From start to finish: A framework for the production of small area official statistics

Nikos Tzavidis, Li-Chun Zhang, Angela Luna Hernandez
(University of Southampton)
Timo Schmid, **Natalia Rojas-Perilla**[1]
(Freie Universität Berlin)

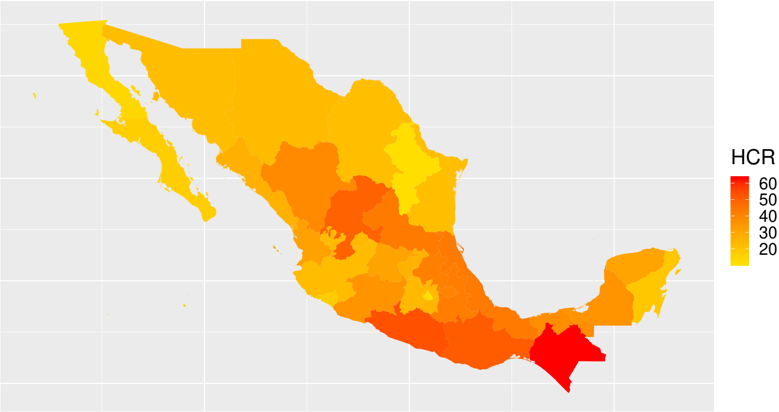
Taller Regional sobre desagregación de estadísticas sociales mediante metodologías de estimación en áreas pequeñas

November 2018
Santiago de Chile

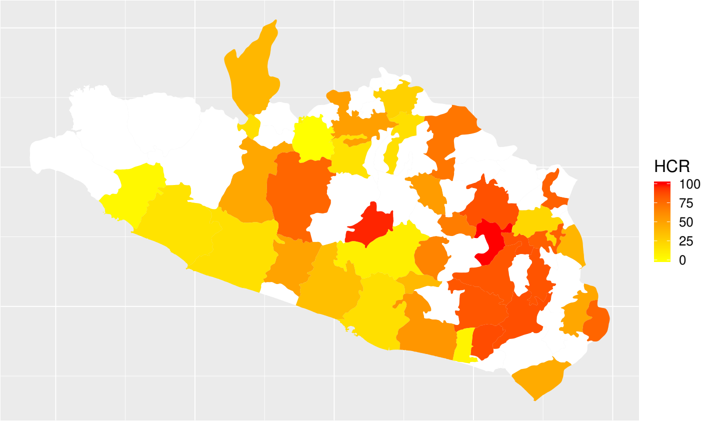
[1]Presenting author

Contents

Poverty in Mexico (regional level)



Poverty in Guerrero (municipal level)



Why do we need disaggregated prediction?

- Reliable **knowledge** on the **socio-demographic indicators** of a country's population **is essential** for sound evidence-based **policymaking**.
- Traditionally, this knowledge is **collected via household surveys** and is provided by National Statistical Institutes.
- The surveys are generally designed to provide **reliable estimates** for the indicators only **for larger domains** like national or regional level.

Introduction to Small Area Estimation

- *Direct estimators* are often not reliable for some domains/ areas of interest
- In these cases we have two choices:
  1. Oversampling over that domains
  2. Applying statistical techniques that allow for reliable estimates in that domains ✓

Small Domain or Small Area Geographical area or domain where direct estimators do not reach a minimum level of precision.

Small Area Estimator (SAE) An estimator created to obtain reliable estimates in a Small Area.

Combining different data sources

- In order to provide reliable estimates in all subdomains, efficient ways of *combining* information are required
- **Survey Data:** Available for y and for x related to y
- **Census/Administrative Data:** Available for x but not for y

SAE in 3 Steps:

1. Use survey data to estimate **models** that link y to x
2. Combine the estimated model parameters with x, for out of sample units, to form predictions
3. Estimate **finite population target parameters**

SAE methods and their practicioners

- Increasing interest on small area estimates due to their potential to inform policy decisions
- Increasing complexity of **targets of estimation**
- SAE is an area with rapid methodological development
- Productive cooperation between academia and practitioners
- Significant progress with uptake of SAE methods by National Statistical Institutes (NSIs)
- Not yet part of the standard set of **official statistics outputs**

Motivation

Methodological overview

Emdi

What are the inputs for SAE practitioners?

Aim[2]: To propose practical guidelines for the production of official small area statistics:

- To provide a framework based on three broadly defined stages:

Specification → Analysis & Adaptation → Evaluation

- To use **parsimony** as the core principle for decisions

[2] Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T., and Rojas-Perilla, N., (2018). **From Start to Finish: A Framework for the Production of Small Area Official Statistics.** *Journal of the Royal Statistical Society, Series A*

(I) General SAE specifications

Specification ⟶ Analysis & Adaptation ⟶ Evaluation

Define targets of estimation:

- Existing estimates and experience with particular methods?
- Linear and non-linear indicators
- Target geography

Determine data availability and geographical coverage:

- Methods and data requirements
- Unrealistically granular geographies

Guerrero case study: (I) Specification

Case study: Estimation of poverty (non-linear) indicators for municipalities in Guerrero (by law)

- Data: Survey microdata (ENIGH 2010), census microdata (NPHC 2010)
- 81 municipalities in state of Guerrero. Only 40 are included in the survey
- Appropriate for **unit-level models**

	Min.	Median	Max.
Sample	**9**	24	**511**
Census	**1097**	4922	**208910**

- Outcome: Total household per-capita income from work
- Explanatory variables: Socio-demography

(II) Analyze and adapt SAE methods

Specification ⟶ Analysis & Adaptation ⟶ Evaluation

Initial simple estimates:

1. Direct estimator ✓
2. Synthetic estimator
3. Composite estimator

More complex SAE methods (area- & unit-level):

- Following the principle of parsimony
- Simultaneously estimating linear and non-linear indicators

