Utilizing Alternative Data Sources for Official Statistics

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors/einer Doktorin der Wirtschaftswissenschaft des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin

vorgelegt von Till Koebe

aus Berlin

Berlin, 2022

Dekan: Prof. Dr. Dr. Giacomo Corneo
Erstgutachter: Prof. Dr. Timo Schmid
Zweitgutachter: Prof. Dr. Jan Marcus
Tag der Disputation: 7. November 2022

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Dr. Timo Schmid (Otto-Friedrich-Universität, Bamberg). His patience and support has been greatly appreciated throughout the journey.

I am also thankful to Prof. Dr. Jan Marcus (Freie Universität, Berlin) for helping me finalize the thesis in one way or the other.

Special thanks go to the State of Berlin for supporting this thesis through the Elsa-Neumann scholarship.

Furthermore, I am very grateful to all others who have accompanied me on the way to this thesis, especially my co-authors, my friends, my wife and my family.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, ”Utilizing Alternative Data Sources for Official Statistics”

1. Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 1163-1190.

2. Koebe, T. (2020). Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. *PloS one*, 15(11), e0241981.

3. Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., & Schmid, T. (2022). Intercensal updating using structure-preserving methods and satellite imagery. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1– 27.

4. Koebe, T., & Arias-Salazar, A. (2022). Releasing survey microdata with exact cluster locations and additional privacy safeguards. *arXiv preprint arXiv:2205.12260.*

Contents

Introduction

> To this end, we acknowledge the importance of developing sound statistical infrastructures, including through dedicated statistical surveys, appropriate domestic, national and international legal and technical frameworks for data access and use, while protecting personal data and privacy, strengthening of NSOs' capabilities in using linked data, increased availability of open data, and enhanced collaboration with the private sector and relevant stakeholders, including in exploring alternative sources of data and data collection practices. (Declaration of the G20 Digital Ministers, 2021)

To ensure that knowledge is not exclusive to a few, but available for the many, public agencies publish statistics as a public good - also called official statistics. Facing increasing competition about the prerogative of 'facts', statistical agencies are expected to provide more disaggregated, frequent, granular and reliable statistics in a timely manner (MacFeely, 2016). Already today, enormous amounts of data are generated around us frequently as more and more parts of our lives move into the digital realm. Much of this information on human action is linked to place and time. Exploiting those spatio-temporal patterns may help to better understand underlying dependencies and trends in the socio-economic fabric of our society. Consequently, can official statistics leverage this new "data deluge" (Vale, 2011) and live up to the promises of better, more relevant statistics?

In attempts to do so, the use of satellite imagery, mobile phone data, social network data and other new data sources for demographic or socio-economic mapping has drawn much attention in recent years (e.g. Pokhriyal and Jacques (2017); Blumenstock et al. (2015); Leyk et al. (2019) and Fatehkia et al. (2020)). The advantages are obvious: data is already collected at a higher frequency or better geographical coverage than traditional statistical methods such as household surveys. On the other hand, limitations are multiple, too: the data generating processes underlying new data sources are usually neither controlled by national statistical offices nor do they adhere to statistical data collection standards. The consequences are, among others, selection biases that are hard to address, complex error structures that are difficult to pin down and conceptual frameworks that are demanding to reconcile (Pestre et al., 2020). While acknowledging these methodological and political challenges, where can new data sources provide an actual value added to official statistics then?

The thesis contributes to this field of research on applied statistics in two ways: Part I showcases how new data sources can be utilized to improve traditional statistical data collection techniques, notably censuses and surveys, especially in settings with weak national statistical systems. Specifically, Chapter 1 investigates whether the extensive coverage of mobile networks can be used during household survey operations a) to estimate socio-demographic key performance indicators – exemplified using the literacy rate in Senegal – for small areas not considered in the sampling process, and b) to drive down sampling errors for the in-sample areas by ‘borrowing strength‘ from mobile network data. Chapter 2 turns an eye on the often called ‘gold standard‘ in official statistics: the census. It explores whether census updates can overcome the risk of outdated census data in between the decennial collection cycle by utilizing annual, fine-granular subnational population estimates derived from satellite imagery in combination with recent survey data.

In contrast to the application-driven first part of the thesis, Part II focuses on overcoming methodological challenges in augmenting official statistics with new data sources, especially when combining these disparate data sources in the first place. For example, official statistics are usually collected for administrative areas whereas mobile networks work on the level of network cells and satellite imagery on the level of pixels. Chapter 3 therefore looks at different strategies to geographically link statistical data with data from mobile networks on an area-level. A major shortcoming in most area-level matching strategies in official statistics is the uncertainty of the data collection locations in survey data – the survey clusters. Many household survey programs displace the true locations of the survey clusters to protect the respondents’ privacy with the consequence that the more granular the matching level, the noisier the area-level survey estimates – an example of the privacy-utility trade-off official statisticians regularly face when publishing data. Chapter 4 addresses that trade-off by proposing an alternative microdata dissemination strategy that uses two datasets: the original survey microdata with geographical identifiers for large areas only, and a fully synthetic dataset with the true cluster locations.

The presented use cases and methodological contributions are just a few among many steps necessary to pave the way for alternative data sources to be utilized in the business processes of official statistics. But as enterprise data became an important pillar for measuring modern-day economic activity a century ago, as administrative data overhauled long-time census practices only recently, eventually (privately-held) non-statistical data from mobile and social networks, from satellite imagery or other sensors may find their way into mainstream official statistics one day to produce statistics in a more disaggregated, frequent, granular, reliable and timely manner after all.

Part I

Use Cases for Improving Official Statistics with New Data Sources

Chapter 1

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

1.1 Introduction

> If you can't measure it, you can't manage it. (Michael Bloomberg, former Mayor of New York City)

A country's budget can hardly be allocated efficiently, if the country does not know where the money is needed the most. Reliable knowledge on the socio-demographic indicators of a country's population is essential for sound evidence-based policymaking. For instance, the geographic distribution of wealth is used to make decisions regarding the allocation of resources. Traditionally, this knowledge is collected via household surveys and is provided by National Statistical Institutes (NSI). The surveys are generally designed to provide reliable estimates for the indicators only for larger domains such as the national or the regional level. One possible way to derive estimates on spatially disaggregated levels, like municipalities or communes, is by using small area methods (Rao and Molina, 2015). During the last decade there has been a substantial growth in the development and application of model-based small area methods for the estimation of indicators. Examples are manifold in literature: Elbers et al. (2003), Molina and Rao (2010) and Pratesi (2016) used small area techniques for the estimation of poverty indicators and, recently, Lopez-Vizcaino et al. (2015) and Chambers et al. (2016) investigated the estimation of labour force indicators. For a comprehensive review we refer to Pfeffermann (2013) and Rao and Molina (2015). However, the production of precise small area estimates of indicators relies on the availability of predictive auxiliary variables like census or register information. In many countries successive census and national surveys are conducted with long lag times. Both require a well-functioning infrastructure, starting from cars for the interviewers to computers and well-trained personnel for the analysis. With national statistical systems in developing countries often being subject to unstable funding and a lack of human resources, the collection and processing of relevant data imposes a great challenge or often does not exist (Ghosh and Rao, 1994). For instance, in Angola the most recent census before 2014 was conducted in 1970 and the official population grew by more than 400% in that period (Blumenstock et al., 2015).

An alternative to the usage of census information for small area estimation is to investigate different sources of passively collected data like social media sources (e.g. Facebook, Twitter etc.) or mobile phone data. Eagle et al. (2010) used recently social network data to measure economic growth in the UK. Nevertheless, social media data are rare in developing countries whereas mobile phone data are a remarkable exception. The unique subscriber penetration is between $40\% - 55\%$ in developing countries with a share of around 40% in Sub-Saharan Africa (GSMA, 2015).

In this paper we investigate how mobile phone data (in combination with survey data) can be used to predict socio-demographic indicators at regionally disaggregated levels when census information is not available. The motivation is that mobile phone data are collected as a by-product and include valuable information on the timing and frequency of communication events and patterns of location and travel choices (Blumenstock et al., 2015). Eagle et al. (2010) and Deville et al. (2014) showed that spatially aggregated measures of mobile phone usage and penetration have a high correlation with spatially aggregated statistics from censuses. At this point we should make clear that the paper does not discuss whether the socio-demographic indicators can be directly estimated using only the mobile data. We are aware of some important recent work by Blumenstock et al. (2015). The authors predict poverty and wealth by using an individual's past history of mobile phone usage in combination with a phone survey. In our paper we had access to the Demographic and Health Survey (DHS) 2011 and mobile phone data covering the year 2013 in Senegal.

The Republic of Senegal is located in West Africa at the Atlantic Ocean between Mauritania to the North and Guinea-Bissau to the South. At the most Western tip lies Dakar, the country's capital and also the largest city. The set-up of administrative areas in Senegal is complex, but can be divided into four different levels: 14 regions, 45 departments, 123 arrondissements and 431 communes. The total population is estimated at about 13.5 million (2013) and consists of several ethnic groups, e.g. the Wolof or the Serer.

From a methodological point of view the present article uses area-level small area models (Fay and Herriot, 1979) in combination with covariates from alternative data sources. The resulting estimates are benchmarked such that the aggregated small area estimates produce the official national estimate for the country. We also apply transformation to restrict the indicator of interest, for instance the literacy rate, to particular intervals when necessary. However, the idea of alternative covariates is not new in literature. Porter et al. (2014) applied functional covariates extracted from Google in spatial Fay-Herriot models (Pratesi and Salvati, 2009). Recently, Marchetti et al. (2015) give a comprehensive overview how alternative data sources can be used in the context of small area estimation. Nevertheless, none of these papers considered in detail the usage of mobile phone data. To the best of our knowledge, this paper is the first attempt to provide an easily applicable approach for NSIs to model a basket of regionally disaggregated socio-demographic indicators using survey data in combination with mobile phone data. In particular, the paper investigates the usability of mobile phone data, in this case tower-to-tower traffic in Senegal from 2013, for constructing fine granular indicators, like literacy and poverty rates, access to electricity and safe water or religious affiliations. The application here aims at estimating the socio-demographic indicator *literacy rate* for women and men for regionally disaggregated areas because it is a common problem across Africa. From an applied point of view, the paper also discusses the processing, cleaning and handling of the mobile phone data used as additional source of information.

