Unit level small area models for business survey data

Chiara Bocci*, Paul A. Smith°

* *Dept. of Statistics, Computer Science, Applications "G. Parenti"*
University of Florence, Italy

° *S3RI and Department of Social Statistics & Demography*
University of Southampton, UK

AUEB - 14 May 2024

UNIVERSITÀ DEGLI STUDI FIRENZE

Outline

- Small area estimation problem
- Characteristics of business surveys
- Unit level modelling strategies for small area estimation
- Previous research on Dutch structural business survey (SBS) dataset (Smith *et al.* 2021)
  - Robust estimation approaches
- AIDA - Italian business dataset
  - Repeatability of robust estimation results
  - Extension to transformation-based approaches
- Conclusions

Small area estimation (1)

Sample survey are widely used in practice to provide estimates not only for the total population of interest but also for a variety of subpopulations (domains):

- Geographic domains (area) like regions, municipalities, school districts, health service areas, …
- Socio-demographic domains like specific age-sex-race groups, …
- Economic sectors, firm industry sectors, ...

Small area typically denotes any domain for which the specific **sample is not large enough** to support direct estimates of adequate precision.

Having only a small (possibly empty) sample in a given area, the only possible solution to the estimation problem is to borrow information from other related data sets.

UNIVERSITÀ DEGLI STUDI FIRENZE

Small area estimation (2)

Data requirements:

- **Survey data**: available for the target variable y and for the auxiliary variable *X*, related to *Y*
- **Census/Administrative data**: available for *X* but not for *Y*

SAE in 3 steps:

1. Use survey data to estimate a model that link *Y* to *X*
2. Combine the estimated model parameters with *X* for out of sample units, to form predictions
3. Use these predictions (jointly with the survey data) to estimate the target parameters

Business surveys

Characteristics

- Skewed populations
- Likely to include outliers
- Presence of good auxiliary information from registers

Small area estimation for business surveys

- Skewed variables
- Detailed stratification
- Non-negligible sampling fractions/informative sampling
- Large variation in sampling weights

MSE estimates are often very large, we want to reduce them

Unit-level modelling strategies (1)

1. Standard multilevel model + EBLUPs

2. Robust models
   - M-regression: robust estimators
   - M-quantile regression
     - Robust projective/naïve – no outliers in predicted part
     - Robust predictive – accounts for some outliers in predicted part

3. Transformations
   - EBP approach – with data-driven transformation
   - Estimators on log scale
     - Require bias-correction for back transformation

Unit-level modelling strategies (2)

4. Models with non-Normal errors
   - GB2
   - Gamma
   - Skew normal
   - Mixtures of normal distributions
   - Other potential approaches (empirical distribution quantiles, GLMMs, …)

… work in progress …

Non-ignorable sampling - compensation strategies

- Including variables predicting selection probabilities in model $\Rightarrow$ sampling $\approx$ ignorable

- Include sampling weights in models
  - Pseudo EBLUP
  - weighted naïve M-quantile
  - weighted bias-corrected M-quantile
  - Pseudo EBP
  - weighted EB (SWEE)
  - …

UNIVERSITÀ DEGLI STUDI FIRENZE

Dutch structural business survey (SBS) (Smith et al. 2021)

- Example dataset derived from Dutch SBS (survey) and tax administrative data (known)
  - retail industry
  - exclude largest businesses – but a few take-all strata remain

- Two years of tax data
  - Year 1
    - Register information for sample selection according to SBS design
    - Auxiliary information for model fitting
  - Year 2
    - Proxy for survey responses
    - (±) Whole population known – we assess repeated sampling properties of different estimators

SBS sampling design

- **Stratified design**, with strata defined by a combination of NACE1 industrial classification (**20 classes** in the retail sector) and **9 size classes**.
- The largest businesses (in size classes 6–9, with employment 50 or greater) are **completely enumerated**.
- The sample sizes in other strata determined by Neyman allocation with some additional constraints on subpopulations (including for the retailing sector)
- **We use the design of the 2009 SBS excluding the completely enumerated strata.**
- Within strata, samples are selected by **SRS without replacement**.

Pseudo-SBS dataset

- We obtain:
  - **Population** size N = 63,981, 5 size classes and 20 industries
  - Total **sample** size n = 5,074, with sample sizes per industry varying from 21 to 769.
- Response variable: tax-turnover for 2007, ***tto***
- Small areas: 20 **industry classes**
- Auxiliary variables:
  - tax-turnover from 2006, ***tax1***
  - industrial classification, ***ind***
  - size class, based on the working persons in the business in bands 1, 2–4, 5–9, 10–19, 20–49, ***sc***
  - employment, measured as the number of working persons, ***wp***
- Model for simulations specified as (following Krieg et al., 2012):

$$tto_{i,ind} = \beta_0 + \beta_1 tax1_{i,ind} + \boldsymbol{\beta}_2 \mathbf{sc}_{i,ind} + \beta_3 wp_{i,ind} + \beta_4 \left(tax1 \times wp\right)_{i,ind} + u_{ind} + e_{i,ind}$$

Direct estimators

- Horvitz-Thompson (HT)

$$\hat{y}_{ind}^{HT} = \sum_{i \in ind \cap s} d_i y_i \qquad d_i = 1/\pi_i$$

- Generalized regression estimator (GREG)

$$\hat{y}_{ind}^{GREG} = \hat{y}_{ind}^{HT} + \hat{\boldsymbol{\beta}}^{GREG} \left[\mathbf{X}_{ind} - \hat{\mathbf{x}}_{ind}^{HT} \right]$$

Using Dutch SBS approach for auxiliary variables:

- *tax1* × *ind* × *size class* (0-9, 10-49)

Non-robust small area estimators (1)

Based on a two-level random effect model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$

with industry-specific random effects $\mathbf{u} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right)$

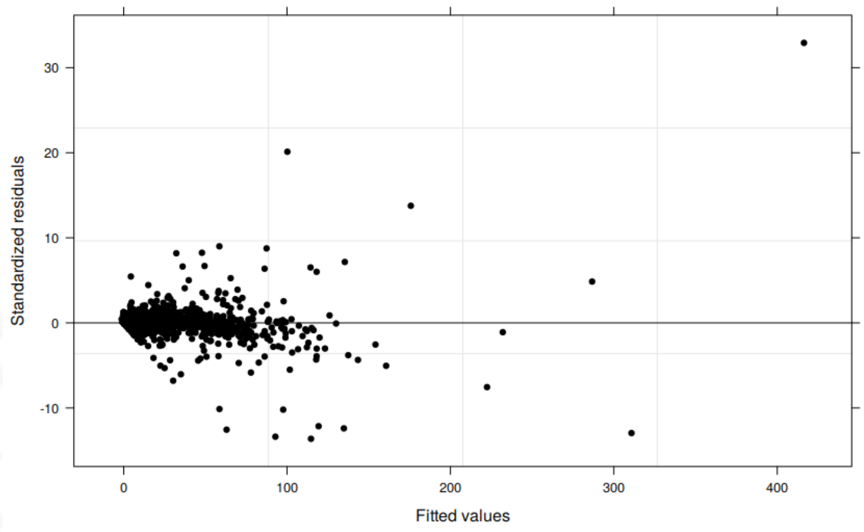
and individual-specific random effects $\mathbf{e} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_e\right)$

