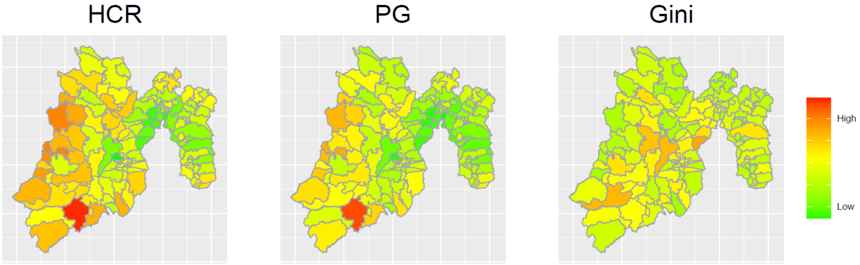
6 – Small Area Estimation of non-linear indicators

Acknowledgements: Thanks to Natalia Rojas (Freie Universität Berlin) for providing useful materials.

Content of the chapter

- Introduction
- Empirical Best Prediction
- Transformations
- Evaluation via simulations

Typical Results from Poverty Mapping



Models for Poverty Estimation Methods

Area-level models (for means):
They can be used if only aggregated data are available over the areas, usually leading to some loss of information

$\rightarrow$ *Fay-Herriot estimator* (Fay and Herriot, 1979)

Unit-level models (for means):
They are established for unit values of a target variable to unit-specific explanatory ones, using more detailed information and leading to improvement of precision

$\rightarrow$ *Battese-Harter-Fuller model* (Battese et al., 1988)

EBLUP Method

Nested error linear regression model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m,$$
$$u_i \overset{iid}{\sim} N(0, \sigma_u^2) \text{ the random area-specific effects,}$$
$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2) \text{ the unit-level error terms, } \; u_i \perp e_{ij}$$

- Target indicator: means of the target variable in each domain: $\bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$
- Fit the model with the sample data, obtaining $\hat{\beta}$, $\hat{\sigma}_u$, $\hat{\sigma}_e$, $\hat{u}_i$

$\Rightarrow$ Empirical Best Linear Unbiased Predictor (EBLUP) of $\bar{y}_i$

$$\hat{\bar{y}}_i = \frac{1}{N_i} \left[\sum_{j \in s_i}^{N_i} y_{ij} + \sum_{j \in r_i}^{N_i} \hat{y}_{ij} \right], \quad \bar{y}_{ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i.$$

Reference: Battese et al. (1988)

Non-linear Income-based Indicators

- Small area estimation methods is mainly focused on estimating means and totals at small areas
- New developments in SAE methodologies in a poverty mapping context are focused on estimating complex non-linear quantities at small area level, e.g FGT indicators

Possible Solutions:
EBP, ELL and MQ methods are the most frequently approaches to estimate averages, quantiles, poverty and inequality indicators for small areas.

Recent Methodologies

- The World Bank method (ELL)
  (Elbers et al., 2003)
- The Empirical Best Predictor (EBP) method
  (Molina and Rao, 2010)
- EBP based on normal mixtures
  (Elbers and van der Weide, 2014)
- Methods based on M-Quantiles
  (Giusti et al., 2012)

Poverty Mapping Framework I

1. Choose the poverty measure(s): As a starting point, the selection of poverty indicators and a poverty line t if required. Moreover, choose $\boldsymbol{y}$, the target variable and $\mathbf{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^T$, the design matrix, containing p appropriate explanatory variables, where y_{ij} is a value of $\boldsymbol{y}$ and $\mathrm{x}_{ij}^T = (\mathrm{x}_{1ij}, \ldots, \mathrm{x}_{pij})$, the values of the variables related to y_{ij}.
2. Select input data:
   - Survey data: available for $\boldsymbol{y}$ and $\mathbf{X}$.
   - Census data: available for $\mathbf{X}$.
3. Choose a poverty estimation method.

Poverty Mapping Framework II

4. Estimation process:
    - Estimate a statistical model on the survey data, that links $\mathbf{y}$ to $\mathbf{X}$ and obtain the corresponding parameters.
    - Obtain a synthetic population of $\boldsymbol{y}$ in the census data, using the previously obtained parameters and the covariates available in the census.
    - Estimate target poverty indicators and their associated precision estimates.
5. Produce maps:

   Obtain a high-resolution map here called *poverty map* by plotting the resulting estimates on the geographical coordinates.

Content of the chapter

- Introduction
- Empirical Best Prediction
- Transformations
- Evaluation via simulations

Notation

- ✓ Assume a finite population of size N
- ✓ D regions of sizes $N_1, \ldots, N_m$
- ✓ $i = 1, \ldots, m$, refers to a ith region
- ✓ $j = 1, \ldots, N_i$, refers to the jth individual
- ✓ n_i, the sample size of region i with $n = \sum_{i=1}^{D} n_i$
- ✓ s_i, the in-sample units in region i
- ✓ r_i, the out-of-sample units in region i
- ✓ Let $\boldsymbol{y}$ be the target variable (welfare measure)
- ✓ $\mathbf{y}_i^T = \left(\mathbf{y}_{is}^T, \mathbf{y}_{ir}^T\right)$
- ✓ Assume $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ p explanatory variables
- ✓ Use a fixed poverty line (e.g. $t = 0.6 \times \text{Median}(\boldsymbol{y})$)

The EBP approach

Point of departure: Random effects model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, D,$$

Estimation process:

1. Use the sample data to estimate $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{u}_i$ and $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.
2. For $l = 1, ..., L$
   - Generate $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$ and $u_i^* \sim N(0, \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_i))$ and obtain a pseudo-population

     $$y_{ij}{}^{*(l)} = \mathbf{x}_{ij}^T \hat{\beta} + \hat{u}_i + u_i^* + e_{ij}^*$$

   - Calculate the poverty measures of interest $\theta_i^{(l)}$.
3. Obtain $\hat{\theta}_i^{EBP} = 1/L \sum_{l=1}^{L} \hat{\theta}_i^{(l)}$ for each region i.