Especially child labour, poverty and poor access to education are common problems across the Africa continent (Ford, 2007). Poverty in developing countries is not only a result of low income, but also of a lack of opportunities to improve the situation (UNESCO, 2015). Literacy is one of the keys to improve people's chances to escape from the lowest poverty levels. Although there are countries with a situation worse than the one of Senegal, the country is only ranked 117th out of 127 countries in the Education for All Development Index (EDI) published by the UNESCO (2012). Especially the literacy rate is quite low compared to other African countries (literacy rate in 2011: 37.8% for women and 60.0% for men (Agence Nationale de la Statistique et de la Démographie, 2012)). The high number of illiterates can be partially explained by historic reason. Senegal was a former French colony until it gained independence from France in 1960. At that point the school attendance of children in the primary school was at 36%, while the country's average literacy rate was around 34% (Schelle, 2013). The origin for this low share of literacy lies in the little interest of the colonial rulers in educating the indigenous people. Other colonial powers in West Africa like Germany (Togoland) or England (Gold Coast, now called Ghana) had a pupils count which was around four times as high as Senegal's count (Schelle, 2013). Concerning the country's literacy rate from 2011, not much has changed in this regard since the withdrawal of the French power in 1960. Another problem of the educational situation is the slow development of a coherent education system due to opposing education concepts with different traditions. The indigenous African concept coexists next to the Islamic and Western concept. Nowadays, if children visit school, they often visit a public school and additionally a Qur'anic school in Senegal. In 2002 a new system emerged, the so called franco-arabic schools. A hybrid form of a bilingual (French and

Arabic) school with a heavy curriculum. Although Villalon and Bodian (2012) predict this franco-arabic schools could be the future and predominant form of public schools, Senegal is after more than 50 years of independence still in the development stage of a coherent education system. The problem is doubtless not only due to a fragmented school system, but also caused by low attendance rates of children at any school. Although primary and secondary education is compulsory and free in Senegal, many parents still do not send their children to school, and drop-out rates are high (Ford, 2007). UNESCO (2012) reported that as the level of education increases especially the enrollment ratios of women in comparison with men strongly decrease. Although Senegal achieved a gender parity in primary education, the disparity for secondary education is even more severe. For every 100 boys attending secondary education in Senegal, only around 79 girls attend (UNESCO, 2012). This is one reason for low literacy rates especially among women. According to UNESCO (2012) more than two million women in Senegal miss skills in basic literacy. Especially in the country's poor regions like Matam and Tambacounda, both located in the East, girls are involved in economic activities and therefore the parents keep the girls out of the school to earn some additional income. Next to economic reasons, gender-based violence, early marriage and pregnancy as well as the traditional role of women in the society are further issues which add to low literacy rates for women (UNESCO, 2012).

The Senegalese government wants to significantly improve the literacy rate, especially for women. For instance, in the early 2000s, the government built community schools and literacy centers for disadvantaged people, like women who missed a basic school education. However, according to the literacy rates for 2011 there is still a large gender disparity and a persisting need to address this issue in Senegal. Organizations like the UNESCO and UNICEF are constantly working on this educational issue and initiated several projects. Currently the Senegalese government and the UNESCO office in Dakar run a project to improve the literacy rate for women (UNESCO, 2015). In particular, the PAJEF project (Projet d'alphabétisation des jeunes filles et jeunes femmes) provides, for instance, access to organized literacy classes and develops training manuals. The project currently runs in seven regions identified by the National Agency of Statistics and Demography (ANSD - Agence Nationale de Statistique et de la Demographie) in Senegal. Further information is available in UNESCO (2015).

So far Senegal belongs to the most successful countries in advancement of gender equality for the enrollment in primary schools, but the national number of illiterate women remains high. All the efforts mentioned above are experimental and not countrywide because of a lack of spatially disaggregated knowledge where more support is needed. To obtain a higher countrywide literacy rate, areas of high illiteracy have to be identified. In this paper we propose an approach for NSIs based on small area estimation for deriving estimates of the share of literates by gender by using mobile phone data for the 431 communes in Senegal. The estimates are used to identify hot spots of illiterate women for the PAJEF project with a need for additional infrastructure.

The structure of the paper is as follows. In Section 1.2 we describe the DHS survey and the mobile phone data including the cleaning and preparation. In Section 1.3 we review small area estimation using Fay-Herriot models. The methodological approach for constructing socio-demographic indicators based on mobile phone data is described and computational details are provided. In Section 1.4 we present the results of the application for the indicator *literacy rate* in Senegal by using the mobile phone data. The performance of the proposed approach is empirically evaluated in a large-scaled design-based simulation in Section 1.5. Finally, in Section 1.6 we conclude the paper with some final remarks and discuss limitations of the proposed approach. Additional results are presented in the supplementary materials.

1.2 Data sources: survey data and mobile phone data

In this section we describe the data sources used in the analysis. In particular, we had access to the Demographic and Health Survey (DHS) 2011 and mobile phone data covering the year 2013 in Senegal. We present details regarding practical implementation of the time-intensive cleaning and preparation of the mobile phone data and discuss the construction of mobile phone covariates.

1.2.1 Demographic and Health Survey

The DHS program collects representative data on population, health, HIV and nutrition in over 90 countries. The data that we use are from the DHS survey 2011 carried out by the ANSD in Senegal. The survey includes a section on the production of socio-demographic indicators on household level and another part on assessing the availability of material and human resources. In particular, the DHS survey consists of three questionnaires: (i) a household questionnaire, (ii) a women's questionnaire and (iii) a men's questionnaire. The household survey collects information on the usual household members including, for instance, gender, age, education, survival of parents, and child labor. Additional information like household characteristics (source of water, availability of electricity, building material and type of toilet), ownership, use of mosquito nets and several health related questions are collected as well. The household survey is also used to identify men and women for the individual questionnaires. The questionnaire for women consists of ten sections covering socio-demographic indicators (like age and date of birth, schooling, literacy, and ethnicity), reproduction, use of contraception, pregnancy, marriage and female genital mutilation. The men's questionnaire is a short version of the questionnaire for women covering socio-demographic characteristics and health related questions. Note as socio-demographic characteristics are only available in the gender-specific questionnaires we focus in the analysis in this paper on the women's and men's questionnaires. For additional information regarding the variables and the questionnaires we refer to Agence Nationale de la Statistique et de la Démographie (2012).

The survey aims to cover the complete country and is based on a stratified two-stage cluster sampling design. The 28 strata are defined by a cross-classification of the 14 regions and rural/urban areas in Senegal. The survey is designed to produce reliable results for most indicators for the 14 regions. In the first sampling stage 391 census districts (147 urban and 244 rural) were drawn with probability proportional to size (number of households in the census districts). In the second sampling stage 21 households were selected with equal probability in each of the 391 census districts which were sampled in the DHS survey. Among the 21 house-

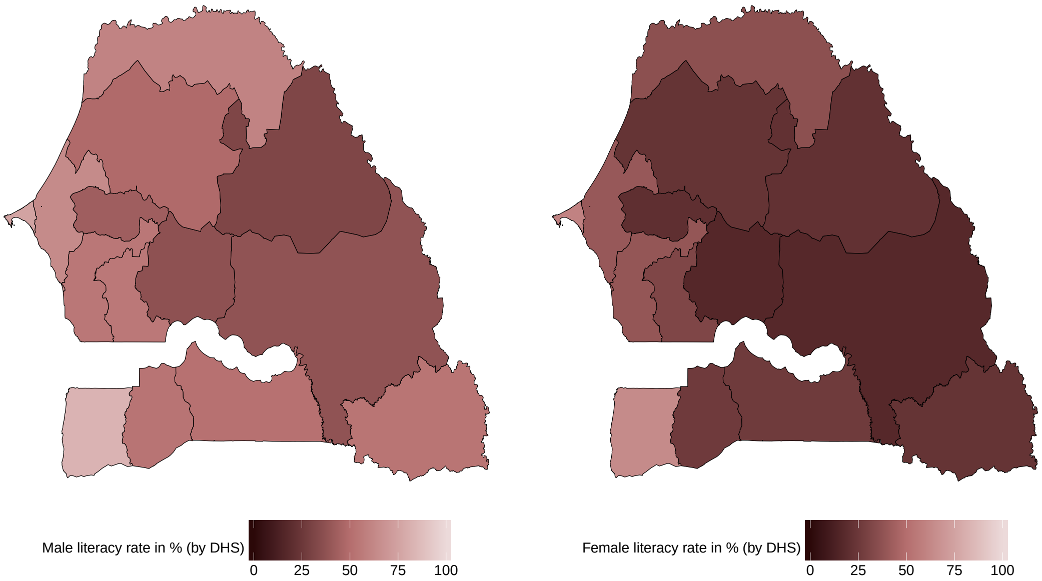

Male literacy rate in % (by DHS)
0
25
50
75
100
Female literacy rate in % (by DHS)
0
25
50
75
100

Figure 1.1: **Estimates for the literacy rate by gender on regional level based on DHS survey 2011.**

holds selected for the women's survey, 8 households were drawn for the men's survey. All men (age between 15-59) and women (age between 15-49) in these households were interviewed. The interview was successfully conducted for 15,688 women (response rate of 92.7 percent) and for 4,929 men (response rate of 87 percent) (Agence Nationale de la Statistique et de la Démographie, 2012).

Figure 1.1 presents results based on DHS survey 2011 of the indicator *literacy rate* by gender for the regions in Senegal. In particular, the variable *literacy* is collected by four different categories in the DHS survey. The categories 'able to read only parts of sentence' and 'able to read whole sentence' are grouped as 'literate'. The answers 'blind/ visually impaired', 'cannot read at all' and 'no card with required language' are categorized 'illiterate'. The initial results indicate that the proportion of *literate women* (37.8%) in Senegal is lower than the proportion of literate men (60.0%). The results are consistent with the official published results of the Agence Nationale de la Statistique et de la Démographie (2012).