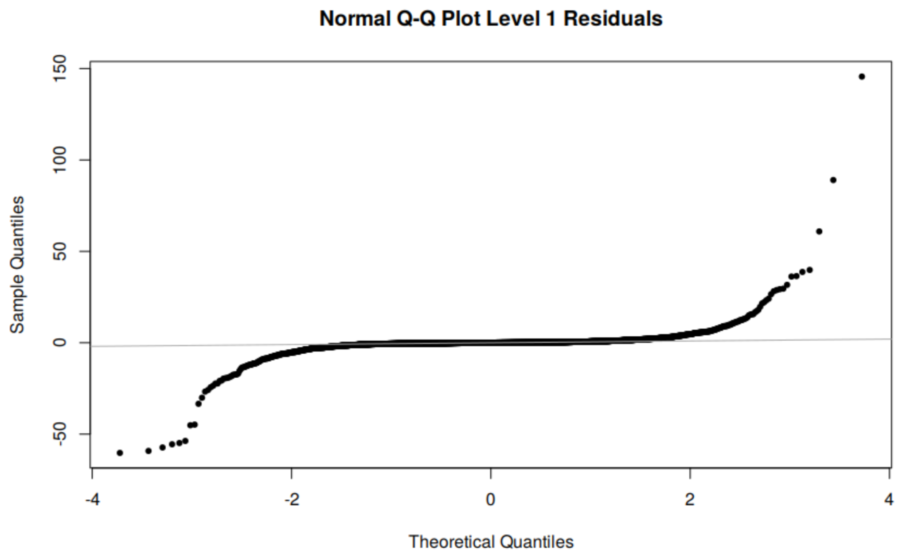
- EBLUP $\hat{y}_{ind}^{EBLUP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \mathbf{z}_i^T \hat{\mathbf{u}}\right)$

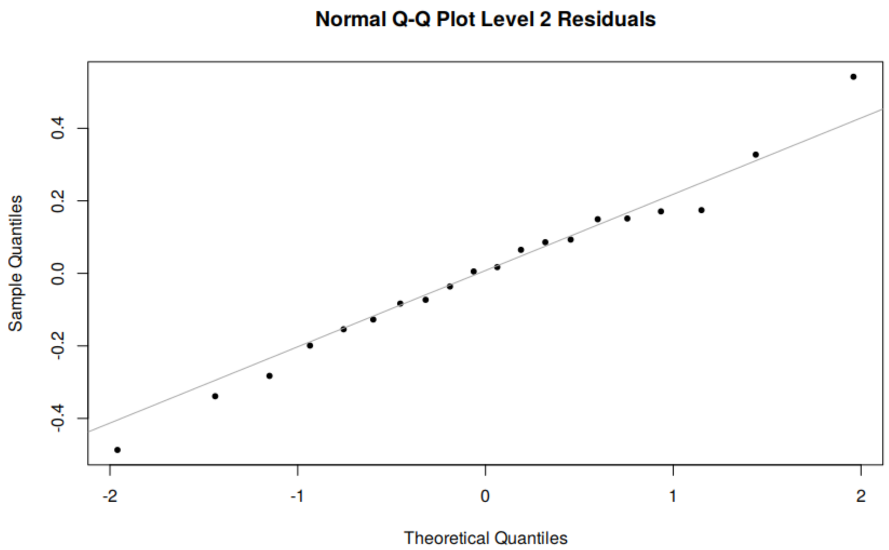
UNIVERSITÀ DEGLI STUDI FIRENZE

Diagnostic plots for multilevel linear model

(one sample – SBS Data)







UNIVERSITÀ DEGLI STUDI FIRENZE

Non-robust small area estimators (2)

- EBLUP with individual-level variance as function of the size class

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad \mathbf{u} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right)$$

$$\mathbf{e}_{sc} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{e,sc}\right)$$

- pseudo-EBLUP (You & Rao 2002, Fabrizi et al., 2014)

$$\hat{y}_{ind}^{pEBLUP} = \hat{\gamma}_{ind,w} \sum_{i \in s_{ind}} w_{ind,i} y_{ind,i} + \left(\mathbf{X}_{ind} - \hat{\gamma}_{ind,w} \sum_{i \in s_{ind}} w_{ind,i} \mathbf{x}_{ind,i} \right) \hat{\boldsymbol{\beta}}_w$$

$$\hat{\gamma}_{ind,w} = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2 \delta_i\right) \qquad \delta_i = \sum_{i \in s_{ind}} \left(\frac{w_{ind,i}}{\sum_{i \in s_{ind}} w_{ind,i}} \right)^2$$

USE SAMPLING WEIGHTS

Regression coefficients are fitted using the **survey weights** as well

UNIVERSITÀ DEGLI STUDI FIRENZE

Robust estimators

- Robust EBLUP (Sinha & Rao, 2009)
  - M-regression with random effect
  - Huber function $\psi(a) = a \min\left(1, b_\psi / |a|\right)$ applied to the residuals, with tuning constant b_ψ = 1.345, to reduce the outlier effect in the estimation

$$\hat{y}_{ind}^{REBLUP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\psi,SR} + \mathbf{z}_i^T \hat{\mathbf{u}}^{\psi,SR} \right)$$

Note: Robust industry level effect ≈ 0 with the SBS dataset

- Robust synthetic estimator
  M-regression estimation,
  only fixed effects

$$\hat{y}_h^{RSYN} = \sum_{i \in h \cap s} y_i + \sum_{i \in h \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\psi,M}$$

UNIVERSITÀ DEGLI STUDI FIRENZE

Robust projective estimators (1)

- Based on a linear model for the **M-quantile regression** $m_q(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_q^{\psi}$
  - Naïve M-quantile estimator (Chambers & Tzavidis, 2006)
    - Define q_i s.t. $y_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{q_i}^{\psi}$
    - In each domain find mean $\bar{q}_{ind} = \frac{1}{n_{ind}} \sum_{i \in ind} q_i$

$$\hat{y}_{ind}^{MQ} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi}$$

    - Inconsistent. Assumption that all the outliers are observed in the sample, so called **naïve** (Tzavidis et al., 2010)
  - Weighted naïve M-quantile estimator (Fabrizi et al., 2014)

$$\hat{y}_{ind}^{wMQ} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{w,\,\bar{q}_{ind}}^{\psi}$$

Regression coefficients are fitted using the **sampling weights**

Robust projective estimators (2)

- Consistent, bias-corrected M-quantile estimator (Chambers & Dunstan, 1986)

$$\hat{y}_{ind}^{MQCD} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} + \frac{N_{ind} - n_{ind}}{n_{ind}} \sum_{i \in ind \cap s} \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} \right)$$

  - Adds a third term to correct for the potential bias at the cost of allowing the variance to increase.

- Weighted bias-corrected M-quantile estimator (Fabrizi et al., 2014)

$$\hat{y}_{ind}^{wMQCD} = \sum_{i \in ind \cap s} w_i y_i + \left(\sum_{i \in ind \cap U} \mathbf{x}_i^T - \sum_{i \in ind \cap s} w_i \mathbf{x}_i^T \right) \hat{\boldsymbol{\beta}}_{w,\, \bar{q}_{ind}}^{\psi}$$

