Statistical Methods & Applications (2024) 33:205–233
https://doi.org/10.1007/s10260-023-00722-0

ORIGINAL PAPER

Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany

Sandra Hadam[1] · Nora Würz[2] · Ann-Kristin Kreutzmann[3] · Timo Schmid[2]

Accepted: 4 August 2023 / Published online: 18 September 2023

Abstract
The ongoing growth of cities due to better job opportunities is leading to increased labour-related commuter flows in several countries. On the one hand, an increasing number of people commute and move to the cities, but on the other hand, the labour market indicates higher unemployment rates in urban areas than in the surrounding areas. We investigate this phenomenon on regional level by an alternative definition of unemployment rates in which commuting behaviour is integrated. We combine data from the Labour Force Survey with dynamic mobile network data by small area models for the federal state North Rhine-Westphalia in Germany. From a methodical perspective, we use a transformed Fay–Herriot model with bias correction for the estimation of unemployment rates and propose a parametric bootstrap for the mean squared error estimation that includes the bias correction. The performance of the proposed methodology is evaluated in a case study based on official data and in model-based simulations. The results in the application show that unemployment rates (adjusted by commuters) in German cities are lower than traditional official unemployment rates indicate.

Keywords Bias correction · Fay–Herriot model · Mean squared error · Small area estimation · Unemployment rates

1 Introduction

Since jobs are predominantly located in cities, more people move to the cities. For example, the continuous growth of cities is creating shortages on the German housing and real estate markets (Möbert 2018). Most large cities have higher population growth rates than the national average [see e.g., an interactive map of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR 2017)]. Due to urban labour migration, the number of people living in cities is

Sandra Hadam and Nora Würz are contributed equally to this work.

Extended author information available on the last page of the article

steadily increasing nationwide. As Buch et al. (2014), smaller cities recorded less net immigration than large cities, which is caused by the attractiveness of larger cities and the advantages of living in them. These are better infrastructure, more education and job opportunities, an extensive cultural infrastructure, and other location-specific amenities (Buch et al. 2014; Gans 2017).

In contrast to this trend, unemployment rates in Germany are higher in the cities compared to its surroundings. The unemployment rate is one of the most important economic indicators. Unemployment has far-reaching indirect effects on the respective region: It favours the decline of wage levels, educational activities within companies, population mobility, life and health satisfaction, intelligence, and school performance, as well as rising right-wing extremism (Grözinger 2009). The persistence of spatial disparities in unemployment in an economy is also shown by Elhorst (2003). He points out that regional unemployment is influenced by labour supply (affected by changes in the labour force, such as migration and commuting), labour demand, and wage-setting factors. Kosfeld and Dreger (2006) conduct a spatial analysis of the German regional labour market, showing that strong spatial dependencies can distort the relationship between employment and unemployment. Also Patuelli et al. (2011) include spatial linkages to effectively predict regional economic variables and to uncover spatial patterns. Particularly, there are strong relationships of dependence between cities and their surrounding areas. Identifying the cities as job magnets and finding high unemployment rates at the same time seems contradictory. According to Grözinger (2018), this phenomenon is a ’false’ effect and can be explained by the common definition of unemployment. Traditional unemployment rates are defined by the International Labour Organization (ILO) as the number of unemployed persons counted at their place of residence divided by the total number of persons in the labour force who are resident in the target area. This definition includes only the place of residence as a focal point for calculating these rates. In contrast to traditional unemployment rates, an alternative definition using the workplace as a focal point enables other insightful interpretation possibilities. Following Grözinger (2018), this alternative definition puts the resident unemployed of an area in relation to the labour force of the same area counted at the workplace. The alternative unemployment rate include commuters at their workplace and thus reflect the difference in the supply of jobs. This definition provides valuable information on missing workplaces in regional areas and support policy decisions in urban planning. Thereby, policymakers can identify regions where it might be useful to promote the settlement of companies to lower their unemployment rate and shorten commuter movements. For cities, lower alternative unemployment rates are assumed compared to the traditional definition. Low alternative unemployment rates contribute to the attractiveness of cities and the moving and commuting behaviour towards urban areas. Grözinger (2018) investigates this difference, among others, for regional areas in the German federal states Bavaria and Schleswig-Holstein.

Furthermore, the comparison of both rates also provides valuable information on commuting behaviour in regional areas.

For analysing unemployment rates in the context of commuter behaviour, we look at the regional level of Functional Urban Areas (FUAs). For member countries of the Organisation for Economic Co-operation and Development (OECD), FUAs have been created as harmonised geometries describing urban areas (Dijkstra and Poelman 2011). These regional areas are composed of city cores and their commuting zones. In this application, we use the FUAs in particular to include commuters and commuter areas to a greater extent. Hence, we are interested in considering only the city core and commuter zone separately, which is a spatial level underneath the FUA. We refer to our regional target level in the following as the FUA sublevel. This spatial level is particularly suitable for comparing the two unemployment rates described above, which differ in the spatial reference of the working population. Since this regional level is available for all OECD countries, our comparison of unemployment rates is transferable to other OECD countries and does not represent a purely German phenomenon. Furthermore, due to data availability, we only consider Germany and particularly the federal state of North Rhine-Westphalia (NRW) which is the federal state with the highest number of commuters in Germany (Bundesagentur für Arbeit 2022b).

To estimate unemployment rates, our primary data source is the European Union Labour Force Survey (LFS). The LFS enables the estimation of both unemployment rates. The survey is designed on the governmental regions level, which is a higher regional level than the FUA sublevel (Eurostat 2019b). According to the Nomenclature of Territorial Units for Statistics (NUTS) of the European Union, the German governmental region correspond to the NUTS 2-level and the districts to the NUTS 3-level. The FUA sublevel can be composed from the NUTS 3-level. Estimates on the spatial fine FUA sublevel that are only based on survey data (direct estimates) are likely to have large variances due to relatively small sample sizes. To increase the accuracy of the direct estimates on lower spatial levels, small area estimation (SAE) methods can be used [see e.g., Rao and Molina (2015), Tzavidis et al. (2018)]. SAE methods generally combine survey data with other data sources. For example, Costa et al. (2006), Pereira et al. (2011), and Martini and Loriga (2017) estimate unemployment rates using SAE methods by using administrative data as auxiliary information. Molina and Strzalkowska-Kominiak (2020) discuss different types of SAE estimators to calculate the percentage of people in the labour force for Swiss communes out of the LFS. They use administrative data that are provided at unit-level as auxiliary information. Similarly, Marino et al. (2019) propose semi-parametric empirical best prediction for unemployment rates that requires unit-level information. For many research questions, appropriate register or administrative data is not available. In particular, unit-level data is strictly protected. Furthermore, aggregated data is often not available at spatial finer resolutions, so that information at the target level is missing. One possibility is to use alternative data sources as covariates. Toole et al. (2015) and Steele et al. (2017) propose the usage of passively collected mobile phone data as auxiliary information, as they have a finer spatial resolution, high timeliness, and are available in real time. Basically, mobile network data can serve as a basis for producing statistics with a very high level of spatial, temporal and population coverage. For example, Steele et al. (2017) use Call Detail Records (CDRs) from the mobile network and remote sensing data for estimating poverty indices in developing countries. Toole et al. (2015) estimate changes in unemployment rates after shocks in the economy in case of mass layoffs at a plant by using mobile phone data. Moreover, Marchetti et al. (2015) have investigated solutions for a broad range of applications in using new digital data. They suggest three ways to use new digital data together with SAE techniques and show the potential of these data sources to mirror aspects of well-being and other socio-economic phenomena.

Our analyses are based on dynamic mobile network data, which is more widely available and has more information content than mobile phone data. This data source validly reflects actual commuting behaviour as well as time of day and residential population, which is important for providing auxiliary information. Since commuters and daytime population affect unemployment rates, the usefulness of these covariates for estimating the traditional and alternative unemployment rates becomes apparent. Our application combines mobile network data with data from the LFS to improve the estimation of both unemployment rates on the FUA sublevel. The aim is to compare both definitions of unemployment rates at the level of interest, thus highlighting the influence of commuters. As sample sizes are small at the FUA sublevel SAE methods are needed. From a methodological perspective, we consider the Fay–Herriot (FH) model (Fay and Herriot 1979) using mobile network data as auxiliary information. The inverse sine transformation of the dependent variable is used frequently in literature to estimate proportions when applying the FH model (Casas-Cordero et al. 2016; Burgard et al. 2016; Schmid et al. 2017). The transformation offers the advantage of stabilization of the sampling variances and helps to approximate better the normality assumptions of the model. Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) apply a naive back-transformation to obtain FH estimates and their confidence intervals on the original scale. In contrast, we use a bias corrected back-transformation following Sugasawa and Kubokawa (2017) while using as well the inverse sine transformation. To measure the uncertainty of these specific FH estimates, we propose a parametric bootstrap procedure orientated on González-Manteiga et al. (2008) to receive not only confidence intervals but also estimates for the mean squared error (MSE). The methodology is validated with official rates based on the Urban Audit. In a model-based simulation study, we show the benefit of a bias corrected back-transformation compared to a naive one.

The paper is structured as follows: Sect. 2 defines both types of unemployment rates and explains how they deal differently with commuters. Subsequently, this section introduces the data sources for constructing these indicators. Section 3 describes the statistical methodology. The SAE methods and the corresponding MSE estimation is applied in Sect. 4 to estimate both unemployment rates for the German federal state NRW on FUA sublevel. Section 5 investigates the methodology on German data for estimating the traditional unemployment rates and compares the results with official data. Furthermore, in Sect. 6, we conduct a model-based simulation study to assess the quality of the proposed estimator. Section 7 discusses further research potential.