Initial triplet of estimates

1. Direct estimator

- Hajek-Brewer Ratio estimator (No auxiliary information)

$$\hat{\bar{\theta}}_k^{Direct} = \left(\sum_{i=1}^{n_k} y_{ik}/\pi_{ik}\right) / \left(\sum_{i=1}^{n_k} 1/\pi_{ik}\right)$$

- GREG estimator (Auxiliary information)

$$\hat{\bar{\theta}}_{k,GREG}^{Direct} = \frac{1}{N_k}\sum_{i=1}^{n_k} w_{ik} y_{ik},\, w_{ik} = g_{ik}/\pi_{ik},$$

$$g_{ik} = 1 + (X - \sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik}/\pi_{ik})^T (\sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik}\mathbf{x}_{ik}^T/\pi_{ik})^{-1}\mathbf{x}_{ik}$$

Initial triplet of estimates (Cont)

2. Synthetic estimator

- No auxiliary information

$$\hat{\theta}_k^{Synthetic} = \hat{\theta}_g$$

- Auxiliary information

$$\hat{\theta}_k^{Synthetic} = \bar{\boldsymbol{x}}_k^T \hat{\boldsymbol{\beta}},$$

$$\hat{\boldsymbol{\beta}} = (\sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik} \boldsymbol{x}_{ik}^T / \pi_{ik})^{-1} (\sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik} y_{ik} / \pi_{ik})$$

3. Composite estimator

$$\hat{\theta}_k^{Composite} = \alpha_k \hat{\theta}_k^{Direct} + (1 - \alpha_k) \hat{\theta}_k^{Synthetic}$$

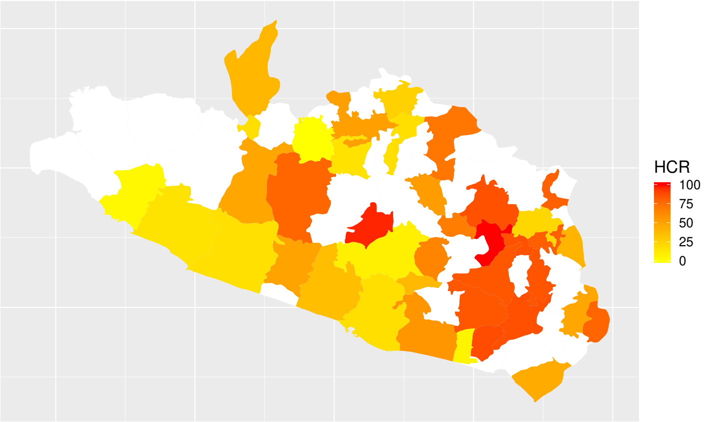
Direct estimation based only on the survey

Weighted Head Count Ratio (HCR)

$$\hat{\theta}_i^{HCR} = \frac{1}{\sum_{j=1}^{n} w_{ij}} \sum_{j=1}^{n} w_{ij} I(y_{ij} < t), \quad j = 1, \ldots, n, \quad i = 1, \ldots, D,$$

where y_{ij} is the variable of interest, w_{ij} are the sampling weights for domain i and individual j, t is a selected threshold and I an indicator variable that is 1 if y_{ij} is below the threshold

Guerrero case study: (II) Analysis & Adaptation (Direct estimation)



Guerrero case study: Use of unit-level models

Modeling of a welfare variable using unit-level models. Some alternatives:

- The World Bank Approach (Elbers et al., 2003)
- The Empirical Best Predictor (EBP) Approach (Molina & Rao, 2010) ✓
- M-Quantile methods (Chambers & Tzavidis, 2006)
- EBP with normal mixtures (Elbers and van der Weide, 2014)
- Microsimulation via quantiles - MvQ (Weidenhammer et al., 2016)

The EBP approach (Molina & Rao, 2010)

Nested error regression model (Battese et. al, 1988)

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_j + e_{ij}, \quad i = 1, ..., n_j, \quad j = 1, ..., D$$

1. Use sample data to fit the model and obtain $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and predict $\bar{u}_j = E(u_j|y_s)$ for in-sample domains
2. Use census data to micro-simulate L synthetic populations by

$$y_{ij} = \mathbf{x}_{ij}^T \hat{\beta} + \bar{u}_j + u_j + e_{ij},$$

where $u_j \sim N(0, \hat{\sigma}_u^2 * (1 - \gamma_j))$ and $e_{ij} \sim N(0, \hat{\sigma}_e^2)$

3. Calculate the target linear and non-linear indicators in each replication and average over L

Target model assumptions

- EBP relies on Gaussian assumptions :
  - $u_k \overset{iid}{\sim} N(0, \sigma_u^2)$, the random area-specific effects ✓
  - $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, the unit-level error terms, $u_k \perp \epsilon_{ik}$ ✓
- Model assumptions
  - Normality ✓
  - Homoscedasticity
  - Linearity
- Residual diagnostics
  - Q-Q plots of residuals at different hierarchical levels
  - Influence diagnostics
  - Plots of standardised residuals

How to deal with model assumption violations?