Reference: Molina and Rao (2010).

Parametric Bootstrap: MSE Estimation

- Fit the random effects model to the original sample
- Generate $u_i^* \sim N(0, \hat{\sigma}_u^2)$, $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$
- Construct B bootstrap populations

$$y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^* + e_{ij}^*$$

- For each b population compute the population value θ_i^{*b}
- From each bootstrap population select a bootstrap sample
- Implement the EBP with the bootstrap sample, get $\hat{\theta}_i^{*b}$

$$\widehat{MSE}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_i^{*b} - \theta_i^{*b})^2$$

- Use $\widehat{MSE}(\hat{\theta}_i)$ to compute estimated coefficients of variation (CVs)

Reference: González-Manteiga et al. (2008).

Using R-package emdi: EBP method

- The R package **emdi** provides two synthetic data sets to show its applicability in order to receive e.g. empirical best predictions.
  - `eusilcS_HH`: sample data from Austrian regions about income and demographics on the household level
  - `eusilcP_HH`: corresponding population data for the Austrian regions

  $\hookrightarrow$ requirement that both data sets contain the same covariates is given

- The aim of the following application is the prediction of the equivalized income for household in the nine Austrian regions. Using the AIC criteria a model that includes gender, the equivalized household size and several income and benefit variables is appropriate to predict the income:

$$eq\_income_{ij} = c + gender_{ij} + eq\_hhs_{ij} + \sum_{k=1}^{12} inc\_bene_{ijk} + u_i + e_{ij}$$

Reference: Kreutzmann et al. (2016).

Using R-package emdi: EBP method

The function that conducts the EBP approach in the R package **emdi** is the function `ebp()`. For the example the call looks as follows:

```
# EBP estimation function
ebp_au <- ebp(fixed = eqIncome ~ gender + eqsize +
                        py010n + py050n + py090n +
                        py100n + py110n + py120n +
                        py130n + hy040n + hy050n +
                        hy070n + hy090n + hy145n,
              pop_data = eusilcP_HH,
              pop_domains = "region",
              smp_data = eusilcS_HH,
              smp_domains = "region",
              pov_line = 0.6*median(eusilcS_HH$eqIncome
                 ),
              transformation = "no",
              L=50,
              MSE = T,
              B = 50)
```

Reference: Kreutzmann et al. (2016).

Using R-package emdi: EBP method - Summary output

In order to receive a first overview, the summary command can be used.

```
# Summary for the EBP method
  > summary(ebp_au)
```

```
  Out-of-sample domains:  0
  In-sample domains:  9

  Sample sizes:
  Units in sample:  503
  Units in population:  25000
                  Min. 1st Qu. Median   Mean  3rd Qu. Max.
  Sample_domains   16   26       43     55.9    94     101
  Pop_domains     799 1671     1889     2778  4071    5857
```

Reference: Kreutzmann et al. (2016).

Using R-package emdi: EBP method - Summary output

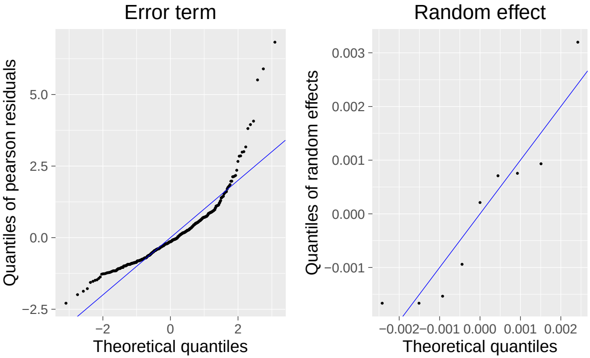
```
Explanatory measures:
Marginal_R2 Conditional_R2
0.5198029      0.5198029

Residual diagnostics:
             Skewness Kurtosis Shapiro_W    Shapiro_p
Error         2.17646 12.5925  0.8551573   4.0933e-21
Random_effect 0.64311  2.6048  0.8870226   1.8589e-01

ICC:  2.610126e-08
```

Reference: Kreutzmann et al. (2016).

Graphical Investigation of Normality Assumptions

Q-Q plots can help to assess the normality assumptions and it belongs to one of the plots that are automatically provided when applying the function `plot` to an `emdi` object.

```
# Residual diagnostics
    > plot(ebp_au)
```



Motivating Alternative Methods

- EBP relies on Gaussian assumptions :
  - ✓ $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, the random area-specific effects
  - ✓ $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, the unit-level error terms, $u_i \perp e_{ik}$

Model Checking (Residual diagnostics)

- Q-Q plots of residuals at different levels
- Influence diagnostics: Plots of Cook's distances
- Plot standardised residuals vs fitted values - Heteroscedasticity
- Plot standardised residuals vs design weights - Informative sampling

Model Adaptations

- Use a EBP formulation under an alternative distribution (Graf et al., 2015)
- Use robust methods as an alternative to transformations (Chambers and Tzavidis, 2006; Ghosh, 2008; Sinha and Rao, 2009; Chambers et al., 2014; Schmid et al., 2016).
- Use non-parametric models (Opsomer et al., 2008; Ugarte et al., 2009).
- Elaborate the random effects structure e.g. include spatial structures (Pratesi and Salvati, 2008; Schmid and Münnich, 2014).
- Consider extensions to two-fold models

$\Rightarrow$ Explore the use of transformations

Content of the chapter

- Introduction
- Empirical Best Prediction
- Transformations
- Evaluation via simulations

Use of Transformations

$\Rightarrow$ Explore the use of one-to-one transformations $T(y_{ij})$ for improving the validity of the model assumptions and the precision of small area prediction in poverty mapping

> *Even in those cases where a normal distribution is not accomplished, the use of a transformation often lead to a symmetric distribution, in which certain restrictions on the first four moments are satisfied.*
> *(Draper and Cox, 1969)*

Why are Transformations Useful?

