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

1. INTRODUCTION

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

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

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)

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

» Unbiased estimator for any design
Starting point: Employment rate

CV = coefficient of variation

© Federal Statistical Office (Destatis) | Institute for Research and Development in Federal Statistics
Starting point: Unemployment rate
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
    \[
    \hat{\theta}_i^{FH} = x_i^T \hat{\beta} + \hat{u}_i + \hat{\varepsilon}_i = \gamma_i \hat{\theta}_i^{direct} + (1 - \gamma_i)x_i^T \hat{\beta},
    \]
    Where \( \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{\varepsilon_i}^2} \) denotes the shrinkage factor for area \( i \).
  - Transformed FH estimator: \( \hat{\theta}_i^{FH} = f^{-1}(\hat{\theta}_i^{FH}) = \sin^2(\hat{\theta}_i^{FH}) \)
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
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