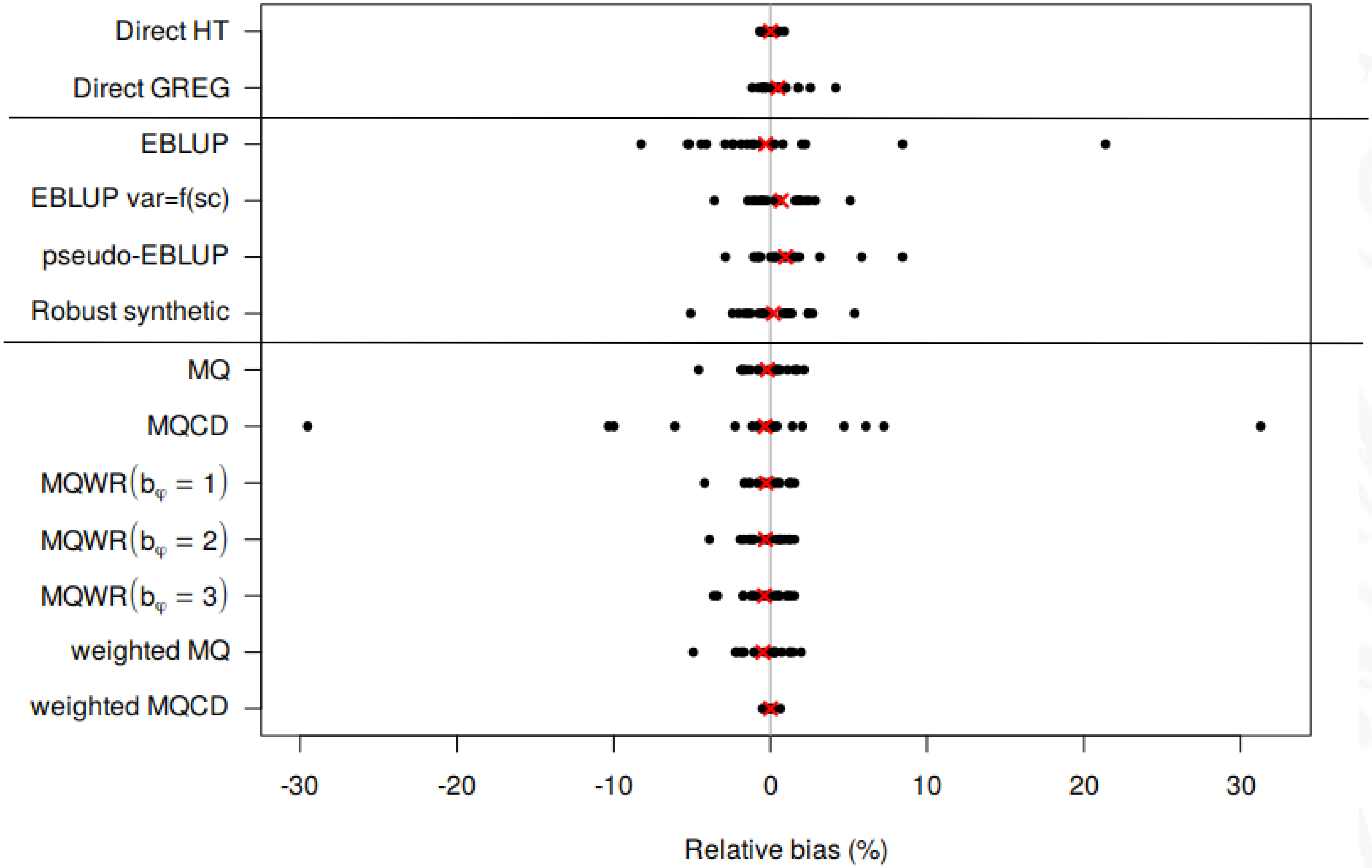
Robust predictive estimators

- Robust predictive bias-adjusted M-quantile estimator (Welsh & Ronchetti, 1998)

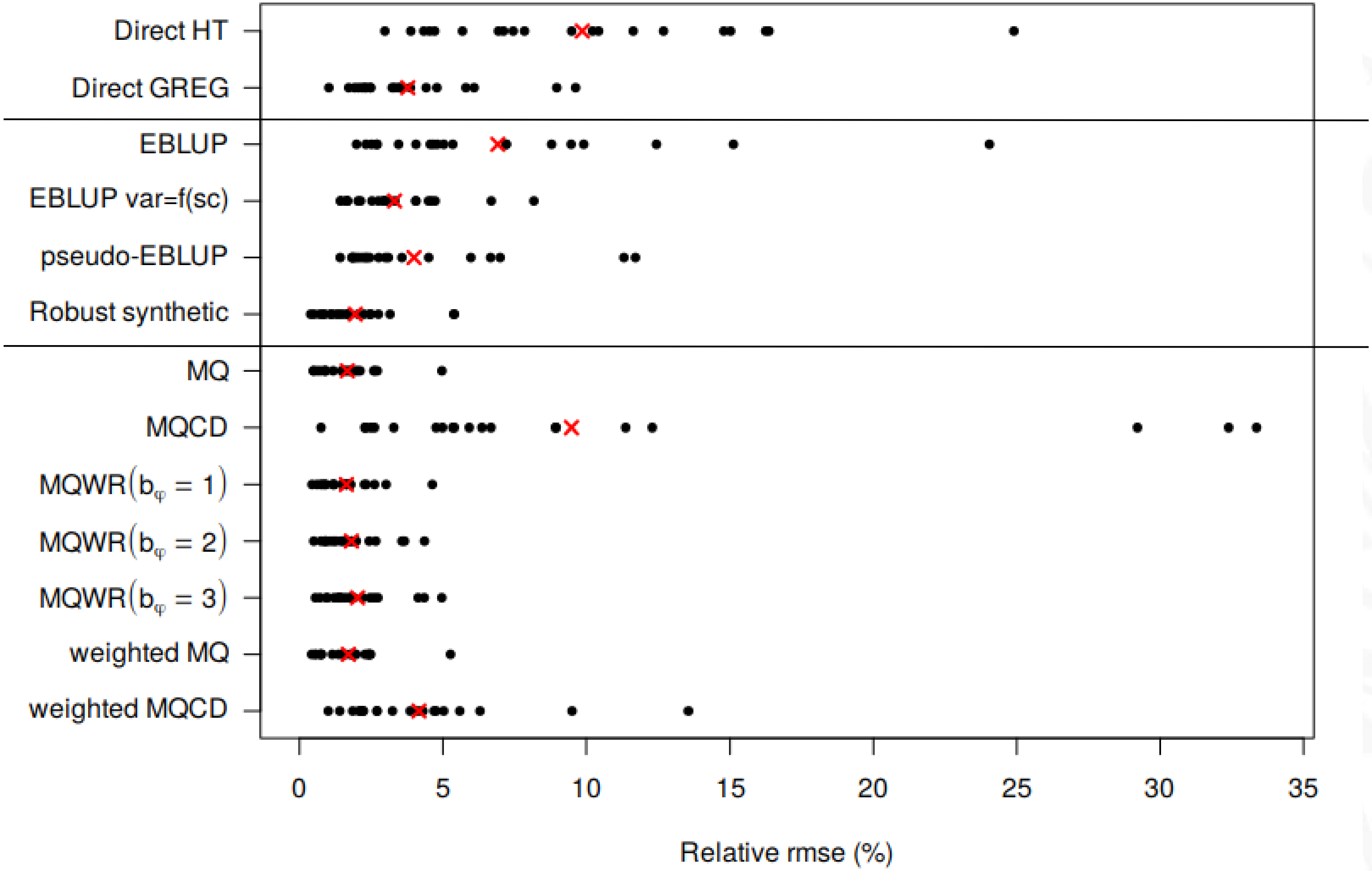
$$\hat{y}_{ind}^{MQWR} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} + \frac{N_{ind} - n_{ind}}{n_{ind}} \sum_{i \in ind \cap s} \omega_{ind}^{MQ} \phi \left\{ \frac{\left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\bar{q}_{ind}}^{\psi} \right)}{\omega_{ind}^{MQ}} \right\}$$

- Second Huber function, tuning parameter b_{φ}, robust estimate of scale of the residuals ω
  - b_{φ} codes bias-variance trade-off
  - expect $b_{\varphi} > b_{\psi}$

Relative Bias (%) of industry estimates - SBS



Relative RMSE (%) of industry estimates - SBS



Further explorations

- Dutch SBS data was available for specific research project only
- Need alternative source of population data for evaluation of methods

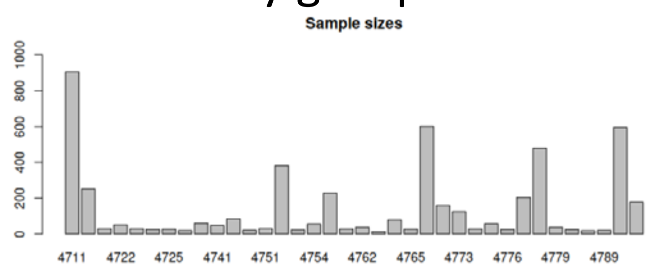
Objectives:

- Assess robust estimation approaches from Smith *et al*. (2021) on a second dataset
- Assess different transformation-based approaches

Faced with a new business dataset, what is the best approach to small area estimation?