2 Data sources and definitions for regional unemployment rates

In this section, we first introduce the two definitions of unemployment rates each dealing differently with commuters as well as our regional target level: the FUA sub-level (Sect. 2.1). Subsequently, our two data sources are described: the LFS data (Sect. 2.2) and mobile network data (Sect. 2.3).

2.1 Traditional and alternative definition of unemployment rates

The unemployment rate according to the definition of the ILO provides an international comparable indicator (ILO 2018). Following the ILO-definition, the traditional unemployment rate $\theta_{\mathrm{UR}_1,i}$ for regional area i is given by

$$\theta_{\mathrm{UR}_1,i} = \frac{\mathrm{N}_{i,\text{unempl. (residence)}}}{\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (residence)}}}. \tag{1}$$

This unemployment rate is defined by the number of unemployed persons living in area i ($\mathrm{N}_{i,\text{unempl. (residence)}}$) divided by the labour force of area i. The labour force is composed of the number of unemployed and employed persons living in area i ($\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (residence)}}$). For traditional unemployment rates, the focal point for counting employed and unemployed persons is their place of residence, where persons aged 15 to 74 are considered in the ILO-definition (ILO 2018; Eurostat 2018a). Please note that for reasons of comparability with German official statistics, we use the age range of 15–64 years throughout the analysis. In contrast to the traditional definition, the second definition proposed by Grözinger (2018) uses the workplace as a focal point and thus counts employed persons at the area i where their workplace is located. Since unemployed persons have no place of work, they count at area i where they live. The definition changes to

$$\theta_{\mathrm{UR}_2,i} = \frac{\mathrm{N}_{i,\text{unempl. (residence)}}}{\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (workplace)}}}. \tag{2}$$

We refer to $\theta_{\mathrm{UR}_2,i}$ as alternative unemployment rate for area i. It is composed by the number of unemployed persons ($\mathrm{N}_{i,\text{unempl. (residence)}}$) divided by the labour force aged 15–64 ($\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (workplace)}}$). Comparing alternative unemployment rates to traditional ones, the denominator changes as employed persons count in the area i where they work. Overall, both unemployment rates treat commuters differently. If commuting is not exactly balanced, the two unemployment rates differ, and this difference reveals the influence of commuters. If the traditional unemployment rate in area i ($\theta_{\mathrm{UR}_1,i}$) is higher than the alternative one ($\theta_{\mathrm{UR}_2,i}$), there is a stronger commuter movement from other areas to area i than the other way around which is assumed for larger cities.

We focus on the alternative definition of unemployment rates as defined in Eq. (2). However, there are, other alternative definitions such as those of the Federal Labour Office in Germany (Bundesagentur für Arbeit 2022a) or the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics 2021), which take into account a more socio-political perspective and the relative underutilisation of the labour supply. In contrast to the alternative definition according to Grözinger (2018) used here, the labour force remains the same as in the traditional unemployment rate, while the numerator changes.

In this study, the geographical target level for investigating unemployment rates is the FUA sublevel which is particularly suitable to illustrate the difference in both definitions of unemployment rates caused by commuter flows. To the best of our knowledge, the FUA sublevel is the only OECD harmonised geometry that allows a distinction between city cores and their commuter zones. City cores are urban centres with at least 50 000 inhabitants. The commuting zone contains the surrounding travel-to-work areas of the city core where at least 15% of their employed residents are working in the respective city core (Eurostat 2018b). Please note that the FUA sublevel as well as the FUA do not cover the whole territory of a country. Germany has in total 208 units, which are relevant for determining FUAs. These are composed of 125 city cores and 83 commuting zones. Since some commuting zones can be assigned to several city cores, there are fewer commuting zones than city cores.

2.2 Labour force survey

The LFS (Eurostat 2019b) enables the estimation of the traditional and alternative unemployment rates introduced in Sect. 2.1. It is a household survey conducted in 35 countries including all 27 EU member states and the UK, which provides information about the labour market participation. In Germany, the LFS is part of the German Microcensus, which is a one-percent sample of the population and collected annually. All inhabitants who have their main or secondary residence in Germany and live in private or collective households are included. The sampling design corresponds to a stratified single-stage cluster sample, where neighbouring buildings are sampled and all households and persons within this cluster are surveyed. The sample districts are stratified according to region and size of the buildings (Eurostat 2019c). In the used LFS data, regional disaggregation is carried out using the EU-harmonised NUTS classification (Eurostat 2018c). In Germany, the NUTS 1-level corresponds to the 16 federal states, the NUTS 2-level to 38 governmental regions, and the NUTS 3-level to the 401 administrative districts (European Parliament and

Council 2003). Traditional unemployment rates using LFS data are published on the 38 governmental regions level (NUTS 2-level). However, our target level is the smaller FUA sublevel which can be composed from the NUTS 3-level in Germany. As all LFS observations contain information about the NUTS 3-level and even finer, we can use the individual information of the LFS participants to match (a) the place of residence and (b) the place of work to the corresponding FUA sublevel.

In addition to the FUA sublevel, there are other possible spatial levels that are suitable to examine unemployment rates. The so-called Labour Market Areas (LMAs) are functional spatial areas that capture regional labour market structures based on commuting flows (Franconi et al. 2017). In principle, both territorial structures pursue the same goal. Nevertheless, there are practical reasons and advantages to prefer the FUA sublevel in the context of this work: First, the LMAs are compiled from commuter statistics using a specially developed algorithm. In contrast, FUAs are based on territorial structure, are already harmonised, and comparable across countries. Second, the separation of the city cores and commuter zones is an advantage of FUAs versus LMAs, which is fundamental for our analysis. Third, Germany provides indicators for the Urban Audit, which is an official statistic and publishes labour market indicators, including traditional unemployment rates, at the level of the entire FUA (one level above our target level) which we can use for external validation. All in all, the FUAs are more suitable for our analyses than the LMAs, since they fit better to the research question and are easier to handle.

In this work, we consider the year 2016 with an overall sample size of 369,986 observations in the LFS. Since the FUA sublevel does not cover the whole territory, the sample size decreases to 271,587 observations. Due to known gender differences in employment, the following analyses are carried out separately by sex. Men work more often full-time, while the proportion of women employed part-time has increased in recent years, so that overall fewer women than men are unemployed (Klammer and Menke 2020; Statistisches Bundesamt 2021). Due to the still existing classical gender role model, women commute fewer and shorter distances than men. These differences in behaviour justify why it is meaningful to examine unemployment separately by sex (Augustijn 2018). Table 1 represents the sample sizes in the LFS based on the published NUTS 2-level and on the FUA sublevel by sex. It can be seen that the sample sizes are smaller in case of the FUA sublevel. On average, the sample sizes decrease by a factor of 7.3. Since the LFS was designed to produce

Table 1 Distribution of sample sizes in the LFS on NUTS 2-level and FUA sublevel in Germany by sex

Sex	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NUTS 2-level						
Female	1162	2916	4104	4623	5521	10,684
Male	1318	3304	4565	5114	6108	11,675
FUA sublevel						
Female	100	216	368	635	646	7 973
Male	97	244	421	702	749	8 559

reliable estimates on NUTS 2-level, the challenge of this work is to estimate reliable unemployment rates on the smaller FUA sublevel. Even if the sample sizes for FUA sublevel appear to be rather high, with a median of 368 and 421 for men and women, respectively, results in Sect. 4 show that the coefficient of variation (CV) of the direct estimates often exceeds the threshold of 20% which specifies reliable estimates at Eurostat (Eurostat 2019a). SAE methods are discussed to obtain more reliable model-based estimates on the FUA sublevel. Since SAE methods take advantage of auxiliary variables from other data sources the auxiliary information used here is described in more detail in the next Sect. 2.3.

2.3 Mobile network data

To estimate unemployment rates on FUA sublevel using SAE methods, we take advantage of suitable auxiliary information. Many SAE applications are based on register data as a second data source. These data sources are not timely or are aggregated to higher (regional) levels. Alternative data sources have the potential to overcome these disadvantages. For example, Toole et al. (2015) or Steele et al. (2017) have used mobile phone data for SAE. Mobile network data are explored to estimate daytime population, commuter patterns or tourism behaviour [see e.g., De Meersman et al. (2016), Galiana et al. (2018)]. Mobile network data represent mobile activities or signals from the mobile network of the respective mobile network operator. A mobile activity is defined as an event caused by a length of stay in a specific geometry without movement (also known as dwell time). Signalling data are produced automatically, regularly and only register the location of the cell tower to which a mobile device is connected at a specific time. Therefore, they are collected as a by-product and tend to be less costly compared to official survey data. The major advantage of these data sources are their real-time availability, high temporal actuality, nationwide availability, and their finer spatial resolution. Mobile activities can be obtained at the spatial resolution of cities, communities or grid cells, so that a simple assignment to other spatial levels such as the FUA sublevel is possible. This spatial flexibility and high resolution are not feasible with register or administrative data. In many countries, like Germany, register data are strictly protected and thus not available at high resolution or on specific regional levels. In addition, mobile network data are dynamic, so that the movement of activities can be observed over the course of the day as well as daily, during a week or a month. Previous analyses in Germany have shown that mobile network data correlate strong with register-based census data like population figures and with the population mobility, more precisely commuter movements (Hadam 2018, 2021). This is, among other things, due to the high penetration rate of mobile devices in the German population (Statistisches Bundesamt 2022). Accordingly, mobile network data provide a reliable picture of the real physical locations of the German daytime and night-time population or with other words the resident and working population compared to official statistics with a fixed reporting date. Since our aim is to estimate an alternative unemployment rate accounting for commuters mobile network data reflecting resident and working population are especially suitable auxiliary information [cf. Hadam (2021)].

