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

Sandra Hadam*, Nora Würz**, and Ann-Kristin Kreutzmann**

*Federal Statistical Office of Germany, Wiesbaden, Germany
**Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany

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 (LFS) 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 (MSE) 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: Commuting zones, Fay-Herriot model, Signalling data, Small area estimation, Unemployment rates

1 Motivation

Since jobs in Germany are predominantly located in cities, more people move to the cities (see e.g. an interactive map on growing and shrinking German cities and communities (Federal Institute for Research on Building, Urban Affairs and Spatial Development, 2017)). In Germany, the continuous growth of cities is creating shortages on the housing and real estate markets (Möbert, 2018). Most large cities have higher population growth rates than the national average. Due to the comparatively high urban labour migration, the number of people living in cities is steadily increasing nationwide. As Buch *et al.* (2014) showed, smaller cities recorded less net immigration than large cities, which in turn is characterized by the attractiveness of larger cities and the benefits of living in them. This is mainly due to the urbanization advantages, which are reflected in better infrastructure, more extensive education and work opportunities, a more 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. 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. Thus, this definition provides valuable information on missing workplaces in regional areas and therefore support policy decisions in urban planning. Thereby policymakers can identify regions where it might be useful to promote the settlement of companies in order to lower their unemployment rate and shorten commuter movements. For cities, lower alternative than traditional unemployment rates are assumed. Low alternative unemployment rates contribute to the attractiveness of cities and the moving and commuting behaviour towards these urban areas. Grözinger (2018) investigates this difference, among others, for regional areas in the German federal states Bavaria and Schleswig-Holstein.

For analysing unemployment rates in the context of commuter behaviour, the regional target areas are city cores and their commuting zones. For European countries and other member countries of the Organisation for Economic Co-operation and Development (OECD), Functional Urban Areas (FUA) have been created as harmonised geometries describing urban areas (Dijkstra and Poelman, 2011). A FUA is composed by a city core and its commuter zone. However, we are interested in the spatial level below, which considers the city core and commuter zone separately. We therefore refer to our regional target level in the following as the FUA sublevel. This regional level is available for all OECD countries. Thus, our comparison of unemployment rates presented here for the German federal state of North Rhine-Westphalia (NRW) is transferable to other OECD countries. Using the FUA sublevel as regional layer, the observation area is divided into city cores and commuting zones and does not consider remaining regions, as the FUA sublevel does not cover the whole territory of a country. This spatial level is particularly suitable for comparing both unemployment rates, which differ in whether commuters are included or not.

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 district level, which is a higher regional level than the FUA sublevel (Eurostat, 2019b). Hence, estimates on this spatially finer sublevel that are only based on survey data (called direct estimates throughout the paper) 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) or 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 small area 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 on unit level as auxiliary information. For other studies, appropriate register or administrative data is not always available. Especially, the access to unit-level data is mostly not possible due to data protection regulations. In the case of suitable aggregated data, the problem often occurs that these data are only available on more aggregated spatial areas than the desired target level. Due to a lack of finer spatial register or administrative data, one possibility is to use alternative data sources as covariates. Toole *et al.* (2015) or Steele *et al.* (2017) propose the usage of passively collected mobile phone data as auxiliary information. The advantages of mobile phone data or mobile network data are their finer spatial resolution and their timeliness, as they are available in real time. 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. 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 socioeconomic phenomena.

Our analyses are based on dynamic mobile network data. It is advantageous that this data source reflects commuting behaviour as well as daytime and resident population. 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 our regional target area, the FUA sublevel. The aim is to compare both definitions of unemployment rates at the level of interest, thus highlighting the influence of commuters. From a methodological perspective, we consider the Fay-Herriot (FH) model (Fay and Herriot, 1979) using mobile network data as auxiliary information. In general, the FH model produces estimates on a continuous range. However, the unemployment rate should be within the interval $[0, 1]$. One possible solution is to use a transformation. In particular, we transform the dependent variable with an inverse sine transformation following Casas-Cordero *et al.* (2016), Burgard *et al.* (2016), and Schmid *et al.* (2017). They all 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). To measure the uncertainty of the bias corrected back-transformed 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: Section 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 Section 4 to estimate the difference of both unemployment rates for the German federal state NRW on FUA sublevel. For evaluation reasons, Section 5 investigates the methodology on German data for estimating the traditional unemployment rates and compares the results with official data. Furthermore, in Section 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, and the FUA sublevel as our geographical unit of interest for the analysis of these indicators (Subsection 2.1). Subsequently, our two data sources are described (Subsection 2.2 and 2.3) in particular, the LFS survey data and mobile network data. Subsection 2.3 emphasizes the advantages of dynamic mobile network data as an alternative data source to create covariates.

2.1 Traditional and alternative definition of unemployment rates

The Federal Statistical Office of Germany (Destatis) publishes traditional unemployment rates according to the definition of the International Labour Organization (ILO), which provides an international comparable indicator (International Labour Organization, 2018). Following the ILO-definition, the unemployment rate is given by

$$\theta_{\mathrm{UR}_1,i} = \frac{\mathrm{N}_{i,\text{unemployed}}}{\mathrm{N}_{i,\text{unemployed}} + \mathrm{N}_{i,\text{ employed counted at residency}}}, \tag{1}$$

where $\theta_{\mathrm{UR}_1,i}$ is the traditional unemployment rate for the regional area i. This unemployment rate is defined by the number of unemployed persons living in area i ($\mathrm{N}_{i,\text{ unemployed}}$) 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 ($\mathrm{N}_{i,\text{ unemployed}} + \mathrm{N}_{i,\text{employed counted at residency}}$). 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 (International Labour Organization, 2018; Eurostat, 2018a). Please note that for reasons of comparability we use the age range of 15-64 years. Therefore, the traditional unemployment rates are estimated with a modified age group throughout the analysis. For traditional unemployment rates, employed persons are counted at area i where they live. However, there is another possible reference point for employed persons: Employed persons can be counted in area i where their workplace is located. Unemployed persons are counted exclusively in area i where they live in because they have no place of work. Thus, the second definition proposed by Grözinger (2018) uses the workplace of employed persons as a focal point and thus the definition changes to

$$\theta_{\mathrm{UR}_2,i} = \frac{\mathrm{N}_{i,\text{ unemployed}}}{\mathrm{N}_{i,\text{ unemployed}} + \mathrm{N}_{i,\text{ employed counted at workplace}}}. \tag{2}$$

$\theta_{\mathrm{UR}_2,i}$ is the alternative definition of unemployment rates for the regional area i. In (2), the number of unemployed persons ($\mathrm{N}_{i,\text{unemployed}}$) is divided by the labour force aged 15 to 64. In contrast to traditional unemployment rates, the labour force of alternative unemployment ratse is composed of the number of unemployed persons living in area i and the number of employed persons having their workplace in area i ($\mathrm{N}_{i,\text{ unemployed}} + \mathrm{N}_{i,\text{employed counted at workplace}}$).

In comparison, both unemployment rates treat commuting behaviour differently. A traditional unemployment rate of 5% would mean that in an area i with a labour force of 100 persons 5 persons are unemployed. If in contrast the alternative unemployment rate of the same area i is considered, then, according to Grözinger (2018), the commuter flows of the labour force are taken into account and consequently, the unemployment rate of area i should change. Based on the scenario above, the following changes occur: 60 people work and live in the same area i, 35 people commute to area i where they work and 25 persons commute from area i to other areas for work. This changes the number of labour force at the place of work from 100 to 110 persons due to the inclusion of commuters and leads to a lower alternative unemployment rate of 4.55%. Therefore, the difference of both unemployment rates reveals the influence of commuters. If $\theta_{\mathrm{UR}_1,i}$ is higher than $\theta_{\mathrm{UR}_2,i}$, then there is a stronger commuter movement from other areas to area i.