As the ANSD aims to estimate socio-demographic indicators for the 431 communes in Senegal, we allocated the information of the DHS survey to the administrative areas (communes). In particular, we had access to the geographical coordinates of the centroids of the 391 census districts. As the actual coverage of the census districts was not available, we matched the centroids of the census districts with the geographical boundaries of the 431 communes. Six out of the 391 census districts were excluded from the analysis because the coordinates of the centroids were missing. Direct survey estimates are only available for 242 out of the 431 communes given the data from the DHS survey 2011. A summary of the commune specific sample sizes for the women's and men's questionnaires is provided in Table 1.1. Figure 1.2 shows direct estimates for the literacy rate by gender on commune level for the capital Dakar (right panel) and for the rest of Senegal (left panel). Communes filled with white color represent areas with zero sample size, so direct estimates based on the DHS survey 2011 are

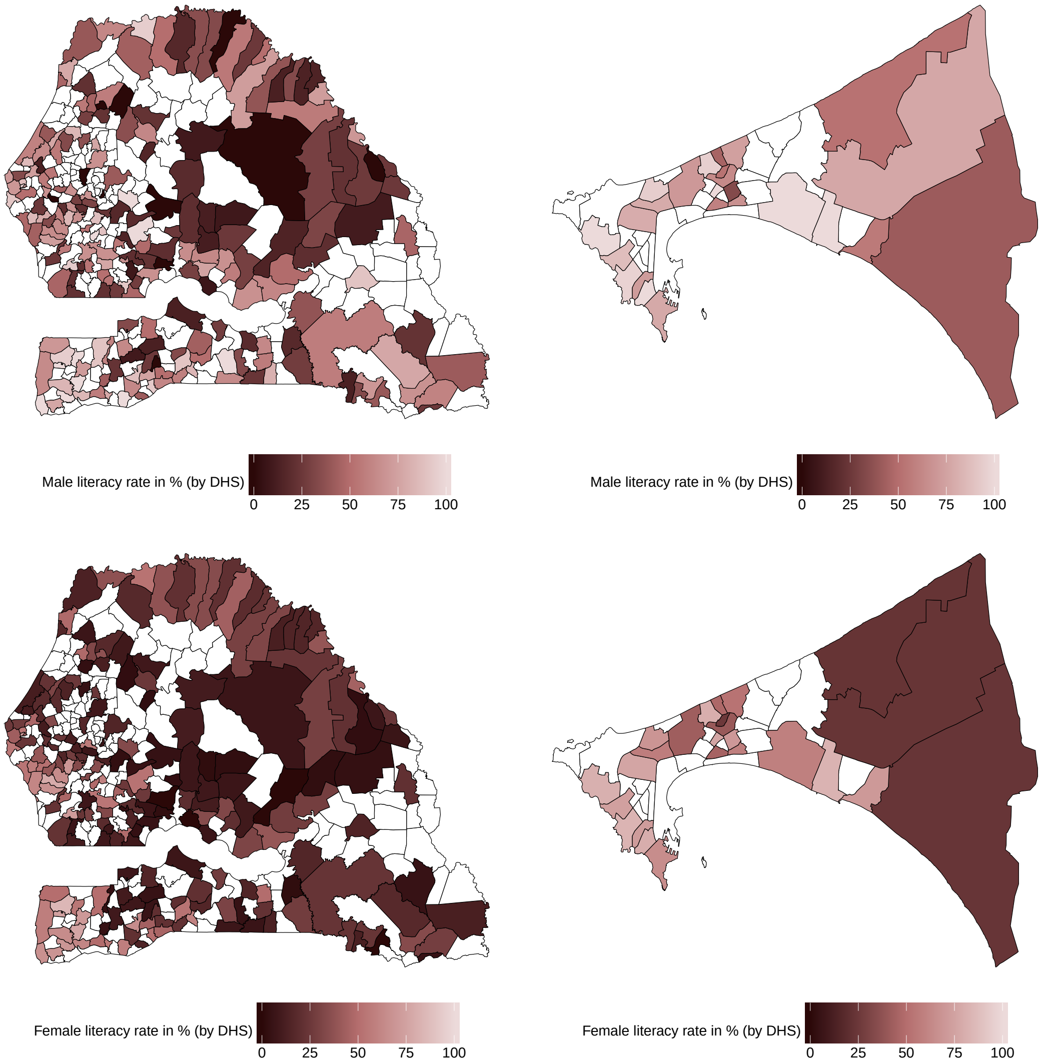

Male literacy rate in % (by DHS)
0
25
50
75
100
Male literacy rate in % (by DHS)
0
25
50
75
100
Female literacy rate in % (by DHS)
0
25
50
75
100
Female literacy rate in % (by DHS)
0
25
50
75
100

Figure 1.2: **Estimates for the literacy rate by gender on commune level based on DHS survey 2011:** Senegal (left panel) and Dakar (right panel).

Table 1.1: **Sample sizes over communes.**

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA
Women's questionnaire	15	35	44	63.90	61	756	189
Men's questionnaire	2	10	14	19.98	20	160	189

not available. The spatial distribution of literacy on commune level is not clearly visible and the identification of hot spots of illiterates with a need for additional infrastructure might be difficult.

The application of small area methods could significantly improve the interpretation of Figure 1.2 by providing results for the communes with zero sample size. This requires fitting of an appropriate model to the survey data. The estimated model parameters are then combined with known population information. The reason that we relied on mobile phone data for predicting socio-demographic indicators is twofold: first, the predictive power of the covariates for socio-demographic indicators from the Senegalese census is limited and second, the ANSD is interested in a widely applicable approach based on the DHS survey for disaggregated indicators independent of census data.

1.2.2 Mobile Phone Data

The mobile phone data used in this paper consist of anonymized call detail records (CDR) from the Senegalese telecommunication company Sonatel covering the year 2013. The dataset is based on more than 9 million unique mobile phone numbers and represents a market share of around 60%. In particular, we had access to the tower-to-tower traffic of all 1666 mobile phone towers in Senegal. In the following we discuss the practical implementation of the processing of the mobile phone data and present details regarding the construction of the mobile phone covariates.

Data processing and cleaning

The preprocessing of the mobile phone raw data is essential and accounts for a considerable amount of time in the whole analysis. The dataset is not *dirty* or *noisy* in the sense of an excessive amount of missing values or illogical recorded values. The data is collected automatically by machines and not gathered by human hand. This means errors in the data are more likely a consequence of machine breakdowns than of human failure.

The traffic of all 1666 towers in Senegal for 2013 is about 1.1 Terabyte of data stored in a cloud system. Because of the massive amount of data, the mobile phone records need to be preprocessed directly in the cloud system. In particular, the raw data is organized in a Hadoop cluster with one separate file by hour per day per month. Hadoop is an open-source software for storing and handling massive data. Each single row contains an interaction and has several characteristics. For example indicating if it is an incoming or outgoing interaction, if it is a phone call or short message service (SMS), which tower received and sent the interaction, or simply the duration of a call in minutes. To process these data we used Apache Hive (Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization) and its SQL logic. MapReduce is applied to create daily, monthly and yearly aggregates of the variables of interest on the cluster. The programming model MapReduce is an implementation for processing large datasets with parallel algorithms on a cluster.

For instance, the aggregated dataset for SMS usage includes the number of incoming and outgoing SMS for every tower on an hourly basis for the year 2013. Table 1.2 shows the head of a preprocessed dataset for the usage of SMS. The first column is the observation indicator which reaches in January 2013 alone around 50 million rows. Variable *DH* tracks the day and hour of a sent SMS; *TO* and *TI* are the tower numbers corresponding to outgoing and incoming, respectively; *E* gives the number of events happening, i.e. SMS being sent. So the first row says that on the 1st of January at midnight there was sent 1 SMS from tower 1 to tower 61. We also had access to the exact geo-coordinate (longitude and latitude) of the towers provided by Sonatel.

Table 1.2: **Structure of the call detail records for SMS.**

	DH	TO	TI	E
1	2013-01-01 00	1	61	1
2	2013-01-01 00	1	340	1
3	2013-01-01 00	1	419	1
4	2013-01-01 00	1	420	1
5	2013-01-01 00	1	447	2
6	2013-01-01 00	1	495	1

Construction of mobile phone covariates

Mobile phone data are measured on tower level on an hourly basis with an excessive amount of observations over the year. To construct variables which can be used as covariates for a statistical model for estimating indicators on commune level, the data needs to be aggregated by two dimensions: time and geographic level. First, in order to reduce the amount of data, the aggregation was done up to the whole year 2013 for each tower. Annual aggregates may disregard sub-annual trends, but since most of the socio-demographic indicators, especially the literacy rate, are time insensitive variables, this fact can be neglected. Second, for having the covariates on the same geographical level like the DHS survey, we used the aggregated (by time) covariates on tower level and averaged them for higher geographic levels like communes or regions. Note as the actual coverage of the mobile towers are unknown, we matched the geo-coordinate of the tower with the geographical boundaries of the 431 communes.

In total we constructed around 70 mobile phone covariates on commune level based on the call detail records. The aggregation routine is done in `R` by using the package `data.table`. The package extends `data.frames` in R based on SQL logic and focuses on fast aggregation of large data (Dowle et al., 2014). For instance, we construct the sum of the number of calls starting from/ending in a specific tower and denote these variables as *outgoing calls* / *incoming calls*, respectively. In addition, we also build the variable *call volume* which sums up the minutes of calls. In the following we label SMS and phone calls together as *events*. For each event we also calculated the *ratios* of the number of outgoing events divided by incoming events. The variable *mean distance* is defined as the average distance in kilometers for an event. In particular, the distance is computed on the tower level by taking the distance of the outgoing tower to the incoming tower for each event and dividing it by the amount of events between the two towers. The covariate *distance-to-dakar* measures the distance from each tower to a centroid of the region Dakar. In addition we construct the variable *isolation* which quantifies the diversity of interactions by users of a tower. The variable is defined for an outgoing tower t_i by

$$Isolation(t_i) = \sum_{\substack{j \neq i \\ j=1}}^{1666} \mathbf{I}_{E(t_i,t_j)}, \tag{1.1}$$

where the indicator function $\mathbf{I}$ is 1 if the condition $E(t_i, t_j)$ is true, i.e. an event happened between the towers t_i and t_j, and 0 otherwise. The variable ranges between 0 and 1666 (total number of towers). We measure the average amount of information an event contains by the variable *Entropy* (de Montjoye et al., 2014). The intuition behind Entropy is that the more unlikely an event is to happen, the more information it contains once it happens. Entropy for a tower t_i is defined by

$$Entropy(t_i) = -\sum_{\substack{j \neq i \\ j=1}}^{1666} p(t_i,t_j) \cdot log\big[p(t_i,t_j)\big], \tag{1.2}$$

where $p(t_i, t_j)$ is the probability of an event between the towers t_i and t_j. In addition, we calculated the *monthly growth* and the *variation* (i.e. variance) of monthly aggregates for the number and volume of events respectively. Variables *Calls-to-dakar* and *sms-to-dakar* reflect the amount of calls or SMS for each tower that were directed to towers located in the capital Dakar. A complete list and description of the covariates is provided in the supplementary materials.

Additionally to the variables described above and in the supplementary materials, we created behavioral indicators based on the mobile phone data with the open-source python toolkit bandicoot (Montjoye et al., 2013). A list of these variables can be found at the bandicoot website. As the bandicoot indicators are constructed for analyzing individual patterns based on the mobile behavior of each single user, we summarized the information to tower level. In particular, a bandicoot indicator on tower level is calculated as a weighted average of all individuals' indicators where this tower was part of the interaction. The steps are as follows: first, we calculated the bandicoot indicators on a monthly level for all single users. Second, we extracted the number of interactions (calls and SMS) during that month for each user and tower combination from the call detail records. Third, we used the number of interactions as a weight to average the individuals' indicators on tower level for each month. Finally, we averaged the monthly values to obtain a yearly indicator for each tower.