AIDA dataset (1)

- Multi-year database of Italian businesses (from Bureau van Dijk)
  - Information on all Italian companies required to file their accounts (joint-stock companies, excluding banks, insurance companies and public bodies)
- We extract data similar to SBS
  - retail* businesses in Italy: 36 codes
  - auxiliary information from 2018 and target variable from 2020
- Stratified design, Neyman allocation
- We obtain:
  - **Population** size N = 71,568, 5 size classes and 36 industry groups
  - Total **sample** size n = 5,000, with sample sizes per industry varying from 12 to 905.



**retail excluding petrol stations*

AIDA dataset (2)

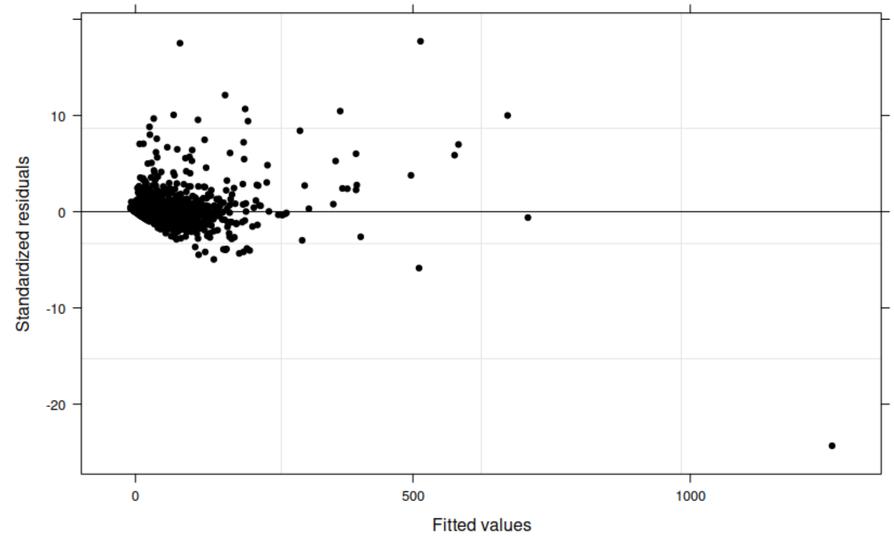
- Response variable: turnover from 2020, t^{2020}
- Small areas: **36 industry groups**
- Auxiliary variables:
  - turnover from 2018, t^{2018}
  - industrial classification, ***ind***
  - size class, based on the working persons in the business in bands 1, 2–4, 5–9, 10–19, 20–49, ***sc***
  - employment, measured as the number of working persons, ***wp***

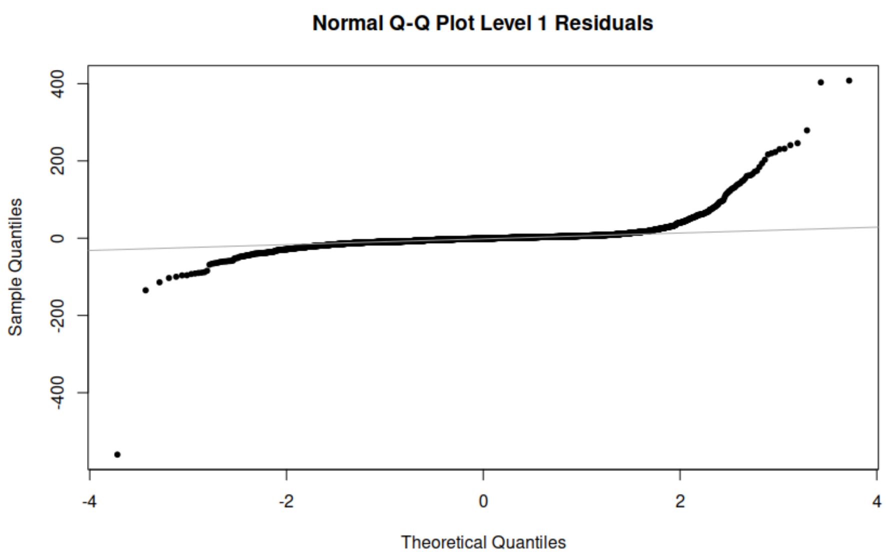
- Model for simulations
  - Reproduced model from Smith et al. (2021) (but only size, not size class)

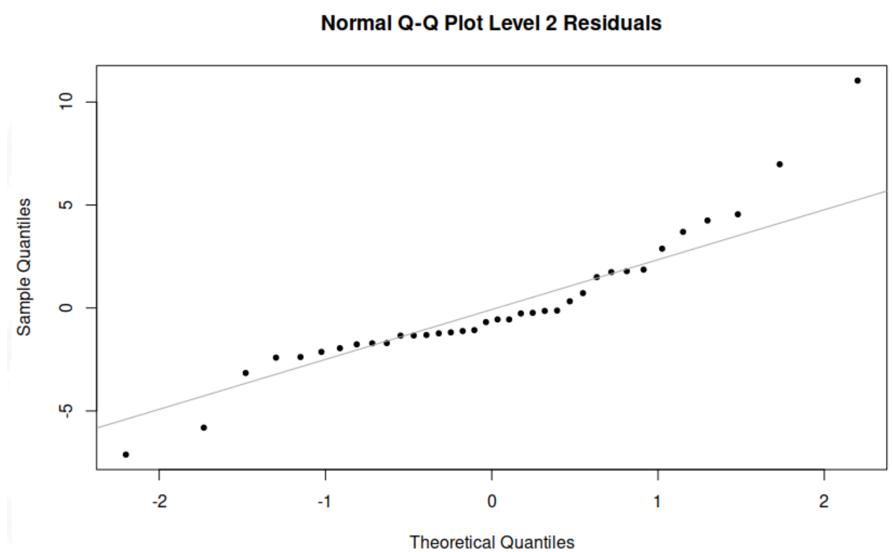
$$\mathbf{X\beta} : \beta_0 + \beta_1 t_{i,ind}^{2018} + \beta_3 wp_{i,ind} + \beta_4 \left(t_{i,ind}^{2018} \times wp \right)_{i,ind}$$

Diagnostic plots for multilevel linear model

(one sample – AIDA Data)

