boldsymbol{b}_i\right\}.\end{aligned}$$

Under the model assumptions, it can be concluded that

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

which is then predicted by

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

using the model parameter estimates obtained from the coordinate descent algorithm.

3. Regional Hypertension Prevalence Estimation

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 hyper parameters $\alpha^u = \alpha^a = 0.5$ 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. As the elastic net is a sparsity-inducing penalty, an automatic variable selection is conducted in the estimation process. On the unit-level, demographic, comorbidity-related and lifestyle-related variables are selected. On the area-level, mainly socio-economic and labour market variables are selected. The methodology obtains the following results.

Figure 3-1
Estimated Hypertension Prevalence for 2010, Federal States

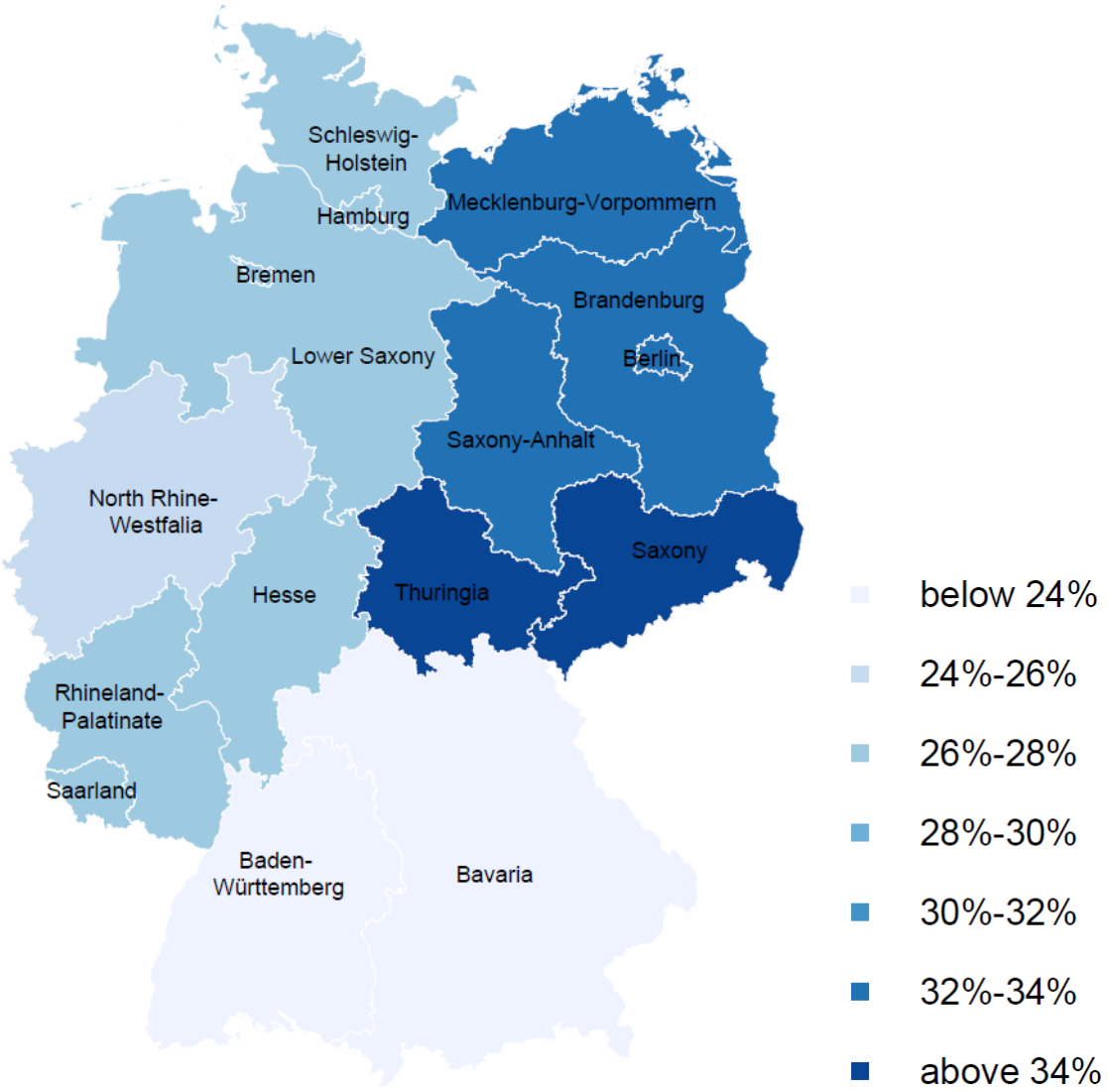

Figure 3-1 is a heat map of Germany in which the estimated hypertension prevalence per federal state are displayed. The nationwide hypertension prevalence is at 26.8%. This is consistent with the results of the Robert Koch Institute (2012), which calculated a survey-based 95%-confidence interval of [25.9%; 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).

4. Summary and 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. (2019a) 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), pp. 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, pp. 931-942.

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

Burgard, J. P., J. Krause, and R. Münnich (2019b), "Regularized Small Area Models for the Combination of Unit- and Area-Level Data", 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), pp. 428-438.

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), pp. 269-277.

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

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

Münnich, R. T., J. P. Burgard, S. Gabler, M. Ganninger, and J. P. Kolb (2016), "Small area estimation in the German census 2011", *Statistics in Transition and Survey Methodology*, 17(1), pp. 25-40.

Rao, J. N. K., and I. Molina (2015), *Small area estimation (2 ed.)*, Hoboken: Wiley.

Robert Koch Institute (2012), *Daten und Fakten: Ergebnisse der Studie Gesundheit in Deutschland aktuell 2010*, Berlin: RKI.

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, pp. 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, pp. 1088-1095.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267-288.

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

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), pp. 301-320.