First descriptive statistics

Figure 1.3 gives a first impression of the spatial distribution of the 1666 mobile phone towers (red points) in Senegal. The towers are spread over the whole country with higher densities in

Table 1.3: **Mobile phone towers over communes.**

	Min.	1st Qu.	Median	Mean	3rd Qu.	90%	Max.	NA
Number of towers	1	1	2	4.11	4	9	60	30

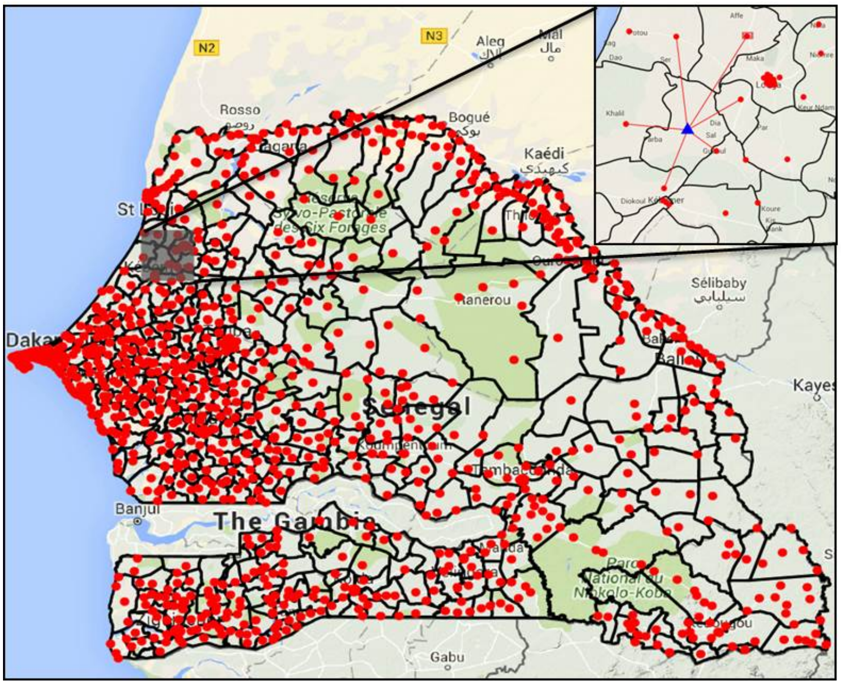


Figure 1.3: **Location of mobile phone towers in Senegal.**

regions with higher population densities. For instance, most of the towers are located in the region of the capital Dakar which itself is located on the Cap-Vert Peninsula on the Atlantic coast in the West. Table 1.3 shows summary statistics of the number of mobile phone towers over the communes. The mean number of towers per commune is 4.1 with a maximum of 60. Although Figure 1.3 suggests a good coverage of the country by mobile phone towers, there are 30 communes without mobile phone towers. Most of these communes are quite small and they are mainly covered by towers which are close-by. For instance, the map at the top on the right of Figure 1.3 shows the area around the commune Badegne Ouolof without tower information. Badegne Ouolof is located in north-western Senegal within the Louga Region on a total of around 300 square kilometers. The centroid of Badegne Ouolof is represented by a blue triangle. In order to apply small area estimation methods for the *out-of-covariate* communes, the covariates are constructed by inverse distance weighting from neighboring mobile towers. In particular, the assigned covariates to *out-of-covariate* communes are calculated by a weighted average of the covariates available at known tower locations. We used the Euclidian distance function and a power parameter of 2 in the weighting.

1.3 Description of the small area estimation method

In this section we describe the methodological approach for constructing socio-demographic indicators based on mobile phone data. Since our aim is to provide an easy-applicable approach for the production of official statistics, especially for the ANSD in Senegal, we apply relatively simple small area estimation methods and correct for misspecifications by adjustments. The implemented approach should meet three conditions:

1. the method should provide estimates for all 431 communes in Senegal;
2. the estimates should be *close* to the direct estimators for communes with *large* sample sizes;
3. the aggregated estimates for the communes should produce the official national estimate for the country.

Note that the Ministry of Chile recently conducted a small area project for the estimation of poverty in Chile based on similar guidelines (Casas-Cordero et al., 2016). In addition the mobile phone covariates are only available on area-level (communes) and it is not possible to link the individuals in the survey with the mobile phone numbers because of confidentiality constraints. Based on the mentioned conditions and available data we considered a benchmarked transformed Fay-Herriot estimator in this paper.

1.3.1 Fay-Herriot estimator

We assume that the population U, consisting of N units, is divided into m disjoint small areas. The sample s is selected from the population by using a complex sampling design. The population is separated into n sampled and $N-n$ non-sampled units, indexed by s and r, respectively. We use the subscript i to indicate the restriction to the area i, for instance, n_i and N_i denote the sample size and the population size in area i, respectively. Let y denote a continuous variable of interest and y_{ij} the response value of unit j in area i and ω_{ij} are the corresponding sampling weights. An estimator for the population mean θ_i of the variable of interest y in area i is given by

$$\hat{\theta}_i^{direct} = \sum_{j=1}^{n_i} \omega_{ij} y_{ij} \Big/ \sum_{j=1}^{n_i} \omega_{ij}. \tag{1.3}$$

The area level model proposed by Fay and Herriot (1979) (hereafter FH model) links the direct estimates with area-level covariates. The FH model is based on two stages:

$$\text{Sampling model (first stage)}: \quad \hat{\theta}_i^{direct} = \theta_i + \varepsilon_i \tag{1.4}$$

$$\text{Linking model (second stage)}: \quad \theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i, \tag{1.5}$$

where $\boldsymbol{x}_i^T$ and $\boldsymbol{\beta}$ denote the $(k \times 1)$ vectors of area-level covariates and regression parameters, respectively. The sampling errors are assumed to be normally distributed and independent with $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon_i})$. The random effects u_i are assumed to be independently normally distributed with $u_i \sim N(0, \sigma^2_u)$. For additional details we refer to Rao and Molina (2015). The combination of both models leads to an area-level linear mixed model given by

$$\hat{\theta}_i^{direct} = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i + \varepsilon_i. \tag{1.6}$$

Let $\hat{\boldsymbol{\beta}}$ define the empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$ and $\hat{u}_i$ the empirical best linear unbiased predictor (EBLUP) of u_i (Henderson, 1950; Searle, 1971), where the variance component σ_u^2 can be estimated by maximum likelihood or residual maximum likelihood (Datta and Lahiri, 2000; Rao and Molina, 2015). The EBLUP of θ_i under the FH model is obtained by

$$\begin{aligned} \hat{\theta}_i^{FH} &= \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i && (1.7) \\ &= \hat{\gamma}_i \hat{\theta}_i^{direct} + (1 - \hat{\gamma}_i) \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}, && (1.8) \end{aligned}$$

where $\hat{\gamma}_i = \hat{\sigma}_u^2 (\hat{\sigma}_u^2 + \sigma_{\varepsilon_i}^2)^{-1}$ denotes the shrinkage factor for area i and $\hat{u}_i = \hat{\gamma}_i (\hat{\theta}_i^{direct} - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. In practice, many of the small areas may have zero sample sizes, so a direct estimator is not available. In this case we rely on synthetic estimation as follows (Rao and Molina, 2015):

$$\hat{\theta}_{i,out}^{FH} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}. \qquad (1.9)$$

The MSE of the EBLUP in (1.7) can be obtained by analytic solutions following Prasad and Rao (1990) and Datta et al. (2005). Note that Li and Lahiri (2010) pointed out that standard estimation methods of the variance component in the Fay-Herriot model can produce zero estimates of the strictly positive model variance. Standard methods for the estimation of the variance component considered in the literature are, for instance, the Prasad-Rao method-of-moments estimator (Prasad and Rao, 1990), the Fay-Herriot method-of-moments estimator (Fay and Herriot, 1979), the maximum likelihood estimator or the residual maximum likelihood estimator. As a consequence, the EBLUP estimator (1.7) can reduce to a regression estimator, which can have an overshrinking problem. Li and Lahiri (2010) propose an adjusted maximum likelihood estimator of the variance component. In particular, an adjusted likelihood of σ_u^2 is defined by

$$L_{adj}(\sigma_u^2) = \sigma_u^2 \times L(\sigma_u^2), \qquad (1.10)$$

where $L(\sigma_u^2)$ can be either the profile likelihood function or the residual likelihood function. Under certain regularity conditions, the adjusted maximum likelihood estimator of σ_u^2 is consistent for a large number of areas m and the shrinkage factors, γ_i, are all strictly greater than 0, even for small m, and are also consistent for large m (Li and Lahiri, 2010). From a Monte-Carlo simulation study carried out by Li and Lahiri (2010) results that in terms of bias and mean squared error, the adjusted maximum profile likelihood method turns out to be better than the adjusted maximum residual likelihood approach. For this reason, we use the adjusted profile likelihood function for estimating the value of σ_u^2 in the paper. Note that Yoshimori and Lahiri (2014) recently proposed an improvement to the adjusted likelihood estimators of Li and Lahiri (2010), showing better performance in a simulation study when σ_u^2 is small relative to the sampling variance $\sigma_{\varepsilon_i}^2$.

1.3.2 Transformed Fay-Herriot estimator

Some socio-demographic indicators are restricted to a specific range. For instance, the share of literates in an area i should be within the interval $[0, 1]$. However, there is no guarantee that the FH estimates produces estimates in a particular range. In the context of estimating small area proportions Jiang and Lahiri (2001), Liu et al. (2014), Bell and Franco (2015), and others present different modeling options for the linking and sampling distribution for area-level models. In particular, Ha et al. (2014) propose a normal-logistic model (NL) with a logistic distribution for the linking model. In addition, they extend the model by Liu et al. (2014) to general complex survey designs and denote it as a normal-logistic random sampling variance model (NLRS). The NLRS model captures parts of the uncertainty due to the estimation of the small area sampling variance. For additional details we refer to Ha et al. (2014). Following Carter and Rolph (1974), Jiang et al. (2001) and Raghunathan et al. (2007) we use in this paper an arcsine transformation for the modeling. Let y now denote a binary variable of interest and y_{ij} is the 0-1 response value of unit j in area i. The steps of the estimation are as follows:

1. Transform the direct estimator via $\vartheta_i = f(\hat{\theta}_i^{direct}) = \arcsin\sqrt{\hat{\theta}_i^{direct}}$.
2. The sampling variance of ϑ_i is approximated by $\sigma^2_{\varepsilon_i} = 1/(4\tilde{n}_i)$, where $\tilde{n}_i$ stands for the effective sample size (Jiang et al., 2001). In particular, the effective sample size is the sample size divided by an estimate of the design effect.
3. Estimate $\hat{\theta}_i^{FH}\{\vartheta_i, 1/(4\tilde{n}_i)\}$ according to (1.7). $\hat{\theta}_i^{FH}$ is truncated to the interval $[0, \pi/2]$ if necessary .
4. Back-transform the estimator $\hat{\theta}_i^{FH}$ to the original scale via

$$\hat{\theta}_i^{FH,trans} = f^{-1}(\hat{\theta}_i^{FH}) = \sin^2(\hat{\theta}_i^{FH}) \quad \text{for } i = 1, ..., m, \tag{1.11}$$

where $\hat{\theta}_i^{FH,trans}$ denotes the transformed FH estimator.

