Data-driven Transformations for the Estimation of Small Area Means

Nora Würz[1], Nikos Tzavidis[2],
Timo Schmid[1]

[1] Freie Universität Berlin
[2] University of Southampton

ITACOSM 2019
06.06.2019

Outline

Importance of transformations for SAE models

- Small sample sizes within (certain) subpopulations lead to unreliable direct estimators. Small area estimation is a powerful tool to overcome this problem.
- Small area models rely on linear mixed models, so the Gaussian assumption of the error terms must hold.
- For many variables, like income, often the Gaussian assumption is not satisfied in applications.
- Transforming the dependent variable helps to meet these assumptions.

Research gap: Estimating SAE Means

		census data	
		unit-level	area-level
sampled data	unit-level	log transformation for the Battese-Harter-Fuller model **Molina & Martín (2018)**	log and log-shift transformation for the Battese-Harter-Fuller model
		general transformations for the Battese-Harter-Fuller model	
	area-level	not relevant for applications	log transformation for the Fay-Herriot model **Slud & Maiti (2006)** general transformation for the Fay-Herriot model **Sugasawa & Kubokawa (2017)**

Estimating non-linear indicators: General transformations for the EBP **Rojas-Perilla et al. (2017)**

Outline

The Battese-Harter-Fuller model (Battese et al. 1988)

The **Battese-Harter-Fuller model** (BHF) is a small area model based on (area level) covariates and a area-specific random effect (u_d).

$$y_{di} = x_{di}^t \beta + u_d + e_{di} \qquad u_d \sim N(0, \sigma_u^2) \qquad e_{di} \sim N(0, \sigma_e^2)$$

Where $d = 1, ..., D$ indicates areas or domains including $i = 1, ..., N_d$ individuals. The auxiliary variables x_{di}^t are linear related to the dependent variable y_{di}.

The empirical best linear unbiased predictor (**EBLUP**) for the area mean ($\overline{y}_d$) is:

$$\tilde{\overline{y}}_d = \frac{1}{N_d}\Big(\sum_{i \in s_d} y_{di} + \sum_{i \in \overline{s}_d} \underbrace{x_{di}^t \tilde{\beta} + \tilde{u}_d}_{=\tilde{y}_{di}}\Big).$$

s_d describes the observed individuals of area d and $\overline{s}_d$ the unobserved individuals.

Transformations within the BHF model

If the normality assumptions of the error terms within the model do not hold, a (data-driven) transformation $h()$ can help to meet these assumptions.

$$h(w_{di}) := y_{di} = x_{di}^t\beta + u_d + e_{di}$$

EBLUP of the back-transformed area mean of interest $\overline{w}_d$:

$$\tilde{\tau}_d = \tilde{\overline{w}}_d = \frac{1}{N_d}\left(\sum_{i\in s_d} w_{di} + \sum_{i\in \overline{s}_d} \tilde{w}_{di}\right)$$

Calculating $\tilde{w}_{di}$: For strict convex (concave) functions $h^{-1}()$, the naive back-transformed variable underestimates (overestimates) $\tilde{w}_{di}$:

$$\underbrace{h^{-1}(\tilde{y}_{di}) = h^{-1}(E[y_{di}|y_s])}_{\text{naive transformation}} \quad \underbrace{\leq}_{\textbf{Jensen inequality}} \quad \underbrace{E[h^{-1}(y_{di}|y_s)] = \tilde{w}_{di}}_{\text{empirical best prediction}}$$

Log and Log-shift transformation

The **log-transformation** is used in many applications.

$$h(w_{di}) = log(w_{di}), \qquad h^{-1}(y_{di}) = exp(y_{di})$$

The **log-shift transformation** (Yang, 1995) extends the log transformation by including a transformation parameter λ.

$$h(w_{di}) = log(w_{di} + \lambda), \qquad h^{-1}(y_{di}) = exp(y_{di}) - \lambda$$

The transformation parameter λ is estimated with the REML method like in Rojas-Perilla et al. (2017) and Kreutzmann et al. (2018)