In Germany, there are three mobile network operators: Deutsche Telekom, Vodafone, and Telefónica Deutschland, with a respective market share of one-third each. The data records available to Destatis and used for this work contain mobile activities of Deutsche Telekom customers. In compliance with data protection rules, the mobile activities are anonymised and aggregated. Regionally fluctuating market shares of each mobile network operator are adjusted regionally as part of the extrapolation procedure at the respective operator. Thus, the estimated local market shares of each operator are used as weights to adjust the mobile network data. The data records include contract, prepaid, and further customers. In addition, mobile network data contain information on socio-demographic characteristics of mobile device users, such as age group, sex, and nationality of the SIM card owner. However, the characteristics are only available for contract customers. Furthermore, the following assumptions were made in the data provider's data generation process: since the number of mobile activities depends on the dwell time of mobile devices, long mobile device activities are counted and included in the data record according to the length of the dwell time, while short mobile activities are not considered. The dwell time in the data record available is two hours to filter out short mobile device activities (for example, quick movements between the grid cells). Finally, only values based on a minimum number of 30 activities per geometry were provided due to data protection reasons.

Our aim is to analyse the effect of commuters on the two proposed unemployment rates. Since we use a model-based method, suitable covariates are crucial. We only use mobile network data for this purpose and no further covariates. As we will show in Sect. 4.1, our models with only mobile network covariates lead to high coefficients of determination, so mobile network data are sufficient as SAE covariates in our case.

We define from the mobile network data 27 auxiliary variables. Between 7 and 16 auxiliary variables are chosen by model selection procedure (cf. Section 4.1). The data contains mobile activities for a statistical week that consists of 24-h days. These were selected from the months April, May, and September in 2017 without school or public holidays to avoid distortions in the representation of commuters. The mobile activities comprise the average activities on the selected weekdays. The weekdays are categorised according to five types of days, with the days from Tuesday to Thursday being grouped together. Since the counted activities of mobile devices alone are not meaningful enough, further covariates are constructed from the available mobile network data at the FUA sublevel. The aim in creating the covariate is to highlight the differences between the daily and resident population and thus the commuters themselves. This is particularly reflected in the changes in the intensities of mobile activities. Based on this, covariates are calculated in the form of ratios, shares, and change values which reflect exactly these differences. Since it is assumed that the unemployed persons are more likely to stay at home during the day and the employed are more likely to stay at the place of work, the rate and change of activities in the morning and evening hours are calculated. This means, that the change from place of work to the place of residence and vice versa is modelled. This includes the change in mobile activities of working hours and hours spent at home as well as the change in activities of potential commuters. In addition, the change in activities during the day is calculated and the differences in core times or peaks in mobile activities are determined. The core times are based on the usual working times in Germany (7 am to 4 pm). Furthermore, differences in mobile network activities among to socio-demographic characteristics such as age, nationalities (summarised by continent), and sex can also be considered. These characteristics also have an influence on commuting behaviour. An overview of the selected mobile network covariates can be found in the supplementary material in Table 6.

3 Small area method

In this section, the statistical methodology for estimating unemployment rates on FUA sublevel is described. As the LFS is designed for higher regional levels, a model-based approach enriched by auxiliary variables from mobile network data is used. We use the FH model (Fay and Herriot 1979), an area-level model that links direct estimates to area level covariates. The FH model is especially useful in countries with strict data protection requirements like Germany, as the auxiliary variables and the direct estimates only need to be available on an aggregated level. As in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) we use the inverse sine transformation on the dependent variable to estimate proportions using area-level models. Following Sugasawa and Kubokawa (2017), we derive the inverse sine transformed FH model including a bias correction for the back-transformation. A parametric bootstrap, which incorporates the bias correction, is proposed.

3.1 Fay–Herriot estimates

In the following, we assume a finite population of size N, which is divided into d areas. The present sample consists of areas with different sample sizes $n_1, ..., n_d$ drawn by a complex design from the population. To refer to the actual area, we use the subscript i. The population size and sample size of this area is indicated with N_i and n_i, respectively. The FH model is composed of two levels. The model for the first level is the sampling model

$$\hat{\theta}_i^{\text{direct}} = \theta_i + e_i \quad \text{with } i \in 1, ..., d.$$

$\hat{\theta}_i^{\text{direct}}$ is an unbiased direct estimator for a population indicator of interest θ_i. The sampling errors $e_i \sim N(0, \sigma_{e_i}^2)$ are independent and normal distributed and their variance $\sigma_{e_i}^2$ is assumed to be known. In applications, the sampling variance is typically supplied by the data provider or estimated from unit-level sample data. We use the survey package from R (Lumley 2004; R Core Team 2022) and consider the sampling design of the LFS and the survey weights to estimate $\hat{\theta}_i^{\text{direct}}$ and the respective sampling variances.

The second stage of the FH model links a vector with p area-specific covariates $\mathbf{x}_i$ (aggregates, e.g. area-level means) to the indicator of interest using an area-specific random effect u_i for each area $i \in 1, ..., d$:

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2),$$

where β is a vector of unknown fixed-effects parameters.

Combining both levels results in the FH model:

$$\hat{\theta}_i^{\text{direct}} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e_i}^2).$$

The model assumes that the random effects u_i are identically independently normally distributed and the sampling errors e_i are independently normally distributed. The regression parameters $\hat{\boldsymbol{\beta}}$ can be estimated as best linear unbiased estimator of $\boldsymbol{\beta}$ and the random effect $\hat{u}_i$ as empirical best linear unbiased predictor of u_i (Rao and Molina 2015). For the estimation of the variance of the random effects σ_u^2, several approaches are available: The FH method of moments, the maximum likelihood method (ML), and the restricted maximum likelihood method (REML) among others (Rao and Molina 2015). For our analysis, we use the REML method.

Through this combination, we obtain the resulting FH estimator, which is an empirical best linear unbiased predictor of θ_i. It is as a weighted combination of the direct estimator $\hat{\theta}_i^{\text{direct}}$ and the synthetic estimator $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ for each area i:

$$\begin{aligned} \hat{\theta}_i^{\text{FH}} &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \hat{\gamma}_i \hat{\theta}_i^{\text{direct}} + (1 - \hat{\gamma}_i) \mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \end{aligned} \tag{3}$$

where the shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ defines the weight on both parts for each area i. Whenever the variance of the sampling errors is relatively small for a specific area i, more weight is assigned on its direct estimator.

3.2 Back-transformed Fay–Herriot estimates

As unemployment rates are a percentage, we transform the dependent variable to profit from the variance stabilization of the sampling variance. Thus, we use the inverse sine transformation $h(x) = \sin^{-1}(\sqrt{x})$ as in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017). Note that Schmid et al. (2017) compared in a design-based simulation study the inverse sine transformation with alternative modelling options, for instance an estimator based on a normal-logistic distribution. Both estimators lead to very similar results regarding MSE and bias. Raghunathan et al. (2007) defends the choice of the inverse sine transformation for estimating cancer risk factors rates against generalized linear models with their higher complex design features and computational tasks. While they all use a naive back-transformation $h^{-1}(x) = \sin^2(x)$, we transform the FH estimator back to the original level with consideration to the back-transformation bias. Burgard et al. (2016) mentioned the methodology for a bias corrected back-transformation. We derive the back-transformation following Sugasawa and Kubokawa (2017), who introduce the FH model for general transformations on the dependent variable. Following Jiang et al. (2001), we approximate the sampling variances of the transformed direct estimates by $\tilde{\sigma}^2_{e_i} = 1/4\tilde{n}_i$, where $\tilde{n}_i$ denotes the effective sample size. The design effects and thus the effective sample size can also be estimated with the *survey* package (Lumley 2004; R Core Team 2022). For the model on the transformed scale, we consider the assumptions of the FH model

$$\sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \tilde{\sigma}^2_{e_i}). \tag{4}$$

Out of the FH model on transformed scale in Eq. (4), $\hat{\boldsymbol{\beta}}$ and $\hat{u}_i$ can be estimated, as described in the previous Sect. 3.1. Replacing the model parameters with their estimates leads to the FH estimator on the transformed level:

$$\hat{\theta}_i^{\text{FH}^*} = \hat{\gamma}_i \sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) + (1 - \hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}.$$

However, the goal is to get the FH estimator on the original scale $\left(\hat{\theta}_i^{\text{FH, trans}}\right)$. For this reason, $\hat{\theta}_i^{\text{FH}^*}$ must be back-transformed. According to the Jensen-inequality (Jensen 1906), a naive back-transformation $\left(\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right)$ leads to biased results due to the non-linearity of the transformation. To avoid this bias, the following formula using the known distribution of the FH estimator on the transformed level $\hat{\theta}_i^{\text{FH}^*} \sim \mathcal{N}\left(\hat{\theta}_i^{\text{FH}^*}, \frac{\hat{\sigma}_u^2\tilde{\sigma}^2_{e_i}}{\hat{\sigma}_u^2+\tilde{\sigma}^2_{e_i}}\right)$ is used

$$
\begin{aligned}
\hat{\theta}_i^{\text{FH, trans}} &= E\left\{\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right\} \\
&= \int_{-\infty}^{\infty} \sin^2(t) f_{\hat{\theta}_i^{\text{FH}^*}}(t)dt \\
&= \int_{-\infty}^{\infty} \sin^2(t) \frac{1}{\sqrt{2\pi \frac{\hat{\sigma}_u^2 \tilde{\sigma}_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}}} \exp\left(-\frac{\left(t - \hat{\theta}_i^{\text{FH}^*}\right)^2}{2\frac{\hat{\sigma}_u^2 \sigma_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}}\right)dt,
\end{aligned}
\tag{5}
$$

where $\hat{\theta}_i^{\text{FH, trans}}$ denotes the transformed FH estimator. To solve this integral, numerical integration techniques are applied. In Sect. 6, the proposed bias corrected FH estimator ($\hat{\theta}_i^{\text{FH, trans}}$) is evaluated in a close to reality model-based simulation study.