For constructing the confidence intervals for θ_i we use a parametric bootstrap procedure following Casas-Cordero et al. (2016). See also Chatterjee et al. (2008) and Li and Lahiri (2010). The steps are as follows:

1. For given $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\gamma}_i$ estimated with the transformed direct estimator ϑ_i, sampling variance $1/(4\tilde{n}_i)$ and covariates $\boldsymbol{x}_i$, we generate u_i^* from $N(0, \hat{\sigma}_u^2)$ and ε_i^* from $N(0, 1/(4\tilde{n}_i))$.
2. Using u_i^* and ε_i^* to generate the bootstrap sample,

$$\hat{\theta}_i^{*,(b)} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + u_i^* + \varepsilon_i^* \tag{1.12}$$

and the corresponding bootstrap population parameter

$$\theta_i^{*,(b)} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + u_i^*. \tag{1.13}$$

3. Using the bootstrap sample, we estimate the model parameters in (1.6). Based on the estimated model parameters from the bootstrap sample, we compute the corresponding FH estimator (1.7) in area i, $\hat{\theta}_i^{FH,(b)}$.

4. Calculate the following pivotal quantity:

$$t_i^{(b)} = \frac{\theta_i^{*,(b)} - \hat{\theta}_i^{FH,(b)}}{\sqrt{\hat{\gamma}_i^{(b)}/(4\tilde{n}_i)}} \tag{1.14}$$

5. Repeat steps 1-4 B times.

6. For each area i, calculate the $100\alpha/2$ quantile q_{1i} and $100(1-\alpha/2)$ quantile q_{2i} of $\{t_i^{(b)},\ b = 1,\ldots,B\}$.

7. An approximate $100(1-\alpha)$ confidence interval for θ_i is defined as: (lo_i, up_i), where $lo_i = \hat{\theta}_i^{FH} + q_{1i}\sqrt{\hat{\gamma}_i/(4\tilde{n}_i)}$, and $up_i = \hat{\theta}_i^{FH} + q_{2i}\sqrt{\hat{\gamma}_i/(4\tilde{n}_i)}$. If lo_i is negative, it is truncated to 0 and if up_i is greater than $\pi/2$, it is truncated to $\pi/2$. This truncated confidence interval is defined (lo_i^*, up_i^*). Back-transform the lower and the upper limits (lo_i^*, up_i^*) for each area to obtain the approximate $100(1-\alpha)$ confidence interval: $(\sin^2(lo_i^*), \sin^2(up_i^*))$.

Note that the back-transformed confidence interval can be obtained because the function $\sin^{-1}$ and $\sin^2$ are monotonically increasing functions of the parameters in the ranges of interest (Casas-Cordero et al., 2016). In addition, as the upper and lower bound of the confidence interval depends on the effective sample size $\tilde{n}_i$ in area i, we can only estimate the confidence interval for the in-sample areas. In order to obtain a second-order correct confidence interval for out-of-sample areas, one has to replace Equation 1.14 by the following pivotal quantity:

$$t_{i,out}^{(b)} = \frac{\theta_{i,out}^{*,(b)} - \boldsymbol{x}_{i,out}^T \hat{\boldsymbol{\beta}}^{(b)}}{\hat{\sigma}_u^{(b)}}, \tag{1.15}$$

where $\hat{\boldsymbol{\beta}}^{(b)}$ and $\hat{\sigma}_u^{(b)}$ are estimates of $\boldsymbol{\beta}$ and σ_u based on in-sample b-th parametric bootstrap replicate (Chatterjee et al., 2008). Following Chatterjee et al. (2008), the boundaries of the confidence interval for the out-of-sample areas are given by $lo_{i,out} = \boldsymbol{x}_{i,out}^T\hat{\boldsymbol{\beta}} + q_{1i,out}\hat{\sigma}_u$, and $up_{i,out} = \boldsymbol{x}_{i,out}^T\hat{\boldsymbol{\beta}} + q_{2i,out}\hat{\sigma}_u$, where $q_{1i,out}$ and $q_{2i,out}$ are the $100\alpha/2$ and $100(1-\alpha/2)$ percent quantiles of $\{t_{i,out}^{(b)}, b = 1, ..., B\}$, respectively.

An alternative approach is to apply a jackknife method on the transformed scale proposed by Jiang et al. (2001). In particular, Jiang et al. (2001) consider an arcsine transformation and show that the bias of the jackknife MSE estimator is of order $o(m^{-1})$. Then, the authors approximate the MSE in the original scale for $\hat{\theta}_i^{FH,trans}$ cFHback) by

$$mse[\hat{\theta}_i^{FH,trans}] = f^{-1\prime}(\hat{\theta}_i^{FH})mse(\hat{\theta}_i^{FH}),$$

where $f^{-1\prime}$ denotes the derivative of f^{-1} (defined in Equation 1.11) with respect to $\hat{\theta}_i^{FH}$. $mse(\hat{\theta}_i^{FH})$ is an estimate of the MSE obtained by the jackknife method proposed by Jiang et al. (2001). Future work can be devoted to compare the width of the confidence interval and the coverage rate obtained with the parametric bootstrap (Casas-Cordero et al., 2016) and the jackknife procedure (Jiang et al., 2001).

1.3.3 Benchmarked Fay-Herriot estimators

Although the model-based estimator in (1.11) provides estimates for all communes (small areas) in Senegal, the aggregated estimates on national level can differ from the corresponding direct estimator. Following Datta et al. (2010) we use a benchmark approach to achieve the internal consistency with the direct estimator on national level.

For the FH model proposed in Section 1.3.1, we seek for a benchmarked FH estimator $\hat{\theta}_i^{FH,bench}$ such that

$$\sum_{i=1}^{m} \xi_i \hat{\theta}_i^{FH,bench} = \tau,$$

where

$$\tau = \sum_{i=1}^{m} \xi_i \hat{\theta}_i^{direct}.$$

We define the weights by $\xi_i = N_i/N$. We define the benchmarked FH estimator (Datta et al., 2010) by

$$\hat{\theta}_i^{FH,bench} = \hat{\theta}_i^{FH} + \left(\sum_{i=1}^{m} \frac{\xi_i^2}{\phi_i} \right)^{-1} \left(\tau - \sum_{i=1}^{m} \xi_i \hat{\theta}_i^{FH} \right) \frac{\xi_i}{\phi_i} \quad \text{for } i = 1, ..., m. \tag{1.16}$$

There are several ways to define the weight ϕ_i (Datta et al., 2010). For instance, $\phi_i = \xi_i / \hat{\theta}_i^{FH}$ leads to a ratio adjustment of the FH estimator, where small areas with larger estimates will receive a larger adjustment and vice versa. Another option is to define the weights by $\phi_i = \xi_i / \widehat{\text{MSE}}(\hat{\theta}_i^{FH})$. That means that small areas with higher variability in terms of MSE will receive a larger adjustment.

For the benchmarked transformed FH estimator proposed in Section 1.3.2, we use a *naive* approach where the weights are given by $\phi_i = \xi_i$. Thus, the benchmarked transformed FH estimator results as a constant shift from the transformed FH estimator 1.11 and is given by

$$\hat{\theta}_i^{FH,trans,bench} = \hat{\theta}_i^{FH,trans} + \left(\tau - \sum_{i=1}^{m} \xi_i \hat{\theta}_i^{FH,trans} \right) \quad \text{for } i = 1, ..., m. \tag{1.17}$$

1.4 Application: estimating literacy rates in Senegal

In this section the benefits of using the presented Fay-Herriot-type estimators in combination with mobile phone covariates for the estimation of socio-demographic indicators are illustrated in an application which uses the data from the DHS survey 2011 and the mobile phone data we described in Section 1.2. The application aims at estimating the literacy rate by gender on commune level in Senegal. The analysis is carried out by using the variables *literacy women* and *literacy men* from the gender-specific questionnaires introduced in Section 1.2. The estimates are used to identify hot spots of illiterate women for the PAJEF project with a need for additional infrastructure and financial support from the government.

1.4.1 Model selection and model checking

Before proceeding with the analysis of literacy in Senegal, we discuss the model selection and present some diagnostic plots. As discussed in Section 1.2.2 there are various commune specific covariates available from the mobile phone data. To select reasonable covariates in the context of Fay-Herriot models we followed an approach taken by several authors (Jiang et al., 2001; Ha et al., 2014; Casas-Cordero et al., 2016). In particular, we used the Bayesian information criterion (BIC) based on a linear regression model with $\arcsin\sqrt{\hat{\theta}_i^{direct}}$ as dependent variable and considered only data from the 50% largest communes in terms of sample size for the model selection. The reasons for choosing this particular model selection technique are threefold: First, we implicitly assume that for these communes the sampling variability of the direct estimators is reduced, so standard selection techniques are applicable. Second, we apply the BIC to penalize the model complexity more heavily compared to the Akaike information criterion (AIC) in order to enhance the interpretability of the final model. Third, although we are aware of more complex methods for Fay-Herriot model selection discussed in Marhuenda et al. (2014) we use a simple approach which is efficiently implemented by automatic stepwise selection procedures in standard statistical software. The final model on commune level for the variables *literacy women* and *literacy men* include 7 and 8 mobile phone covariates with an adjusted R^2 of 82% and 76% respectively. Note that we report an adjusted R^2 here proposed by Lahiri and Suntornchost (2015) that accounts for the sampling error variability.

Based on the transformed direct estimates from the DHS survey 2011 and the set of selected mobile phone covariates on commune level we fitted area level mixed models (1.6) by gender. As discussed in Section 1.3 the sampling variances of the direct estimates are approximated by $1/4\tilde{n}_i$ where $\tilde{n}_i$ denotes the sample size divided by the design effect. Following Casas-Cordero et al. (2016), we used the design effect on regional level as an approximation for the design effect on commune level. The reason here is that the variance estimation of the direct estimator is unstable because of a low number of cluster or even not directly possible because only one cluster is nested in some communes. We refer to Opsomer et al. (2012) for a recent discussion on this issue in the context of forestry data.