Outline

Outline

Bias-corrected estimation using unit-level census data: General transformations

For a **general transformation** $h()$, a bias-correction can be performed via numerical integration:

$$\tilde{w}_{di} = E[h^{-1}(y_{di})|y_s] = \int_{-\infty}^{+\infty} h^{-1}(x) f_{y_{di}|y_s}(x) dx$$

$$y_{di}|y_s \sim N(\mu_{di|s}, v_{di|s})$$

For the **log transformation**, this integral can be solved analytically (cf. Molina & Martín, 2018):

$$\tilde{w}_{di} = E[h^{-1}(y_{di})|y_s] = exp(\mu_{di|s} + \underbrace{\frac{\sigma_u^2(1-\gamma_d)+\sigma_e^2}{2}}_{=\alpha_d \text{ (bias-correction)}})$$

Outline

BHF model with totals

For the log and log-shift transformation, the EBLUP-formula can be expressed with totals of the back-transformed auxiliary variable

log transformation

$$\tilde{\tau}_d \approx \frac{1}{N_d}\left(E_d exp(\tilde{u}_d + \alpha_d)\right)$$

log-shift transformation

$$\tilde{\tau}_d \approx \frac{1}{N_d}\left(E_d exp(\tilde{u}_d + \alpha_d)\right) - \lambda$$

with

$$E_d = \sum_{i=1}^{N_d} exp(x_{di}^t \tilde{\beta})$$

Remaining Question: How can we estimate these totals?

Estimation of the totals (1/3)

Datasources:

- **Census:** Aggregated data for the auxiliary variables
  - Area-level mean
  - Area-level variation (and covariance)
- **Sample:** Unit-level data

Step 1: Estimating the density of $x_{dj}^t \tilde{\beta}$ for each area

- No parametric assumptions
- Small sample sizes: Use additionally sampled data from other areas

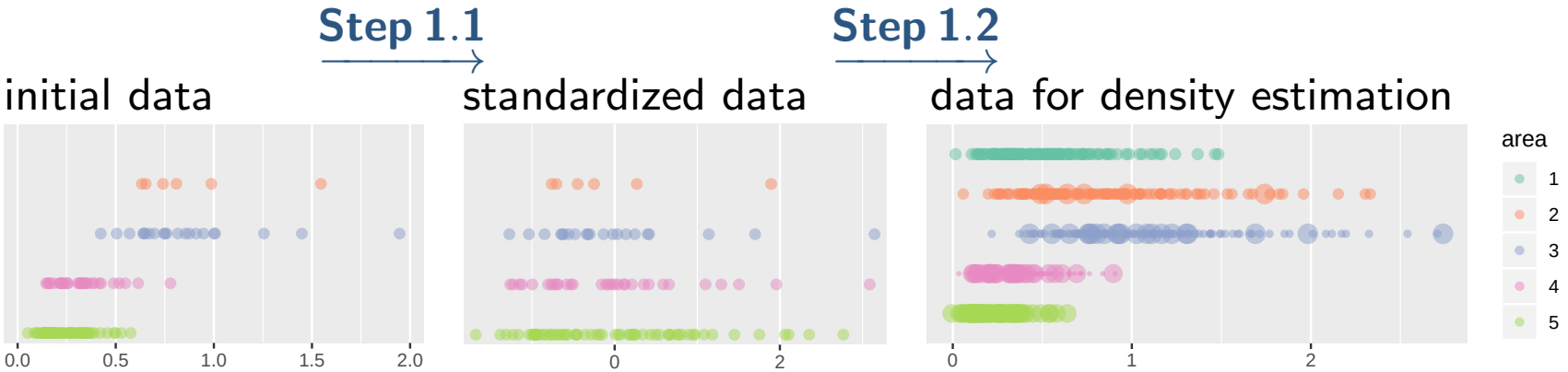
Estimation of the totals (2/3)

Step 1.1: Standardization with the observed area mean and standard deviation in the sample.