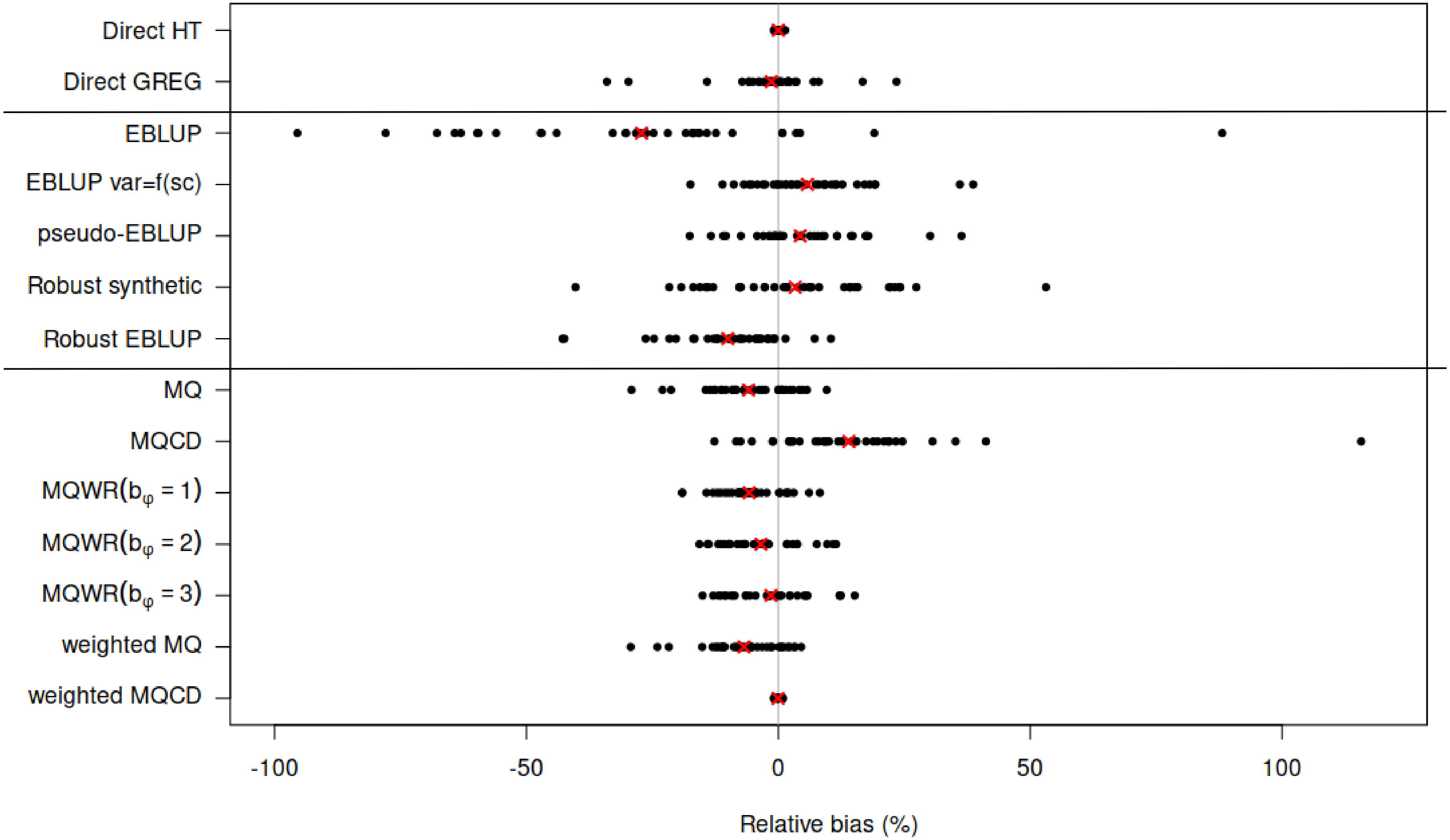




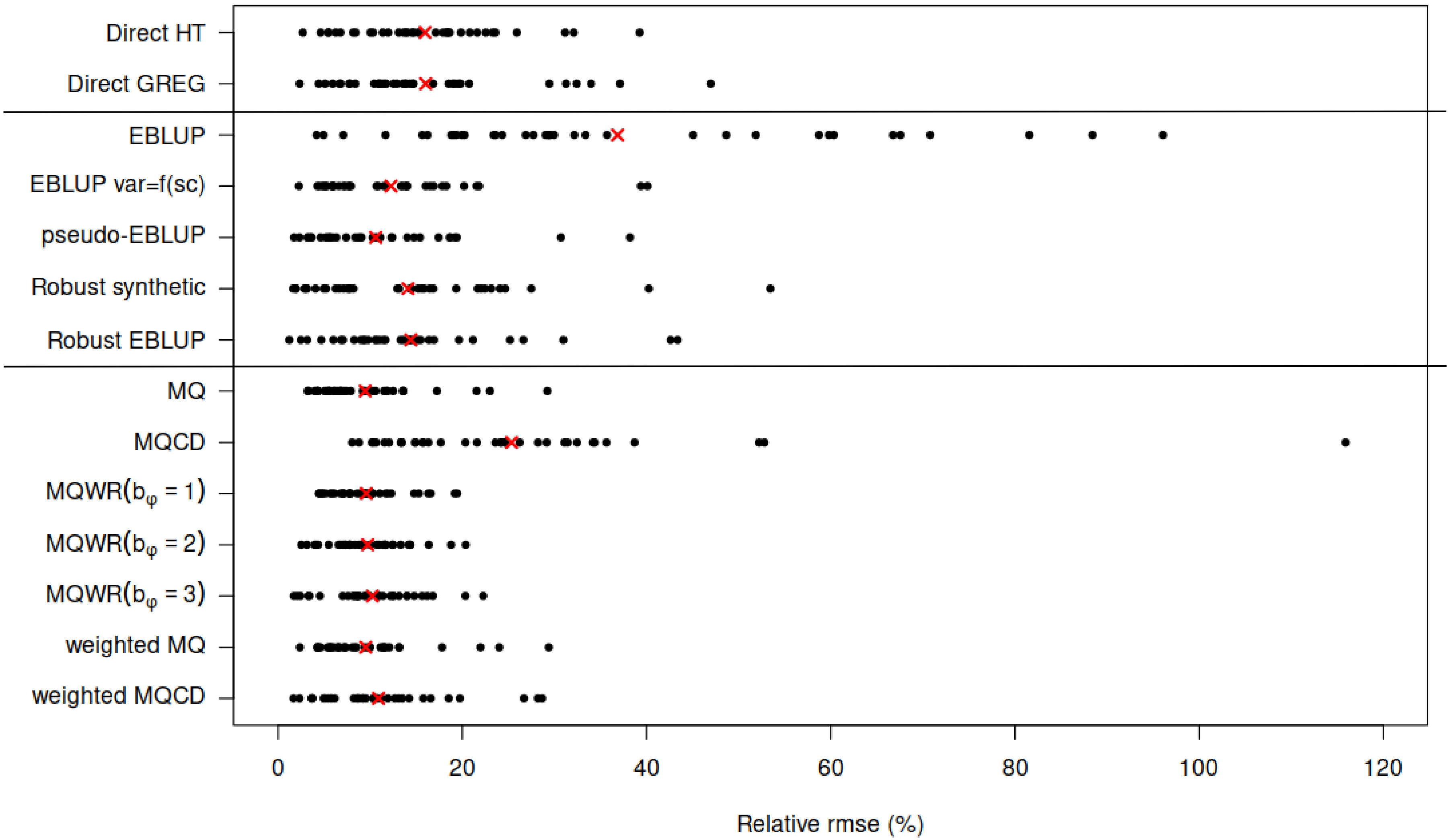


UNIVERSITÀ DEGLI STUDI FIRENZE

Relative Bias (%) of industry estimates - AIDA



Relative RMSE (%) of industry estimates - AIDA



Transformation estimators (1)

- Empirical Best Prediction (EBP) approach by Molina & Rao (2010) (Rojas-Perilla et al. 2019)
  - Generate transformed data y*
  - Use y* to fit multilevel model and obtain estimates of parameters
  - Generate pseudo-populations from original data by repeated sampling of residuals from model
  - Back transform pseudo-population values and calculate indicator in each pseudo-population
  - Average pseudo-population indicators

Predictions of the target outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data.

