CITY DATA FROM LFS AND BIG DATA

BigSurv18: Big Data meets Survey Science

27th October 2018, Barcelona

Sandra Hadam
Federal Statistical Office, Germany

Link between Mobile Phone Data with LFS Indicators

Ref. Ares(2017)4591837 - 20/09/2017

Eurostat and DG Regional and Urban Policy grants for 2017

Call for proposals

City data from LFS and big data (ESSNET)

CCI 2016.CE.16.BAT.107

1. INTRODUCTION

In accordance with Article 128 of the Financial Regulation we are pleased to invite y
submit your application for the award of a grant in the framework of **Cross border city statistics**.

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A
J. R. Statist. Soc. A (2017)

Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal

Timo Schmid and Fabian Bruckschen,
Freie Universität Berlin, Germany
Nicola Salvati
Università of Pisa, Italy
and Till Zbiranski
Freie Universität Berlin, Germany

[Received April 2016. Final revision June 2017]

Summary. Modern systems of official statistics require the accurate and timely estimation of sociodemographic indicators for disaggregated geographical regions. Traditional data collection methods such as censuses or household surveys impose great financial and organizational burdens on national statistical institutes. The rise of new information and communication technologies offers promising sources to mitigate these shortcomings. We propose a unified approach for national statistical institutes in developing countries based on small area estimation that allows for the estimation of sociodemographic indicators by using mobile phone data. In particular, the methodology is applied to mobile phone data from Senegal for deriving subnational estimates of the share of illiterates disaggregated by gender. The estimates are used to identify hotspots of illiterates with a need for additional infrastructure or policy adjustments. Although we focus on literacy as a particular sociodemographic indicator, the approach proposed is applicable to indicators from national statistics in general.

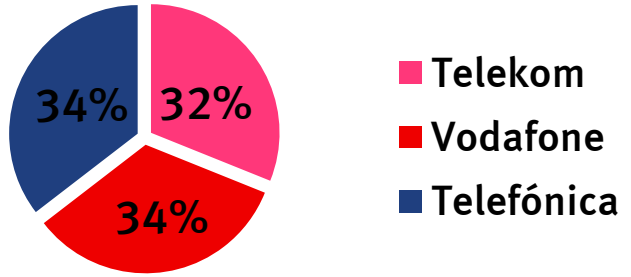
Keywords: Indicators; Model-based estimation; Official statistics; Small area estimation

Literature: Schmid, T, F. Bruckschen, N. Salvati and T. Zbiranski (2017). Constructing socio demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A (Statistics in Social Sciences) 180, 1163–1190.

Mobile Phone Data

- Mobile Providers in Germany
  - Market share of 1/3 (state 4th quarter 2017)
  - Cooperation Agreement with T-Systems
- Mobile Phone Data
  - Specific geometry
  - Signaling data: anonymized and aggregated
- Minimum number of counts per area ≥ 30

Market shares of German mobile providers (4th quarter 2017).

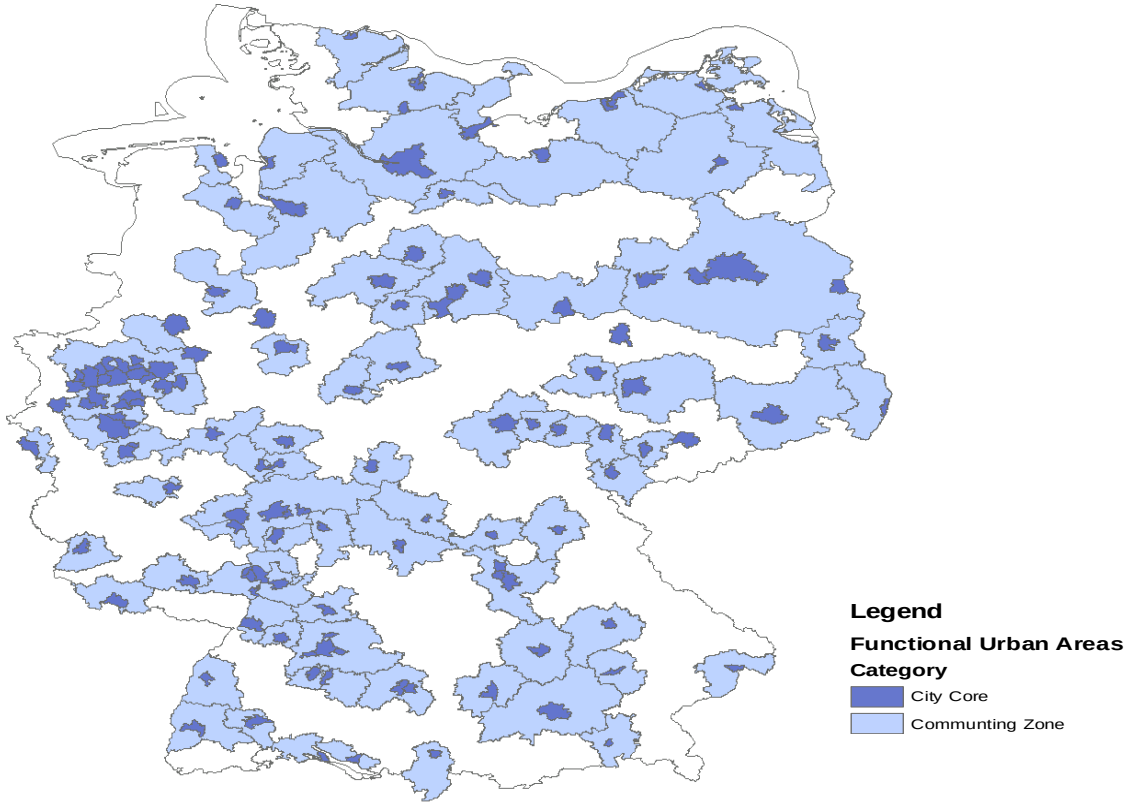


See Bundesnetzagentur:
https://www.bundesnetzagentur.de/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Marktbeobachtung/Deutschland/Mobilfunkteilnehmer/Mobilfunkteilnehmer_node.html