- Rescaling or standardization
- For fulfilling assumptions:
  - Normality: Reducing skewness and controlling kurtosis
  - Homoscedasticity: Variance-stabilization
  - Linearity: linearizing dependence between two variables

What Influences the Choice of a Transformation?

- Kurtosis
- Skewness
  - Positively skewed
  - Negatively skewed
- Heterogeneity
- Data scale and range
  - Zero values
  - Negative values
- Contamination due to outliers

Use of Transformations in SAE in Poverty Mapping

Specific problems:

- Highly positive unimodal skewed and leptokurtic data sets
- Extensions of the transformations to the mixed model
  - For which source of randomness?
- Invertibility on $\mathbb{R}$
- Appropriate for handling with zero and negative values
- Target indicator
  - Poverty gap, head count ratio
  - Gini coefficient, quantile share ratio

First Selection of Transformations

- Shifted transformations
  - Log-shift
- Power transformations
  - Box-Cox
  - Exponential
  - Sign power
  - Modulus
  - Dual power
  - Folder power
  - Convex-to-concave
- Multi-parameter transformations
  - Johnson
  - Sinh-arcsinh

Transformations for the EBP Method I

Log-Shift Transformation (λ) (Royston and Lambert, 2011)

$$T_\lambda(y_{ij}) = \log(y_{ij} + \lambda),$$

Box-Cox Transformation (λ) (Box and Cox, 1964)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y_{ij} + s), & \lambda = 0 \end{cases},$$

Dual Power Transformation (λ) (Yang, 2006)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \log(y_{ij} + s) & \text{if } \lambda = 0. \end{cases}$$

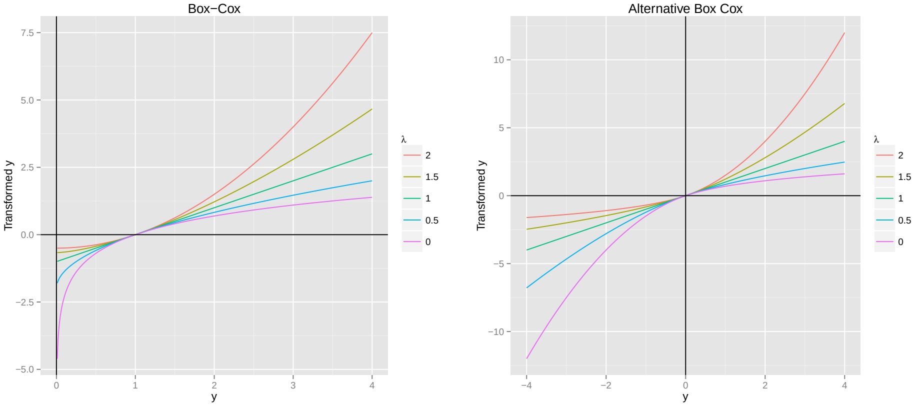
with λ the transformation parameter and $y_{ij} + s > 0$

Transformations for the EBP Method II

Convex-to-Concave Transformation (λ) (Yeo and Johnson, 2000)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y_{ij} \geq 0; \\ \log(y_{ij}+1) & \text{if } \lambda = 0, y_{ij} \geq 0; \\ \frac{(1-y_{ij})^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y_{ij} < 0; \\ -\log(1-y_{ij}) & \text{if } \lambda = 0, y_{ij} < 0. \end{cases}$$

Convex-to-Concave Transformation



Estimation Methods (λ)

- **Which estimation procedure already exist?**
  - Skewness minimization
  - Divergence minimization
  - Maximum Likelihood Estimation (MLE)

- **And for linear mixed models?**
  - Skewness minimization
  - Divergence minimization
  - REML optimization

Minimization of the Pooled Skewness Estimation

Let e_λ be the unit-level error terms and u_λ the random area-specific effects of the linear mixed model. Their corresponding skewness and variances are denoted by S_{e_λ}, S_{u_λ}, $\sigma^2_{e_\lambda}$ and $\sigma^2_{u_\lambda}$.

$$w = \frac{\sigma^2_{e_\lambda}}{\sigma^2_{e_\lambda} + \sigma^2_{u_\lambda}}$$

$$\mathbb{PS} = w\left|S_{e_\lambda}\right| + (1-w)\left|S_{u_\lambda}\right|$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \quad \mathbb{PS}$$

Divergence (Kolmogorov-Smirnof) minimization

$$\hat{F}_{e_\lambda}(t) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{1}\{e_{ij,\lambda} \leq t\}$$

$$\mathbb{D}(t) = \sup_{t \in \mathbb{R}} |\hat{F}_{e_\lambda}(t) - \Phi(t)|$$

$$\hat{\lambda} = \underset{\lambda}{\text{argmin}} \quad \mathbb{D}(t)$$

Let e_λ be the standardized unit-level error terms of the linear mixed model. $\Phi(t)$ denotes the standard normal distribution.

Other divergences (Kullback-Leibler, Cramer-von Mises)

Residual Maximum Likelihood (REML)

- Using a scaled version of the transformation
- This allows for applying standard maximum likelihood theory

$$
\begin{aligned}
L_{\mathrm{REML}}(T_\lambda, \lambda|\phi) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\
&- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{V}_i^{-1}\boldsymbol{X_i}^{\boldsymbol{T}}\right| \\
&- \frac{1}{2}\sum_{i=1}^{D}[T_\lambda(\boldsymbol{y}_i) - \boldsymbol{X_i}\beta]^T\boldsymbol{V}_i^{-1}[T_\lambda(\boldsymbol{y}_i) - \boldsymbol{X_i}\beta]
\end{aligned}
$$

Reference: Gurka et al. (2006).