- There is no back-transformation bias

Transformations for EBP

- Linear
- Log transformation:

  $\log(y_{ij} + s)$

- Log-shift transformation (Yang 1995)

  $\log(y_{ij} + \lambda)$

  - λ fitted

- With s deterministic, default is $s = 1$

- Box-Cox transformation (Box & Cox 1964)

$$\begin{cases} \dfrac{\left(y_{ij}+s\right)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log\left(y_{ij}+s\right) & \text{if } \lambda = 0 \end{cases}$$

- Dual power transformation (Yang 2006)

$$\begin{cases} \dfrac{\left(y_{ij}+s\right)^{\lambda}-\left(y_{ij}+s\right)^{-\lambda}}{2\lambda} & \text{if } \lambda \neq 0 \\ \log\left(y_{ij}+s\right) & \text{if } \lambda = 0 \end{cases}$$

EBP steps in more details

1. select a transformation, fitting the shift parameter to obtain $\hat{\lambda}$ if necessary, and obtain $y_i^* = T_{\hat{\lambda}}(y_i)$
2. use the transformed data in the unit level model (5) to estimate $\hat{\boldsymbol{\beta}}^*$, and the variance components $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$; calculate $\hat{\gamma}_{ind} = \hat{\sigma}_u^2 / \left(\hat{\sigma}_u^2 + n_{ind}^{-1}\hat{\sigma}_e^2\right)$
3. for l in $1, \ldots, L$
   - take random draws v_{ind} from $N\left(0, (1-\hat{\gamma}_{ind})\hat{\sigma}_u^2\right)$ for each value of ind included in the sample, or from $N\left(0, \hat{\sigma}_u^2\right)$ for any out-of-sample values of ind
   - take random draws of e_i from $N\left(0, \hat{\sigma}_e^2\right)$ for each value of i
   - obtain pseudopopulation l as $y_i^{*(l)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^* + \hat{u}_{ind}^* + v_{ind} + e_i^*$ when ind is included in the sample, choosing ind such that $i \in ind$, and as $y_i^{*(l)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^* + v_{ind} + e_i^*$ when i belongs to an out-of-sample value of ind
   - back-transform the pseudopopulation values to obtain $y_i^{(l)}$
   - calculate the estimate of interest for each ind with pseudopopulation l, $\hat{y}_{ind}^{(l)}$
4. take the average of the statistic of interest for each ind over the l replicates, $\hat{y}_{ind}^{EBP} = \frac{1}{L}\sum \hat{y}_{ind}^{(l)}$.

Transformation estimators (2)

- Kalberg-type synthetic predictor (Chandra & Chambers, 2011)

$$\hat{y}_{ind}^{SYN-EP} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\hat{c}_i^{SYN-EP}\right)^{-1} exp\left(x_i^{\mathrm{T}}\hat{\beta} + \frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{2}\right)$$

where $\hat{c}_i^{SYN-EP} = exp\left[\frac{1}{2}x_i^{\mathrm{T}}\hat{V}(\hat{\beta})x_i + \frac{1}{4}\hat{V}(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)\right]$

- Model based direct estimator (MBDE) (Chandra & Chambers, 2011)

$$\hat{y}_{ind}^{MBD} = \sum_{i \in ind \cap s} w_i^{MBD} y_i$$

The weights are function of $\hat{y}_{ind}^{SYN-EP}$

Transformation estimators (3)

- Empirical best (EB) predictor (Berg & Chandra, 2014)

$$\hat{y}_{ind}^{EB} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \exp\left\{ x_i^{\mathrm{T}} \hat{\beta} + \hat{\gamma}_{ind} \left(\frac{1}{n_{ind}} \sum_{i \in ind \cap s} \log(y) - x_s^{\mathrm{T}} \hat{\beta} \right) + \frac{1}{2} \hat{\sigma}_e^2 \left(\frac{\hat{\gamma}_{ind}}{n_{ind}} + 1 \right) \right\}$$

- EB predictor – bias corrected (Berg & Chandra, 2014)

$$\hat{y}_{ind}^{EBbc} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \left(\hat{c}_i^{EB} \right)^{-1} \hat{y}_{ind}^{EB}$$

where $\hat{c}_i^{EB} = \exp\left[\frac{1}{2} \left(\hat{\mathbf{a}}_i + \hat{c}_{1,ind} \hat{V}(\hat{\sigma}_e^2) + \hat{c}_{2,ind} \hat{V}(\hat{\sigma}_u^2) + 2\hat{c}_{3,ind} \hat{Cov}(\hat{\sigma}_e^2, \hat{\sigma}_u^2) \right) \right]$

see Berg & Chandra (2014) for the formulas for $\hat{\mathbf{a}}_i$, $\hat{c}_{1,ind}$, $\hat{c}_{2,ind}$ and $\hat{c}_{3,ind}$

Weighted predictors under transformation

- Pseudo-EBP (Guadarrama et al., 2018)

  Basic procedure as before, but

  - estimates are conditioned on the weighted means;
  - parameters are derived from a weighted unit level model (fitted using maximum likelihood (Pfeffermann & Sverchkov, 2007), or using the method of moments of You & Rao (2002).

- Weighted EB predictor - SWEE (Zimmermann & Münnich, 2018)

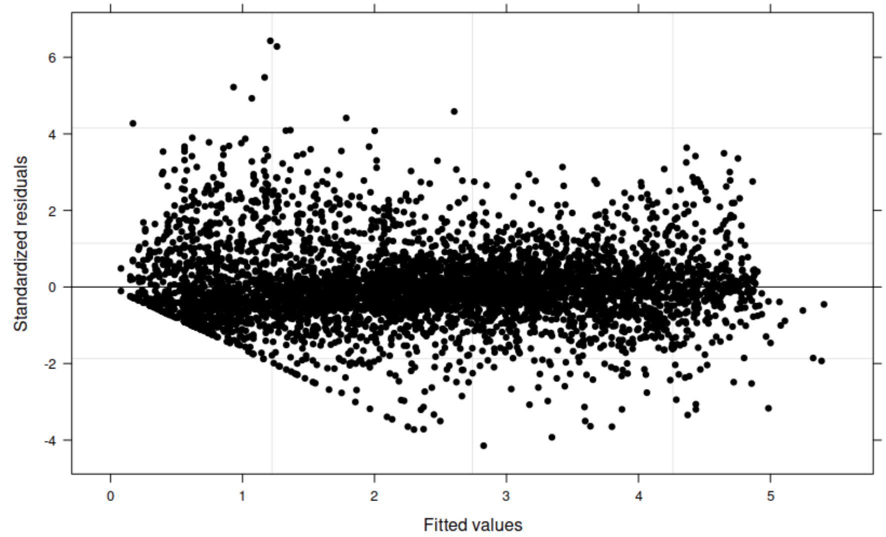
  Parameters are derived from a weighted unit level model, adapting You & Rao (2002) (as in Pseudo-EBLUP) to log-transformed data.