Mobile Phone Data for Small Area Estimation

- Study area: Germany
- Geometry: Communities
- Temporal resolution:
  - Statistical Sunday from 2018
  - Average value of mobile activities from 8 to 11 p.m.
  - Dwell time: 2 hours
- Characteristics:
  - Age groups (20-29, 30-39, 40-49, 50-59, 60-69, 70+ years)
  - Gender (male/female)
  - Crossing of characteristics
  - Mobile Country Code
  - Minimum number of counts per area ≥ 30

Functional Urban Areas (FUA‘s)



- 208 units (FUA‘s)
- 125 City Cores
- 83 Commuting Zones
- Composition based on NUTS 3 Areas and communities

Labour Force Survey (LFS)

- One-Percent-sample population
- Year: 2016
- Sample Size: 725.829 observations/individuals
- Sample Size on FUA Level: 533.356
- Published on NUTS 2 Level
- Individual data on NUTS 3 Level and communities
- Investigation of Employment- and Unemployment rate - ILO

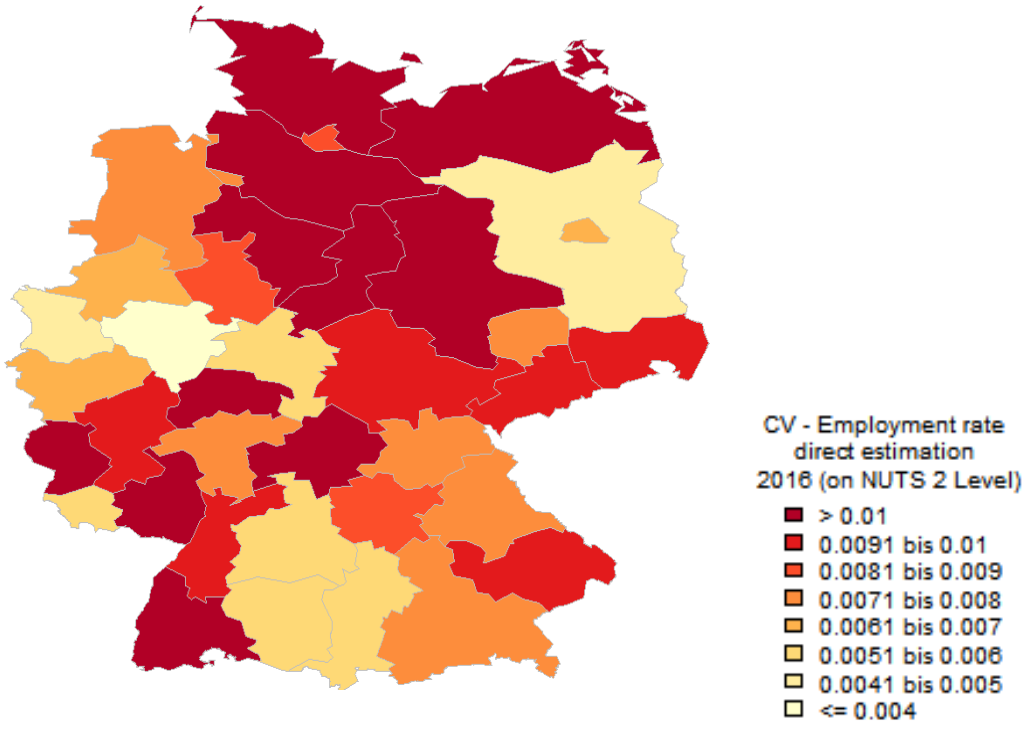
Direct Estimation

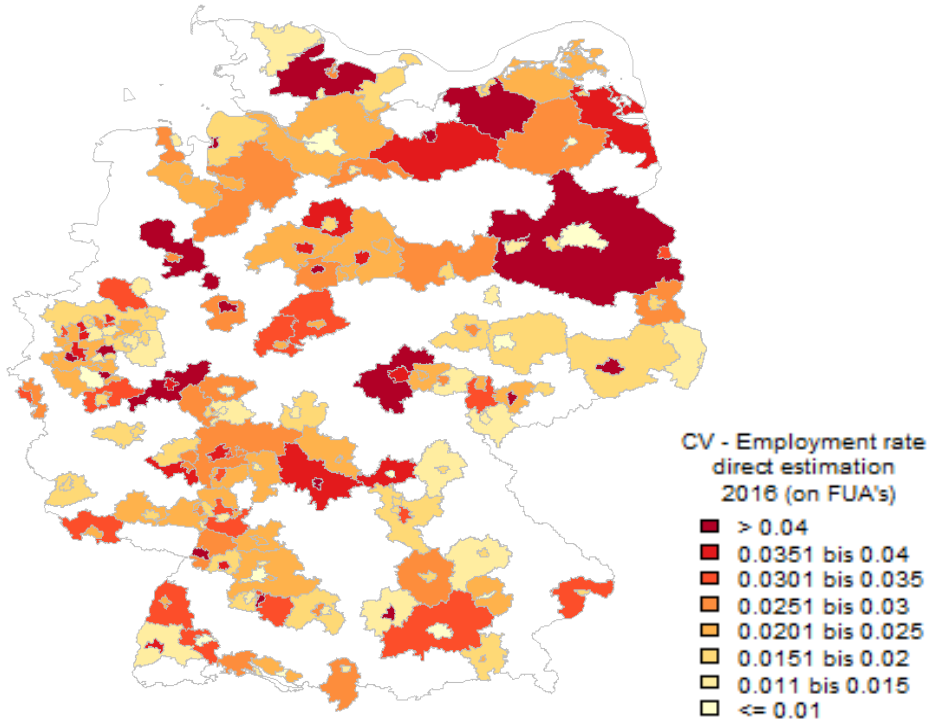
- Estimation of means based on sampling and survey design (Horvitz-Thompson-estimator)
  - One-Stage clustered sample (area sample)