Scaled Transformations I

Scaled Log-Shift Transformation (λ)

$$T_\lambda(y_{ij}) = \alpha \log(y_{ij} + \lambda),$$

Scaled Box-Cox Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ij} + s), & \lambda = 0 \end{cases},$$

Scaled Dual Power Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{2}{\alpha} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \alpha \log(y_{ij} + s) & \text{if } \lambda = 0. \end{cases}$$

Reference: Rojas-Perilla et al. (2016).

Scaled Transformations II

Scaled Dual Power Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+1)^\lambda - 1}{\alpha^{\lambda-1}\lambda} & \text{if } \lambda \neq 0, y_{ij} \geq 0; \\ \alpha \log(y_{ij}+1) & \text{if } \lambda = 0, y_{ij} \geq 0; \\ -\alpha^{\lambda-1} \frac{(1-y_{ij})^{2-\lambda}-1}{2-\lambda} & \text{if } \lambda \neq 2, y_{ij} < 0; \\ -\alpha \log(1-y_{ij}) & \text{if } \lambda = 2, y_{ij} < 0. \end{cases}$$

with α chosen in such a way that the Jacobian of the transformation is 1.

Reference: Rojas-Perilla et al. (2016).

Estimation Algotithm (λ)

REML Algorithm for the EBP Method:

1. Choose a transformation
2. Define a parameter interval for λ
3. Set λ to a value inside the interval
4. Maximize the residual log-likelihood function with respect to ϕ conditional on the fixed λ
5. Repeat 3 and 4 until a maximum $\hat{\lambda}$ is found
6. Apply the EBP method (Montecarlo Approximation)

Reference: Rojas-Perilla et al. (2016)

The EBP approach under transformations

Estimation process:

1. Find a **transformation** ($\hat{\lambda}$) and obtain $T(y_{ij}) = \tilde{y}_{ij}$
2. Use the transformed sample data to estimate $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and $\hat{u}_i$.
3. For $l = 1, ..., L$
   - Generate $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$ and $u_i^* \sim N(0, \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_i))$ and obtain a pseudo-population

     $$\tilde{y}_{ij}^{*(l)} = \mathbf{x}_{ij}^T \hat{\beta} + \hat{u}_i + u_i^* + e_{ij}^*$$

   - Transform $\tilde{y}_{ij}^{*(l)}$ back to the original scale and obtain $y_{ij}{}^{*(l)}$
   -
   - Calculate the poverty measures of interest $\theta_i^{(l)}$.
4. Obtain $\hat{\theta}_i^{EBP} = 1/L \sum_{l=1}^{L} \hat{\theta}_i^{(l)}$ for each region i.

Reference: Rojas-Perilla et al. (2016)

Parametric Bootstrap MSE Estimation

1. For $b = 1, ..., B$
   - Using the already estimated $\hat{\beta}, \hat{\sigma}^2_u, \hat{\sigma}^2_e, \hat{\lambda}$ from the transformed data $T(y_{ij}) = \tilde{y}_{ij}$, generate $u^*_i + e^*_{ij}$ and simulate a bootstrap superpopulation $\tilde{y}^{*(b)}_{ij} = \mathbf{x}^T_{ij}\hat{\beta} + u^*_i + e^*_{ij}$
   - Transform $\tilde{y}^{*(b)}_{ij}$ to original scale resulting in ${y_{ij}}^{*(b)}$
   - For each b population compute the population value θ^{*b}_i
   - Extract the bootstrap sample in $\tilde{y}^{*(b)}_{ij}$ and perform the EBP method on them.
     Note, as the transformed sample data is used, the estimation of λ is skipped and the original $\hat{\lambda}$ is used to transform the data back to the original scale. Obtain $\hat{\theta}^{*b}_i$.
2. $\widehat{MSE}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{*b}_i - \theta^{*b}_i)^2$

Reference: Rojas-Perilla et al. (2016)

Accounting for the λ Estimation in the Parametric Bootstrap MSE

Estimating the uncertainty of small area estimates

1. For $b = 1, ..., B$
   - Using the already estimated $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$ from the transformed data $T(y_{ij}) = \tilde{y}_{ij}$, simulate a bootstrap superpopulation $\tilde{y}_{ij}^{*(b)} = \mathbf{x}_{ij}^T \hat{\beta} + u_i^* + e_{ij}^*$
   - Transform $\tilde{y}_{ij}^{*(b)}$ to original scale resulting in ${y_{ij}}^{*(b)}$
   - For each b population compute the population value θ_i^{*b}
   - Extract the bootstrap sample in ${y_{ij}}^{*(b)}$ and perform the EBP method on them. Note, as the re-transformed sample data is used the estimation of λ is newly done. Obtain $\hat{\theta}_i^{*b}$.
2. $\widehat{MSE}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_i^{*b} - \theta_i^{*b})^2$

Reference: Rojas-Perilla et al. (2016)

Using R-package emdi: EBP method

The function `ebp()` under the R package `emdi` is suitable for selecting a transformation: logarithmic or Box-Cox transformation as follows

```
# EBP estimation function under a Box-Cox
    transformation
ebp_au <- ebp(fixed = eqIncome ~ gender + eqsize +
                      py010n + py050n + py090n +
                      py100n + py110n + py120n +
                      py130n + hy040n + hy050n +
                      hy070n + hy090n + hy145n,
              pop_data = eusilcP_HH,
              pop_domains = "region",
              smp_data = eusilcS_HH,
              smp_domains = "region",
              pov_line = 0.6*median(eusilcS_HH$eqIncome
                 ),transformation = "box.cox",L=50,
               MSE = T,B = 50)
```

Reference: Kreutzmann et al. (2016)

Using R-package emdi: EBP method - Summary output

In order to receive a first overview of EBP under a Box-Cox transformation, the summary command can be again used.