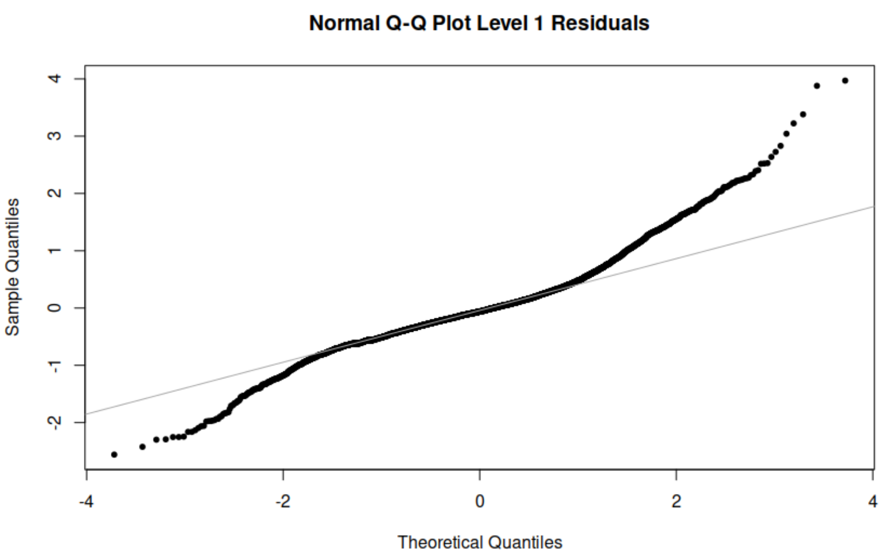
$$\hat{y}_{ind}^{SWEE} = \sum_{i \in ind \cap s} y_i + \sum_{i \in ind \cap r} \exp\left[\mathbf{x}_i^{\mathrm{T}} \hat{\beta}^{*SWEE} + \tilde{u}_{ind}^{*} + \frac{1}{2}\hat{\sigma}_e^{*2}\left(\hat{\gamma}_{ind}^{w}\delta_{ind}^{2} + 1\right)\right]$$

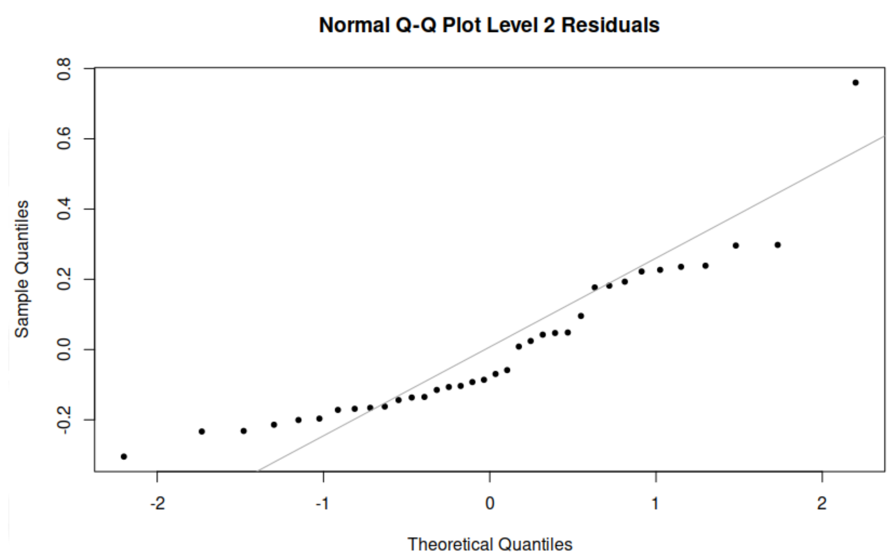
survey weights are used to obtain these

Diagnostic plots for multilevel loglinear model - with log(x)

(one sample – AIDA Data)







UNIVERSITÀ DEGLI STUDI FIRENZE

EBP relative RMSEs (%): model with x

Industry	linear	log (deterministic shift)	log shift	Box-Cox	dual power
…					
4751	35.76	20.69	19.31	18.86	19.26
4752	5.06	176167.13	52060.44	54.92	23.67
4753	45.20	35.65	37.28	38.42	37.86
4754	20.38	14.70	13.89	14.10	14.51
…					
median (rrmse)	29.50	23.02	19.27	14.36	14.28
mean (rrmse)	37.16	10033.93	3291.47	22.15	17.94

Regression model $\mathbf{X\beta} : \beta_0 + \beta_1 t_{i,ind}^{2018} + \beta_3 wp_{i,ind} + \beta_4 \left(t_{i,ind}^{2018} \times wp \right)_{i,ind}$

UNIVERSITÀ DEGLI STUDI FIRENZE

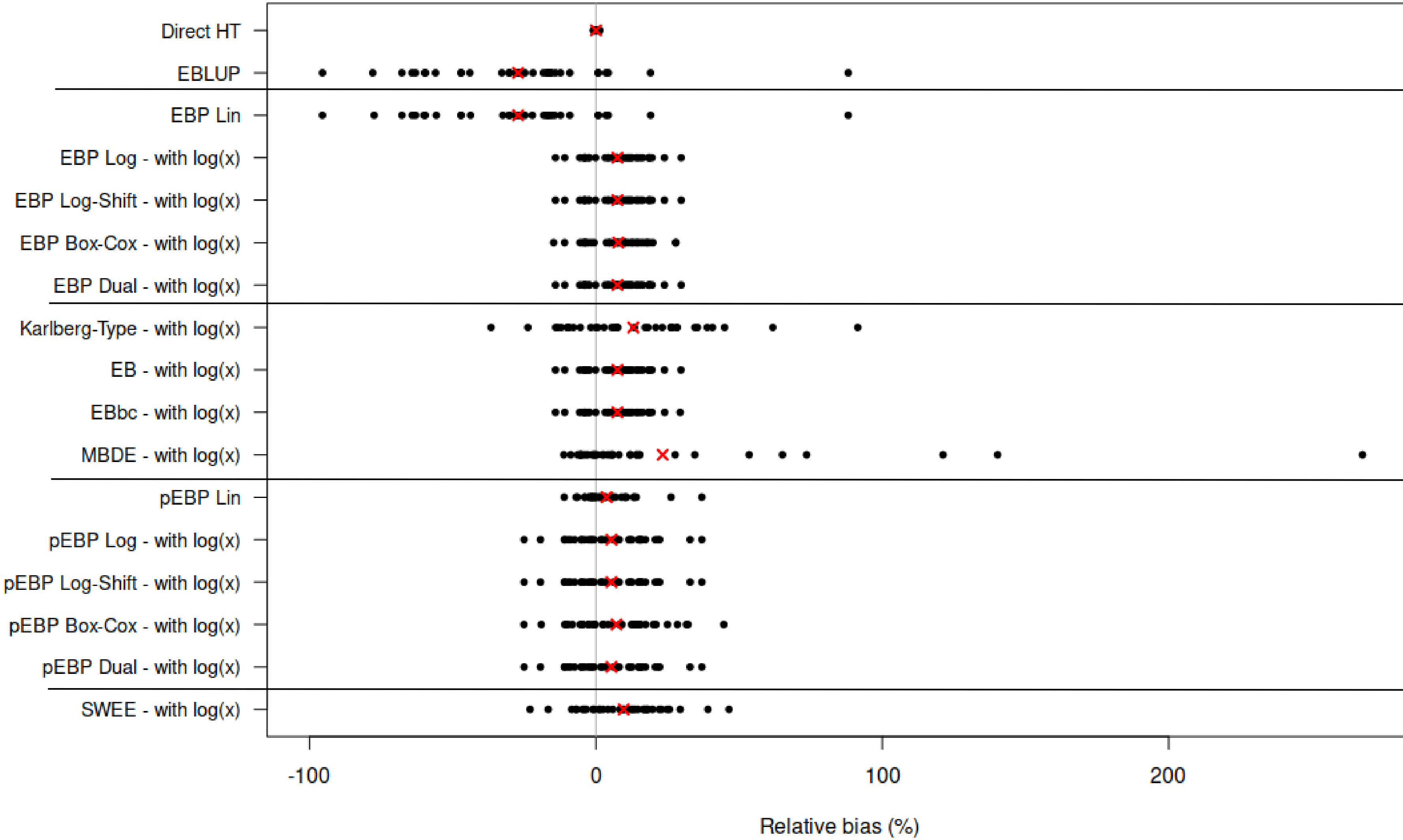
EBP relative RMSEs (%): model with log(x)