- Develop a method that allows for the corresponding violation/new assumptions
- Use a distribution-free/flexible method (robust or non-parametric statistics)
- Redesign the model by transforming the data for satisfying model assumptions
  - **Parsimonious solution**
  - Allows practitioners to apply methods available for parametric statistics

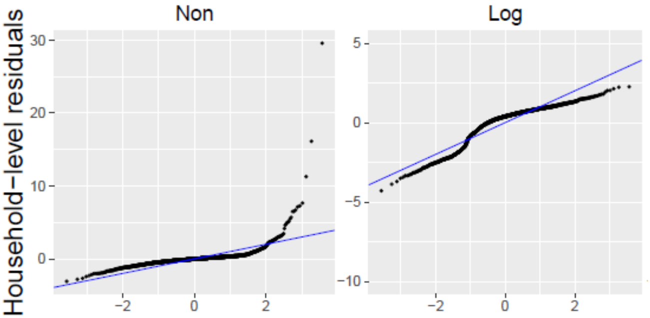
Adapt the model if model assumptions not met

- EBP formulation under an alternative distribution (Molina, I et al., 2015)
- Robust methods (Ghosh et al., 2008; Sinha & Rao, 2009; Chambers et al., 2014; Dongmo-Jiongo et al., 2013)
- Use non-parametric methods (Opsomer et al., 2008; Ugarte et al., 2009)
- Elaborate random effects structure - spatial structures (Pratesi & Salvati, 2009; Schmid et al., 2016)
- Consider extensions to two-fold models (Marhuenda et al., 2017)
- Use of transformations ✓

Guerrero case study: (II) Analysis & Adaptation

EBP - Nested error regression model (Battese et. al, 1988)

$$\log(y_{ij}) = \mathbf{x}_{ij}^T\beta + u_j + e_{ij}, \quad u_i \overset{iid}{\sim} N(0,\sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} N(0,\sigma_e^2)$$



Data-driven transformations (e.g. Box-Cox)

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

Relevant steps of using data-driven transformations?

- Relevant steps:

Assumption → Transformation → Estimation

- Modelling issues under transformations

What is the proposed adaptation gap?

Aim[3]: To tackle model assumption violations by using data-driven transformations as opposed to ad-hoc chosen transformations:

- To **generalize** an ML-based and data-driven approach for estimating the transformation parameter under the linear mixed regression model
- To **incorporate** the ML-based method in the SAE context
- To **propose** MSE estimation approaches that account for the uncertainty due to the estimation of the transformation parameter: Parametric and semi-parametric (wild) bootstrap

[3] Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N., (2018). **Data-Driven Transformations in Small Area Estimation.** *Journal of the Royal Statistical Society, Series A*, under revision

The EBP approach under transformations

1. Select $T_\lambda(y_{ij}) = y_{ij}^*$ **and obtain the transformed sample data**
2. Use **transformed sample data** to fit the model and obtain $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and predict $\bar{u}_j = E(u_j|y_s)$ for in-sample domains
3. Use census data to micro-simulate L synthetic populations by

$$y_{ij}^* = \mathbf{x}_{ij}^T \hat{\beta} + \bar{u}_j + u_j + e_{ij},$$

   where $u_j \sim \ N(0, \hat{\sigma}_u^2 * (1 - \gamma_j))$ and $e_{ij} \sim \ N(0, \hat{\sigma}_e^2)$

4. **Back-transform to the original scale**
5. Calculate the target linear and non-linear indicators in each replication and average over L

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

Selected data-driven transformations

- Shifted transformations
  - **Log-shift**
- Power transformations
  - **Box-Cox**
  - Manly
  - Gpower
  - Modulus
  - **Dual power**
  - Bickel-Docksum
  - Yeo-Johnson
- Multi-parameter transformations
  - Johnson
  - Sinh-arcsinh

Functional form of transformations

Log-Shift transformation (λ) (Royston et al., 2011)

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

Box-Cox transformation (λ) (Box & 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 $y_{ij} + s > 0$

Estimation methods

1. Parametric maximum likelihood-based approaches
   - Rely on parametric assumptions
   - It is very sensitive to outliers
   - The transformations needs to be differentiable to quarantine the existence of the Jacobian
2. Analytical approaches
   - Optimizing distributional moments
   - Minimizing the divergence (distance between two distributions)

Estimation algorithm for the EBP method (λ)

Residual Maximum Likelihood (REML) (Gurka et al., 2006)

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 θ conditional on the fixed λ
5. Repeat 3 and 4 until a maximum $\hat{\lambda}$ is found
6. Apply the EBP method

Scaled transformations

Scaled log-shift transformation (λ)

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

Scaled Box-Cox transformation (λ)

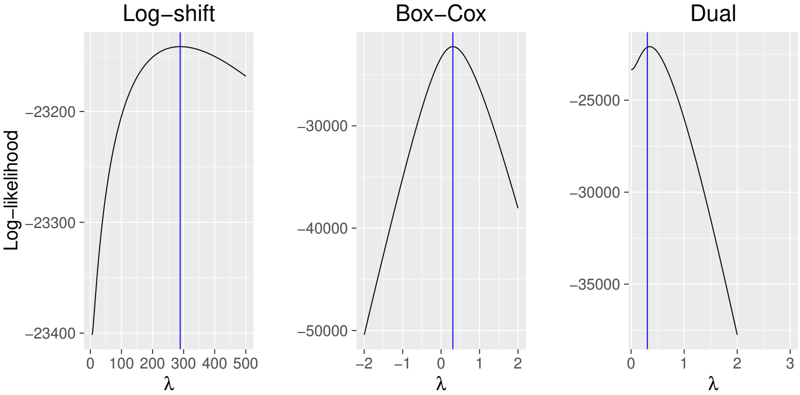
$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda} & \lambda \neq 0; \\ \alpha \log(y_{ij} + s) & \lambda = 0. \end{cases}$$

Scaled dual power transformation (λ)

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

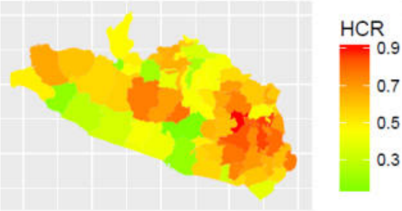
with α chosen in such a way that the Jacobian of the transformation is 1

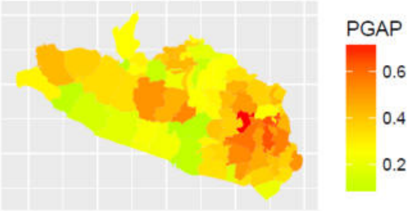
Guerrero case study: (II) Analysis & Adaptation