Step 1.2: Adjust the data with the true variation and mean.

sample size in area d

- low: Use data from all areas
- medium: Use data from all areas (higher weight on data from area d)
- high: Use only data from area d



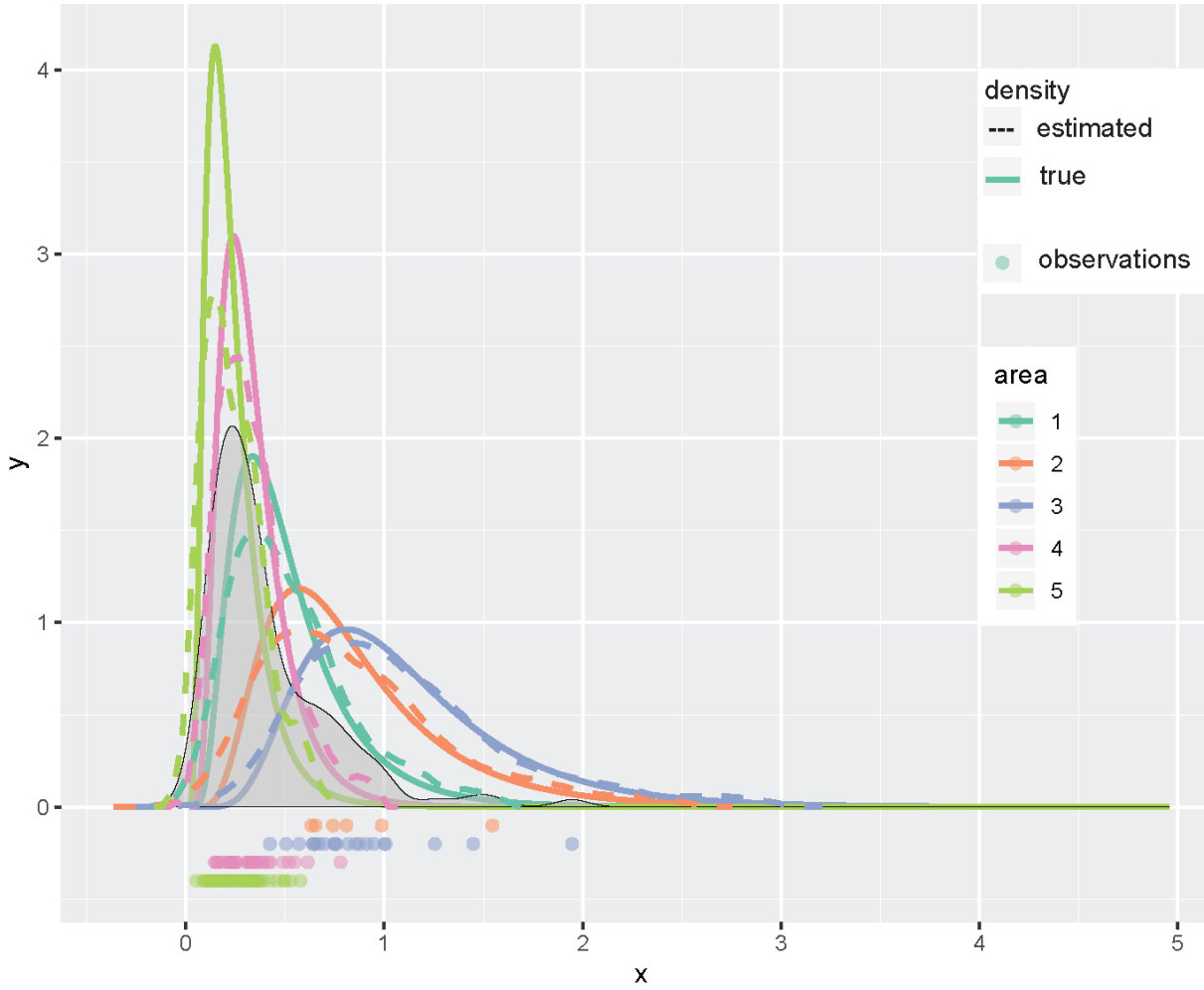
Estimation of the totals (3/3)