3.3 Uncertainty estimation

As a measurement of uncertainty for $\hat{\theta}_i^{\text{FH, trans}}$, a parametric bootstrap MSE as well as parametric bootstrap confidence intervals are constructed. When using a FH model without transformations or with a log transformation, analytical solutions to estimate the MSE are known (Prasad and Rao 1990; Datta and Lahiri 2000; Slud and Maiti 2006). Up to our knowledge, no analytical solution is available in the case of the inverse sine transformation. Bootstrap methods are very promising to estimate the MSE. Casas-Cordero et al. (2016) construct confidence intervals using a parametric bootstrap procedure, in which confidence interval limits are built on the transformed scale with subsequent naive back-transformation for each bootstrap replication. In contrast to this methodology, our goal is to construct confidence intervals and a MSE for FH estimates from a model using the inverse sine transformation. Another difference is that, instead of the naive back-transformed FH estimates, the bias corrected back-transformed FH estimates are included within the bootstrap procedure. Our parametric bootstrap is orientated on the bootstrap procedure of González-Manteiga et al. (2008). In the following, the steps of the used bootstrap method to construct both measurements of uncertainty are shown:

- From the model on the transformed scale [Eq. (4)], take $\tilde{\sigma}_{e_i}^2$ and estimate $\hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ using the sample data.
- For $b = 1, ..., B$
  - Generate area specific random effects $u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and sampling errors $e_i^* \sim \mathcal{N}(0, \tilde{\sigma}_{e_i}^2)$.
  - Bootstrap samples:

    Use u_i^* and e_i^* to construct the bootstrap sample on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}^{\text{direct}}_{i,(b)}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i^* + e_i^*.$$

Use the bootstrap sample to estimate the FH estimator on the transformed scale $\left(\hat{\theta}^{\text{FH}^*}_{i,(b)}\right)$ as described in Sect. 3.2.

Determine the FH estimates on the original scale $\left(\hat{\theta}^{\text{FH, trans}}_{i,(b)}\right)$ using [Eq. (5)] to account for the bias correction.

– Bootstrap population:

Use u_i^* to construct the bootstrap population on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}^{\text{direct}}_{i,(b)}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i^*.$$

For each bootstrap population, calculate the population mean on the original scale

$$\theta^{\text{trans}}_{i,(b)} = \sin^2\left(\mathbf{x}_i^T\boldsymbol{\beta} + u_i^*\right).$$

- Predict the MSE and the 95% confidence intervals

$$\text{MSE}(\hat{\theta}^{\text{FH, trans}}_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{\text{FH, trans}}_{i,(b)} - \theta^{\text{trans}}_{i,(b)}\right)^2 \tag{6}$$

$$\begin{aligned}\text{CI}(\hat{\theta}^{\text{FH, trans}}_i) = \Big[&\hat{\theta}^{\text{FH, trans}}_i + q_{0.025}\left(\hat{\theta}^{\text{FH, trans}}_{i,(b)} - \theta^{\text{trans}}_{i,(b)}\right);\\ &\hat{\theta}^{\text{FH, trans}}_i + q_{0.975}\left(\hat{\theta}^{\text{FH, trans}}_{i,(b)} - \theta^{\text{trans}}_{i,(b)}\right)\Big],\end{aligned} \tag{7}$$

where $q_{0.025}$ is the 2.5% quantile over the bootstrap replications and $q_{0.975}$ respectively the 97.5 % quantile.

The methodology presented above for constructing uncertainty measurements for the back-transformed FH estimates is also evaluated within a simulation study (cf. Section 6).

4 Alternative unemployment rates including commuters in North Rhine-Westphalia

In this section, we determine and discuss traditional and alternative unemployment rates that deal differently with commuters. For this purpose, we use the LFS data from Sect. 2.2 and the mobile network data from Sect. 2.3. Traditional and

Table 2 Measurements to validate the FH models for traditional (UR_1) and alternative unemployment rates (UR_2) separated by sex: this table shows the estimated variance of the random effects ($\hat{\sigma}_u^2$), the Shapiro–Wilks (S.–W.) p-value for level 1 and level 2 error terms as well as the modified R^2

	Men		Women	
	UR_1	UR_2	UR_1	UR_2
$\hat{\sigma}_u^2$	0.000320	0.000361	0.000716	0.000880
S.–W. p-value: level 1	0.308668	0.495064	0.809323	0.866098
S.–W. p-value: level 2	0.695112	0.549476	0.861257	0.901708
Modified R^2	0.772521	0.908642	0.632059	0.575550

alternative unemployment rates have been introduced in Sect. 2.1. The members of the labour force are counted for the two rates at different reference points: At the place of residence (traditional unemployment rates) or at the place of work (alternative unemployment rates). In particular, they assign commuters to different small areas. When using traditional unemployment rates, the contradiction of high unemployment rates in the city cores results from the exclusion of commuting. Alternative unemployment rates are expected to exceed traditional ones in commuter zones and to be lower in city cores. We confirm this empirically. The rates are estimated separately by sex and at the target level of the FUA sublevel.

4.1 Model selection and validation

Four models need to be created and validated. Following Schmid et al. (2017), the Bayesian information criterion for a simple linear regression model is used for the

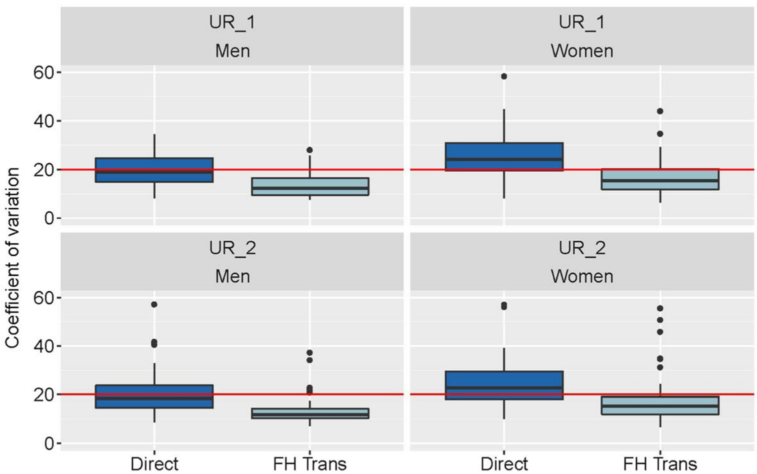

Fig. 1 Reduction of the coefficient of variation by using the transformed FH model instead of direct estimation for estimating unemployment rates in NRW

model selection. As dependent variable, we use the inverse sine transformed direct estimates from LFS and the auxiliary information is mobile network data (cf. Section 2.3). In total, 6 to 16 of 27 potential mobile network covariates are selected depending on the model. The covariates of all four models are listed in Table 6 within the Appendix A. Since the models are built on the transformed scale, the coefficients have no natural interpretation in terms of expected values at the original level, but their direction is directly interpretable. The chosen covariates reflect most likely relationships between working and non-working hours and the changes in mobile activities due to commuting during the day and evening. The latter is represented less strongly in the females model, which is in line with lower commuting patterns of women. An increase of covariates that proxy possible commuter movements generally leads to a decrease of alternative unemployment rates (UR_2). The reverse is the case for traditional ones (UR_1). All models include changes from night to day activities of other nationalities, most likely tourists, which have a positive impact on regional employment. As expected, negative values have been observed for these coefficients.