FUAs are in general composed of city cores and their commuting zones. In this study, however, we are interested in a separate consideration of city core and its commuting zone. In the following, we will refer to this spatial level as FUA sublevel. This regional level 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 centres and their commuter zones. Therefore, these analyses are transferable to other OECD countries. 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 (Eurostat, 2018b). 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 Subsection 2.1. The LFS is a household survey conducted in 35 countries including all 27 EU member states and the United Kingdom, 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.

In the used LFS data, regional disaggregation is carried out using the administrative nomenclature of territorial units for statistics (NUTS) classification. These NUTS-levels are harmonised geometries in Europe to compare specific areas with each other and get geographically finer from NUTS 1-level to NUTS 3-level (Eurostat, 2018c). In Germany, the NUTS 1-level corresponds to the 16 federal states, 38 areas are related to governmental regions (NUTS 2-level) and the smallest unit of this nomenclature, the NUTS 3-level, refers to the 401 administrative districts (Europäisches Parlament und Rat der Europäischen Union, 2003). The next finer resolution are the 11 059 municipalities. Unemployment rates using LFS data are published on governmental regions level (NUTS 2-level). However, our target level is the smaller FUA sublevel. In Germany, the FUA sublevel can be composed directly from the finer geometry of the municipalities or by using the administrative districts (NUTS 3-level) and the administratively independent cities to determine the sublevel. As all LFS observations contain information about the associated administrative districts (NUTS 2-level) and municipalities, we can use the individual information of the LFS participants to allocate them to the FUA sublevel. Thus, we have to match a) the place of residence and b) the place of work to the corresponding sublevel. For the assignment of individuals to the place of work, we use a proportional allocation in a few cases.

In this work, we consider the year 2016 with an overall sample size of 369 986 observations. Since we evaluate the FUA sublevel, the sample size decreases to 271 587 observations. As unemployment rates are different for both sexes, the following analyses are carried out separately by sex. 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 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, we have problems with the reliability of the direct estimators in this application. In particular, Eurostat considers estimators with a Coefficient of Variation (CV) less than 20% to be reliable (Eurostat, 2019a). In Section 4 we will show, that the direct estimators exceed this threshold in many cases. Therefore, we apply SAE methods and we will discuss the reliability, measured

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	1 162	2 916	4 104	4 623	5 521	10 684
	Male	1 318	3 304	4 565	5 114	6 108	11 675
FUA sublevel	Female	100	216	368	635	646	7 973
	Male	97	244	421	702	749	8 559

by the CV, for the direct and the model-based estimates on the FUA sublevel. Since SAE methods require auxiliary variables from other data sources, such as census, register or mobile network data, the auxiliary information used here is described in more detail in the following.

2.3 Mobile network data

To estimate unemployment rates for FUA sublevel using SAE methods, suitable auxiliary information is needed. Many SAE applications are based on register or survey data, which seem to be suitable for most applications due to the extensive information available. Besides numerous possibilities of use, these data sources also have some disadvantages, such as the missing timeliness or the higher regional aggregation. Alternative data sources have the potential to overcome these disadvantages. Toole *et al.* (2015) or Steele *et al.* (2017), for example, have shown that mobile phone data are a promising alternative data source in the context of SAE. Mobile network data are also explored to estimate daytime population, commuter patterns or tourism behaviour (see e.g., De Meersman *et al.*, 2016; Galiana *et al.*, 2018). The major advantage of these data are the availability in real-time and their finer spatial resolution, which means that one can get the latest mobile network data, or more exactly mobile activities, at the spatial resolution of cities, communities or grid cells. This is not always the case for register or administrative data. In addition, mobile network data are dynamic, which means that the movement of activities can be observed over the course of the day as well as daily, during a week or a month. This means in turn that commuter movements can be illustrated and used more quickly and more up-to-date. Even the daily and the resident population can be mapped with these data. Furthermore, the mobile network data reveal rural and urban areas due to the distribution and intensity of mobile activities. This can be used to determine approximately whether the activities were carried out at the workplace or at the place of residence. This information cannot be obtained via administrative or register data, since these data are static and provide only information at a specific time or reporting date.

In Germany, there are three mobile network operators, Deutsche Telekom AG, Vodafone, and Telefónica, 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. This includes contract, prepay, and further customers. A mobile activity is defined as an event caused by a length of stay at a location or in a specific geometry without movement (also known as dwell time), with all (mobile) signalling data being evaluated. Furthermore, 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. In compliance with data protection rules, the mobile activities are anonymised and aggregated. 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 mobile activities are subject to some assumptions or prerequisites in order to obtain the data record. Since the number of mobile activities depends on the dwell time of mobile devices, long mobile device activities corresponding to the length of the dwell time are counted and included in the data record, while short mobile activities are not taken into account. The dwell time in the data record available is two hours in order to filter out short mobile device activities which result, for instance, from quick movements between the grid cells. Finally, only values based on a minimum number of 30 activities per geometry were provided.

Our aim is to analyse the effect of commuters on both unemployment rates. Since we use a model-based method, suitable covariates are crucial. Hence, we create a comprehensive tailor-made data record with dynamic aggregated mobile activities of Deutsche Telekom customers, which we used to construct covariates for the model-based estimation. The mobile activities in this data set refers to the municipality level of NRW and thus can be aggregated up to the FUA sublevel since municipalities are nested within these subareas. The data contains mobile activities for a statistical week that consists of 24-hour days. These were selected from the months April, May, and September in 2017 without school or public holidays in order to avoid distortions in the representation of commuters. The mobile activities comprise the average activities on the weekdays selected. The weekdays are categorized 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 to estimate the alternative unemployment rates and to illustrate the difference between both rates, 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 in the regions under consideration. 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, which are also very well reflected in the mobile activities. Furthermore, when estimating unemployment rates, we distinguish between sex and other socio-demographic characteristics such as age, since different characteristics are assumed to have different influences on commuting behaviour and thus on the rate. Therefore, the shares of women and men in mobile activities were still calculated, as well as persons over and under 50 years old. Furthermore, the shares of young mobile device users as well as the shares of nationalities, summarized by continents, were calculated. Especially in areas with a high volume of tourists, this can have an influence on the unemployment rates. In total, we have 27 predictors that have the potential to explain possible differences and the variation of both unemployment rates in NRW. An overview of the calculated 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 from the LFS on FUA sublevel is described. More generally, our target indicator is a small area mean. For this indicator, we propose an estimator, which helps to increase the accuracy of the direct estimates. As the survey is designed for higher regional levels, auxiliary variables from mobile network data are used. For this purpose, 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 estimators only need to be available on an aggregated level. The FH model produces estimates on a continuous range, but on the opposite the unemployment rate for each area is located within the interval $[0, 1]$. To produce only FH estimators on the desired interval, transformations offer important possibilities. For this reason, we select the inverse sine transformation, which ensures that all estimates are within the mentioned range. Following Sugasawa and Kubokawa (2017) we derive an 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 $n_1, \dots, n_d$ 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 a linear mixed model consisting of covariates $\mathbf{x}_i$, an area-specific random effect u_i, and a sampling error e_i for each area $i \in 1, ..., d$

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

The model assumes that the random effects u_i are identically independently normally distributed and the sampling errors e_i are independently normally distributed. Out of this model, the regression parameters $\hat{\boldsymbol{\beta}}$ can be estimated as best linear unbiased estimator of β and the random effect $\hat{u}_i$ as empirical best linear unbiased predictor of u_i (Rao and Molina, 2015). The variances $\sigma_{e_i}^2$ of the sampling errors are assumed to be known. In many applications, this assumption is not met and the variances must be approximated. Since the variances of the sampling errors are not known in our investigation, the next subsection explains how they are obtained. For the estimation of the variance of the random effect σ_u^2, several approaches are available: The FH method of moments, the maximum Likelihood Method (ML), and the Residual Maximum Likelihood Method (REML) (Rao and Molina, 2015) among others. 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 $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ as follows:

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

where the shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ defines the weights on both parts for each area i. Whenever the variance of the sampling error is relatively small, more weight lies on the direct estimator. To assess the accuracy of the FH estimator, its MSE is determined. If the variance of the random effects u_i is estimated with ML or REML methods, following Prasad and Rao (1990) and Datta and Lahiri (2000), an analytical solution is provided to obtain the MSE of the FH estimator.