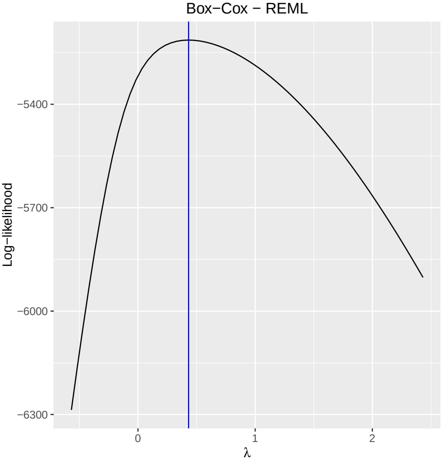
```
# Summary for the EBP method
  > summary(ebp_au)
```

```
Transformation:
 Transformation Method Optimal_lambda Shift_parameter
        box.cox   reml      0.4317972               0

Explanatory measures:
 Marginal_R2 Conditional_R2
   0.4543301      0.4543301

Residual diagnostics:
              Skewness Kurtosis Shapiro_W    Shapiro_p
Error          0.76051   6.3646   0.95643   4.9497e-11
Random_effect 0.58501    2.5533   0.95227   7.1501e-01
```

Reference: Kreutzmann et al. (2016)

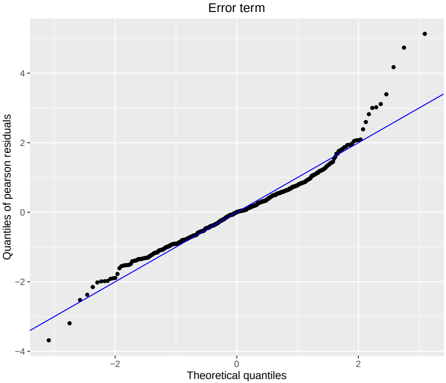
Optimization of Transformation Parameter

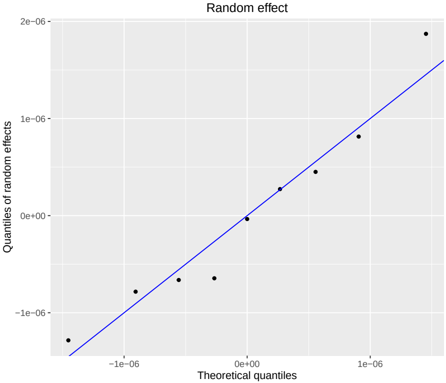
After using the EBP method under a Box-Cox transformation, a graphical representation of the REML optimization algorithm is made by using the function `plot`.



Graphical Investigation of Normality Assumptions

Q-Q plots under a Box-Cox transformation can help to assess the improvement of normality assumptions. It is automatically provided when applying the function `plot`.





Content of the chapter

- Introduction
- Empirical Best Prediction
- Transformations
- Evaluation via simulations

Model- and Design-based Simulations

Complementary Evaluations:

- **Model-based evaluation**
  - Uses synthetic data generated under a model
  - Sampling is performed repeatedly from the population generated in each Monte-Carlo round
  - Useful for evaluating performance and sensitivity of new methods under different assumptions
- **Design-based evaluation**
  - Uses Frame data (census data, for instance) or Synthetic data preserving the survey characteristics
  - Sampling is performed repeatedly from a fixed population
  - Useful for comparing different methods in a particular case

Quality Measures

Root mean square error:

$$RMSE_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{i,r} - \theta_{i,r}\right)^2}$$

Relative bias [%]:

$$RB_i = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\theta}_{i,r} - \theta_{i,r}}{\theta_{k,r}} \cdot 100$$

Absolute bias:

$$Bias_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{i,r} - \theta_{i,r}$$

Model-Based Evaluation

Population data: is generated for $m = 50$ areas with $N = 200$ via

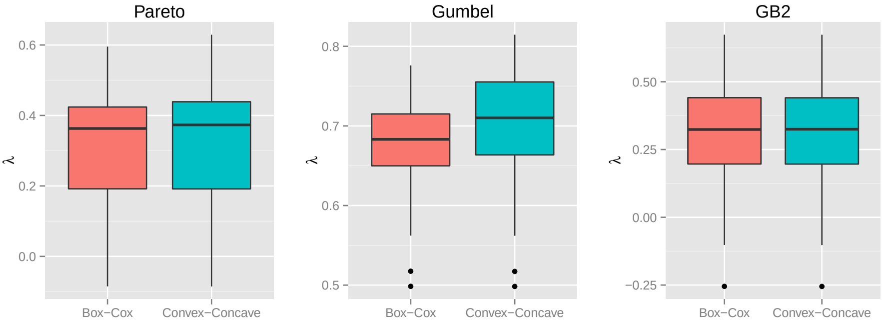
$$y_{ij} = 4500 - 400x_{ij} + u_i + e_{ij}$$

- Covariates $x_{ij} \sim N(\mu_i, 3^2)$ with $\mu_i \sim U(-3, 3)$
- Random effects $u_i \sim N(0, 500^2)$
- Unbalanced design leading to a sample size of $n = 921$ ($min = 8$, $mean = 18.4$, $max = 29$)
- 100 Monte Carlo replicates with L=50 bootstraps