Step 1.3:
Kernel density estimation

kernel:
Epanechnikov

bandwidth:
Crossvalidation

Step 2:
Calculate the total via numerical integration



Outline

Outline

Model-based simulation study

Scenarios

Scenario	Model	x_{di}	z_{di}	μ_d	u_d	e_{di}
Normal	$4500 - 400x_{di} + u_d + e_{di}$	$N(\mu_d, 3)$		$U[-3, 3]$	$N(0, 500^2)$	$N(0, 1000^2)$
Log-scale	$exp(10 - x_{di} - 0.5z_{di} + u_i + e_{di})$	$N(\mu_d, 2)$	$N(0, 1)$	$U[2, 3]$	$N(0, 0.4^2)$	$N(0, 0.8^2)$

50 areas

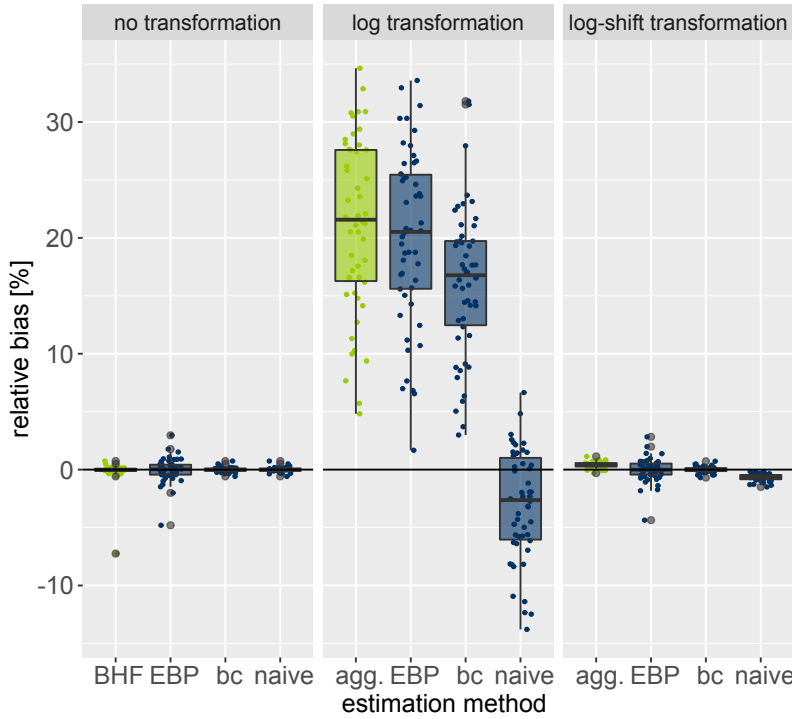
- Population: 200 individuals within each area
- Sample: various sample sizes

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	19.25	40.00	39.82	62.00	79.00

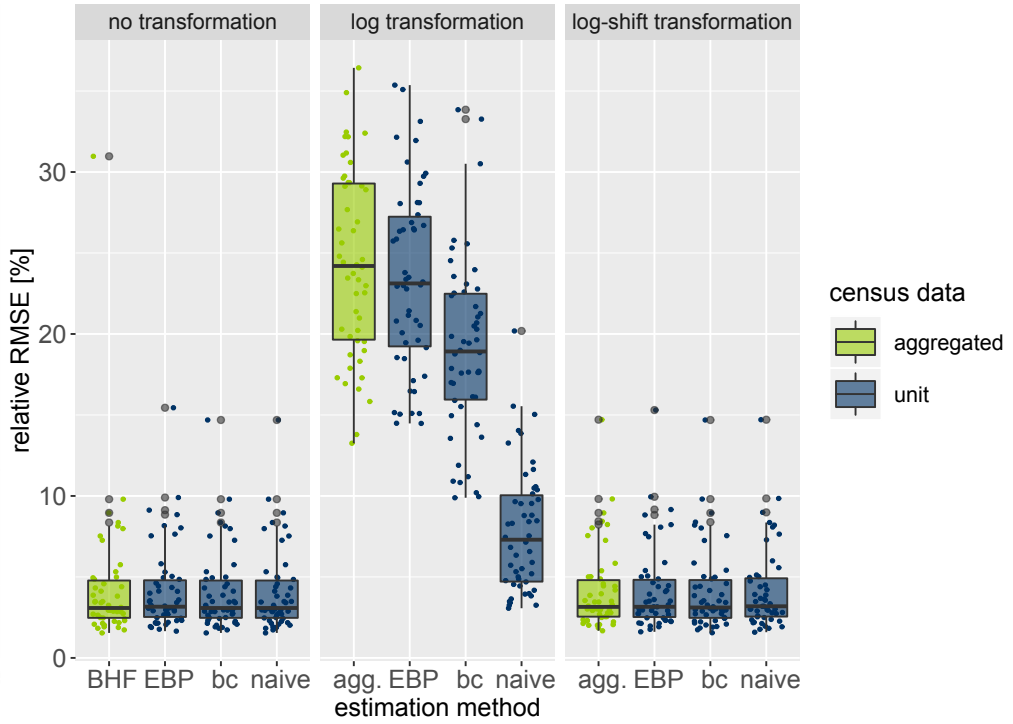
Thresholds for density estimation: $t_{low} = 10$ and $t_{high} = 60$