3.2 Back-transformed Fay-Herriot estimates

Since the FH model produces estimates on a continuous range, whereas the unemployment rate is a percentage, we transform the dependent variable to obtain FH estimators in the desired range. Thus, we use an 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). 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), which introduce the FH model for general transformations on the dependent variable. In our application, we do not have access to the variances of the sampling errors at the target level. Therefore, we have to approximate the variances of the direct estimates. Following Jiang *et al.* (2001), we get the sampling variances for the inverse sine transformed direct estimates $\left(\sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right)\right)$ using the effective sample size $\tilde{n}_i$ and thus obtain $\tilde{\sigma}^2_{e_i} = 1/4\tilde{n}_i$. For the model on the transformed scale, we consider the assumptions of the FH model

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

In applications, the model parameters must be estimated. Out of the FH model on transformed scale (4), $\hat{\boldsymbol{\beta}}$ and $\hat{u}_i$ can be estimated, as described in the previous Subsection 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 *et al.*, 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

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

where $\hat{\theta}_i^{\text{FH, trans}}$ denotes the transformed FH estimator. To solve this integral, numerical integration techniques are applied. In Section 6, the proposed bias correction for the FH model using the inverse sine transformation is evaluated.

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 determine 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 an inverse sine transformation. Therefore, bootstrap methods are very promising to estimate the MSE. Casas-Cordero *et al.* (2016) construct confidence intervals using a parametric bootstrap procedure, in which the thresholds of the confidence intervals from the bootstrap are built on the transformed scale with subsequent naive back-transformation.

In contrast to this methodology, our goal is to estimate 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:

1. Out of the model on the transformed scale (4), estimate $\hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ using the sampled areas. Approximate the sampling variance of the direct estimator on transformed scale $\left(\sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right)\right)$ with $\tilde{\sigma}_{e_i}^2 = 1/4\tilde{n}_i$, where $\tilde{n}_i$ denotes the effective sample size (Jiang *et al.*, 2001).

2. For $b = 1, ..., B$

   - Generate area specific random effects $u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and sampling errors $e_i^* \sim \mathcal{N}(0, \tilde{\sigma}_{e_i}^2)$.
   - Using the bootstrap sample

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

     to determine the FH estimates on the original scale $\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}}\right)$ with (5) to account for the bias correction.
   - For each bootstrap population, calculate the population mean on the original scale:

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

3. Predict the MSE and the 95% confidence intervals

   $$\begin{aligned}
   \text{MSE}(\hat{\theta}_i^{\text{FH, trans}}) &= \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)^2 \\
   \text{CI}(\hat{\theta}_i^{\text{FH, trans}}) &= \left[\hat{\theta}_i^{\text{FH, trans}} + q_{0.025}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right);\right. \\
   &\qquad \left.\hat{\theta}_i^{\text{FH, trans}} + q_{0.975}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)\right],
   \end{aligned}$$

   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 to construct uncertainty measurements for the back-transformed FH estimates is evaluated within a simulation study (cf. Section 6).

4 Alternative unemployment rates including commuters in North Rhine-Westphalia

In this section, we consider and evaluate the influence of commuting behaviour on traditional and alternative unemployment rates. For this purpose, we use the LFS data from Section 2.2 and the mobile network data from Section 2.3. The aim of the application is to illustrate the difference between traditional and alternative unemployment rates by sex for the FUA sublevel in NRW, and thus demonstrating the influence of commuting behaviour on the rates.

4.1 Difference between traditional and alternative unemployment rates

The traditional definition of unemployment rates, as defined in (1), corresponds to the proportion of unemployed persons at the place of residence. As Grözinger (2018) remarks this definition may lead to contradictory results. While the city centres are the job engines, unemployment hotspots are often also located in the city centres. How does this contradiction arise? One reason is the non-inclusion of the increasing commuting behaviour of the working population. The traditional definition of unemployment rates leads to an overestimation of these rates in the city cores and an underestimation in the surrounding travel-to-work areas. This regional distinction is represented at the FUA sublevel that is defined by the commuter behaviour of the employed persons working in the city centres. Is it thus appropriate to calculate the unemployment rates only based on the place of residence, which shows higher unemployment rates in the cities? A possible alternative is to count the working population at their workplaces as in (2). It is assumed that unemployment rates will increase in the commuter zones and decrease - due to the definition of the FUA sublevel - in the city centres. To verify this, we focus on calculating the difference between the traditional and alternative estimated unemployment rates, which can be determined as follows

$$\hat{\theta}_{\Delta,i} = \hat{\theta}^{\text{FH,trans}}_{\text{UR}_1,i} - \hat{\theta}^{\text{FH,trans}}_{\text{UR}_2,i}. \tag{6}$$

$\hat{\theta}^{\text{FH,trans}}_{\text{UR}_1,i}$ is the transformed FH estimate for the traditional unemployment rate and $\hat{\theta}^{\text{FH,trans}}_{\text{UR}_2,i}$ for the alternative unemployment rate. The estimator $\hat{\theta}_{\Delta,i}$ is thus composed of two parts. In order to determine MSEs and confidence intervals for $\hat{\theta}_{\Delta,i}$ out of the MSEs and confidence intervals of both unemployment rates, we estimate conservatively. This means that we do not include the covariances and thus overestimate the MSE of $\hat{\theta}_{\Delta,i}$. To construct the confidence intervals, we use the Mover method (Dormer and Zou, 2002; Newcombe, 2011). The Mover method assumes independence of both FH estimates and thus the covariance is also not taken into account.

4.2 Model selection and validation

In order to calculate the difference between the two unemployment rates, first suitable models are created. For both sexes one model each is needed to obtain $\hat{\theta}^{\text{FH,trans}}_{\text{UR}_1,i}$ as well as one for $\hat{\theta}^{\text{FH,trans}}_{\text{UR}_2,i}$. Thus we have to build four different models. For this purpose, a model selection is performed. Therefore, the Bayesian information criterion on a simple linear regression model on inverse sine transformed dependent variables is constructed following Schmid *et al.* (2017). As dependent variable, we use the direct estimates out of the LFS. By applying an automatic stepwise and backward selection procedure, we obtain the selected models. Note that there are also corrected information criteria for FH models like the corrected Akaike information criterion from Marhuenda *et al.* (2014). Since such criteria are not implemented within the stepwise procedure in R (R Core Team, 2019) we use the simple information criteria. In total, 6 to 16 of 27 potential mobile network covariates are selected depending on the model. Four models are needed to estimate unemployment rates by place of residence and place of work for each sex. 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, values above 0 were estimated in all cases. Thus, the potential problem of negatively estimated variances does not occur in our models. For each FH model on the transformed scale (cf. (4)), the assumptions on the error terms are checked. The normality assumptions of the random effects as well as of the residuals are tested. The p-values of the Shapiro-Wilk test in Table 2 confirm that in all cases the normality assumption for both error terms cannot be rejected.