Guerrero case study: (II) Analysis & Adaptation

	R_m^2	R_c^2	λ	ICC
Non	0.331	0.346	-	0.023
Log	0.263	0.416	-	0.207
Log-Shift	**0.419**	**0.517**	**68.159**	**0.169**
Box-Cox	0.439	0.517	0.263	0.140
Dual	0.443	0.517	0.304	0.132





(III) Evaluate SAE methods

Specification → Analysis & Adaptation → Evaluation

Uncertainty assessment and Method evaluation

- Uncertainty: Area-specific MSE estimation
- Method evaluation: Model vs. design-based evaluations

Uncertainty assessment and Method evaluation

Uncertainty assessment

- Process of obtaining a measure of uncertainty of SA estimates. Often, the chosen measure is an area-specific MSE
- Important for production of official statistics

Method evaluation

- Assessment of an estimator's performance under departures from model assumptions
- Important when deciding which methodology to use for a particular problem

Bootstrap MSE under transformations

- We propose two bootstraps for accounting the uncertainty coming from the estimation of the transformation parameter
- The difference is the mechanism use for generating the bootstrap populations
    1. Parametric: Generates bootstrap populations of the random and unit-level error terms parametrically
    2. Semi-parametric (Wild): The use of a data-driven transformation may reduce deviations from the normality, there may still be departure in the tails of the unit-level distribution.

Parametric bootstrap MSE estimation

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 $T(y_{ij}) = \tilde{y}_{ij}$, generate $u_i^* + e_{ij}^*$ and 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 ${y_{ij}}^{*(b)}$
   - Calculate the poverty measures $F_i(\alpha, t)^{(b)}$
   - Extract the bootstrap sample in $\tilde{y}_{ij}^{*(b)}$ and perform the EBP method on them. Note, as the transformed sample data is used, the estimation of λ is skiped and the original $\hat{\lambda}$ is used to transform the data back to the original scale. Obtain $\hat{F}_i^{BP}(\alpha, t)^{(b)}$
2. $MSE_*\left[F_i^{BP}(\alpha, t)\right] = 1/B \sum_{b=1}^{B} \left[\hat{F}_i^{BP}(\alpha, t)^{(b)} - F_i(\alpha, t)^{(b)}\right]^2$

Accounting for the λ estimation

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 ${y_{ij}}^{*(b)}$
   - Calculate the poverty measures $F_i(\alpha, t)^{(b)}$
   - 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 $\hat{F}_i^{BP}(\alpha, t)^{(b)}$
2. $MSE_* \left[F_i^{BP}(\alpha, t) \right] = 1/B \sum_{b=1}^{B} \left[\hat{F}_i^{BP}(\alpha, t)^{(b)} - F_i(\alpha, t)^{(b)} \right]^2$

Motivation

Methodological overview

Emdi

How can `emdi` support this framework?

Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:

- Estimation → `direct()` and `ebp()`
- Model summary → `summary()`
- Diagnostic plots → `plot()`
- Selection of indicators → `estimators()` and `compare()`
- Visualization → `map_plot()`
- Export → `write.excel()`

[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:

Estimation	`direct()` and `ebp()`
Model summary	`summary()`
Diagnostic plots	`plot()`
Selection of indicators	`estimators()` and `compare()`
Visualization	`map_plot()`
Export	`write.excel()`

[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:

- Estimation → `direct()` and `ebp()`
- Model summary → `summary()`
- Diagnostic plots → `plot()`
- Selection of indicators → `estimators()` and `compare()`
- Visualization → `map_plot()`
- Export → `write.excel()`

[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:

Estimation	direct() and ebp()
Model summary	summary()
Diagnostic plots	plot()
Selection of indicators	estimators() and compare()
Visualization	map_plot()
Export	write.excel()

[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

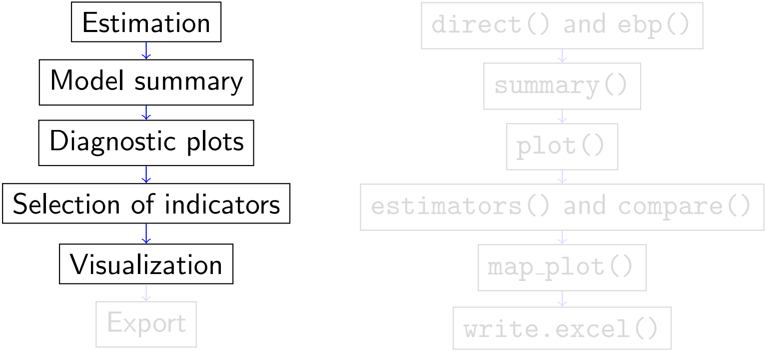
Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:

Estimation	`direct()` and `ebp()`
Model summary	`summary()`
Diagnostic plots	`plot()`
Selection of indicators	`estimators()` and `compare()`
Visualization	`map_plot()`
Export	`write.excel()`

[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

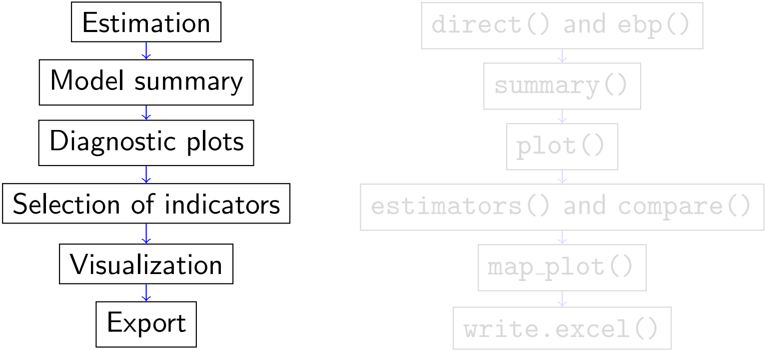
Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:



[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

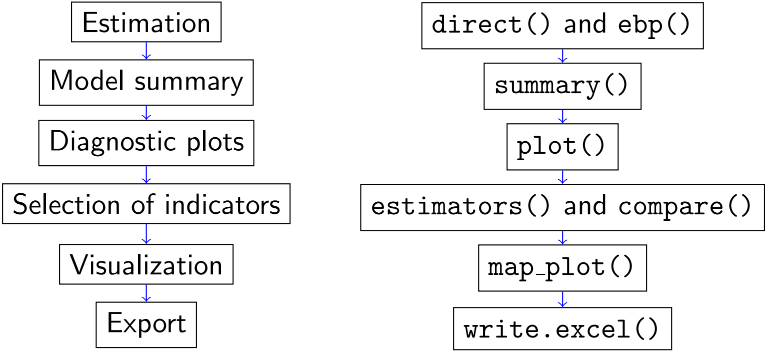
Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:



[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

How can `emdi` support this framework?