$$\hat{\theta}_i^{direct} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$$

- Unbiased estimator for any design

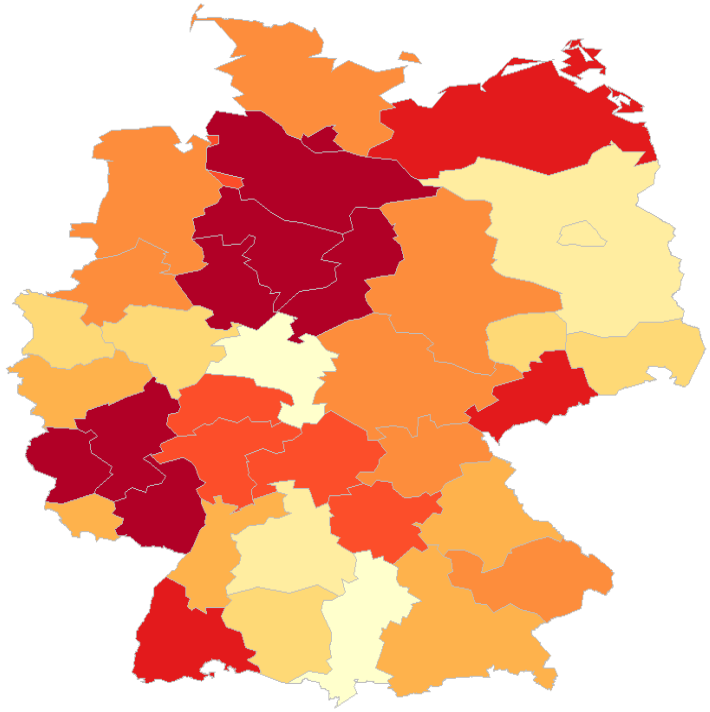
Starting point: Employment rate





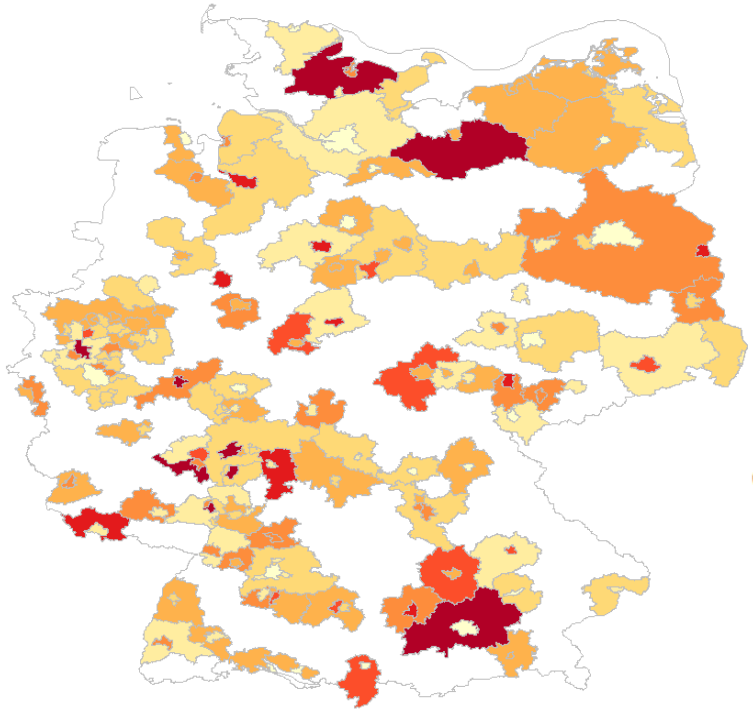
CV = coefficient of variation

Starting point: Unemployment rate



CV - Unemployment rate
direct estimation
2016 (on NUTS 2 Level)

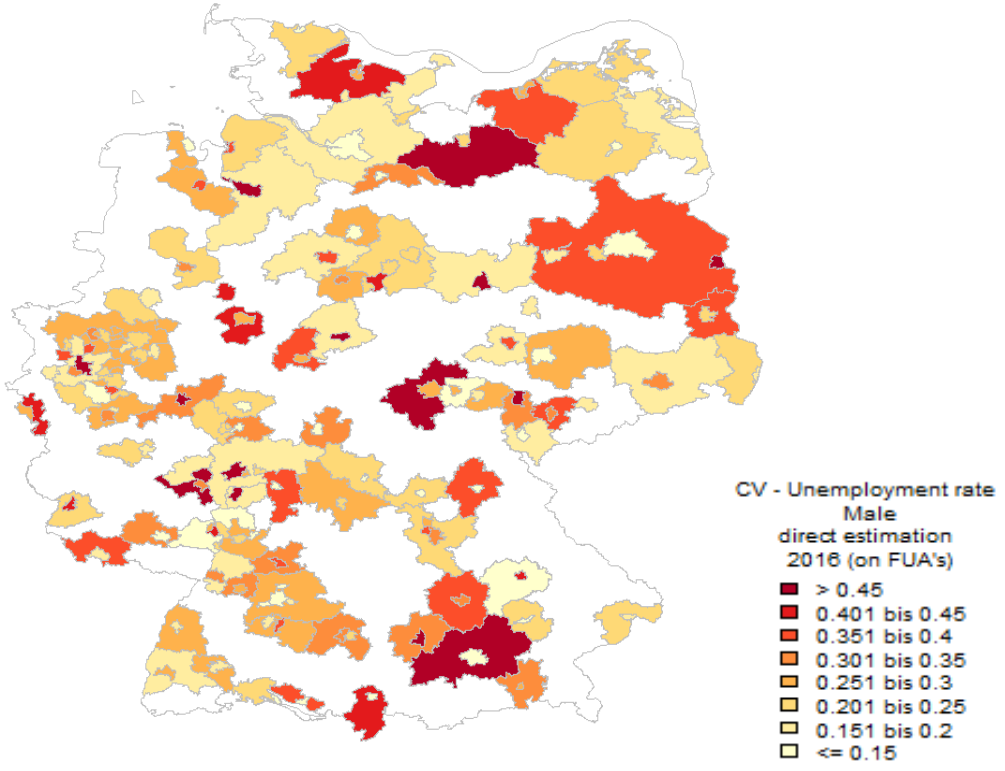
- > 0.1
- 0.091 bis 0.1
- 0.081 bis 0.09
- 0.071 bis 0.08
- 0.061 bis 0.07
- 0.051 bis 0.06
- 0.041 bis 0.05
- <= 0.04