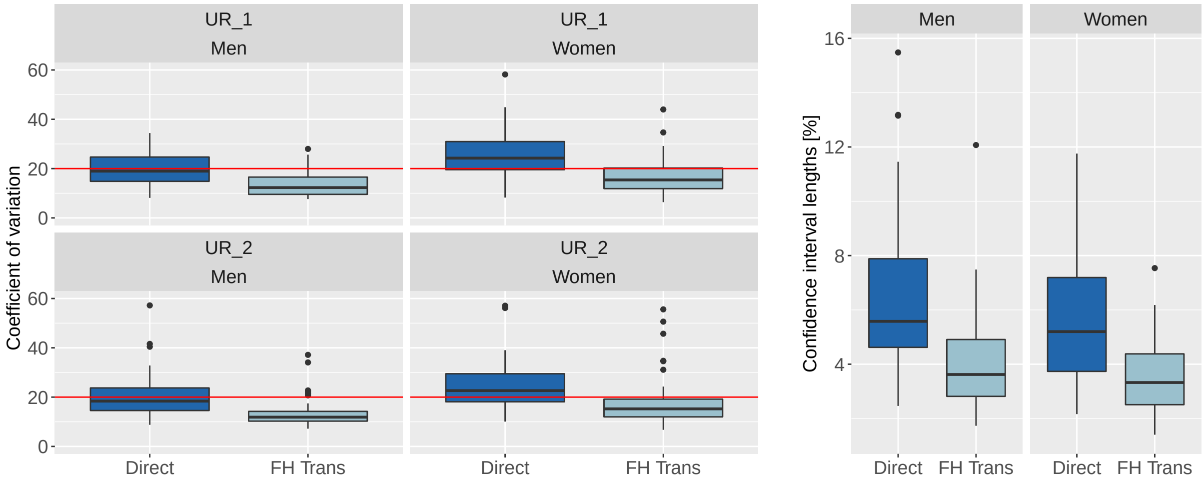
Table 2: Measurements to validate the FH models for the two unemployment rates (UR_1 and UR_2) for both sexes. This table shows the estimated variance of the random effects ($\hat{\sigma}_u^2$), the Shapiro-Wilks (S.-W.) p-value for the random effects (ran. effects) as well as for the standardized residuals (stand. res.) and 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: ran. effects	0.695112	0.549476	0.861257	0.901708
S.-W. p-value: stand. res.	0.308668	0.495064	0.809323	0.866098
modified R^2	0.772521	0.908642	0.632059	0.575550

All four models contain meaningful covariates. Since the models are built on the transformed scale (cf. (4)), the exact values of the coefficients cannot be interpreted. 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. An increase of covariates with regard to possible commuter movements generally lead to a decrease of unemployment rates with focal point place of work. The reverse is the case for unemployment rates with focal point place of residence. All models also 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. Furthermore, all available covariates were included in one of the four models.

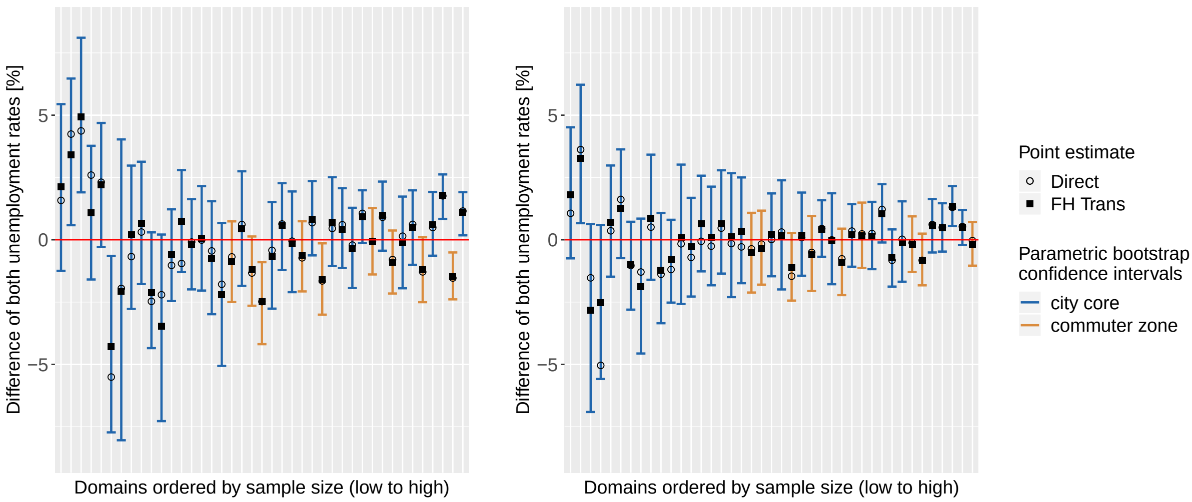
4.3 Gain in accuracy

As the LFS is designed to estimate indicators at higher regional levels, we use the SAE methodology for estimating at the FUA sublevel. To assess the gain in the reliability of the estimators, we first compare the coefficients of variation. Part (a) of 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, place of residence), 29.3% (women, place of residence), 53.7% (men, place of work), and 31.7% (women, place of work) 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, place of residence), 73.2% (women, place of residence), 82.9% (men, place of work), and 78.0% (women, place of work) of the CVs are below 20%. This illustrates that the use of dynamic mobile network data within a small area setting is a powerful tool to heighten the accuracy of the estimated unemployment rates for NRW on FUA sublevel. Another way to assess the accuracy of the estimators is to compare the distribution of the confidence interval lengths in Figure 1 (b). It can be seen that the confidence interval lengths are reduced by applying the transformed FH model compared to the direct estimation. These lengths decrease for all areas. In median, they reduce around 37% for men and 35% for women. This result confirms the necessity for the application of SAE methods.



(a) (b)

Figure 1: Reduction of the uncertainty by using the transformed FH model instead of the direct estimation: Coefficient of variation (a) and confidence interval lengths (b) for estimating unemployment rates in NRW.



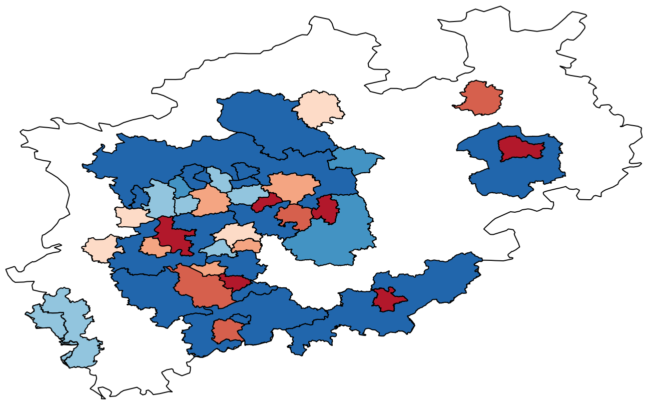
(a) (b)

Figure 2: Confidence intervals for the difference of both definitions of the unemployment rate for men (a) and women (b) in NRW.

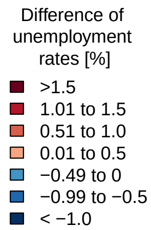
Figure 2 illustrates the estimates for the difference of both unemployment rates in the FUA sublevels of NRW ordered by increasing sample size. In addition to the differences of the estimates from the transformed FH model, the calculated differences of the direct estimates are shown. In most cases, the direct estimates and the FH estimates are really close to each other. However, in particular for regions with smaller samples sizes like Witten and Paderborn, these values can deviate more clearly from each other. For a spatial assignment of city names to corresponding FUA city cores, see Figure 7 in the appendix. Due to the higher uncertainty of the direct estimates for regions with lower sample sizes, the synthetic part within (3) is weighted more highly and bigger differences to the direct estimator appear. However, we are interested, whether both definitions of unemployment rates lead to significant different results in NRW. Figure 2 shows, that both unemployment rates are different for some areas. For the males model, we get significant negative differences of both unemployment rates for the commuter zone belonging to Bonn, Duisburg, and Paderborn as well as for the city core Recklinghausen. These areas are commuter zones that belong to large cities in NRW or, as in the case of Recklinghausen, a smaller city that is close to large cities (Dortmund and Essen). This means that these areas are the place of residence of many employed people who commute from those areas to their workplace. Consequently, the unemployment rates with focal point workplace for the abovementioned areas are higher than the traditional ones. In the opposite direction, we obtain significantly positive values for the cities of Cologne, Düsseldorf, Paderborn, and Siegen. Concerning the females model, we get only two positive significantly estimated differences between the two unemployment rates. This is the case for Düsseldorf and Paderborn, as in the males model, which means that unemployment rates with a focal point workplace are lower than the traditional ones. These areas are city centres and the place of work of many employed women who commute from the surrounding commuter areas to the two city centres. Further comparing the confidence interval plots for men and women, the differences between men's unemployment rates are more significant, which is related to the more pronounced commuting behaviour of men.