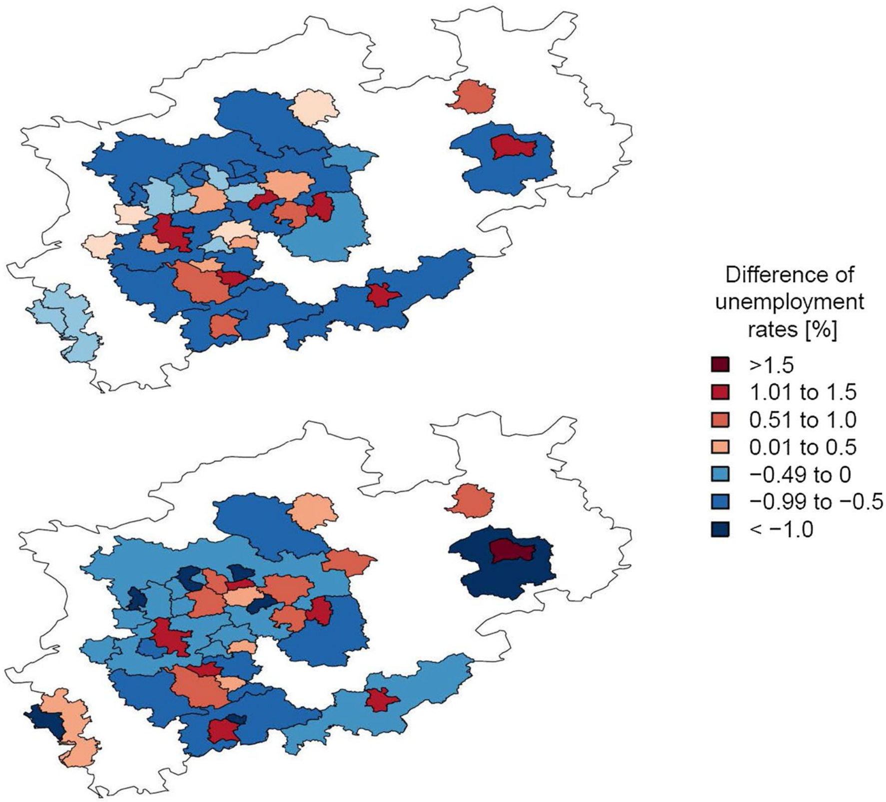

Fig. 2 Difference of unemployment rates due to including commuters for men (above) and women (below). The spatial assignment of city names to the FUA sublevels is shown in the Appendix B

To investigate the explanatory power of the models, we use the modified R^2 from Lahiri and Suntornchost (2015) and obtain values of at least 57% as shown in Table 2. Furthermore, we check whether meaningful results are obtained for estimating the variance of the random effects using REML estimation. As Table 2 shows, positive values were estimated in all cases. Thus, the potential problem of negatively estimated variances does not occur. For each FH model on the transformed scale, the assumptions on the error terms (level 1 and 2) are checked. The normality assumptions of the random effects (level 2) as well as of the residuals (level 1)—obtained from fitting the model [Eq. (4)]—are tested. The p-values of the Shapiro–Wilks test in Table 2 confirm that in all cases the normality assumption for both error terms cannot be rejected. Overall, all four models could be validated and are suitable for subsequent analyses.

4.2 Gain in accuracy

To assess the gain in the reliability of the estimators, we compare the CVs. Figure 1 visualises this measurement for the different methods and definitions of unemployment rates. Eurostat considers estimators with a CV below 20% to be reliable (Eurostat 2019a). If we use direct estimation 53.7% (men; UR_1), 29.3% (women; UR_1), 53.7% (men; UR_2), and 31.7% (women; UR_2) of the CVs are below 20%. The use of the transformed FH model achieves a distinct increase of CVs below this threshold. As a result, 85.4% (men; UR_1), 73.2% (women; UR_1), 82.9% (men; UR_2), and 78.0% (women; UR_2) of the CVs are below 20%. This illustrates that the use of dynamic mobile network data in combination with SAE methodology is a powerful tool to increase the precision of both estimated unemployment rates for NRW on FUA sublevel. If we compare the direct estimates to the estimates from the proposed transformed FH model, both are often close to each other. For regions with smaller samples sizes like Witten and Paderborn, these values can deviate clearly from each other. Due to the higher uncertainty of the direct estimates for regions with lower sample sizes, the synthetic part within Eq. (3) is weighted higher and bigger differences to the direct estimates appear.

4.3 Discussion of the estimated unemployment rates for NRW

Figure 2 illustrates the differences between the alternative and traditional unemployment rates. If the traditional unemployment rates are the same as alternative one, the commuter behaviour is balanced and the calculated difference would be zero. Please note, that the FUA sublevels do not cover the entire federal territory in NRW, these areas are white in Fig. 2. The bluish colors indicate areas where the alternative unemployment rate is higher than the traditional one. Those are mainly the commuter zones in both models, i.e., the commuter flow is directed out of this area. With one exception in the female model, all commuting zones are coloured blue. This means that these areas are the place of residence of many employed people who commute from those areas to their workplace. The reddish areas, however, imply that the alternative unemployment rate is lower than the traditional unemployment rate. This is mainly the case for the city cores of the FUAs. This observation is consistent with Grözinger (2018) motivation for creating an alternative unemployment rate. Nevertheless, a negative value (blue colouring) was detected for a few city cores. This is the case for nine city cores simultaneously in both models. These are the city cores Recklinghausen, Bottrop, Moers, Oberhausen, Duisburg, and Mühlheim an der Ruhr. These six are located in the Ruhr region, which includes the large city cores Essen and Dortmund, to which many people commute from the Ruhr region. Furthermore, this trend was found for the two small city cores (Solingen and Sankt Augustin) and Aachen, which is located directly on the Belgian border. Since most city cores are job engines, many employed people living in the surrounding travel-to-work areas, which is their place of residence, commute into the city cores to work. In the males model, the differences are higher than in the females model, which leads to the conclusion that women are not commuting as often or as far as men (IT.NRW 2019). Possible reasons for this could be the conservative role model of women, the spatial closeness to the family that is guaranteed by the woman (to the school/kindergarten of the children, etc.) or, for example, a work in small, nearby companies/enterprises (Bauer-Hailer 2019).

5 Validity of the proposed method

In the following, we evaluate the methodology used in Sect. 4 to estimate unemployment rates at the FUA sublevel through official data. For Germany, the database Urban Audit provided by Eurostat in cooperation with Destatis and Kommunales Statistisches Informationssystem (KOSIS) is the only source for German unemployment rates at the FUA level (KOSIS-Gemeinschaft Urban Audit 2013; Eurostat 2017, 2019d). This official data source provides traditional unemployment rates, but no alternative unemployment rates for all German FUAs. Thus, the Urban Audit enables a comparison of traditional unemployment rates estimated by using the transformed FH estimator [Eq. (4)] with mobile network data as auxiliary information with the officially published values. As mentioned in Sect. 2.1, we have used the

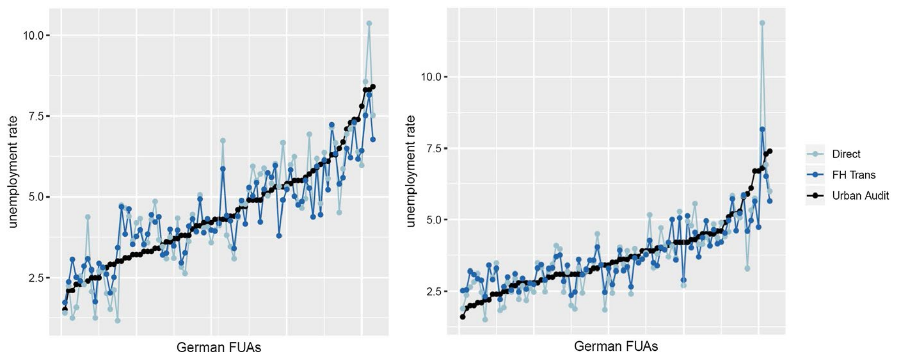

Fig. 3 Comparison of traditional unemployment rates (UR_1) published in Urban Audit (black), estimated with the transformed FH model (dark blue) and the direct estimates from the LFS (light blue) for men (left) and women (right) for all German FUAs (colour figure online)

Table 3 Distribution of the absolute difference to the Urban Audit estimates of the females and males traditional unemployment rates over all German FUAs and in particular over FUAs with small sample sizes below 600

Areas	Sex	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	Female	Direct	0.017	0.246	0.459	0.638	0.800	5.078
		FH Trans	0.005	0.173	0.415	0.512	0.748	1.959
	Male	Direct	0.009	0.202	0.625	0.713	0.998	2.440
		FH Trans	0.008	0.221	0.428	0.573	0.824	1.690
Sample size < 600	Female	Direct	0.030	0.416	0.628	0.930	1.120	5.078
		FH Trans	0.015	0.281	0.516	0.627	0.896	1.959
	Male	Direct	0.068	0.697	1.095	1.129	1.764	2.073
		FH Trans	0.038	0.373	0.676	0.704	1.027	1.690

15–64 age range for the definitions of unemployment rates to ensure comparability with the Urban Audit. Please note, the comparison in this section is made on the entire FUA level and not on the FUA sublevel as in the application in Sect. 4.

For the German federal state NRW, we have an extensive mobile network data record available as auxiliary information. However, we have only limited access to mobile network data and accordingly a data set with less information for the rest of the country. Thus, less covariates are available for the validation. In contrast to Sect. 4, where we use dynamic signalling data, we only have static mobile network activities of a typical Sunday evening for the whole of Germany. We focus on the time period from 8 to 11 pm of the average of eight Sundays of the months April, June, and July in 2018 without school or public holidays. For Sunday evenings, a high correlation has been identified between population figures from the 2011 census and the mobile network activities on the weekend and especially on Sunday evening (Hadam 2018). As traditional unemployment rates are based on the place of residence, it is reasonable to assume that mobile network data of a Sunday evening is suitable as auxiliary variables. In the following, we validate the proposed transformed FH model by comparing the FH estimates with official unemployment rates of the Urban Audit. We use the SAE method and model selection as applied in Sect. 4 with the difference that (a) the regional focus is now FUAs across Germany and (b) we can only use mobile network data from Sunday evening. In the males model, the selected mobile network covariates explain around 47% of the variance in terms of the modified R^2 following Lahiri and Suntornchost (2015) and in the females model around 37%.