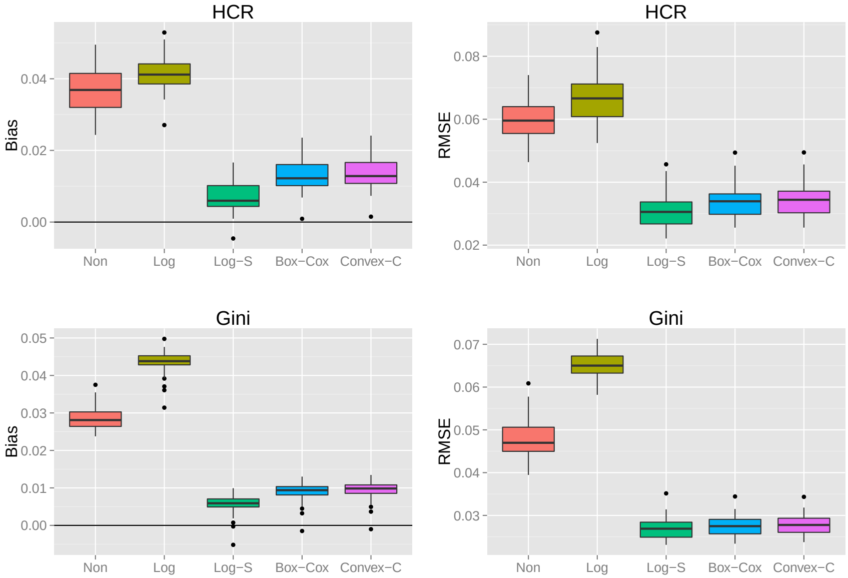
Scenarios: Three different income distribution are investigated:

$$e_{ij} \sim \text{Pareto}(2.5, 100)$$
$$e_{ij} \sim GB2(3, 700, 1, 0.8)$$
$$e_{ij} \sim Gumbel(1, 1000)$$

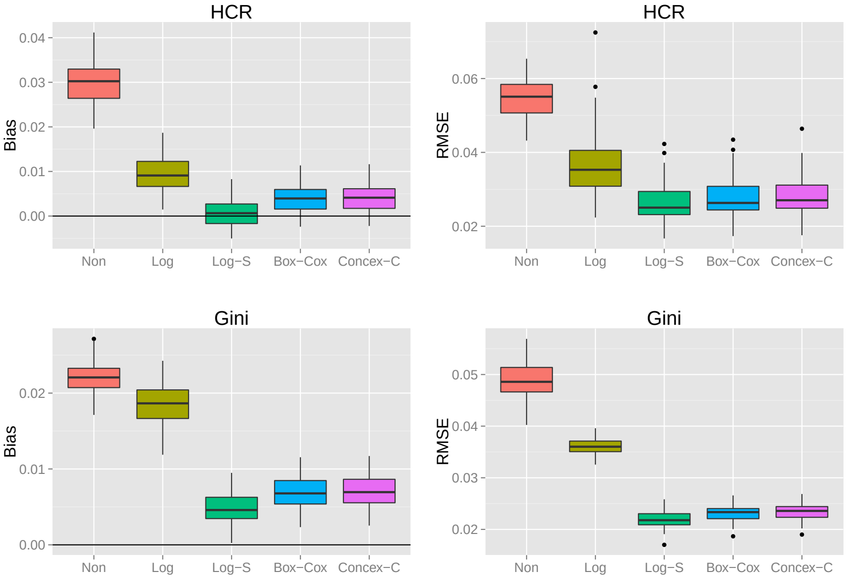
Estimated Transformation Parameters



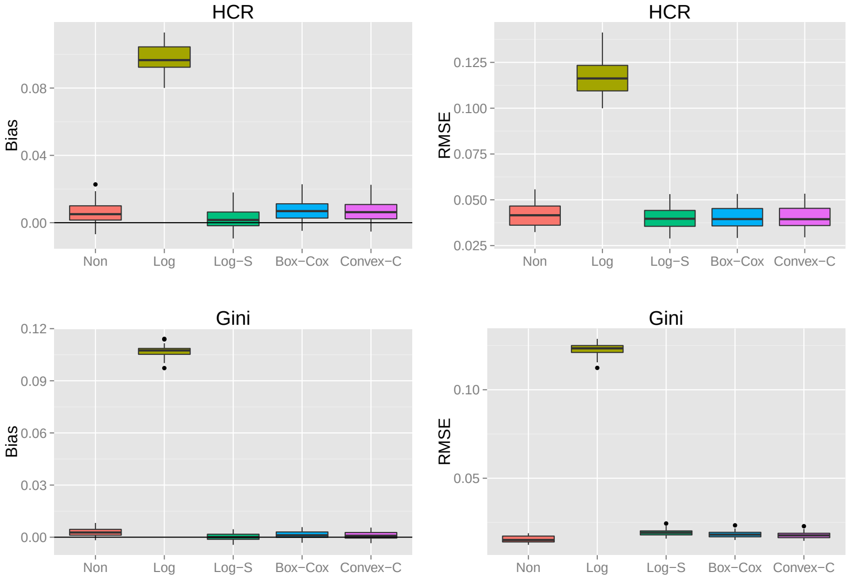
Performance of the Pareto Scenario Using REML



Performance of the GB2 Scenario Using REML



Performance of the Gumbel Scenario Using REML



Design-based Evaluation: State of Mexico (EDOMEX)

- **Sources**: CONEVAL, INEGI, ENIGH
- **Target geography**: State of Mexico is made up of 125 administrative divisions
- **Survey**: 58 are in-sample and 67 out-of-sample
- **Census**: From the 219514 households, there are 2748 in the sample
- **Sample sizes**:

	Min.	Q1.	Median	Mean	Q3	Max.
Survey	3	17	21	47	42	527
Census	650	923	1161	1756	1447	13580

Mexico Case Study: Working Model

Outcome: Two income variables are available in the survey.
The target variable is available only on the survey. Earned per capita income from work is also available on the Census micro data