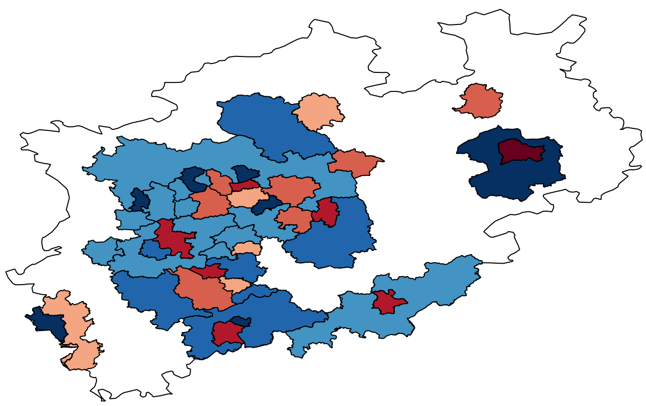
4.4 Discussion of the estimated unemployment rates for NRW

Figure 3 illustrates the calculated differences of the unemployment rate, as defined in (6). If the unemployment rates by the place of residence are the same as by the place of work, the commuter behaviour has no influence on unemployment rates. The calculated difference would be zero. Please note, that the FUA sublevels do not cover the entire federal territory in NRW and are represented as white areas in Figure 3. The bluish areas indicate that the unemployment rate by place of work is higher than the unemployment rate by place of residence. Those are mainly the commuter zones of the FUAs in both models. Blue coloration means that the commuter flow is directed out of this area. The reddish areas, however, imply that the unemployment rate by place of work is lower than the unemployment rate by place of residence. This is mainly the case for the city cores of the FUAs in both models. These results indicate that the alternative unemployment rate in city cores would be generally lower than calculated based on traditional unemployment rates. This in turn implies that many employed people living the surrounding travel-to-work areas, their place of residence, commute into the city centres to work. In the males model, the differences are higher than in the females model, which leads to the assumption that women are not commuting as often or as far as men, just as shown in Landesbetrieb Information und Technik Nordrhein-Westfalen (2019). Possible reasons for this could be the conservative role model of women a 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).



(a)





(b)

Figure 3: Difference of unemployment rates due to including commuters for men (a) and women (b). The spatial assignment of city names to the FUA sublevels is shown in Figure 7.

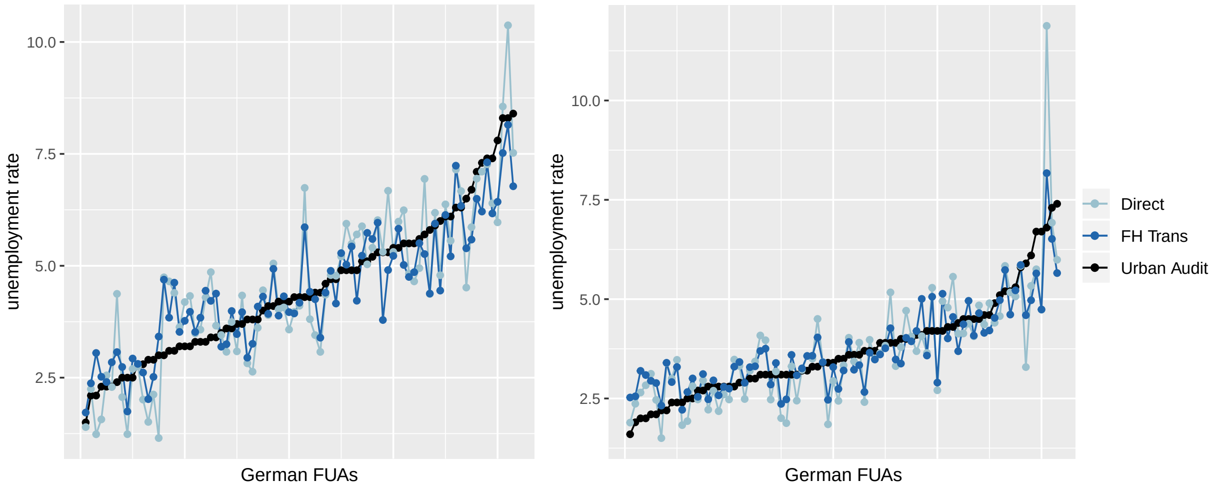
5 Validity of the proposed method

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

For the German federal state NRW, we have an extensive mobile network data record available as auxiliary information. However, for the whole of Germany, we have only limited access to mobile network data and accordingly a data set with less information. Thus, less covariates are available for the validation. In contrast to Section 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 the Sunday evening, it is useful that a high correlation between the population figures from the 2011 census and the mobile network activities on the weekend and especially on Sunday evening (Hadam, 2018) was identified. 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. Since the Urban Audit does not provide alternative unemployment rates and unemployment rates for the commuter zones, we can validate only traditional unemployment rates for the entire FUA. Therefore, we use the SAE method as applied in Section 4 with the difference that a) the regional focus is now the whole of Germany and b) we can only use mobile network data from Sunday evening. Thus, we have built two models at the German level: One for women and one for men. Again, we followed the same model selection procedure as in Section 4.2. 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%. Normality assumptions concerning the random effect and the sampling errors have been checked.

For the validation of the proposed method, Figure 4 shows the estimated unemployment rates using mobile network covariates (FH Trans), the direct estimates, and the official published ones 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. If we compare the direct estimator from the LFS with the FH Trans estimator, it is noticeable that in most cases the FH Trans estimator corrects the direct estimator in such a way that the resulting value is closer to the Urban Audit. This trend is quantified in Table 3. It reports the distribution of the absolute bias of the females and males unemployment rates obtained by the two estimation methods for all FUAs in Germany compared to the Urban Audit. In particular, for almost all distribution values we get a higher absolute bias for the direct estimator compared to the FH Trans estimator. Only in the males model the 25% quantile for the absolute bias is slightly higher for the



(a) (b)

Figure 4: 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 (a) and women (b) for all German FUAs.