Industry	linear	log (deterministic shift)	log shift	Box-Cox	dual power
…					
4751	39.54	17.25	17.25	17.44	17.25
4752	8.85	9.66	9.66	10.45	9.66
4753	37.32	20.69	20.69	19.95	20.69
4754	34.27	10.28	10.28	10.49	10.28
…					
median (rrmse)	32.85	12.41	12.41	12.54	12.41
mean (rrmse)	39.11	13.17	13.17	13.47	13.17

Replace x by log(x) $\mathbf{X\beta}: \beta_0 + \beta_1 \log\left(t_{i,ind}^{2018}\right) + \beta_3 wp_{i,ind} + \beta_4 \left(\log\left(t_{i,ind}^{2018}\right) \times wp\right)_{i,ind}$

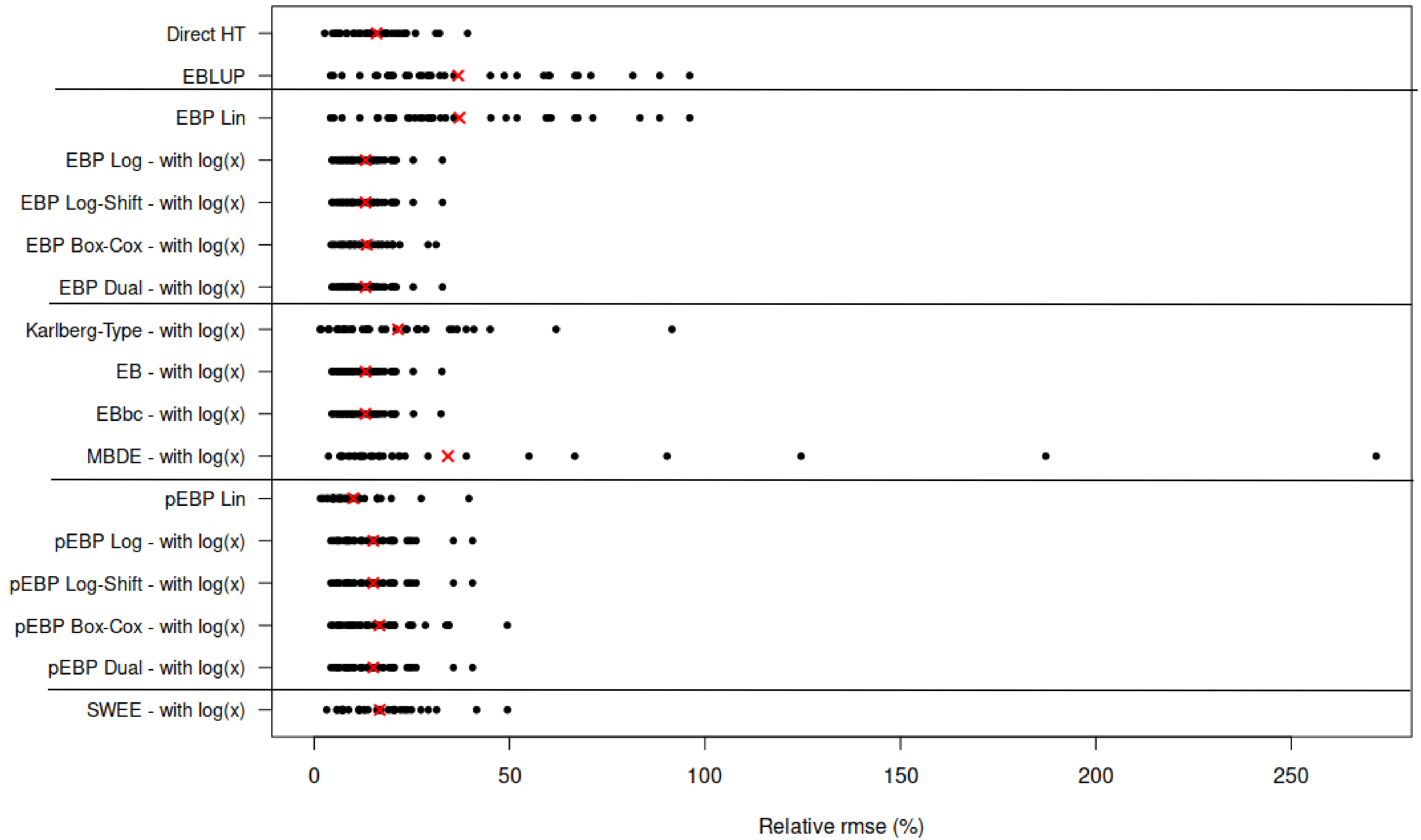
UNIVERSITÀ DEGLI STUDI FIRENZE

Relative Bias (%) of industry estimates - AIDA



UNIVERSITÀ DEGLI STUDI FIRENZE

Relative RMSE (%) of industry estimates - AIDA



UNIVERSITÀ DEGLI STUDI FIRENZE

Transformations vs Robust models

	log shift*	dual power*	EB predictor bias corrected*	M-quantile naïve	weighted M-quantile naïve	bias-adjusted M-quantile ($b\phi = 1$)
median (rrmse)	12.41	12.41	12.41	7.31	7.70	9.14
mean (rrmse)	13.17	13.17	13.13	9.46	9.53	9.59

* with log(x) predictor

UNIVERSITÀ DEGLI STUDI FIRENZE

Conclusions and future work

- Results of simulations on **robust estimators** with Dutch data (Smith et al. 2021) are corroborated with Italian AIDA data
  - **Robust estimators** give substantial rrmse improvement for small areas
  - **Robust bias-adjusted M-quantile** best overall (we need to investigate on methods to set tuning constant in real situation)
- **Transformation** approaches are effective at improving estimates, but less good than (best) robust models.
- Tackle third group of estimators based on **distributions with non-normal errors**

✓ Come to the International Conference on Establishment Statistics (ICES VII) in Glasgow June 2024 for the next thrilling instalment

References

- Berg, E. & Chandra, H. (2014) Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis* **78** 159–175
- Box, G.E.P. & Cox, D.R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B* **26** 211–252
- Chandra, H. & Chambers, R. (2011) Small area estimation under transformation to linearity. *Survey Methodology* **37** 39–51
- Guadarrama, M., Molina, I., & Rao, J.N.K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis* **121** 20–40
- Pfeffermann, D. & Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association* **102**1427–1439
- Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020) Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society, Series A* **183** 121–148
- Smith, P.A., Bocci, C., Tzavidis, N., Krieg, S. & Smeets, M.J.E. (2021) Robust estimation for small domains in business surveys. *Journal of the Royal Statistical Society, Series C* **70** 312–334
- Yang, L. (1995) *Transformation-density estimation*. PhD Thesis. University of North Carolina, Chapel Hill
- Yang, Z. (2006) A modified family of power transformations. *Economics Letters* **92** 14–19
- You, Y.. & Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics* **30** 431–439
- Zimmermann, T. & Münnich, R.T. (2018). Small area estimation with a lognormal mixed model under informative sampling. *Journal of Official Statistics* **34** 523–542

Your questions…

chiara.bocci@unifi.it