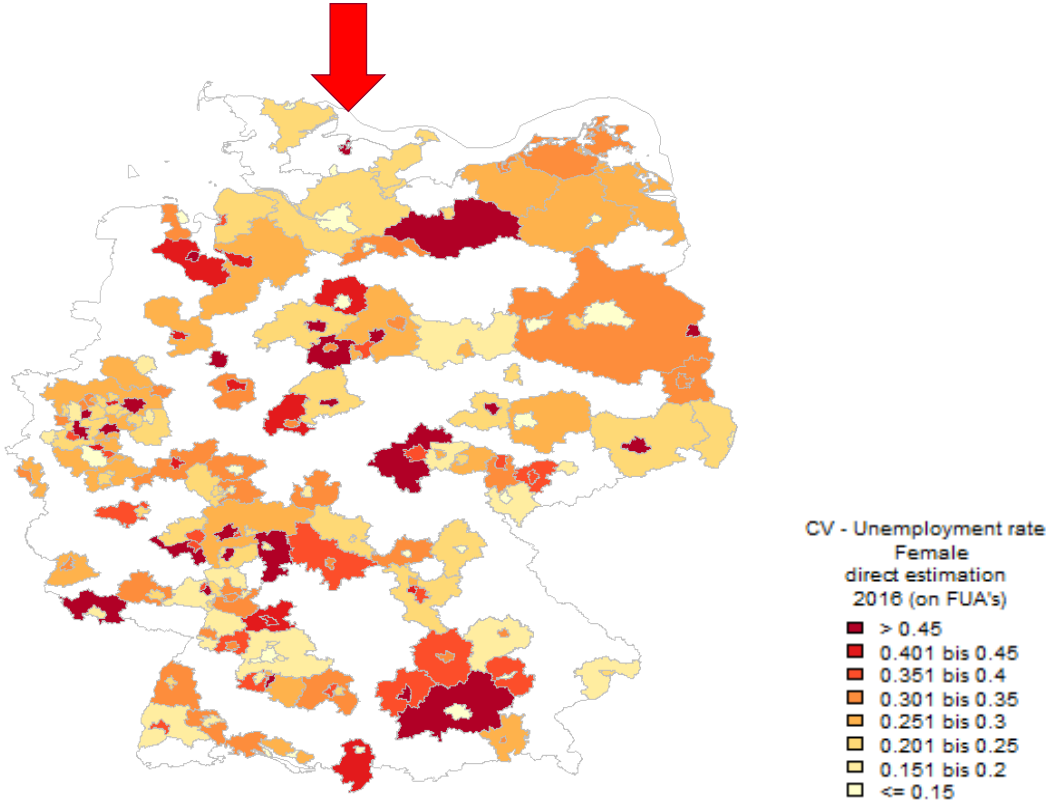


CV - Unemployment rate
direct estimation
2016 (on FUA's)

- > 0.4
- 0.351 bis 0.4
- 0.301 bis 0.35
- 0.251 bis 0.3
- 0.201 bis 0.25
- 0.151 bis 0.2
- 0.101 bis 0.15
- <= 0.1

Unemployment rate by Gender





Small Area Estimation (SAE)

- **Model of Schmid et al. (2017)**
  - **Fay-Herriot estimation**
  - **Area-level linear mixed model:**

$$\hat{\theta}_i^{direct} = \theta_i + \varepsilon_i = x_i^T \beta + u_i + \varepsilon_i,$$

Where $u_i \sim N(0, \sigma_u^2)$ and $\varepsilon_i \sim N(0, \sigma_{\varepsilon_i}^2)$ are normally distributed and independent

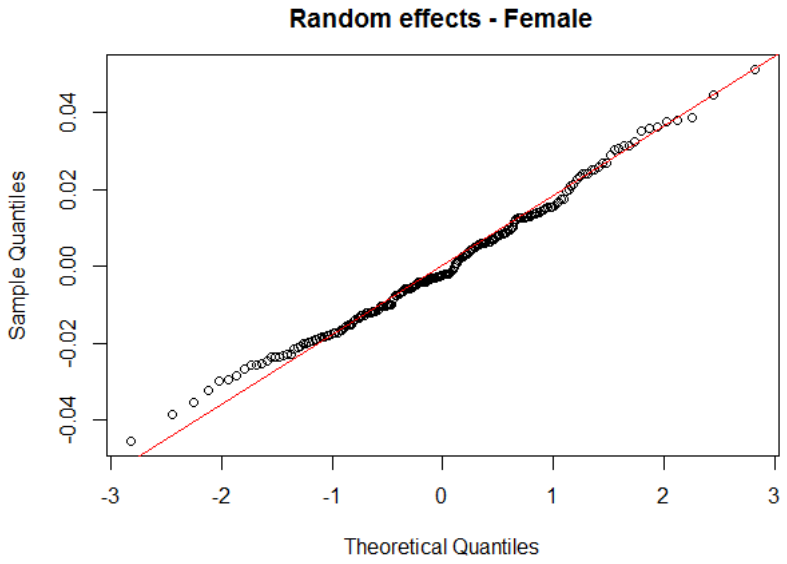
  - x_i **are the mobile phone covariates**
  - **The EBLUP under the Fay-Herriot (FH) model is obtained by**