Table 1.4 reports the design effects of the direct estimators by gender on regional level in Senegal. The estimates are consistent with official results published by the Agence Nationale de la Statistique et de la Démographie (2012) in Senegal and show a high value of the design effect of the direct estimator using DHS survey 2011.

Table 1.4: **Design effects of the direct estimator in Senegal by region.**

Region	Female	Male	Region	Female	Male
Dakar	6.260	2.825	Louga	4.473	3.410
Diourbel	3.186	1.987	Saint Louis	4.584	1.874
Fatick	7.695	2.499	Matam	6.569	3.908
Kaffrine	5.058	2.682	Sedhiou	7.840	3.216
Kaolack	5.153	2.434	Tambacounda	4.281	3.386
Kedougou	2.566	1.962	Thies	5.480	3.227
Kolda	3.434	2.615	Ziguinchor	2.525	2.165

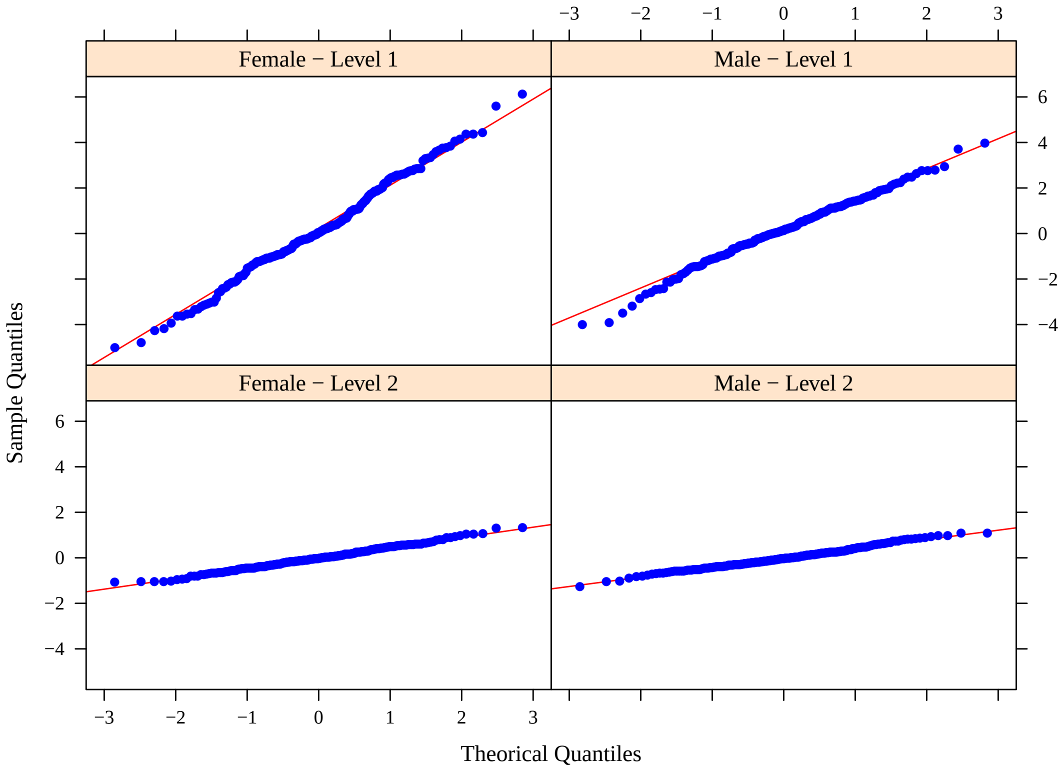


Figure 1.4: **Normal probability plots of standardized residuals (level 1) and the standardized random effects (level 2) for the female model (left panel) and for the male model (right panel).**

Figure 1.4 shows normal probability plots of the standardized residuals (level 1) and the standardized random effects (level 2) obtained from fitting the female model (left panel) and the male model (right panel). The figure indicates some small departures from normality especially in the tails of the distribution. However, the departures are not severe. The Shapiro-Wilk test supports the lack of evidence against the normality assumption for the level 1 standardized residuals (p-values: male model = 0.4453 and female model = 0.4656) and level 2 standardized random effects (p-values: male model = 0.3311 and female model = 0.6059). Using the transformed Fay-Herriot model (1.11) may be advisable for estimating the literacy of women and men.

1.4.2 Small area estimates at commune level

Estimates of the literacy rate by gender for each commune are calculated by using the transformed FH estimator (1.11) (FH Trans) and by the benchmarked transformed FH estimator (1.17) (FH Bench). For constructing the confidence intervals based on the FH Trans we use the parametric bootstrap approach of Casas-Cordero et al. (2016) discussed in Section 1.3. We performed $B = 500$ bootstrap replications. We also include the direct estimator to assess the resulting estimates as the model-based estimators should be consistent with the unbiased direct estimators but with a higher precision. Note that direct estimation is not an option for the DHS survey 2011 on commune level because around 45% of the communes are out-of-sample. The estimators are implemented by computationally efficient algorithms using R. The codes are available from the authors upon request.

Table 1.5 reports the distribution of estimated literacy rates for women and men in the communes in Senegal, split by in-sample, out-of-sample and out-of-covariate communes. Our first observation is that the estimates for female and male literacy rates are higher for the FH Bench compared to the FH Trans, respectively. The reason is that the aggregated FH Trans estimates (women: 36.1% and men: 57.7%) on national level slightly underestimate the national share of literates (women: 37.8% and men: 60.0%) and, thus, need a small adjustment to meet the national estimate for the country.

In order to judge the quality of the model-based FH Trans, we have a closer look to the Figures 1.5 and 1.6. In particular, Figure 1.5 represents the shrinkage factor for the female (left panel) and the male (right panel) model as well as the corresponding sample sizes (dashed lines). On the x-axis, communes are ordered by their sample size (descending order from left to right). We observe that for communes with larger sample size the direct estimator gets substantial weight for both models. In contrast, for communes with a smaller sample size the FH Trans tends to be highly synthetic. Comparing both models we note that the FH Trans for the female model puts in general more weight on the direct estimator than the male model - mean shrinkage factor: 0.262 (female) vs. 0.206 (male) - as a consequence of the larger sample size in the women's questionnaire (cf. Table 1.1). In Figure 1.6 we plot the direct (diamonds) and the FH Trans (dots) estimates of literacy rates against communes ordered by their sample size (descending order from left to right) for the male and female model (top down). For illustration, we show only every fourth commune in the figure. The estimated national literacy rate (based on the DHS survey) is represented by the solid line. The vertical

Table 1.5: **Distribution of the female and male literacy rates over communes in Senegal.**

233 In-sample communes							
Gender	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Female	Direct	0.000	0.105	0.234	0.298	0.474	0.839
	FH Trans	0.002	0.151	0.252	0.296	0.434	0.822
	FH Bench	0.002	0.158	0.264	0.310	0.455	0.861
Male	Direct	0.000	0.250	0.533	0.508	0.720	1.000
	FH Trans	0.062	0.368	0.500	0.516	0.662	0.971
	FH Bench	0.064	0.383	0.520	0.537	0.689	1.000
168 Out-of-sample communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Female	FH Trans	0.005	0.157	0.219	0.271	0.363	0.732
	FH Bench	0.006	0.165	0.230	0.283	0.381	0.766
Male	FH Trans	0.066	0.309	0.468	0.478	0.633	0.960
	FH Bench	0.069	0.322	0.487	0.497	0.659	0.999
30 Out-of-covariate communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Female	FH Trans.	0.130	0.188	0.213	0.227	0.226	0.501
	FH Bench.	0.136	0.198	0.223	0.238	0.237	0.525
Male	FH Trans.	0.346	0.383	0.431	0.444	0.499	0.721
	FH Bench.	0.360	0.398	0.448	0.462	0.519	0.750

lines show the confidence intervals for each commune. Note that we do not report variance estimates for the direct estimator because there was only one sampling cluster nested in most of the communes. We observe that the FH Trans and the direct estimates randomly vary around the national estimate and do not indicate any systematic behaviour to show a possible bias from the modeling. Confirming the findings from Figure 1.5, the direct estimates are very similar to the model-based estimates for communes with a larger sample size. Most of the direct estimates are contained within the confidence intervals for both models. The length of the confidence interval are larger for the male model than for the female model. Due to the shrinkage the FH Trans estimates tend to be more stable around the national estimate than the direct estimator. The variability of the direct estimates and the length of the confidence intervals increase as we move from left to the right side of the figure.

1.4.3 Literacy rates by gender in Senegal

Having assessed the results of the estimators from a statistical perspective, we now discuss the results in the context of female and male literacy in Senegal. As the required approach for the ANSD should meet the third guideline which is that the aggregated estimates for the communes should produce the official national estimate for Senegal we focus in the following only on the benchmarked transformed FH. Figure 1.7 shows the estimates for literacy by gender on commune level for the capital Dakar (right panel) and for the rest of Senegal (left panel).

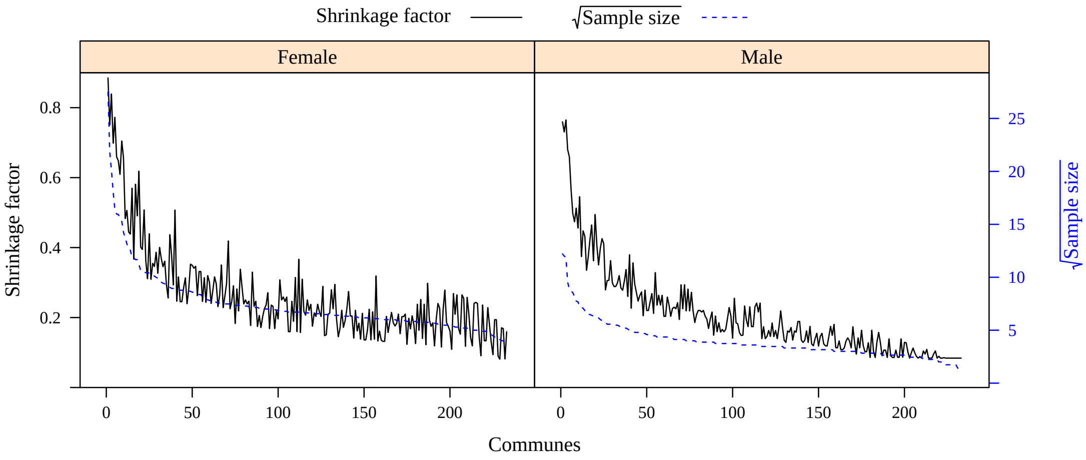


Figure 1.5: **Shrinkage factor for the female model (left panel) and male model (right panel).**