For the validation of the proposed method, Fig. 3 shows the estimated unemployment rates using mobile network covariates (FH Trans), the direct, and the published official estimates from Urban Audit by sex. First, it can be seen that we get similar rates compared to the Urban Audit by using the transformed FH model. Comparing the direct estimator from the LFS with the FH Trans estimator, the FH Trans estimator corrects the direct estimator in such a way that the resulting value is closer

Table 4 Distribution of important parameters in the simulation setting: the sampling error variation σ_{e_i} and the resulting shrinkage factor γ_i coincide with the male model for Germany on FUA sublevel

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
σ_{e_i}		0.0063	0.0202	0.0275	0.0288	0.0366	0.0785
γ_i		0.1199	0.3848	0.5265	0.5355	0.6730	0.9548
$\hat{\theta}_i^{\text{direct}}$	Sim.	0.0000	0.0340	0.0495	0.0538	0.0688	0.2826
	FUA sublevel	0.0054	0.0328	0.0484	0.0508	0.0647	0.1134

The direct estimates $\left(\hat{\theta}_i^{\text{direct}}\right)$ of the simulation study are close to the values for the FUA sublevel

to the Urban Audit. This trend is quantified in Table 3. It reports the distribution of the absolute difference of the females and males unemployment rates obtained by the two estimation methods for all FUAs in Germany compared to the Urban Audit. For almost all distribution values, we get a higher absolute difference for the direct estimates compared to the FH Trans estimates. Only in the males model the 25% quantile for the absolute difference is slightly higher for the FH Trans estimates. As expected, it can be noted that for FUAs with sample size under 600 estimated unemployment rates of both estimation methods show higher values for the absolute difference.

6 Model-based simulation

In the previous two sections, we use the proposed transformed FH model to estimate alternative unemployment rates and subsequently evaluate the suggested methodology with official statistics obtained from Urban Audit. This model-based simulation study is used to investigate how much we benefit from the more complicated transformed FH model with a bias corrected back-transformation compared to the naive back-transformation. According to the Jensen-inequality (cf. Section 3.2), the naive back-transformation is biased under the inverse sine transformation. Furthermore, we want to show,

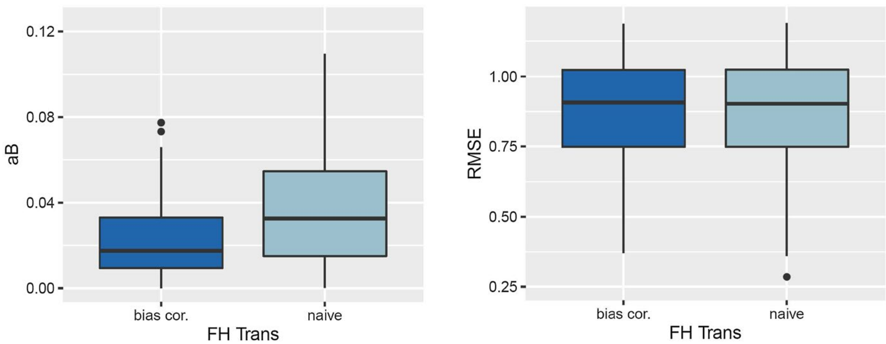

Fig. 4 Distribution of the aB and the RMSE for the transformed FH estimator with bias corrected and naive back-transformation

that the proposed MSE and confidence intervals lead to reasonable results. We investigate these aims in a close to reality environment. The input values of the model-based setting are based on the real data.

The simulation study is implemented with $R = 1000$ Monte-Carlo replications. Within each replication, we generate the covariates ($\mathbf{x}_i$) initially from a lognormal distribution with parameters $(-0.5, 0.04)$. The number of areas is fixed to the number of the FUA sublevels in Germany ($d = 208$). We draw the random effect and the sampling errors from normal distributions: $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and $e_i \sim \mathcal{N}(0, \sigma_{e_i}^2)$. According to the males model for Germany on FUA sublevel, $\sigma_u \approx 0.029$ is defined analogously. In addition, we adopt the variation of the sampling errors σ_{e_i} and keep them constant over the replications. The regression coefficients are set to $\beta_0 = 0.01$ and $\beta_1 = 0.35$. As data generating process, we consider $\hat{\theta}_i^{\text{direct}} = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i + e_i\right)$ to get synthetic direct estimates. The true small area means are $\bar{y}_i = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i\right)$. Table 4 shows the distribution of the variation of the sampling errors and the resulting shrinkage factor as well as the distribution of the direct estimates for the simulation (over all replications) and the actual direct estimated unemployment rates for males in Germany. The distributions are close to each other.

For each replication, we estimate small area means from the transformed FH model: with respect to the back-transformation bias [$\hat{\theta}_i^{\text{FH,trans}}$, cf. Equation (5)] and with naive back-transformation $\left(\hat{\theta}_i^{\text{FH,naive}}\right)$. To assess the quality of the estimates, we obtain for $R = 1000$ Monte Carlo replications the absolute Bias (aB) and the root mean squared error (RMSE) of the estimates, defined as

$$\text{aB}_i = \left|\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)}\right)\right| * 100$$

$$\text{and} \quad \text{RMSE}_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)}\right)^2} * 100,$$

where $\hat{\theta}_i^{\text{FH},(r)}$ is the estimated respective FH value and $\bar{y}_i^{(r)}$ the true value within replication r. Figure 4 shows the reduction of aB. For instance, the median of the aB using a naive back-transformation is 1.86 times higher than with a bias corrected back-transformation. At the same time, we observe nearly the same RMSE (cf. Figure 4) when we use a bias corrected back-transformation instead of a naive

Table 5 Distribution of the quality measurements for the estimated RMSE and the corresponding confidence intervals using the bootstrap procedure as described in Sect. 3.3

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
rB RMSE	−9.34	−2.12	−0.62	−0.55	1.11	7.13
rRMSE RMSE	17.04	18.03	18.53	18.74	18.99	44.97
Coverage	86.70	93.90	94.40	94.34	94.90	96.00

back-transformation. In summary, there is a clear reduction in bias at the cost of a slightly higher RMSE.

We next investigate the properties of the proposed MSE and the confidence intervals. Please note that we compare the bootstrap estimated RMSE [Eq. (6)] to the empirical RMSE, which we treat as the true one. For calculating these uncertainty measurements, we use 1000 bootstrap replications within each Monte Carlo run. As quality measurements, we calculate the relative bias of the uncertainty estimation (rB RMSE) and the relative RMSE of the uncertainty estimation (rRMSE RMSE). They are defined as

$$\text{rB RMSE}_i = \left(\frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \text{MSE}_{\text{est},i}^{(r)}} - \text{RMSE}_{\text{true},i}}{\text{RMSE}_{\text{true},i}} \right) * 100$$

$$\text{and} \quad \text{rRMSE RMSE}_i = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(\text{RMSE}_{\text{est},i}^{(r)} - \text{RMSE}_{\text{true},i} \right)^2}}{\text{RMSE}_{\text{true},i}} * 100,$$

where $\text{RMSE}_{\text{est},i}^{(r)}$ is the estimated RMSE out of the bootstrap procedure (cf. Section 3.3) for each Monte Carlo replication r and $\text{RMSE}_{\text{true},i}$ is the empirical RMSE over the Monte Carlo replications. The relative bias is close to zero as Table 5 shows. On average, we get an underestimation of 0.55% over all areas. The interquartile range goes from −2.12% to 1.11%. In addition, the relative RMSE of the estimated RMSE is important to assess its quality. We get a mean relative RMSE of 18.74% for the estimated RMSE. The low bias and the RMSE show that the proposed MSE estimator yields good results. In addition to the MSE, we can also get bootstrap confidence intervals (cf. Section 3.3). The coverage is defined as the proportion of the time that the estimated confidence interval contains the true value. For the proposed confidence intervals [Eq. (7)], we get in mean a coverage of 94.34%. We can recognize a slight underestimation of the coverage, but the values are close to the target value of 95%. These three measures show that the proposed bootstrap-estimated MSE works.

Overall, our close to reality simulation study shows the reduction of bias while using the transformed FH estimator with bias corrected back-transformation instead of a naive back-transformation. Furthermore it demonstrate the good performance of the newly proposed MSE estimator and confidence intervals for the transformed FH estimator with bias corrected back-transformation.

7 Concluding remarks

The traditional unemployment rate is based on the place of residence of the labour force. Due to the high level of commuting, this may give a distorted impression of regional labour markets. For Germany, traditional unemployment rates show higher rates in city cores compared to its surroundings. For analysing unemployment rates in the context of commuter behaviour, the regional target area are city cores and their commuting zones, which can be extracted from FUAs. In this work, we estimate an alternative unemployment rate, where the focal point of the labour force is their workplace. It adjusts the traditional definition by including commuters. Since the LFS is not designed to produce indicators on smaller areas than NUTS 2-level, a FH approach is used to estimate alternative and traditional unemployment rates on the FUA sublevel. From a methodological point of view we use a bias corrected back-transformed FH estimator and propose a MSE estimator to measure its uncertainty. As the FH approach relies on a model-based method, suitable covariates are required. We select covariates constructed from dynamic mobile network data and validate the selected models. The benefit of dynamic mobile network data is that they represent the changes of the counted aggregated mobile devices during the day and in space. This information can be used to derive the commuting behaviour of the population. The resulting differences between the traditional and the alternative unemployment rates show that the rates in city cores are mainly lower than officially indicated. The assumption that unemployment rates in city cores are lower can be confirmed and thus contributes to the explanation why so many people move to city cores due to more job opportunities. Furthermore, the alternative definition of the unemployment rate removes the static picture of the population, especially of the labour force. The labour force does not necessarily live in the same place where they work. This dynamic cannot be achieved with traditional survey methods and with traditional data. However, exactly this knowledge is necessary to make better decisions regarding urban planning. Moreover, these alternative rates provide potential employers with additional information about the current regional labour market and on missing workplaces. This will help to identify regions for which it might be useful to promote business settlement in order to reduce unemployment rates and shorten commuting distances, as new details of potentially available local workforce are available. The increasing number of commuters should be taken into account in official statistics in the future. Although the application in this paper refers to NRW, the model is also applicable to countries that perform the LFS and have implemented an FUA structure. Thus, this analysis is transferable to at least all European countries. In Germany, we are facing some limitations in mobile network data. We do not have access to individual signalling data or CDRs. No individual activity movements or changes in individual social behaviour can be used for the estimation. For instance, Toole et al. (2015) have shown that unemployed persons have different mobile phone usage profiles than employed ones. This information may increase the explanatory power in estimating unemployment rates compared to the used distribution of mobile activities over time.