Scenario 1: Normal Setting

relative bias

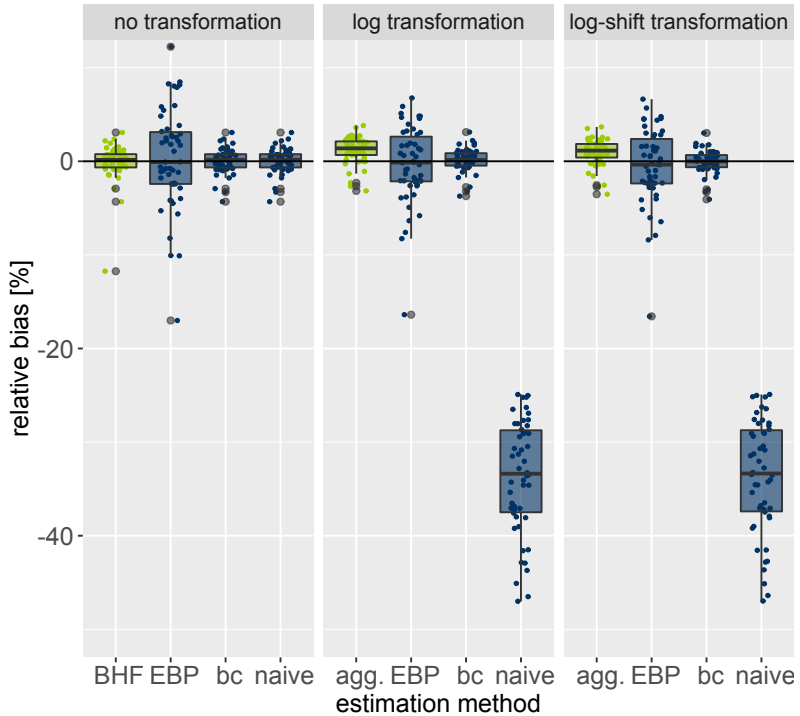


relative RMSE

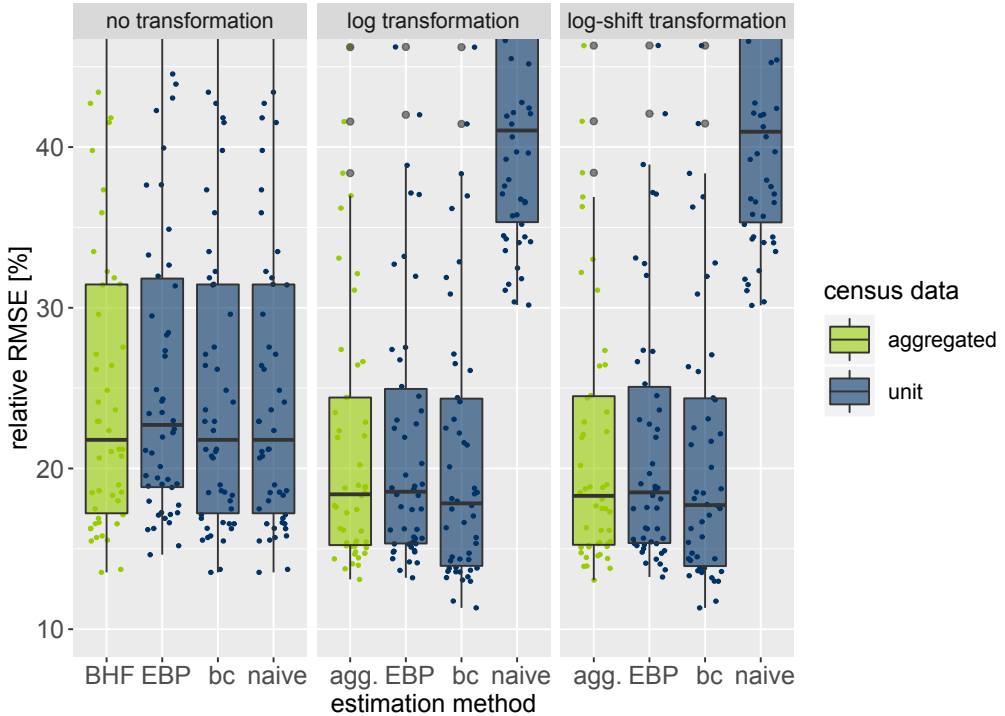


Scenario 2: Log-Scale Setting

relative bias



relative RMSE