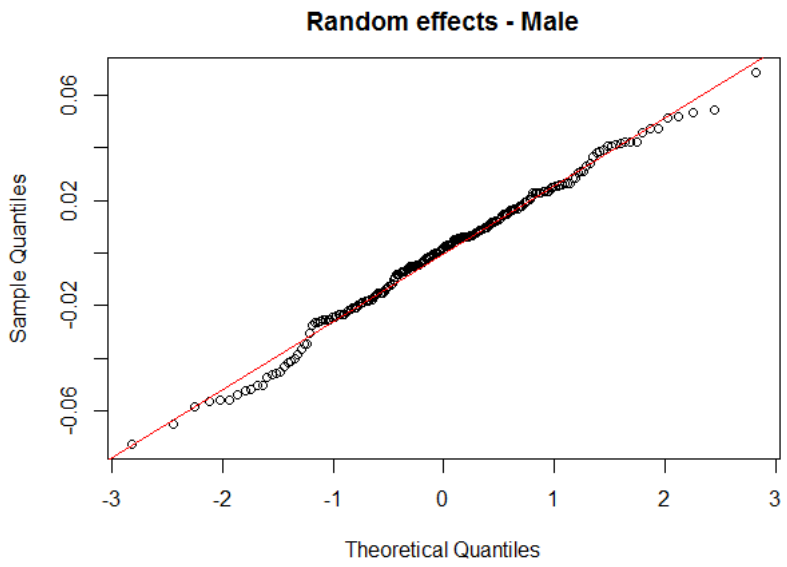
$$\hat{\theta}_i^{FH} = x_i^T \hat{\beta} + \hat{u}_i + \varepsilon_i = \gamma_i \hat{\theta}_i^{direct} + (1 - \gamma_i) x_i^T \hat{\beta},$$

Where $\gamma_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{\varepsilon_i}^2}$ denotes the shrinkage factor for area i .

- **Transfromed FH estimator :** $\hat{\theta}_i^{FH} = f^{-1}\left(\hat{\theta}_i^{FH}\right) = \sin^2(\hat{\theta}_i^{FH})$

Model with Mobile Phone Data: Q-Q Plots





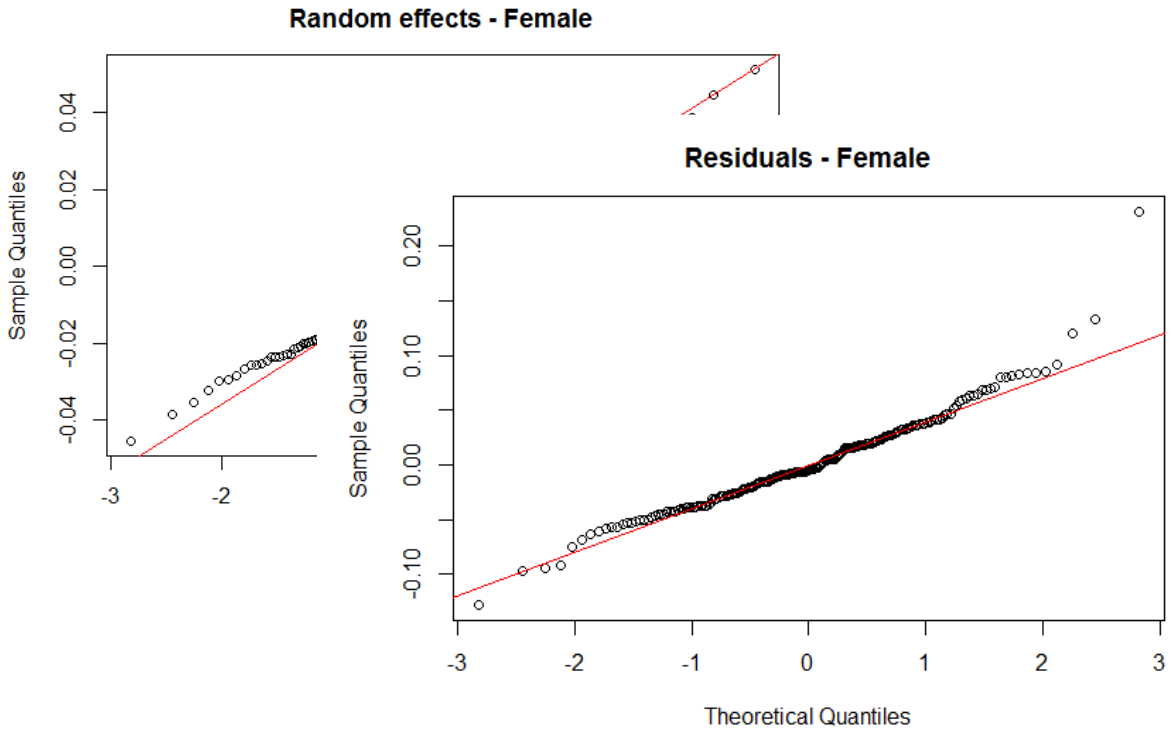
Shapiro-Wilk normality test data: p-value = 0.2414

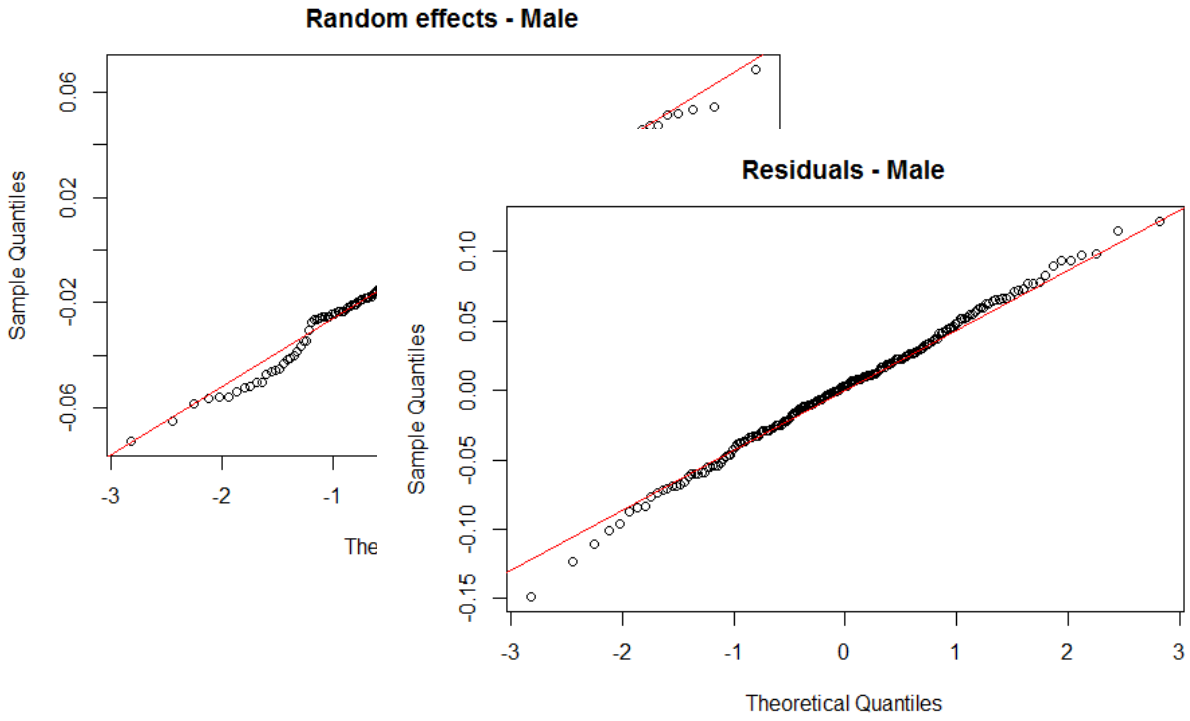
Model : R-squared: 0.1208, Adjusted R-squared: 0.1034