From a methodological point of view, we leave the uncertainty of the difference between the two unemployment rates as further research. So far, we propose an MSE and confidence intervals for each unemployment rate separately. To obtain these two measures for the difference, it is necessary to calculate the covariance between both unemployment rates. For the special case of the difference between a design-based estimator and a FH estimator from the same repeated survey at different points in time, van den Brakel et al. (2016) derives the covariance. It is assumed that the design-based estimator is unbiased and that the covariates for the FH estimator come from the same survey as the design-based estimator. Since these assumptions are not applicable to our case, further research is needed to apply these results to the present case.

In addition, the following research opportunities remain open from an applied perspective. Steele et al. (2017) uses a combination of satellite and mobile phone data to gain more explanatory power in the estimation of poverty indicators. Satellite data include valuable information on a small regional level of building intensities and heights of buildings to differentiate between socially impoverished people, who live in socially weak urban districts, and wealthy people, who are living more likely in less densely populated areas, which could also be suitable for our question. Furthermore, it is of interest to which extend the same differences in unemployment rates also apply to other countries or whether it is a national phenomenon.

Appendix A: Mobile network covariates

See Appendix Table 6.

Table 6 Mobile network covariates: the last four columns refer to the four different models on unemployment rates at the FUA sublevel

Definition of variables		UR_1 male	UR_2 male	UR_1 fem.	UR_2 fem.
Intercept		118.5287	14.8136	0.5674	− 14.3924
Proportion of mobile activities of specific subgroup at defined time					
Central European	7 am to 4 pm	−2.6305	−2.7364	−2.2433	
Central European	5 pm to 11 pm	3.1693	4.0103		
Proportion of mobile activities of specific subgroup at defined time on Sunday					
under 50 s	8 pm to 11 pm		−0.1478	−0.6384	
20 to 30 year olds	8 pm to 11 pm			0.8168	
Change of mobile activities by nationality from night-time (5 pm to 11 pm) to day-time (7 am to 4 pm)					
African		0.0012	0.0013	− 0.0034	− 0.0024
Australia Oceania		− 0.0001	− 0.0001		
Eastern Europe		− 0.0680	− 0.0836		
North American		− 0.0245	− 0.0695	− 0.0273	− 0.0446
Northern Europe		− 0.0176	− 0.0449		
Southeast Europe		− 0.1070	− 0.1654	− 0.1145	− 0.1165
Southern Europe			0.1158	0.1021	
Asia					0.0135
Central Europe					− 5.2348
Relative change of mobile activities between two specific times: (time point 1 – time point 2) / time point 2					
10 am	9 pm	− 3.2224	4.7858		2.4082
8 pm to 10 pm	9 am to 11 am	3.2963	− 3.7265		
4 pm	10 am		− 1.2954	− 1.1901	
9 am to 11 am	3 am to 5 am			2.4185	
Ratio of mobile activities between two specific times: time point 1 / time point 2					
7 am to 4 pm	5 pm to 11 pm	3.0494		5.2589	
5 pm to 5 am	whole day	− 119.1781	− 28.0838		
9 am to 11 am	8 pm to 10 pm	− 3.9171	− 3.5458	− 5.5749	
6 am to 4 pm	whole day	− 111.7206			30.2304
12 pm to 6 am	7 am to 4 pm		2.7691		
3 am to 5 am	9 am to 11 am			2.0348	

The covariates are based on mobile network data of Deutsche Telekom for the years 2017 and 2018 and represent a statistical week. For each selected variable, the regression coefficient is shown

Appendix B: Map of FUA city cores and commuter zones in NRW

See Appendix Fig. 5.

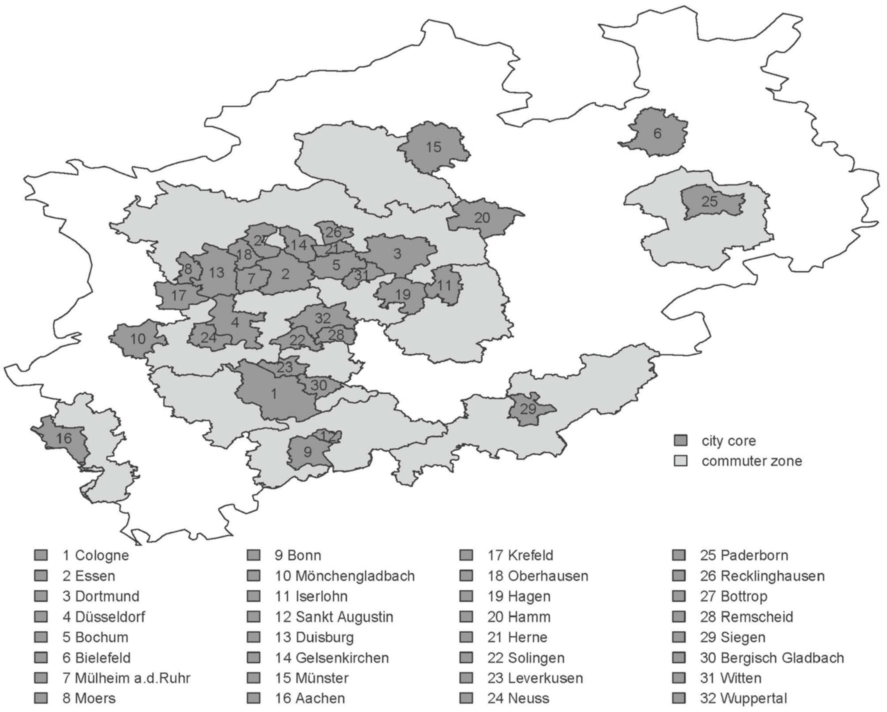

Fig. 5 Assignment of city names to FUA city cores and geographical location of the commuter zones for NRW

Acknowledgements Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin. Finally, the authors are indebted to the Joint Editor, Associate Editor and two referees for comments that significantly improved the paper. Open Access funding enabled and organized by Projekt DEAL.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Augustijn L (2018) Berufsbedingte Pendelmobilität, Geschlecht und Stress. https://www.uni-due.de/imperia/md/content/soziologie/dbsf-2018-02.pdf. Accessed 01 2021

Bauer-Hailer U (2019) Berufspendler im Bundesländervergleich. https://www.statistik-bw.de/Service/Veroeff/Monatshefte/PDF/Beitrag19_02_02.pdf. Accessed 12 2019
BBSR (2017) Wachsen und Schrumpfen von Städten und Gemeinden. https://gis.uba.de/maps/resources/apps/bbsr/index.html?lang=de. Accessed 06 2019
Buch T, Hamann S, Niebuhr A, Rossen A (2014) What makes cities attractive? The determinants of urban labour migration in Germany. Urban Stud. 51(9):1960–1978
Bundesagentur für Arbeit (2022a) Arbeitslosenquote und Unterbeschäftigungsquote. https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Definitionen/Berechnung-der-Arbeitslosenquote/Berechnung-der-Arbeitslosenquote-Nav.html. Accessed 10 2022
Bundesagentur für Arbeit (2022b) Pendlerverflechtungen der sozialversicherungspflichtig Beschäftigten nach Ländern—Stichtag: 30.06.2016. https://statistik.arbeitsagentur.de/SiteGlobals/Forms/Suche/Einzelheftsuche_Formular.html?nn=20934 &topic_f=beschaeftigung-pendler-blxbl &dateOfRevision=201606-202106. Accessed 10 2022
Burgard JP, Münnich R, Zimmermann T (2016) Impact of sampling designs in small area estimation with applications to poverty measurement. In: Pratesi M (ed) Analysis of poverty data by small area estimation. Wiley, Hoboken, pp 85–108
Casas-Cordero C, Encina J, Lahiri P (2016) Poverty mapping for the Chilean comunas. In: Pratesi M (ed) Analysis of poverty data by small area estimation. Wiley, Hoboken, pp 379–403
Costa A, Satorra A, Ventura E (2006) Improving small area estimation by combining surveys: new perspectives in regional statistics. Stat Oper Res Trans 30(1):101–121
Datta GS, Lahiri P (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Stat Sin 10(2):613–627
De Meersman F, Seynaeve G, Debusschere M, Lusyne P, Dewitte P, et al (2016) Assessing the quality of mobile phone data as a source of statistics. European Conference on Quality in Official Statistics, Spanien, Madrid
Dijkstra L, Poelman H (2011) Archive: European cities—the EU-OECD functional urban area definition. https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:European_cities_%E2%80%93_the_EU-OECD_functional_urban_area_definition#A_harmonised_definition. Accessed 06 2019
Elhorst PJ (2003) The mystery of regional unemployment differentials: theoretical and empirical explanation. J Econ Surv 17(5):709–748
European Parliament and Council (2003) Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). Offic J Eur Union 154(1):1–41
Eurostat (2017) City statistics (URB): national reference metadata in euro SDMX metadata structure (ESMS). https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm. Accessed 11 2018
Eurostat (2018a) Dataset details: harmonised unemployment rate by sex. https://ec.europa.eu/eurostat/web/products-datasets/-/teilm020. Accessed 10 2018
Eurostat (2018b) Glossary:Functional urban area. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area. Accessed 05 2018
Eurostat (2018c) NUTS—systematik der Gebietseinheiten für die Statistik: Hintergrund. https://ec.europa.eu/eurostat/web/nuts/background. Accessed 08 2019
Eurostat (2019a) DataCollection: precision level DCF. https://datacollection.jrc.ec.europa.eu/wordef/precision-level-dcf. Accessed 06 2019
Eurostat (2019b) EU Labour Force Survey database user guide. https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf. Accessed 08 2019
Eurostat (2019c) Labour force survey in the EU, candidate and EFTA countries—main characterstics of national surveys. https://ec.europa.eu/eurostat/de/web/products-statistical-reports/-/KS-FT-19-008?inheritRedirect=true. Accessed 01 2021
Eurostat (2019d) Städte (Urban Audit): Datenbank. https://ec.europa.eu/eurostat/de/web/cities/data/database. Accessed 08 2019
Fay RE, Herriot RA (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74(366):269–277
Franconi L, Ichim D, D'Alò M, Cruciani S (2017) Guidelines for Labour Market Area delineation process: from definition to dissemination. https://ec.europa.eu/eurostat/cros/system/files/guidelines_for_lmas_production08082017_rev300817.pdf. Accessed 01 2021