Outline

Design-based simulation study

Data

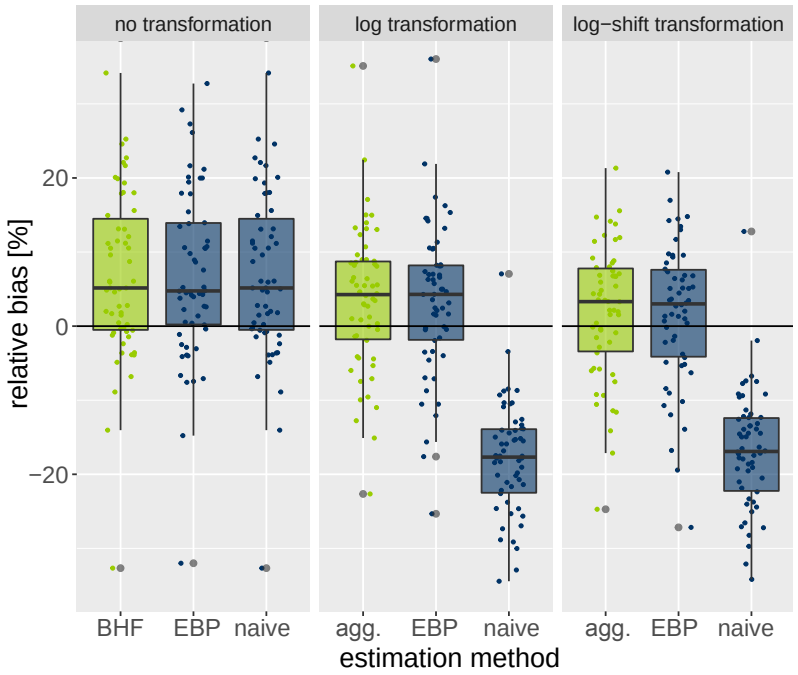
- Based on the Mexican census for the State of Mexico
- Outcome is the earned income from work per capita