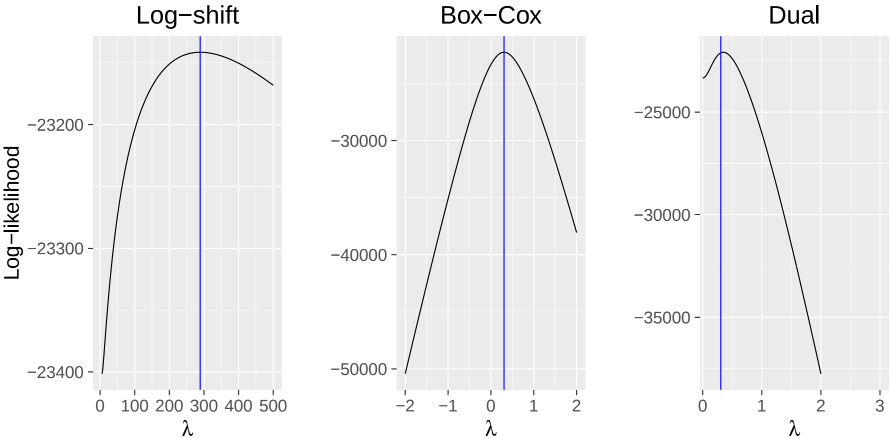
Explanatory variables:

Determinant	Variable
Occupation	Percentage of employees older than 14 years in the household
Education	Highest degree of education completed by the head of household
Household class	Differentiation of households from blood relationship, legal or degree of affinity with the head of household and the other household members
Income earners	Percentage of income earners in a household
Socioeconomic level	Total number of communication assets in the household Total number of goods in the household

Design-based Evaluation: Setup

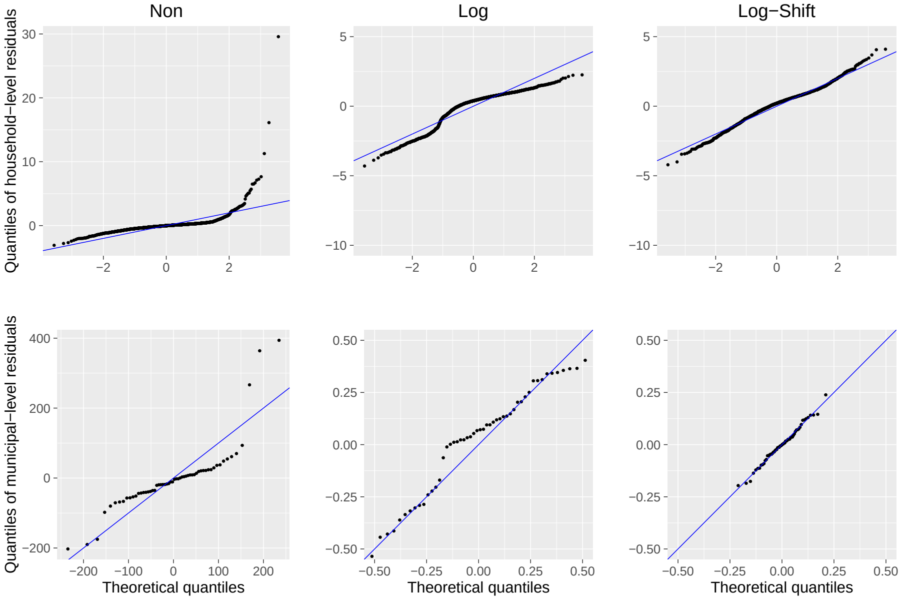
- Design-based simulation with 500 MC-replications repeatedly drawn from EDOMEX Census
- Unbalanced design leading to a sample size of n = 2195 (min = 8, mean = 17.6, max = 50)
- Sampling from each municipality

Mexico Case Study: Parameter Estimation



	Log-shift	Box-Cox	Dual
λ	289.46	0.31	0.35

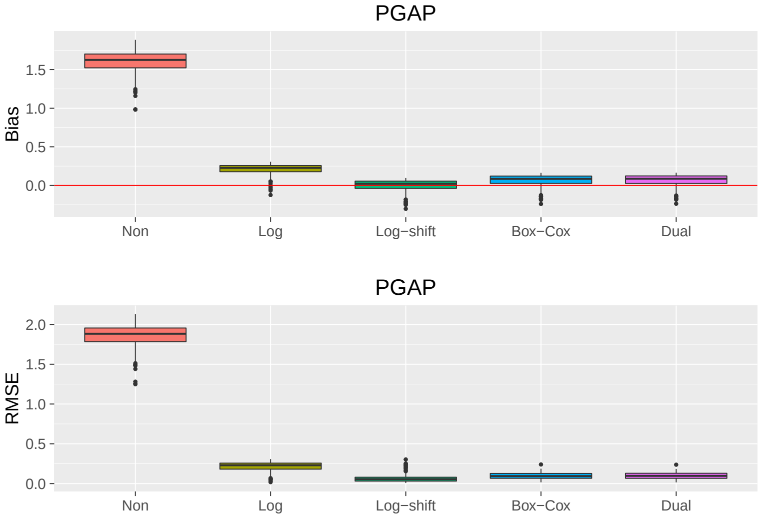
Mexico Case Study: Residual Diagnostics



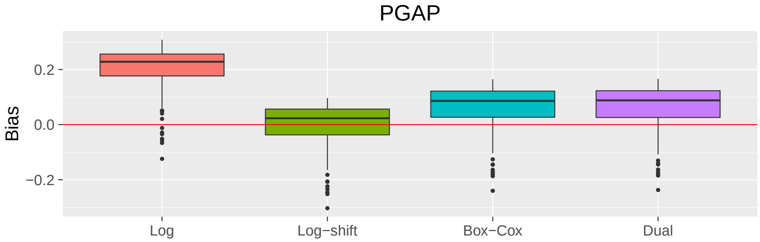
Mexico Case Study: Model Diagnostics

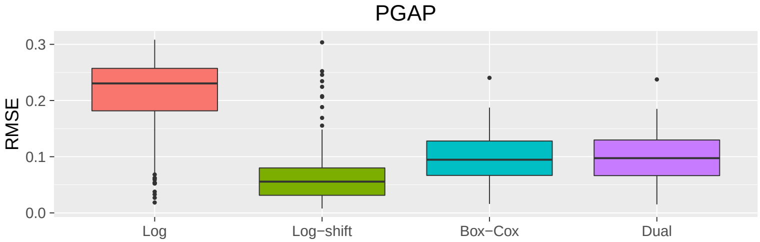
Transformation	No	Log	Log-Shift	Box-Cox	Dual
R^2	30	40	52	48	48
ICC	0.004	0.046	0.032	0.029	0.027