Aim[4]: To create a user-friendly R package for the estimation of regionally disaggregated indicators using SAE methods:



[4] Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package** emdi **for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*, forthcoming

Conclusions and future research directions cont'd

For practitioners of transformations

- Transformations for linear mixed models
- Inference and bias correction procedures
- Both-sides transformation

Transformations in the SAE context

- Extension to the ELL approach and more kinds of models
- Analysis of multi-parameter transformations
- Comparison with other alternatives
- Model selection under the linear mixed model

Essential bibliography

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

ENIGH. (2010) *Encuesta Nacional de Ingresos y Gastos de los Hogares 2010.*

Gurka, M. J., Edwards, L. J., Muller, K. E. & Kupper, L. L. (2006) *Extending the Box–Cox Transformation to the Linear Mixed Model.* Journal of the Royal Statistical Society. Series A (Statistics in Society), 26, 211-252.

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

Rao, J.N.K. (2003) *Small area estimation.* John Wiley & Sons.

Thank you very much for your attention.

Natalia Rojas-Perilla (natalia.rojas@fu-berlin.de)

Relevant Transformations for Normality

It often implies corrections for homoscedasticity and linearity

1. Logarithm: Due to its effectiveness in turning highly left- and right-skewed or log-normal distributions into more symmetrical ones, it is commonly used
2. Tukey, 1957: Proposed a family of power transformations based on monotonic functions y^λ for $\lambda \neq 0$ and $\log(y)$ for $\lambda = 0$. Box-Cox, 1964 modified this to avoid the discontinuity at $\lambda = 0$ (the limit)
    - Negative values
    - Truncation problem
3. Multiparametric Johnson for scale, location, and shape (skewness and tailweight)

Truncation Problem of Box-Cox

The inverse function of the Box-Cox transformation is:

$$f^{-1}(y_i) = (\lambda y_i + 1)^{\frac{1}{\lambda}}.$$

When $\lambda y_i + 1 > 0$?

- $\Rightarrow y_i > \frac{-1}{\lambda}$ for $\lambda > 0$
- $\Rightarrow y_i < \frac{1}{\lambda}$ for $\lambda < 0$

This is a problem for prediction using SAE methods. We are back-transforming an expression based on different sources and the back-transforming data may be out the range

Selected Transformations for Normality

1. Negative values:
   - Manly: Exponential distribution family
   - Modulus: Not used in extreme departures
   - Bickel-Doksum: Kurtosis problem
   - Convex-to-concave principle: Gpower (for picked distributions), Yeo-Johnson (similar to the Box-Cox)
2. Truncation problem:
   - Dual power: Consistency

Transform Both-Sides Model (TBS)

For non-linear regression models

$$T(y_i, \lambda) = T\left[f(x_i, \beta), \lambda\right] + e_i$$

For linear regression models

$$T(y_i, \lambda) = T\left[x_i, \beta, \lambda\right] + e_i$$

- It is assumed that $f(x_i, \beta)$ already fits the data adequately, but residual variation is non constant or there are departures from normality, preserving the relationship
- The classical ML-based method under the linear regression model is used for finding λ by incorporating the transformation on the explanatory variables

TBS for Linearizing Relationships

- TBS model under the Box-Cox transformation is useful for correcting non-linear relationships. The techniques imply the correction of other model assumption violations.
- For finding the different powers for the target and explanatory variables we can use the “ladder of transformations” empirical rule
- The Box and Tidwell method for finding the powers is an iterative method for testing which variable should be transformed or not and optimize individually the transformation parameter.

Interpretation of Transformations

1. Level-level: For a change x by one unit, one expect y to change by β units
2. Log-Level: For a change x by one unit, one expect y to change by 100β percent
3. Level-Log: For a change x by one percent, one expect y to change by $\beta/100$ units
4. Log-Log: For a change x by one percent, one expect y to change by β percent

Transformation Derivations

$$y_i^*(\lambda) = \begin{cases} \frac{(y_i+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \log(y_i + s) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Box-Cox (shift)(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{BC}}$.

Transformation Derivations Cont'd

The definition of the geometric mean is:

$$\bar{y}_{BC} = \left[\prod_{i=1}^{n} y_i + s\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\lambda(y_i + s)^{\lambda-1}}{\lambda} \\ &= \prod_{i=1}^{n} (y_i + s)^{\lambda-1} \\ &= \bar{y}_{BC}^{n(\lambda-1)}. \end{aligned}$$

Transformation Derivations Cont'd

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{(y_i + s)^\lambda - 1}{\lambda} \frac{1}{\bar{y}_{BC}^{\lambda-1}}.$$

The inverse function of the Box-Cox (shift)(A) transformation is:

$$\begin{aligned} f(y_i) &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ x_i &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ y_i &= (\lambda x_i + 1)^{\frac{1}{\lambda}} - s \\ \Rightarrow f^{-1}(y_i) &= (\lambda y_i + 1)^{\frac{1}{\lambda}} - s. \end{aligned}$$