Table 3: Distribution of the absolute bias 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

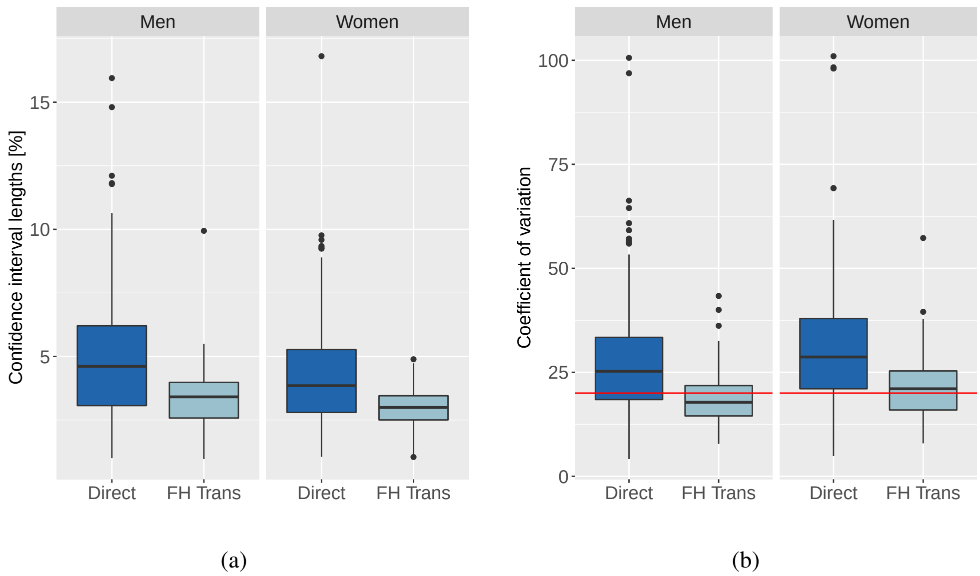


Figure 5: Reduction of the uncertainty by using the transformed FH model instead of direct estimation: Confidence interval lengths (a) and Coefficient of variation (b) for estimating unemployment rates in Germany.

FH Trans estimator. 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 bias.

After examining the deviation of the direct estimator and the FH Trans estimator from the Urban Audit, the aim is now to assess the uncertainty of these estimators. Hence, we compare the accuracy of the direct and the FH Trans estimator. As measurements of uncertainty, we are interested in the confidence interval lengths and the CVs. Figure 5 (a) shows the distribution of the confidence interval lengths of the FH Trans estimates and the direct estimates for the females and the males model. It can be seen that the confidence intervals for the FH Trans estimator are much smaller than those for the direct estimator. On average, reductions of 31.1% (males model) and 30.3% (females model) are achieved in both models. Using the FH Trans estimator also achieves lower CVs (cf. Figure 5 (b)). Therefore, much more CVs fall below the threshold of 20% for reliable estimates according to Eurostat (2019a). For both models, female and male, more than twice as many CVs are below this limit if SAE methods are applied in contrast to the use of direct estimation. The reduction of uncertainty is one strong argument for using the transformed FH model for the prediction. Mobile network data has a great potential as auxiliary data in this context. In contrast to the direct estimation, the transformed FH model with mobile network data leads to estimators, which are closer to those of the Urban Audit. Overall, this subsection shows that the use of SAE techniques using mobile network data improves the estimation of unemployment rates from the LFS.

6 Model-based simulation

In Section 4, we use the proposed transformed FH model to estimate alternative unemployment rates taking commuters into account. We evaluated the suggested methodology in a similar environment where official data (Urban Audit) is available. According to the Jensen-inequality (cf. Subsection 3.2), the naive back-transformation is biased under the inverse sine transformation. The aim of our model-based simulation study is 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 for the conducted analyses. Furthermore, we want to show, that the proposed MSE and confidence intervals for the estimates lead to reasonable results. We investigate these aims in a close to reality environment. Therefore, the input values of the model-based setting are based on the real data.

The simulation study is implemented with $R = 1\,000$ 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 $d = 208$ which correspond to the number of the FUA sublevels in Germany. 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 the German FUA sublevel (cf. Section 5), $\sigma_u = 0.02896193$ is defined analogously. In addition, we adopt the sampling errors $\sigma_{e_i}^2$ and keep them constant over the replications. Table 4 shows the distribution of the sampling errors and the resulting values for the shrinkage factor. 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 reports information about 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.

Table 4: Distribution of relevant parameters in the simulation setting: The sampling error σ_{e_i} and the resulting shrinkage factor γ_i are taken from the underlying FUA sublevel data. The direct estimates $\left(\hat{\Theta}_i^{\text{direct}}\right)$ of the simulation study are close to the values from the 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
	FUAs	0.0054	0.0328	0.0484	0.0508	0.0647	0.1134

For each replication, we compute both estimated small area means from the transformed FH model: With respect to the back-transformation bias $\left(\hat{\Theta}_i^{\text{FH,trans}}, \text{cf. (5)}\right)$ and with naive back-transformation $\left(\hat{\Theta}_i^{\text{FH,naive}}\right)$. To assess the quality of the estimates we obtained over $r = 1, ..., 1\,000$ Monte Carlo replications the absolute Bias (aB) and the Root Mean Squared Error (RMSE) of the estimates, defined as

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

where $\hat{\Theta}_i^{\text{FH},(r)}$ is the estimated respective FH value and $\bar{y}_i^{(r)}$ the true value. Figure 6 shows the reduction of aB. For instance, the median aB using a naive back-transformation is 1.86 times higher than the median aB under a transformed FH model with a bias corrected back-transformation. At the same time, we observe nearly the same RMSE when we use a bias corrected back-transformation instead of a naive back-transformation. Nevertheless, there is a small difference in the RMSE: For the bias corrected back-transformed estimator, the RMSEs are in median 0.07% higher.

We next investigate the properties of the proposed MSE and the confidence intervals. For calculating these uncertainty measurements, we use 1 000 bootstrap replications within each Monte Carlo run. As quality measurements, we use the relative bias of the RMSE (rB RMSE) and the relative RMSE of the

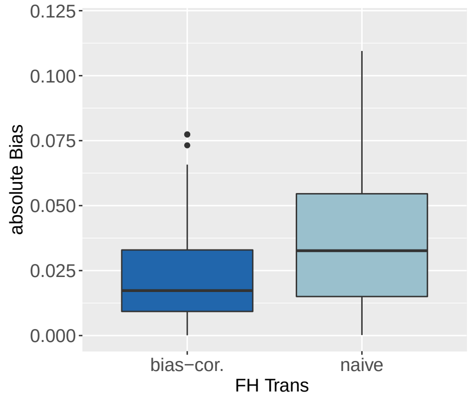


Figure 6: Distribution of the aB for the transformed FH estimator with bias corrected $\left(\hat{\Theta}_i^{\text{FH,trans}}\right)$ and naive $\left(\hat{\Theta}_i^{\text{FH,naive}}\right)$ back-transformation.

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

	Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
rB	-9.34	-2.12	-0.62	-0.55	1.11	7.13
rRMSE	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

RMSE (rRMSE RMSE). They are defined as

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

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

where $\text{RMSE}_{\text{est},i}^{(r)}$ is the estimated root MSE out of the bootstrap procedure (cf. Subsection 3.3) for each Monte Carlo replication r and $\text{RMSE}_{\text{true},i}$ is the empirical root 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 interquantile range goes from -2.12% to 1.11%. In addition to the relative bias, 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. Subsection 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, 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%.

Overall, our simulation study shows advantages for the bias of the transformed FH estimator with bias corrected back-transformation over a naive back-transformation for the used setup, which was adopted to the underlying real data. We can also show a good performance of the newly proposed MSE estimator for the FH estimator with bias corrected back-transformation and the confidence intervals.

7 Concluding remarks

The traditional unemployment rate is based on the place of residence of the labour force by using data from the LFS. Due to the high level of commuting, however, the picture is reversed and gives a distorted impression of the regional labour market. For Germany, traditional unemployment rates show higher rates in cities 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 and adjust this by including commuters in the calculation. 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. Therefore, we select covariates constructed from dynamic mobile network data and validate the selected models. The benefit of dynamic mobile network data is that they can 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. Therefore, the assumption that the unemployment rate in cities is lower can be confirmed and thus contributes to the explanation why so many people move to cities 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 rates provides potential employers with additional information about the current regional labour market and on missing workplaces in regional areas. Thus, it can be identified for which regions it might be useful to promote the settlement of companies in order to lower their unemployment rate and shorten the commute, as new details of potentially available local workforce are available. The increasing number of commuters should therefore 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 Call Detail Records. That means there is no chance to use individual activity movements for the estimation or even to take changes in individual social behaviour into account. 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 the unemployment rate compared to the used distribution of mobile activities over time.