Shapiro-Wilk normality test data: p-value = 0.2802

Model : R-squared: 0.2841, Adjusted R-squared: 0.2591

Model with Mobile Phone Data: Q-Q Plots





Shapiro-Wilk normality test data: p-value = 0.2414 (left) and 7.936e-06 (right)

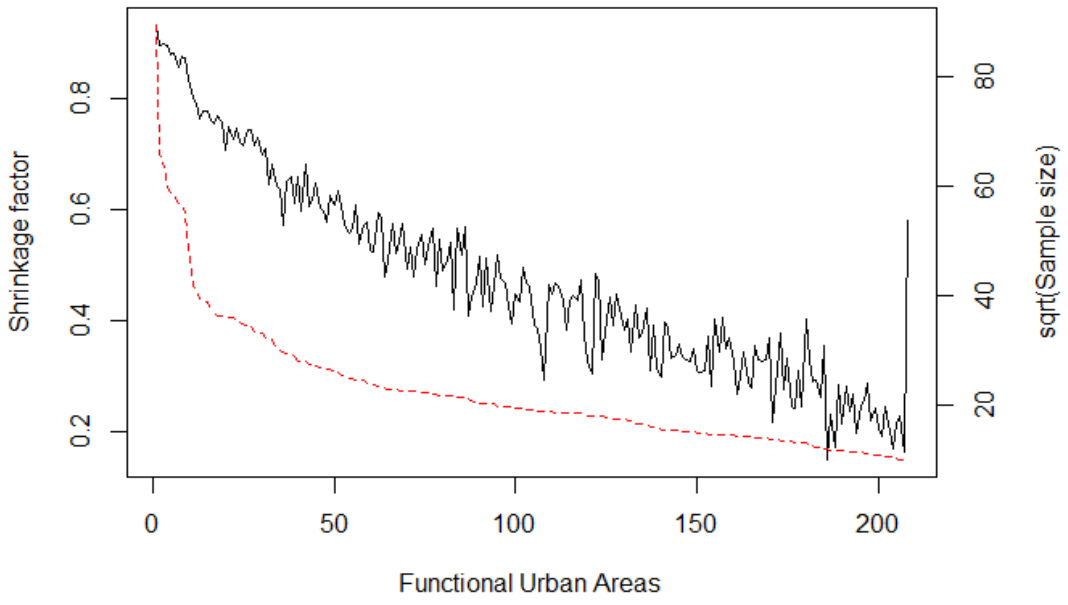
Model : R-squared: 0.1208, Adjusted R-squared: 0.1034

Shapiro-Wilk normality test data: p-value = 0.2802 (left) and 0.9047 (right)

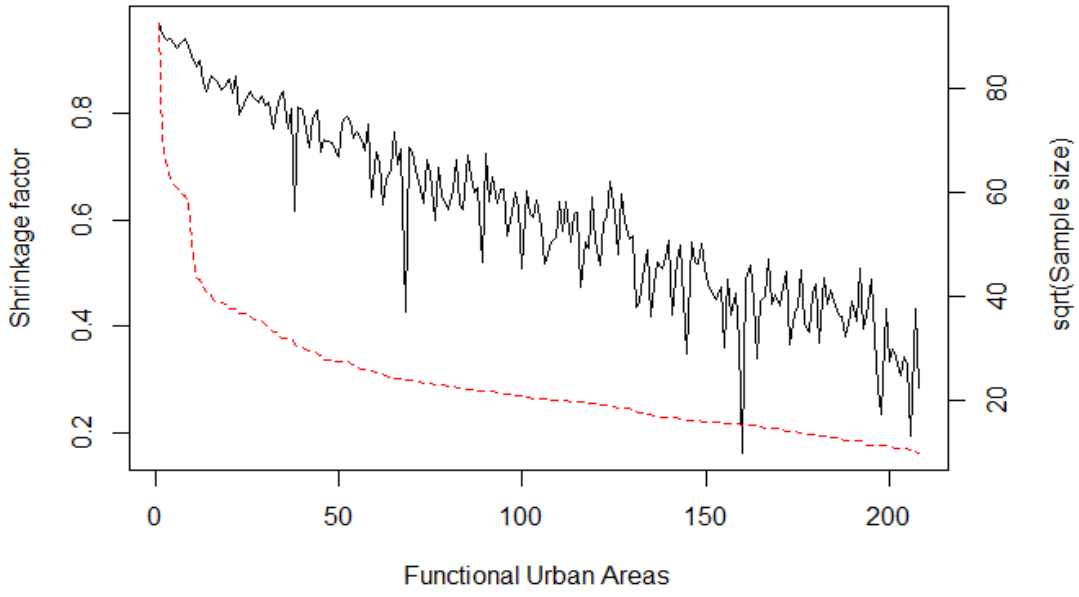
Model : R-squared: 0.2841, Adjusted R-squared: 0.2591