Estimation Methods

1. Parametric maximum likelihood-based approaches
    - Rely on parametric assumptions
    - It is very sensitive to outliers
    - The transformations needs to be differentiable to quarantine the existence of the Jacobian
2. Analytical approaches
    - A non-parametric alternative to estimate λ focus directly on optimizing the form of the distribution of the error terms (3,4 moments optimization)
    - Considering only the skewness or kurtosis may ignore some properties of the distribution. Therefore, we propose to minimize the divergence (distance between two distributions)

Estimation Methods: Moments optimization

- The kurtosis and skewness are crucial features for defining the shape of a normal distribution
- To find a transformation parameter under which the empirical distribution of residuals has skewness and kurtosis as close as possible to zero and three respectively

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\operatorname{argmin}} \, |S_{e_\lambda}|,$$

where S_{e_λ} is the skewness and $\sigma^2_{e_\lambda}$ denotes the variance of the unit-level error terms.

- The index λ is used to emphasize that the skewness and the variance parameters depend on the transformation parameter.

Estimation Methods: Moments optimization

- In the context of linear mixed regression models, there are two independent error terms to be considered.
- We propose a pooled skewness approach that uses a weight w to ensure that the larger the error term variance $\sigma^2_{e_\lambda}$ is, the more weight its skewness will have in the minimization.
- Let S_{u_λ} be the skewness and $\sigma^2_{u_\lambda}$ be the variance of the area-specific random effects u_i. The pooled skewness approach is defined as follows:

$$\hat{\lambda}_{\text{poolskew}} = \underset{\lambda}{\operatorname{argmin}} \left(w|S_{e_\lambda}| + (1-w)|S_{u_\lambda}| \right),$$

$$\text{where } w = \frac{\hat{\sigma}^2_{e_\lambda}}{\hat{\sigma}^2_{u_\lambda} + \hat{\sigma}^2_{e_\lambda}}.$$

Estimation Methods: Divergence Optimization

- Considering only the skewness may ignore other properties of the distribution.
- Two divergence, the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) are used,

$$\hat{\lambda}_{KS} = \underset{\lambda}{\text{argmin}} \ \sup \ |F_n(\cdot) - \Phi(\cdot)|,$$

$$\hat{\lambda}_{CvM} = \underset{\lambda}{\text{argmin}} \int_{-\infty}^{\infty} [F_n(\cdot) - \Phi(\cdot)]^2 \phi(\cdot),$$

- $F_n(\cdot)$ is the empirical cumulative distribution function estimated by using the normalized residuals
- $\Phi(\cdot)$ is the theoretical distribution function of a standard normal distribution and $\phi(\cdot)$ its density.

Estimation Methods: Robust and Bayesian Box-Cox

All are developed under the linear model for the Box-Cox transformation

- ML-based methods are not robust to outliers (in the outcome variable) and depend on parametric assumptions
- Caroll et. al. 1988 generated a robust estimation used for heavy tailed distributions by applying the Huber estimator
- By using the classical profile likelihood, the Huber function is used
- Bayesian approaches use the ML method and applying a non-informative a priori distribution

Why Should We Use REML?

The likelihood is the function of model parameters given specific observed data. In many cases the likelihood depends on different parameters and therefore there are different likelihood types to eliminate effect of nuisance parameters (conditional, marginal, profile).

1. ML (Profile)
   - It does not take the degrees of freedom lost by estimating the fixed effects. It assumes they are known
   - It is biased for variance estimates (underestimation)
2. REML
   - It focuses on variance estimation and the other models parameters are estimated in a second step. It maximizes only the portion of the likelihood that does not depend on the fixed effects.
   - It is a restricted version of ML for eliminating (improving) the bias, leading to less unbiased estimation of the variance components
   - It takes the degrees of freedom into account
   - It is less sensitive to outliers in comparison to ML
   - For small sample sizes it is preferred

REML Derivations

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the Box-Cox transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$J(\lambda, \boldsymbol{y}) = \prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| = \prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}.$$

The log-likelihood function can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) \quad = \quad & -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ - \quad & \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ - \quad & \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda)-\boldsymbol{X_i}\hat{\beta}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda)-\boldsymbol{X_i}\hat{\beta}] + n(\lambda-1)\log\underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)\right)^{\frac{1}{n}}}_{=\bar{y}}. \end{aligned}$$

REML Derivations Cont'd

In order to obtain the scaled transformation of the Box-Cox family, $z^*_{ij}(\lambda)$, the denominator of the term $\frac{y^*_{ij}(\lambda)}{J(\lambda,\mathbf{y})^{1/n}}$ is given by:

$$1/J(\lambda,\mathbf{y})^{\frac{1}{n}} = J(\lambda,\mathbf{y})^{-\frac{1}{n}} = \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}\right]^{-\frac{1}{n}}$$
$$= \bar{y}^{-(\lambda-1)}.$$

Therefore, the scaled Box-Cox transformation is defined as follows:

$$z^*_{ij}(\lambda) = \frac{y^*_{ij}(\lambda)}{J(\lambda,\mathbf{y})^{1/n}} = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\bar{y}^{\lambda-1}\lambda}, & \lambda \neq 0, \\ \bar{y}\log(y_{ij}+s), & \lambda = 0, \end{cases}$$

for $y_{ij} > -s$.