Galiana L, Sakarovitch B, Smoreda Z (2018) Understanding socio-spatial segregation in French cities with mobile phone data. https://dgins2018.statisticsevents.ro/wp-content/uploads/2018/10/08-FR-dgins_segregation_1800905.pdf. Accessed 06 2022
Gans P (2017) Urban population development in Germany (2000–2014): the contribution of migration by age and citizenship to reurbanisation. Comp Popul Stud 42:319–352
González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008) Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. Comput. Stat. Data Anal. 52(12):5242–5252
Grözinger G (2009) Achtung lebensgefahr! indirekte effekte regionaler arbeitslosigkeit auf lebensweise und -qualität. Eur J Econ Econ Polic: Interv 6(1):12–24
Grözinger G (2018) Regionale arbeitslosigkeit: falsche eindrücke von stadt-land-differenzen. Wirtschaftsdienst 98(1):68–70
Hadam S (2018) Use of mobile phone data for official statistics. Methods—Approaches—Dev: Inf Ger Federal Stat Off 1(2):6–9
Hadam S (2021) Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik. AStA Wirtsch- Sozialstat Arch 15(3):197–235
ILO (2018) Unemployment rate. https://ilostat.ilo.org/resources/concepts-and-definitions/description-unemployment-rate/. Accessed 06 2022
IT.NRW (2019) Berufspendler 2011–2018 nach Pendlerart, Beschäftigungsumfang und Geschlecht. https://www.it.nrw/statistik/eckdaten/berufspendler-2011-2018-nach-pendlerart-beschaeftigungsumfang-und-geschlecht. Accessed 12 2019
Jensen JLWV et al (1906) Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math 30:175–193
Jiang J, Lahiri P, Wan S-M, Wu C-H (2001) Jackknifing in the Fay–Herriot model with an example. In: Proceedings of the seminar on funding opportunity in survey research, Bureau of Labor Statistics, pp 75–97
Klammer U, Menke K (2020) Gender-Datenreport. https://www.bpb.de/izpb/307413/geschlechterdemokratie. Accessed 01 2021
Kosfeld R, Dreger C (2006) Thresholds for employment and unemployment. A spatial analysis of German regional labour markets 1999–2000. Pap Reg Sci 85:523–542
KOSIS-Gemeinschaft Urban Audit (2013) Das deutsche urban audit—Städtevergleich im Europäischen statistischen system. https://ec.europa.eu/eurostat/cache/metadata/Annexes/urb_esms_de_an4.pdf. Accessed 06 2022
Lahiri P, Suntornchost J (2015) Variable selection for linear mixed models with applications in small area estimation. Sankhya B—Indian J Stat 77(2):312–320
Lumley T (2004) Analysis of complex survey samples. J Stat Softw 9(1):1–19
Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Gabrielli L (2015) Small area model-based estimators using big data sources. J Off Stat 31(2):263–281
Marino MF, Ranalli MG, Salvati N, Alfò M (2019) Semiparametric empirical best prediction for small area estimation of unemployment indicators. Ann Appl Stat 13(2):1166–1197
Martini A, Loriga S (2017) Small area estimation of employment and unemployment for local Labour Market Areas in Italy. https://www.dst.dk/ext/4299453410/0/formid/4-2-Small-Area-Estimation-of-employment-and-unemployment-for-Local_Labour-Market-Areas-in-Italy--pdf. Accessed 11 2018
Möbert J (2018) The German housing market in 2018. https://www.dbresearch.com/PROD/RPS_EN-PROD/PROD0000000000460528/The_German_housing_market_in_2018.pdf. Accessed 06 2022
Molina I, Strzalkowska-Kominiak E (2020) Estimation of proportions in small areas: application to the labour force using the Swiss census structural survey. J R Stat Soc A Stat Soc 183(1):281–310
Patuelli R, Griffitha DA, Tiefelsdorf M, Nijkamp P (2011) Spatial filtering and eigenvector stability: space-time models for German unemployment data. Int Reg Sci Rev 34(2):253–280
Pereira LN, Mendes J, Coelho PS (2011) Estimation of unemployment rates in small areas of Portugal: a best linear unbiased prediction approach versus a hierarchical Bayes approach. In: 17th European young statisticians meeting
Prasad NGN, Rao JNK (1990) The estimation of the mean squared error of small-area estimators. J Am Stat Assoc 85(409):163–171
Raghunathan TE, Xie D, Schenker N, Parsons VL, Davis WW, Dodd KW, Feuer EJ (2007) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. J Am Stat Assoc 102(478):474–486
Rao JNK, Molina I (2015) Small area estimation, 2nd edn. Wiley, Hoboken

R Core Team (2022) R: a language and environment for statistical computing [Computer software manual], Vienna. https://www.R-project.org
Schmid T, Bruckschen F, Salvati N, Zbiranski T (2017) Constructing sociodemographic indicators for national statistical institutes using mobile phone data: estimating literacy rates in Senegal. J R Stat Soc A Stat Soc 180(4):1163–1190
Slud EV, Maiti T (2006) Mean-squared error estimation in transformed Fay–Herriot models. J R Stat Soc: Ser B (Stat Methodol) 68(2):239–257
Statistisches Bundesamt (2021) Registered unemployed, unemployment rate by sex. https://www.destatis.de/EN/Themes/Labour/Labour-Market/Unemployment/Tables/lrarb002.html. Accessed 01 2021
Statistisches Bundesamt (2022) Equipment of households with information and communication technology (Germany). https://www.destatis.de/EN/Themes/Society-Environment/Income-Consumption-Living-Conditions/Equipment-Consumer-Durables/Tables/liste-equipment-households-information-communication-technology-germany.html#55714. Accessed 10 2022
Steele JE, Sundsøy PR, Pezzulo C, Alegana VA, Bird TJ, Blumenstock J, Bengtsson L (2017) Mapping poverty using mobile phone and satellite data mapping poverty using mobile phone and satellite data. J R Soc Interface 14(127):20160690
Sugasawa S, Kubokawa T (2017) Transforming response values in small area prediction. Comput Stat Data Anal 114:47–60
Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D (2015) Tracking employment shocks using mobile phone data. J R Soc Interface 12(107):20150185
Tzavidis N, Zhang L-C, Luna A, Schmid T, Rojas-Perilla N (2018) From start to finish: a framework for the production of small area official statistics. J R Stat Soc A Stat Soc 181(4):927–979
U.S. Bureau of Labor Statistics (2021) Labor force statistics from the current population survey: concepts and definitions. https://www.bls.gov/cps/definitions.htm#:~:text=The%20unemployment%20rate%20represents%20the,%C3%B7%20Labor%20Force)%20x%20100. Accessed 10 2022
van den Brakel JA, Buelens B, Boonstra H-J (2016) Small area estimation to quantify discontinuities in repeated sample surveys. J R Stat Soc A Stat Soc 179(1):229–250

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sandra Hadam[1] · Nora Würz[2] · Ann-Kristin Kreutzmann[3] · Timo Schmid[2]

✉ Nora Würz
nora.wuerz@uni-bamberg.de

Sandra Hadam
sandra.hadam@destatis.de

Ann-Kristin Kreutzmann
ann-kristin.kreutzmann@fu-berlin.de

Timo Schmid
timo.schmid@uni-bamberg.de

[1] Federal Statistical Office of Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Hesse, Germany

[2] Institute of Statistics, Otto-Friedrich-Universität Bamberg, Feldkirchenstraße 21, 96045 Bamberg, Bavaria, Germany

[3] Institute of Statistics and Econometrics, Freie Universität Berlin, Garystr. 21, 14195 Berlin, Berlin, Germany