For further research, 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. Furthermore, a SAE approach with mobile network data and additional data sources should be considered. Satellite data for example might include valuable information 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. For instance, Steele *et al.* (2017) use a combination of satellite and mobile network data to obtain more explanatory power.

Acknowledgements

Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes.

Appendix

Table 6 describes the mobile network covariates used in the paper. Only aggregated signals of mobile devices originating from the mobile network of Deutsche Telekom for the years 2017 and 2018 are used. Since no individual data are available, further covariates are created from the aggregated mobile activities, based on mobile network data on the basis of aggregates of a statistical week. The variables are categorized according to the absolute values delivered by T-System and according to additionally calculated covariates. As we also have information on Customer Relationship Management (CRM) data from the billing system in addition to the number of mobile activities, we can therefore distinguish between women (F) and men (M), age group (20-29; 30-39; 40-49; 50-59; 60-69; 70-79 years) and nationality of the SIM card owner.

Table 6: Mobile network covariates.

Name	**Covariate**	**Description**
Based on absolute activities		
count_a_g	absolute activities (age, sex)	The absolute numbers of activities by age group (a) and sex (g).
share_a	share activities age group	The proportion of age group (a)
share_a_Sunday evening	share activities age group Sunday evening	The proportion of age group (a) on a Sunday evening
share_g_Sunday evening	share activities sex Sunday evening	The proportion of sex (g) on a Sunday evening
share_n	share activities nationality SIM card (n)	The proportion of the nationality of SIM card owner (n).
share_n_night	share activities nationality SIM card (n) night-time	The proportion of the nationality of SIM card owner (n) at the night-time (5 to 11 pm).
share_n_day	share activities nationality SIM card (n) day-time	The proportion of the nationality of SIM card owner (n) at the day-time (7 am to 4 pm).
Additionally calculated covariate		
ratio_morning_day	ratio activities morning day-time	The ratio of mobile activities between 12 pm to 6 am (morning time) over mobile activities between 7 am to 4 pm (day-time).
ratio_day_evening	ratio activities day-time evening	The ratio of mobile activities between 7 am to 4 pm (day-time) over mobile activities between 5 to 11 pm (night-time).
ratio_early_peak	ratio activities early peak	The ratio of mobile activities between 3 to 5 am (early peak) over mobile activities between 9 to 11 am, which is the peak of the middle of the day.

ratio_late_peak	ratio activities late peak	The ratio of mobile activities between 9 to 11 am over mobile activities between 8 to 10 pm, which is the late peak.
ratio_home	ratio activities home	The ratio of mobile activities between 5 pm to 5 am, where the persons are more likely to stay at home, over all mobile activities.
ratio_work	ratio activities work	The ratio of mobile activities between 6 am to 4 pm, which are typical working hours in Germany, over all mobile activities.
change_work_to_home	change activities work to home	The change of mobile activities of working hours to hours spent at home.
change_home_to_work	change of	The change of mobile activities of hours spent at home to working hours.
change_commuter	change activities commuter	The change of mobile activities of potential commuters between 10 am and 4 pm.
change_top_work_home	change activities peak work to home	The change of mobile activities of the peaks value of working hours at 10 am to hours spent at home at 9 pm.
change_n_night_day	change activities country night to day	The change of mobile activities by nationality of SIM card owner (n) by night- and day-time.

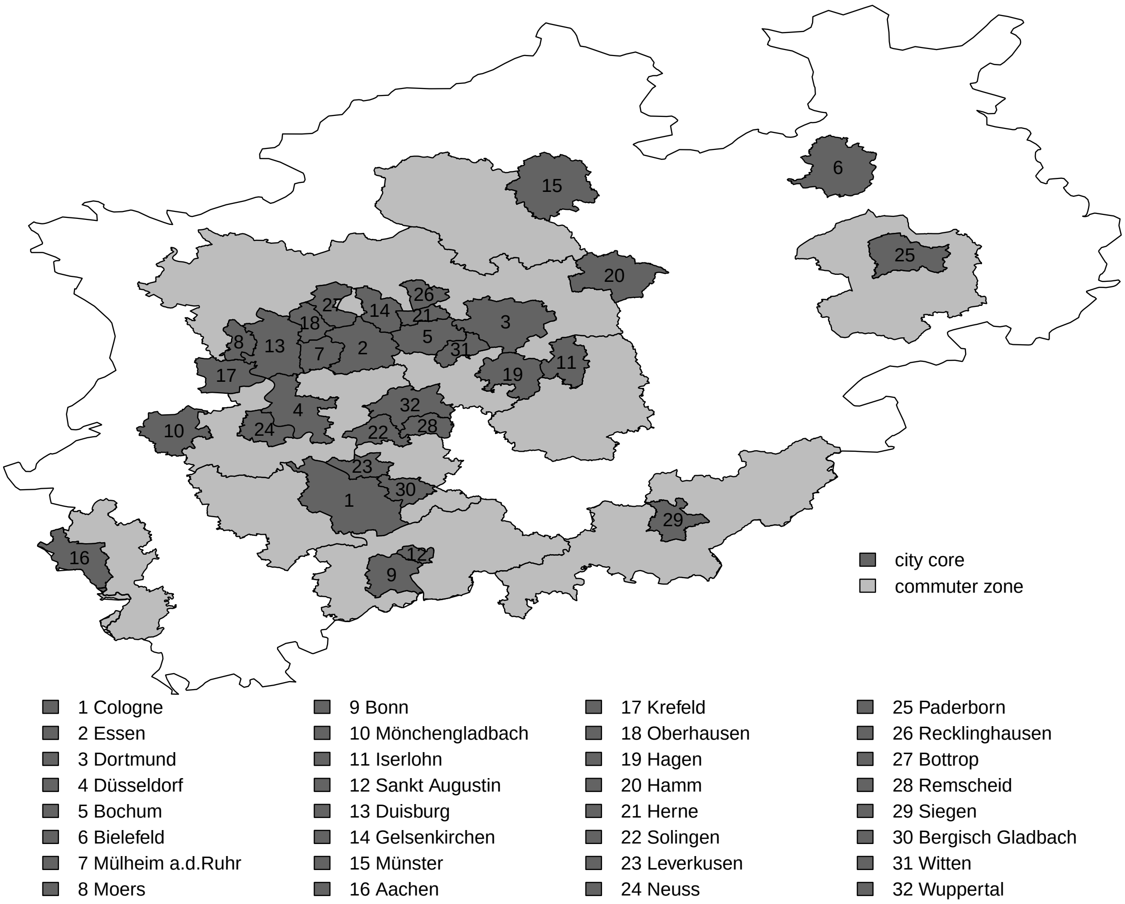


Figure 7: Assignment of city names to FUA city centres.

References

Bauer-Hailer, U. (2019) Berufspendler im Bundesländervergleich. Statistische Landesamt Baden-Württemberg, Stuttgart. https://www.statistik-bw.de/Service/Veroeff/Monatshefte/PDF/Beitrag19_02_02.pdf. [accessed on 07.12.2019].

Buch, T., Hamann, S., Niebuhr, A. and Rossen, A. (2014) What makes cities attractive? The determinants of urban labour migration in Germany. *Urban Studies*, **51**, 1960–1978.

Burgard, J. P., Münnich, R. and Zimmermann, T. (2016) Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement. In *Analysis of Poverty Data by Small Area Estimation* (ed. M. Pratesi), 85–108. Hoboken: John Wiley & Sons.