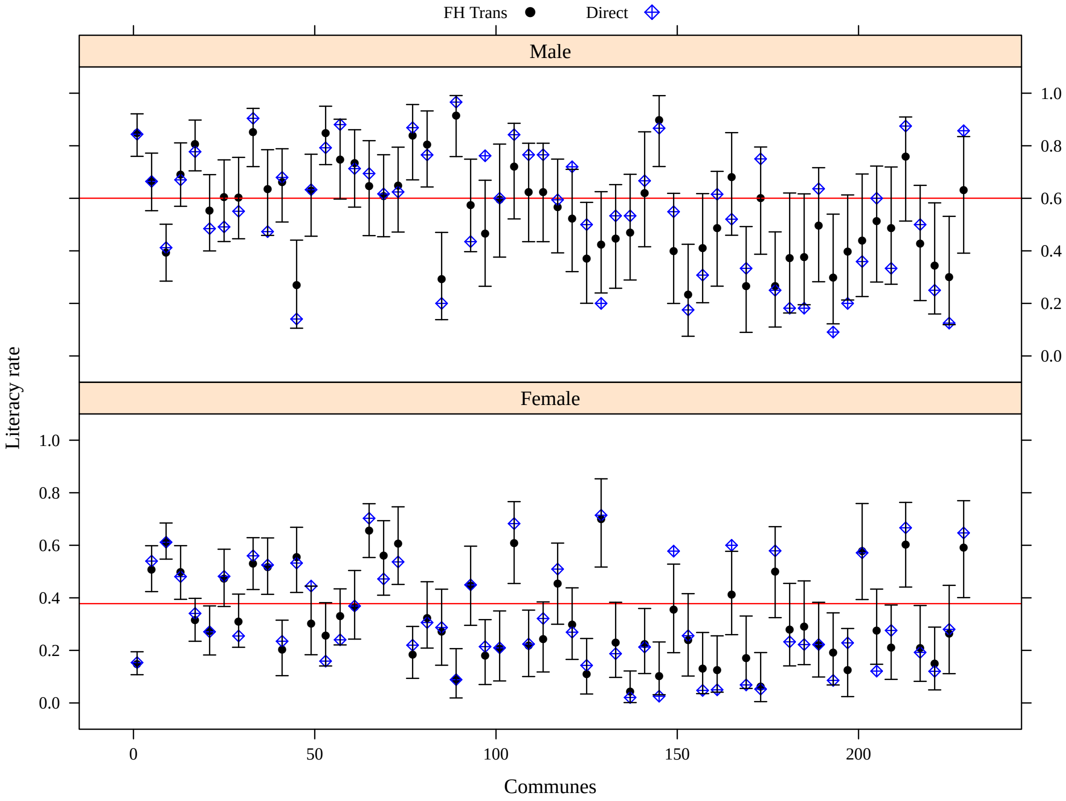


Figure 1.6: **Coverage for the FH Trans for the male and female model (top down).**

In order to simplify the interpretation of the results, Figure 1.7 presents geographical maps for Dakar and for Senegal which are extracted from Google Maps. As a first comment, we note that the relative spatial distribution of male and female literacy rates are very similar in the Dakar region and in the rest of Senegal.

Having a closer look to the Dakar region (right panel) we observe that the coastal area, where the city of Dakar and its harbor are located, shows a very high rate of literates for male and female. This trend continues by moving from the peninsula closer to the main land and is only interrupted by a pocket of lower literacy around the district of Pikine (located to the east of the lake in the middle of Dakar). The district was founded in 1952 by the French colonial government for the former residents of the coastal area around the harbor. Since 1967, it is forbidden by law to build houses on this land because of problems with flooding. Today, however, illegal housings of migrant workers and refugees dominate this area, reflected in remarkably low literacy rates. Moving further into the interior of the country, the area gets more rural and the literacy rate shrinks.

We now turn to the estimated literacy rates for the rest of Senegal in Figure 1.7 (left panel). Next to the Dakar region, the region around Ziguinchor below Gambia reveals a high literacy rate for men and women. The high literacy rates can be explained by the strategic position between the countries Guinea-Bissau and Gambia as well as to its closeness to the Atlantic Ocean. Ziguinchor is Senegal's second largest city and it is also the trade center of the Casamance region (area of Senegal south of Gambia including the Casamance river). Another reason is that the Casamance region is ethnically different from the other parts of Senegal. The region consists mainly of Jola people with a strong influence of Christianity whereas the Islam is the predominant religion in most other parts of the country (Heil, 2014). Another finding is that communes closer to the ocean and to borders in the North to Mauritania and in the South to Guinea-Bissau have higher literacy rates for men and women. In contrast, communes located on the boarders to Mali (South-East) and to Gambia tend to have lower ones. As expected, the density of mobile phone towers in Figure 1.3 is higher in communes with higher literacy rates. Rural communes with a low coverage of mobile phone towers seem to have a lower literacy rates in general. Especially the central part of Senegal in the Matam and Tambacounda region reveals high shares of illiterate men and women.

Although the relative distribution is very similar in Senegal, Figure 1.7 reveals clear differences in terms of absolute values. The literacy rate for women is around 20% lower compared to men. Reasons are manifold in Senegal: Especially in poor regions of the country like Matam and Tambacounda in the eastern part of the country, girls are involved in economic activities and therefore the parents keep the girls out of the school to earn some additional income. Next to economic reasons, unsafe and long roads to school, gender-based violence, early marriage and pregnancy, the traditional role of women in the society and the low quality of the education system are further issues which add to low literacy rates for women. The PAJEF project, already mentioned in the introduction, aims to boost literacy among women in Senegal is currently conducted by UNESCO Dakar and the government of Senegal (UNESCO, 2015). The project runs in the seven regions (Dakar, Diourbel, Fatick, Kedougou, Matam, Saint-Louis and Tambacounda) with the lowest literacy rate identified by the ANSD based on the DHS survey.

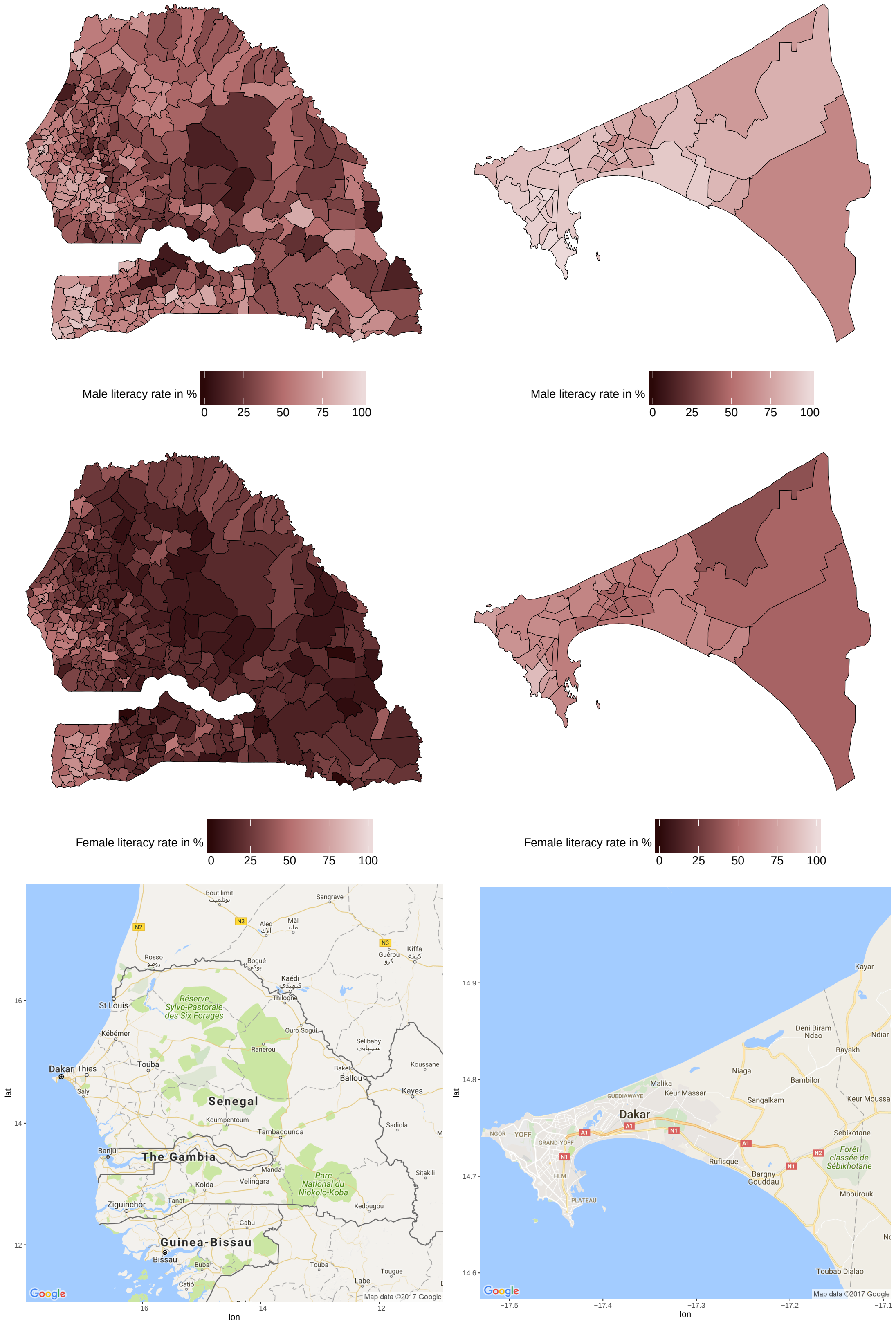


Figure 1.7: **Estimates for the literacy rate by gender on commune level based on a benchmarked FH model:** Senegal (left panel) and Dakar (right panel).