Setup

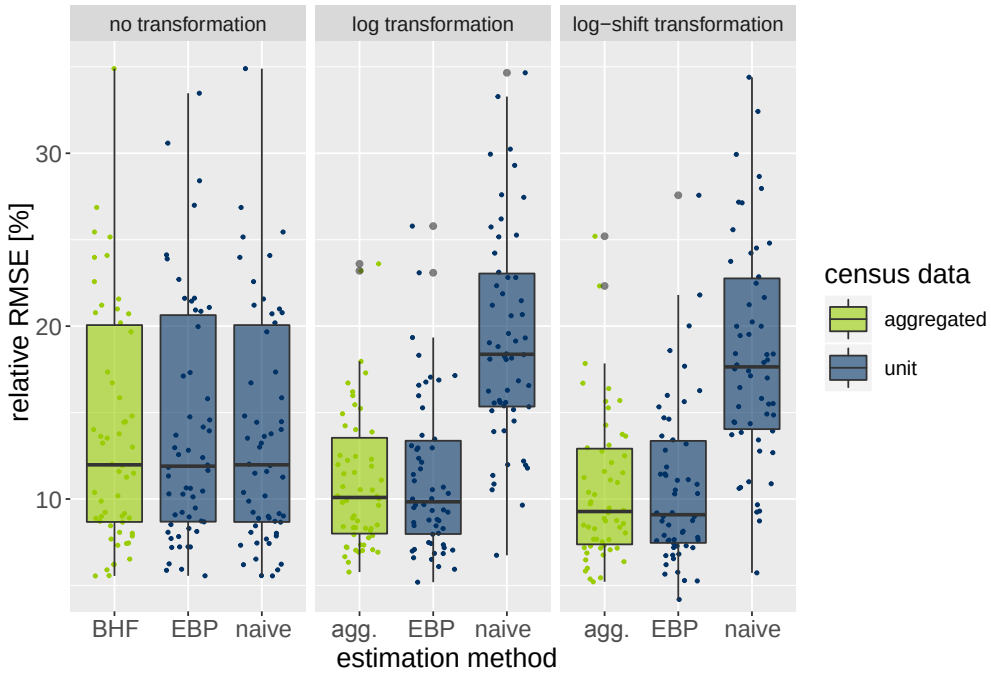
- Design-based simulation: 500 independently drawn samples
  - sampling design following the ENIGH survey (n = 2748)
  - 67 out-sample and 58 in-sample municipalities
  - sample size: Min.: 3, Median: 21, Max.: 527
- 4 covariates leading to a R^2 of around $40 - 50\%$
  - Employees older than 14 years (pct. in the household)
  - Income earners older than 14 years (pct. in the household)
  - Total number of communication assets in the household
  - Total number of goods in the household

Bias and Efficiency of in-sample areas

relative bias



relative RMSE



Bias and Efficiency

Table: Median of the relative Bias/RMSE for the in- and out-sample areas

	in-sample areas				out-sample areas			
	rBias		rRMSE		rBias		rRMSE	
	agg.	EBP	agg.	EBP	agg.	EBP	agg.	EBP
no trafo	5.15	4.75	11.98	11.90	13.08	12.75	17.28	17.72
log trafo	4.26	4.28	10.09	9.84	13.06	13.73	14.66	15.59
log-shift trafo	**3.31**	**3.00**	**9.28**	**9.09**	**12.56**	**12.62**	**14.57**	**14.98**

- In all cases, the log-shift transformation leads to the results with lowest bias and highest efficiency.
- Our method (agg.) using only aggregated covariate data provides comparably good estimates than the EBP method.

Outline

Conclusion and Further Research

Conclusion:

- Estimating means based on aggregated census data with the proposed method leads to as good estimates as methods relying on unit level data.
- The log-shift transformation worked as well as the suitable transformation in the model-based simulations.
  Due to flexibility, the use of data-driven transformations is recommended for working with real data.

Further research:

- Extend these approaches to general transformations within the BHF model using aggregated census data.
- Develop a MSE estimator for the proposed method.

The research is funded by a scholarship of Studienstiftung des deutschen Volkes.

Thank you very much for your attention.

Nora Würz (nora.wuerz@fu-berlin.de)

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (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.

Kreutzmann, A. K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. and Tzavidis, N. (2018). The R package **emdi** for the estimation and mapping of regional disaggregated indicators.

Journal of Statistical Software.

Molina, I. & Martín, N. (2018). Empirical best prediction under a nested error model with log transformation.

The Annals of Statistics, 46(5), 1961-1993.

References

Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2017). Data-Driven Transformations in Small Area Estimation.

Discussion Paper 30/2017, School of Business and Economics, Freie Universität Berlin.