Shrinkage factors

Females model

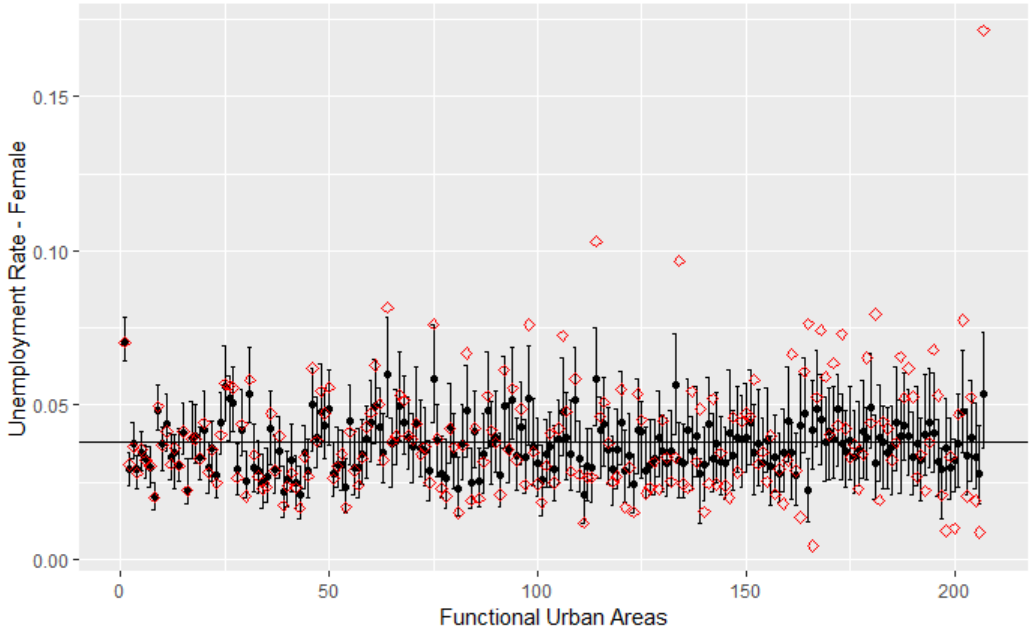


Males model

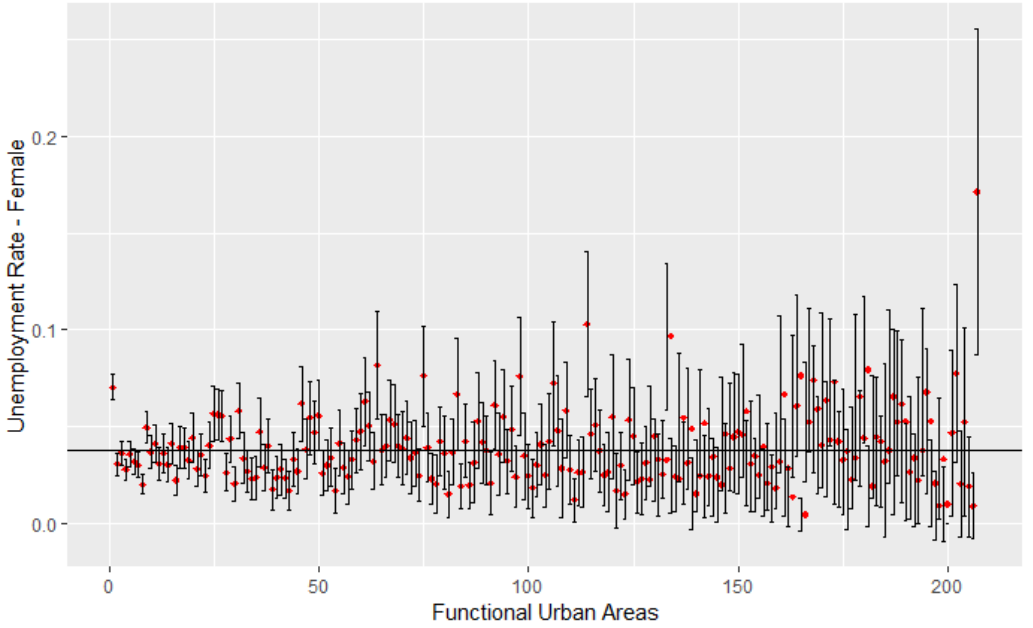


Comparison of direct estimators and Trans FH estimators – Females Model

Trans FH Bench -, direct estimator and confidence interval

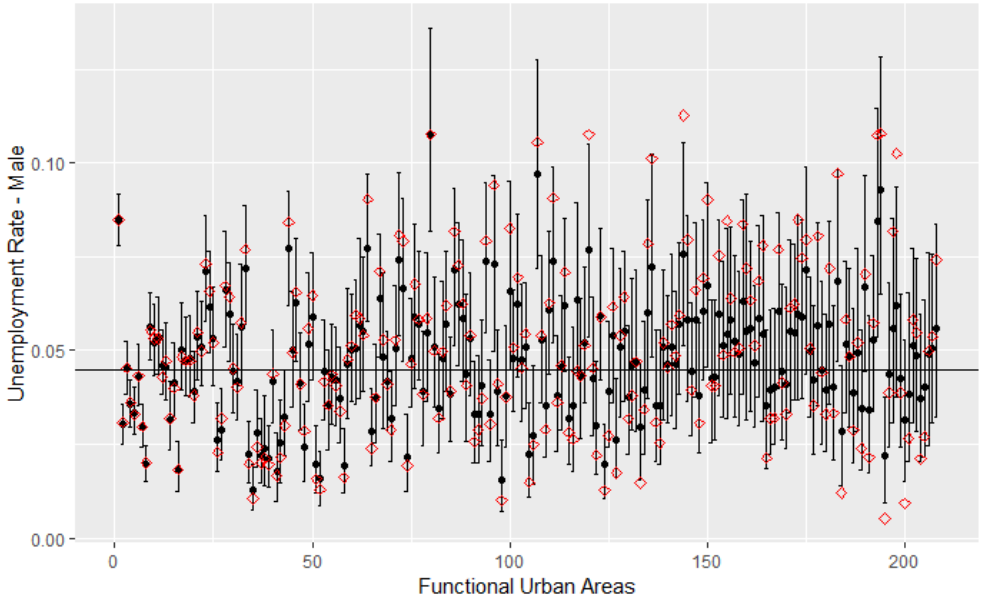


Direct estimator and confidence interval

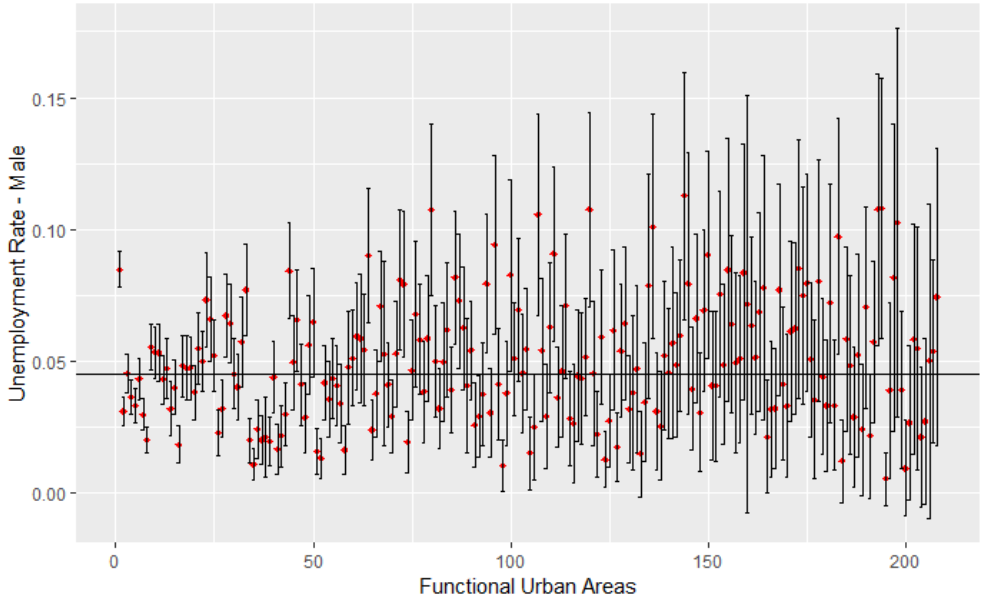


Comparison of direct estimators and Trans FH estimators – Males Model

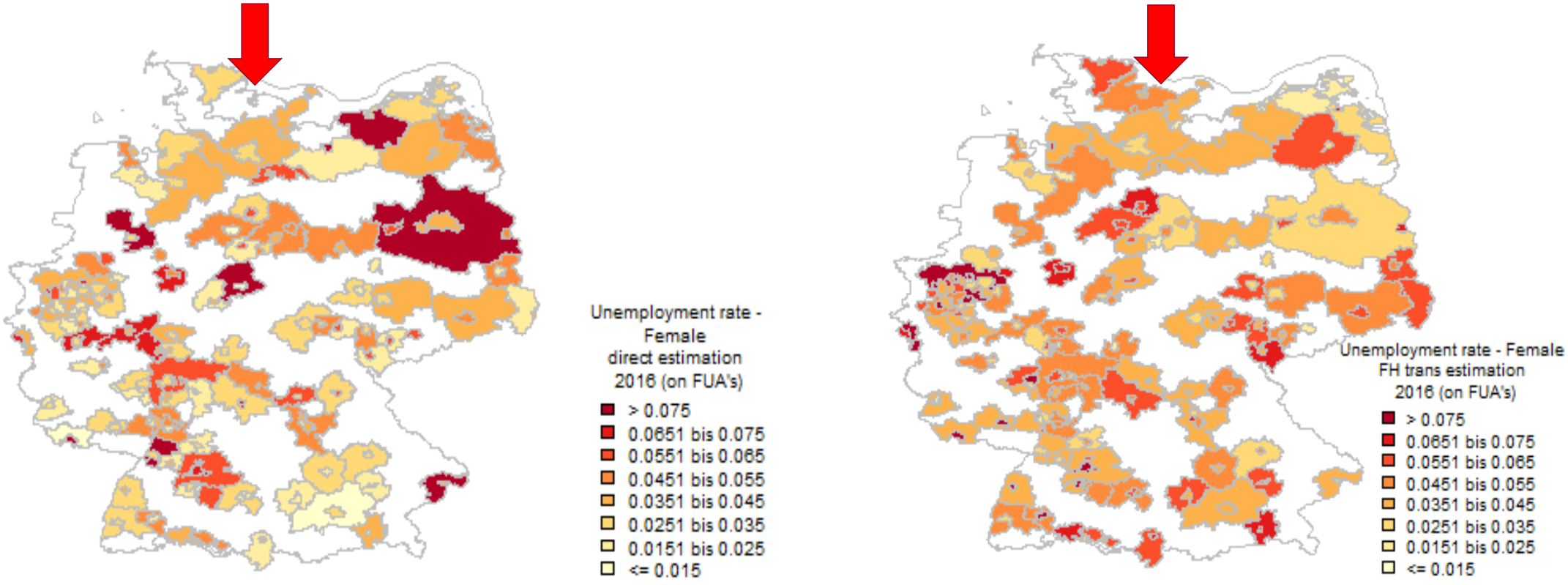
Trans FH Bench -, direct estimator and confidence interval



Direct estimator and confidence interval



Unemployment rate by Gender– Female



Conclusion

- More weight on direct estimators with higher sample sizes
- Smaller confidence intervals (indicates smaller variances)
- Obtaining of EBLUP estimators for out-of-sample domain

- Further research regarding on robust Small Area Estimation
- Alternative big data data sources with more explanatory power

THANK YOU FOR YOUR ATTENTION!

Institute for Research and Development in Federal Statistics

Sandra Hadam

+49 (0) 611 / 75 3452

sandra.hadam@destatis.de

© Fancy by Veer/The World from Above/FAN2043621