Transformation Theorem

Let y be a continuous random variable with density function $f(y)$, taking values in $\mathbb{R}^n$. Let $T(y) = y^*$ a continuous transformation $T(y) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, for which the inverse $T^{-1}(y^*)$ is also continuous. Suppose that the inverse of the transformation is differentiable for all values of $\mathbb{R}^n$ and the Jacobian is not equal to zero. Then $f_{T(y)}(y)$, the density function of the transformed target variable, is given by:

$$f_{T(y)}(y) = f\left[T^{-1}(y^*)\right]|J(y)|$$

Inference

1. Transformation parameter
   - General modeling: In average unbiased under the linear mixed model with REML
   - EBP: Same findings
2. Model parameters
   - General modeling: In average β is slightly biased under the linear mixed model with REML for a non-fixed λ
   - EBP: Focus was not paid on inference. The Gurka correction does not solve the bias problem
3. Prediction
   - General modeling: A general bias is generated
   - EBP: We are interested in the whole population

Bias Corrections by Gurka

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$

Estimation made with Z and for making inferences uses the original model back-transforming the parameters

$$\mathbf{Z} = \mathbf{X}\beta^* + \mathbf{Z}\mathbf{u}^* + \epsilon^*$$

$$\hat{\boldsymbol{\beta}} = \alpha^{-1}\hat{\boldsymbol{\beta}}^*$$

$$\hat{\mathbf{u}} = \alpha^{-1}\hat{\mathbf{u}}^*$$

$$\hat{\epsilon} = \alpha^{-1}\hat{\epsilon}^*$$

The calculation of the variacnes for these estimators brakes the parsimony core principle. $\hat{\boldsymbol{\beta}}$ is not an unbiased estimator and he proposed $\hat{\boldsymbol{\beta}}_{adj}$

Bias Correction in General

For all non-linear transformations applied to the target variable

$$\text{Model: } T(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

$$\text{Interest: } E(\mathbf{Y}|\mathbf{X}) = E\Big[T^{-1}(\mathbf{X}\boldsymbol{\beta} + \epsilon)\Big]$$

We do not know this expression, therefore:

$$\begin{aligned}\text{What we do: } T^{-1}\Big[E(\mathbf{X}\boldsymbol{\beta} + \epsilon)\Big] &= T^{-1}\Big[E(\mathbf{X}\boldsymbol{\beta})\Big] \\ &= T^{-1}\Big[E(\mathbf{Y}|\mathbf{X})\Big] \\ &\neq E\Big[T^{-1}(\mathbf{X}\boldsymbol{\beta} + \epsilon)\Big]\end{aligned}$$

By using $T^{-1}\Big[E(\mathbf{Y}|\mathbf{X})\Big]$ we lost the error component when back-transforming

Bias Correction for EBP?

$$\text{Model} \;\; T(\mathbf{Y}) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}^* + \epsilon^*$$

$$\text{Interest:} \;\; E\left\{\mathrm{I}\left[T^{-1}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}^* + \epsilon^*)\right]\right\}$$

- $\mathrm{I}\left[T^{-1}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}^* + \epsilon^*)\right]$: We know this in every pseudo-population. We back-transform the whole and we do not lost the error components
- $E\left\{\mathrm{I}\left[T^{-1}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}^* + \epsilon^*)\right]\right\}$: We obtain this by averaging over the replications

Consistency

Lemma composed by the following requirements in $T(y)$. It was used as a proof of consistency of the λ estimator under the Dual power transformation for a FH model:

1. Differentiable
2. Monotone with full range
3. Integrable

MSE

- We propose two bootstraps for accounting the uncertainty coming from the estimation of the transformation parameter
- The difference is the mechanism use for generating the bootstrap populations
  1. Parametric: Generates bootstrap populations of the random and unit-level error terms parametrically
  2. Semi-parametric (Wild): The use of a data-driven transformation may reduce deviations from the normality, there may still be departure in the tails of the unit-level distribution.
     - Aims to protect against departures from model assumptions of the unit-level error term
     - Unit-level terms are generated by using the empirical distribution of suitable scaled unit-level residuals

Parametric bootstrap MSE estimation

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 $T(y_{ij}) = \tilde{y}_{ij}$, generate $u_i^* + e_{ij}^*$ and 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 ${y_{ij}}^{*(b)}$
   - Calculate the poverty measures $F_i(\alpha, t)^{(b)}$
   - Extract the bootstrap sample in $\tilde{y}_{ij}^{*(b)}$ and perform the EBP method on them. Note, as the transformed sample data is used, the estimation of λ is skiped and the original $\hat{\lambda}$ is used to transform the data back to the original scale. Obtain $\hat{F}_i^{BP}(\alpha, t)^{(b)}$
2. $MSE_*\left[F_i^{BP}(\alpha, t)\right] = 1/B \sum_{b=1}^{B} \left[\hat{F}_i^{BP}(\alpha, t)^{(b)} - F_i(\alpha, t)^{(b)}\right]^2$

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 ${y_{ij}}^{*(b)}$
   - Calculate the poverty measures $F_i(\alpha, t)^{(b)}$
   - 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 $\hat{F}_i^{BP}(\alpha, t)^{(b)}$
2. $MSE_*\left[F_i^{BP}(\alpha, t)\right] = 1/B \sum_{b=1}^{B} \left[\hat{F}_i^{BP}(\alpha, t)^{(b)} - F_i(\alpha, t)^{(b)}\right]^2$

Semi-parametric Wild Bootstrap

1. Fit model using transformed **y** to obtain estimates $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$.
2. Calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} - \hat{u}_i$.
3. Scale and center these residuals using $\hat{\sigma}_e$. The scaled and centered residuals are denoted by $\hat{\epsilon}_{ij}$.
4. For $b = 1, ..., B$

Semi-parametric Wild Bootstrap

4.1 Generate $u_i^{(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$.

4.2 Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} + u_i^{(b)}$.

4.3 Match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | k \in n\}$ from the sample by using

$$\min_{k \in n} \left| \eta_{ij}^{(b)} - \hat{\eta}_k \right|$$

and define $\tilde{k}$ as the corresponding index.

4.4 Generate weights w from a distribution where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$, respectively.

4.5 Calculate the bootstrap population as $T(y_{ij}^{(b)}) = \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} + u_i^{(b)} + w_{\tilde{k}} |\hat{\epsilon}_{\tilde{k}}^{(b)}|$.