PG Performance under Different Transformations

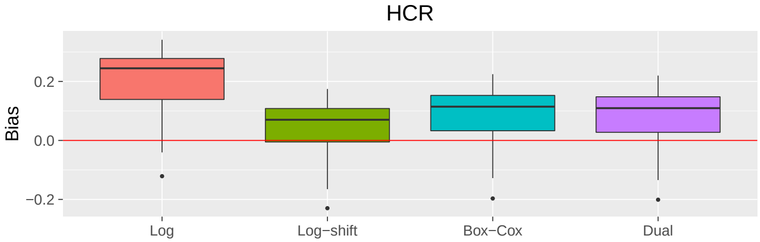


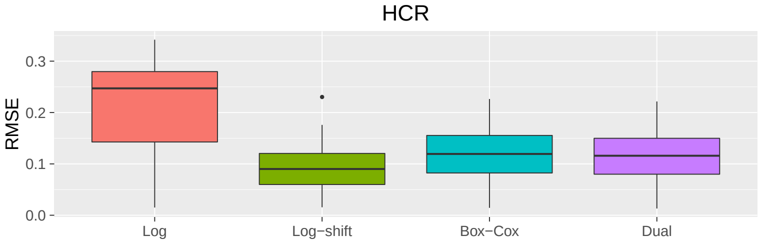
PG Performance under Different Transformations



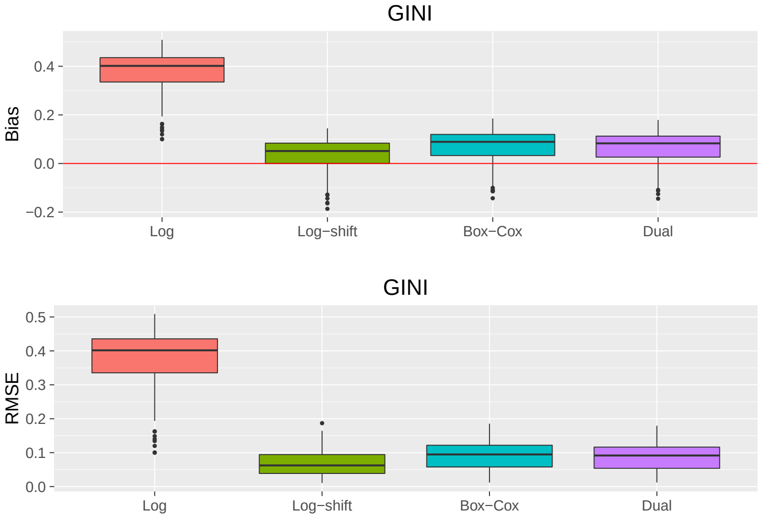


HCR Performance under Different Transformations





Gini Performance under Different Transformations



References I

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

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B 26*, 211–252.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B 76*, 47–69.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. *Biometrika 93*, 255–268.

Draper, N. R. and D. R. Cox (1969). On distributions and their transformation to normality. *Journal of the Royal Statistical Society. Series B (Methodological) 31*, 472–476.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica 71*, 355–364.

Elbers, C. and R. van der Weide (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *World Bank Policy Research Working Paper No. 6962.*.

Fay, R. E. and R. A. Herriot (1979). Estimation of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74*, 269–277.

Ghosh, M. (2008). Robust estimation in finite population sampling. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, pp. 116–122.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati (2012). Semiparametric fay-herriot model using penalized splines. *Journal of the Indian Society of Agricultural Statistics 66*, 1–14.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation 78*, 443–462.

Graf, M., J. Marin, and I. Molina (2015). Estimation of poverty indicators in small areas under skewed distributions. In *Proceedings of the 60th World Statistics Congress of the International Statistical Institute*, The Hague, Netherlands.

References II

Gurka, M. J., L. J. Edwards, K. E. Muller, and L. L. Kupper (2006). Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society Series A 169*, 273–288.

Kreutzmann, A.-K., S. Pannier, N. Rojas, T. Schmid, N. Tzavidis, and M. Templ (2016). *emdi: An R package for Estimating and Mapping regional disaggregated indicators*. R package version 0.7.0.

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

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B 70*, 265–283.

Pratesi, M. and N. Salvati (2008). Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods & Applications 17*, 113–141.

Rojas-Perilla, R., T. Schmid, N. Tzavidis, and S. Pannier (2016). Transformations of small area estimation methods for poverty mapping. Working paper.

Royston, P. and P. C. Lambert (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. StataCorp LP.

Schmid, T. and R. Münnich (2014). Spatial robust small area estimation. *Statistical Papers 55*, 653–670.

Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). Outlier robust small area estimation under spatial correlation. *Scandinavian Journal of Statistics 43*, 806–826.

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics 37*, 381–399.

Ugarte, M., T. Goicoa, A. Militino, and M. Durban (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis 53*, 3616–3629.

Yang, Z. (2006). A modified family of power transformations. *Economics Letters 92*, 14–19.

Yeo, I.-K. and R. A. Johnson (2000). A new family of power transformations to improve normality or symmetry. *Biometrika 87*, 954–959.