Slud, E. V. & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 239-257.

Sugasawa, S. & Kubokawa, T. (2017). Transforming response values in small area prediction.

Computational Statistics & Data Analysis, 114, 47-60.

Yang, L. (1995). Transformation-density estimation.

Ph. D. thesis, University of North Carolina, Chapel Hill.

Bias-corrected estimation using unit-level census data: General transformations

For a **general transformation** $h()$, a bias-correction can be performed via numerical integration:

$$\tilde{w}_{di} = E[h^{-1}(y_{di})|y_s] = \int_{-\infty}^{+\infty} h^{-1}(x) f_{y_{di}|y_s}(x)dx$$

$$\overset{y_{di}|y_s \sim N(\mu_{di|s}, v_{di|s})}{=} \int_{-\infty}^{+\infty} h^{-1}(x) \frac{1}{\sqrt{2\pi v_{di|s}}} exp\left(-\frac{(x-\mu_{di|s})^2}{2v_{di|s}}\right) dx$$

with $\mu_{di|s} = x_{di}^t\beta + \gamma_d(\frac{1}{n_d}\sum_{i\in s_d} y_{di} - x_{di}^t\beta)$ and $v_{di|s} = \sigma_u^2(1-\gamma_d) + \sigma_e^2$

BHF model with totals

For the log and log-shift transformation, the BHF-formula for the area means can be expressed with totals ($E_d = \sum_{i=1}^{N_d} exp(x_{di}^t \tilde{\beta})$):

log transformation

$$\begin{aligned}\tilde{\bar{y}}_d &= \frac{1}{N_d}\left(\sum_{i\in s_d} exp(y_{di}) + \sum_{i\in r_d} exp(\tilde{y}_{di} + \alpha_d)\right) \\ &\approx \frac{1}{N_d}\left(\sum_{i\in s_d} exp(x_{di}^t\tilde{\beta})exp(\tilde{u}_d + \alpha_d) + \sum_{i\in r_d} exp(x_{di}^t\tilde{\beta})exp(\tilde{u}_d + \alpha_d)\right) \\ &= \frac{1}{N_d}\left(E_d exp(\tilde{u}_d + \alpha_d)\right)\end{aligned}$$

log-shift transformation

$$\tilde{\bar{y}}_d \approx ... = \frac{1}{N_d}\left(E_d exp(\tilde{u}_d + \alpha_d)\right) - \lambda$$

Question: How can we estimate the totals?

Estimation of the totals

Step 1.1: Standardization of each observed $x_{di}^t\tilde{\beta}$ with the observed area mean and standard deviation in the sample:

$$z_{di} = \frac{x_{di}^t\tilde{\beta} - mean(x_d^t\tilde{\beta})}{sd(x_d^t\tilde{\beta})}$$

Step 1.2: Adjust the data with true variation and mean:

$$r_{di*} = z_{i*}\underbrace{sd(x\tilde{\beta})_d^{TRUE}}_{=sd(x)_d^{TRUE}\tilde{\beta}_1} + \tilde{\beta}_0 + \underbrace{mean(x\tilde{\beta})_d^{TRUE}}_{=\tilde{\beta}_0+mean(x)_d^{TRUE}\tilde{\beta}_1}$$

Define two thresholds t_{low} and t_{high} for the sample sizes.

thresholds	define z_{i*}	weights for concerning r_{di*}
$n_d < t_{low}$	z_{di} from all areas	equal weights
$t_{low} < n_d < t_{high}$	z_{di} from all areas	higher weight for r_{di*} from d-th area data
$n_d > t_{high}$	z_{di} from area d	equal weights

Estimation of the totals

Step 2: Numerical integration

For calculating the total out of the estimated density (step 1), we perform a numerical integration with $\mathrm{integrate.xy}()$.

$$E_d = \sum_{i=1}^{N_d} exp(x_{di}^t \tilde{\beta}) = N_d E[exp(x_{di}^t \tilde{\beta})]$$
$$= N_d \int_{-\infty}^{+\infty} exp(x) f_{x_{di}^t \tilde{\beta}}(x) dx$$