4.6 Back-transform $T(y_{ij}^{(b)})$ to the original scale and compute the bootstrap population value $I_{i,b}$.

4.7 Select the bootstrap sample and use the EBP method as described above.

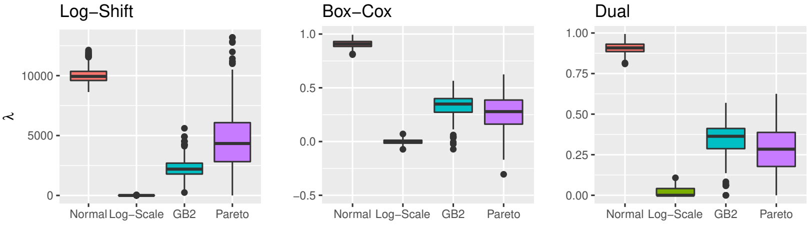
4.8 Obtain $\hat{I}_{i,b}^{EBP}$.

5 $\widehat{MSE}_{Wild} \left(\hat{I}_i^{EBP} \right) = B^{-1} \sum_{b=1}^{B} \left(\hat{I}_{i,b}^{EBP} - I_{i,b} \right)^2.$

Simulation Study: Setting

- 500 finite populations U of size $N = 10000$, partitioned into $D = 50$ areas $U_1, U_2, \ldots, U_D$ of sizes $N_i = 200$.
- With a sample size of $n = \sum_{i=1}^{D} n_i = 921$ whereby the area-specific sample sizes n_i vary between 8 and 29. To assess the data-driven transformations under extreme but realistic cases. Second, the sample sizes are similar in the case study.
- Income distributions: GB2, Pareto, Gumbel, Dagum, Dala, Fisk, Gamma, Weibull
- Extreme value distributions: Weibull, Gumbel
- Four scenarios, denoted by *Normal*, *Log-scale*, *Pareto* and *GB2*, are considered.

Simulation Study: Data-driven transformations



- Data-driven transformations adapt to the shapes of the data distributions in all the scenarios.
- Normal: λs of the Box-Cox and dual power transformations are close to one indicating that no transformation is needed.
- Log-scale: Normality may be achieved by applying the logarithmic transformation. The λs are close to zero
- GB2 & Pareto: The λs are between 0.25 and 0.5

Simulation Study: EBP under Transformations

- Poverty and inequality measures are evaluated. The EBP and MSE estimators are used using $L = 100$ and $B = 500$
- Normal: EBP without transformation is the gold
- Log-scale: EBP with a logarithmic transformation is the gold

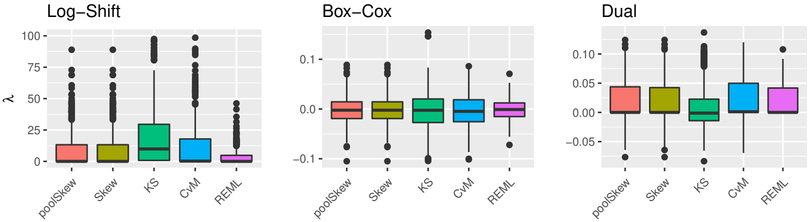
$\rightarrow$ EBP under transformation performs similar in terms of RMSE and bias

- GB2 & Pareto: EBP under transformation performs better
- Inequality measures are more sensitive to the tails of the distribution. In general the precision of SAE estimates are improved and the log-shift performs slightly better for inequality indicators

Simulation Study: MSE comparison

- Under the Box-Cox and for the four scenarios.
- To assess the performance of the two proposed MSE estimators and
- To investigate the ability of the wild bootstrap to protect against departures from the assumptions of the unit-level error term
- In general they show small RMSE and bias. However, wild bootstrap improves MSE estimation with smaller RMSE and bias
- The wild bootstrap offers protection against misspecification for indicators that depend on the tails of the distribution

Simulation Study: Estimation methods comparison



- Log-scale scenario
- The five methods provide similar estimates of λ. REML has smaller variability and appears to be the most stable method
- The impact of the estimation method on point and MSE estimation is only marginal

Empirical Bayes vs. Hierarchical Bayes

- Empirical Bayes may be viewed as an approximation to a fully Bayesian treatment of a hierarchical model wherein the parameters at the highest level of the hierarchy are set to their most likely values.
- The prior distribution is estimated from the data. This approach stands in contrast to standard Bayesian methods, for which the prior distribution is fixed before any data are observed.
- The hierarchical Bayes predictors are obtained by specifying prior distributions for model parameters and computing the posterior distribution of the parameter of interest given all the sample observations in all areas.

Poverty Indicators

Foster-Greer-Thorbecke (FGT) indicators depend on a poverty line t which is equal to 0.6 times the median of the target variable. The FGT index of type α for an area i is defined by

$$F_i(\alpha, t) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{t - y_{ij}}{t} \right)^{\alpha} \mathbb{I}(y_{ij} \leq t), \quad \text{for } \alpha = 0, 1, 2,$$

- When $\alpha = 0$, $F_i(\alpha, t)$ is the HCR and represents the proportion of the population whose income is below the poverty line t.
- Taking $\alpha = 1$, $F_i(\alpha, t)$ defines the PGAP which is a measure of poverty intensity and quantifies the degree, to which the average income of people living under the poverty line differs from the poverty line.

Inequality Indicators

Next to the two deprivation indicators we investigate inequality by a modified QSR -suitable for developing countries with high unemployment rates- defined by

$$\text{QSR}_i = \frac{\sum_{j=1}^{N_i} \mathbb{I}\left(y_{ij} \geq \mathbf{y}_{0.6}\right) y_{ij}}{\sum_{j=1}^{N_i} \mathbb{I}\left(y_{ij} \leq \mathbf{y}_{0.4}\right) y_{ij}},$$

where $\mathbf{y}_{0.6}$ and $\mathbf{y}_{0.4}$, denote the 60% and 40% quantiles of the target variable respectively.