Casas-Cordero, C., Encina, J. and Lahiri, P. (2016) Poverty Mapping for the Chilean Comunas. In *Analysis of Poverty Data by Small Area Estimation* (ed. M. Pratesi), 379–403. Hoboken: John Wiley & Sons.

Costa, A., Satorra, A. and Ventura, E. (2006) Improving small area estimation by combining surveys: New perspectives in regional statistics. *Statistics and Operations Research Transactions*, **30**, 101–121.

Datta, G. S. and Lahiri, P. (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 613–627.

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F. and Reuter, H. I. (2016) Assessing the quality of mobile phone data as a source of statistics. European conference on quality in official statistics (Q2016). https://ec.europa.eu/eurostat/cros/system/files/assessing_the_quality_of_mobile_phone_data_as_a_source_of_statistics_q2016.pdf. [accessed on 09.11.2018].

Dijkstra, L. and Poelman, H. (2011) Archive:European cities - the EU-OECD functional urban area definition. Eurostat, Luxembourg. 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 on 11.06.2019].

Dormer, A. and Zou, G. (2002) Interval estimation for a difference between intraclass kappa statistics. *Biometrics*, **58**, 209–215.

Europäisches Parlament und Rat der Europäischen Union (2003) Verordnung (EG) Nr. 1059/2003 des Europäischen Parlaments und des Rates vom 26. Mai 2003 über die Schaffung einer gemeinsamen Klassifikation der Gebietseinheiten für die Statistik (NUTS). *Amtsblatt der Europäischen Union*, **Nr. L 154**, 1–41.

Eurostat (2017) City statistics (urb): National reference metadata in euro SDMX metadata structure (ESMS). Eurostat, Luxembourg. https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm. [accessed on 09.11.2018].

Eurostat (2018a) Dataset details: Harmonised unemployment rate by sex. Eurostat, Luxembourg. https://ec.europa.eu/eurostat/web/products-datasets/-/teilm020. [accessed on 29.10.2018].

Eurostat (2018b) Glossary: Functional urban area. Eurostat, Luxembourg. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area. [accessed on 13.05.2018].

Eurostat (2018c) NUTS - Systematik der Gebietseinheiten für die Statistik: Hintergrund. Eurostat, Luxembourg. `https://ec.europa.eu/eurostat/web/nuts/background`. [accessed on 21.08.2019].

Eurostat (2019a) DataCollection: Precision level DCF. Eurostat, Luxembourg. `https://datacollection.jrc.ec.europa.eu/wordef/precision-level-dcf`. [accessed on 06.06.2019].

Eurostat (2019b) EU labour force survey database user guide. Eurostat, Luxembourg. `https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf`. [accessed on 22.08.2019].

Eurostat (2019c) Städte (urban audit): Datenbank. Eurostat, Luxembourg. `https://ec.europa.eu/eurostat/de/web/cities/data/database`. [accessed on 21.08.2019].

Fay, R. E. and Herriot, R. A. (1979) Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Federal Institute for Research on Building, Urban Affairs and Spatial Development (2017) Wachsen und Schrumpfen von Städten und Gemeinden. Federal institute for research on building, urban affairs and spatial development (BBSR), Bonn. `https://gis.uba.de/maps/resources/apps/bbsr/index.html?lang=de`. [accessed on 11.06.2019].

Galiana, L., Sakarovitch, B. and Smoreda, Z. (2018) Understanding socio-spatial segregation in French cities with mobile phone data. `http://www.dgins2018.ro/wp-content/uploads/2018/10/08-FR-dgins_segregation_1800905.pdf`. [accessed on 09.11.2018].

Gans, P. (2017) Urban population development in Germany (2000-2014): The contribution of migration by age and citizenship to reurbanisation. *Comparative Population Studies*, **42**, 319–35.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008) Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics & Data Analysis*, **52**, 5242–5252.

Grözinger, G. (2018) Regionale Arbeitslosigkeit: Falsche Eindrücke von Stadt-Land-Differenzen. *Wirtschaftsdienst*, **98**, 68–70.

Hadam, S. (2018) Use of mobile phone data for official statistics. *METHODS-APPROACHES-DEVELOPMENTS: Information of the German Federal Statistical Office*, **1**, 6–9.

International Labour Organization (2018) Unemployment rate. International Labour Organization, Genève. `http://www.ilo.org/ilostat-files/Documents/description_UR_EN.pdf`. [accessed on 29.10.2018].

Jensen, J. L. W. V. *et al.* (1906) Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, **30**, 175–193.

Jiang, J., Lahiri, P., Wan, S.-M. and Wu, C.-H. (2001) Jackknifing in the fay-herriot model with an example. Technical Report, Department of Statistics, University of Nebraska, Lincoln.

KOSIS-Gemeinschaft Urban Audit (2013) Das deutsche Urban Audit - Städtevergleich im Europäischen Statistischen System. KOSIS-Gemeinschaft Urban Audit, Mannheim. `http://www.staedtestatistik.de/fileadmin/urban-audit/UA_Broschuere_2013.pdf`. [accessed on 21.08.2019].

Lahiri, P. and Suntornchost, J. (2015) Variable selection for linear mixed models with applications in small area estimation. *The Indian Journal of Statistics*, **77**, 312–320.

Landesbetrieb Information und Technik Nordrhein-Westfalen (2019) Berufspendler 2011 - 2018 nach Pendlerart, Beschäftigungsumfang und Geschlecht. Landesbetrieb IT.NRW Statistik und IT-Dienstleistungen, Düsseldorf. `https://www.it.nrw/statistik/eckdaten/berufspendler-2011-2018-nach-pendlerart-beschaeftigungsumfang-und-geschlecht`. [accessed on 07.12.2019].

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015) Small area model-based estimators using big data sources. *Journal of Official Statistics*, **31**, 263 – 281.

Marhuenda, Y., Morales, D. and del Carmen Pardo, M. (2014) Information criteria for fay–herriot model selection. *Computational Statistics & Data Analysis*, **70**, 268–280.

Martini, A. and Loriga, S. (2017) Small area estimation of employment and unemployment for local labour market areas in Italy. Istituto Nazionale di Statistica, Rome. `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: 10.11.2018].

Möbert, J. (2018) The german housing market in 2018. Deutsche Bank Research, Frankfurt am Main. `https://everydaypoint.com/wp-content/uploads/2019/04/The_German_housing_market_in_2018-1.pdf`. [accessed on 03.09.2019].

Molina, I. and Strzalkowska-Kominiak, E. (2020) Estimation of proportions in small areas: Application to the labour force using the swiss census structural survey. *Journal of the Royal Statistical Society: Series A*, **183**, 281–310.

Newcombe, R. G. (2011) Propagating imprecision: Combining confidence intervals from independent sources. *Communications in Statistics-Theory and Methods*, **40**, 3154–3180.

Pereira, L. N., Mendes, J. and Coelho, P. (2011) Estimation of unemployment rates in small areas of Portugal: A best linear unbiased prediction approach versus a hierarchical bayes approach. 17th European Young Statisticians Meeting.

Prasad, N. N. and Rao, J. N. (1990) The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

R Core Team (2019) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org`.

Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*. Hoboken: John Wiley & Sons.

Schmid, T., Bruckschen, F., Salvati, N. and Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A*, **180**, 1163–1190.

Slud, E. V. and Maiti, T. (2006) Mean-squared error estimation in transformed fay–herriot models. *Journal of the Royal Statistical Society: Series B*, **68**, 239–257.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J. and Bengtsson, L. (2017) Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, **14**, 20160690.

Sugasawa, S. and Kubokawa, T. (2017) Transforming response values in small area prediction. *Computational Statistics & Data Analysis*, **114**, 47–60.

Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C. and Lazer, D. (2015) Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface*, **12**, 20150185.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. and Rojas-Perilla, N. (2018) From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A*, **181**, 927–979.