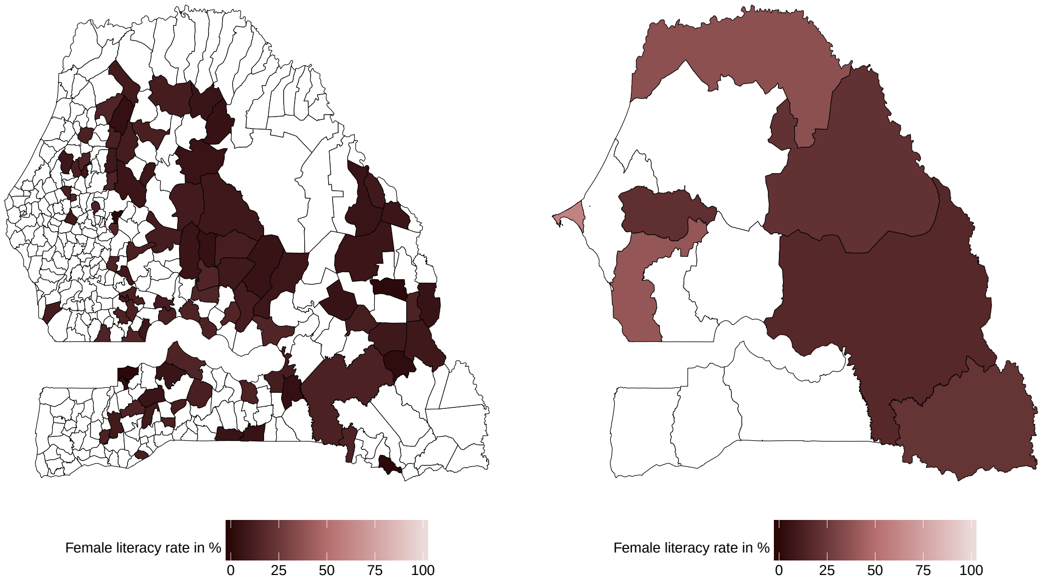


Figure 1.8: **Estimates for the literacy rate for women:** 20% of the communes with the lowest literacy rate (left panel) and seven regions in Senegal identified by the ANSD for the PAJEF project (right panel).

The seven regions and the corresponding literacy rates for women are displayed in Figure 1.8 (right panel). The regions cover around 50% of the country. The left figure shows the literacy rate for women on commune level held by the lowest 20% estimated by using the DHS survey 2011 in combination with mobile phone covariates. There are some hotspots for example in the region around Gambia in the Ziguinchor region or in the Western part of Senegal, with low literacy rates for women but without any financial support. In contrast, the PAJEF project provides financial support to the Saint-Louis region in the north of Senegal or to Dakar where the female literacy rates are above average.

Hence, the use of the proposed approach may enable NSIs and governmental organisations to make sound strategic decisions regarding the best places for investing in creating infrastructure for education. Figures for the indicators *no school education* or *secondary school education or higher* are available from the authors upon request.

1.5 Design-based simulation for unemployment

The analysis of literacy rates by gender in Section 1.4 was sample specific which makes conclusions about efficiency and bias difficult. In this section, we present results from a design-based simulation study that was carried out for assessing the performance of the introduced methodology we discussed in Section 1.3. The aim of the design-based simulation is to investigate the behaviour of the Fay-Herriot type models for estimating socio-demographic indicators based on mobile phone covariates in a controlled environment. In particular, for the evaluation of the approach we had access to the variable *unemployment* from the Senegalese register by collaborating with the staff of the ANSD.

The *pseudo* population in the design-based simulation is based on data collected from a sample of around 1 million individuals in Senegal. The data was collected by ANSD as part of the census 2013 and is spread across the 431 communes. The *pseudo* population reflects around 10% of the population in Senegal. The variable of interest is defined by 0 = employed and 1 = unemployed. Summaries of the population sizes and unemployment rate over communes are given in Table 1.6. Given the fixed *pseudo* population we independently drew $T = 500$ samples following a sampling design similar to the one of the DHS survey. The design is a stratified two-stage cluster sampling design, with the 431 communes as primary sampling units (PSUs). Similar to the DHS survey, we used 14 strata corresponding to the 14 regions of Senegal. In the first sampling stage we selected communes within each stratum with a probability proportional to their size. Around 2% of the individuals within each selected commune are drawn using equal probability systematic sampling. This leads to a sample size of around 15,543 individuals with 196 in-sample communes and 235 out-of-sample communes similar to the women's questionnaire ($n = 15,688$) in the DHS survey (cf. Table 1.1). The summary statistics of the sample sizes over communes are also provided in Table 1.6.

Table 1.6: **Summary statistics over communes.**

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA
Population size	82	717	1303	2257.0	2373	56670	-
Unemployment rate	0.274	0.488	0.550	0.555	0.617	0.898	-
Sample size	3	28	48	79.3	81	1448	235

We investigate the estimators presented in Section 1.3 under repeated sampling performance for the unemployment rate on commune level in Senegal using aggregated mobile phone covariates. To do so, we used an area-level linear mixed model (1.6). The covariates were selected by using the Bayesian information criterion (BIC) and held fixed for the simulation study. In particular, we considered only data from communes with a sample size of more than 30. Like in the application, we implicitly assume that for these communes the sampling variances of the direct estimators are negligible and standard regression model selection tools are applicable. We refer to Jiang et al. (2001) and Ha et al. (2014) for a similar approach for the model selection. The adjusted R^2 by Lahiri and Suntornchost (2015) was on average around 47% depending on the selected sample. We evaluate four estimators for the unemployment rate in the communes in the simulation. These are the direct estimator (1.3), the transformed FH estimator based on an arcsine transformation (1.11) (FH Trans) as well as the normal-logistic model (NL) and the normal-logistic random sampling variance model (NLRS) proposed by Ha et al. (2014). The direct estimator and FH Trans are implemented by computationally efficient algorithms using R. The NL and NLRS are implemented by using JAGS with three parallel chains, each with 20000 iterations, a burn-in of 10000 and the samples were thinned by a factor of two (Liu et al., 2014). The codes are available from the authors upon request. Additionally, we also assess the benchmarked transformed FH estimator (1.17) (FH Bench), the benchmarked NL estimator (NL Bench) and the benchmarked NLRS estimator (NLRS Bench). Note that we also use the *naive* benchmarking approach introduced in Section 1.3.3 for the NL and NLRS estimators.

The performance of the estimators is assessed by the bias (Bias) and root mean squared errors (RMSE) given by

$$\text{Bias}(\hat{m}_i) = \frac{1}{T}\sum_{t=1}^{T}(\hat{m}_{ti} - m_i)$$

$$\text{RMSE}(\hat{m}_i) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{m}_{ti} - m_i)^2},$$

where $\hat{m}_i$ is a generic notation to denote an estimator of the share in commune i and m_i denotes the true population share in commune i.

The results presented in Table 1.7 are splitted by the 191 in-sample, the 210 out-of-sample and the 30 out-of-covariate communes. The table reports summary statistics of the RMSE and Bias of the estimators (FH Trans, NL, and NLRS) over communes. The results confirm our expectations regarding the performance of the estimators. The direct estimator is almost unbiased but suffers from a higher RMSE compared to the model-based approaches (FH Trans, NL, and NLRS) for the sampled communes. The performance of the FH Trans and NLRS is very comparable regarding Bias and RMSE for the in-sample and out-of-sample communes and outperforms the NL estimator in this particular simulation study. For the out-of-covariate communes, where the covariates are obtained by geographically weighting as described in Section 1.2, all model-based estimators (FH Trans, NL, and NLRS) reveal on average a small positive bias.

In order to save space, the corresponding results for the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) are only reported in the supplementary materials. However, the results of the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) and the non-benchmarked estimators (FH Trans, NL, and NLRS) are close in terms of Bias and RMSE because the average of the commune level estimates required only a small adjustment to meet the national estimate for the country. However, note that the benchmarked and the non-benchmarked results are not directly comparable as the FH Bench, NL Bench, and NLRS Bench fulfil the benchmarking constraint.

The results from the study indicate that combining mobile phone covariates with survey data based on model-based estimators can lead i) to gains in efficiency compared to the direct estimator and ii) to reasonable results for communes with zero sample sizes.

1.6 Concluding remarks

Modern systems of official statistics require reliable statistics on socio-demographic indicators on regionally disaggregated levels. These statistics are essential for sound evidence-based policymaking. In this paper we have discussed an easy-applicable approach for NSIs for estimating these indicators by small area methods based on survey data and covariates from alternative data sources. The motivation is to reduce the dependence on census or register information for the NSIs. In particular, we used in this paper passively collected mobile phone data in combination with survey data to predict socio-demographic indicators. Although the paper focuses on literacy rates as specific socio-demographic indicator, the proposed approach is applicable

Table 1.7: **Performance of predictors over communes in design-based simulations.**

191 In-sample communes							
Indictor	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	Direct	0.014	0.053	0.068	0.077	0.091	0.287
	FH Trans	0.016	0.030	0.042	0.053	0.071	0.248
	NL	0.013	0.040	0.049	0.055	0.061	0.260
	NLRS	0.014	0.030	0.041	0.054	0.070	0.251
Bias	Direct	-0.019	-0.002	-0.000	0.000	0.002	0.019
	FH Trans	-0.203	-0.029	0.001	0.001	0.029	0.247
	NL	-0.106	-0.011	0.003	0.009	0.026	0.169
	NLRS	-0.210	-0.029	0.001	0.002	0.029	0.250
210 Out-of-sample communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Trans	0.012	0.030	0.053	0.073	0.104	0.349
	NL	0.011	0.035	0.062	0.076	0.103	0.325
	NLRS	0.010	0.030	0.057	0.073	0.101	0.349
Bias	FH Trans	-0.349	-0.044	0.011	0.007	0.062	0.245
	NL	-0.324	-0.043	0.012	0.013	0.065	0.281
	NLRS	-0.349	-0.043	0.008	0.008	0.059	0.247
30 Out-of-covariate communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Trans	0.010	0.033	0.058	0.079	0.098	0.276
	NL	0.011	0.046	0.075	0.085	0.098	0.280
	NLRS	0.010	0.034	0.059	0.080	0.094	0.279
Bias	FH Trans	-0.173	-0.003	0.043	0.036	0.090	0.276
	NL	-0.153	-0.003	0.048	0.044	0.093	0.280
	NLRS	-0.170	-0.002	0.038	0.036	0.083	0.279

to general indicators. For instance, we can provide results for two other indicators for women in Senegal: i) *Body mass index below 18.5* and ii) *Current usage of any contraception method.* One interesting approach for further research would be to predict the indicators purely on the mobile phone data and to further reduce the dependency of NSIs on actively collected data like survey or register data. For instance, Blumenstock et al. (2015) predicted poverty by using an individual's past history of mobile phone usage in combination with a phone survey. One could extend these results to different indicators. Furthermore, mobile phone data can be used to update the small area estimates in the period between surveys. This would save considerable money, but would require additional assumptions about the model remaining constant between surveys.

For the combination of the survey data and the mobile phone covariates we used an easy-applicable FH small area method for the modeling. Additionally, we have investigated more complex extensions like the spatial FH (Pratesi and Salvati, 2009), the non-parametric FH (Giusti et al., 2012) and the spatial non-stationary FH (Chandra et al., 2015), but the results were comparable. One limitation of our modeling is the approximation of the sampling variance by $1/(4\tilde{n}_i)$ for the transformed FH estimator. As the derivations of the arcsine transformations are based on large-sample theory, we might expect some deficiencies, especially for communes with small sample sizes. Another limitation of our modeling is a potential back-transformation bias in (1.11) due to the nonlinearity of the arcsine transformation. We noted in Section 1.4.2 that the aggregation of $\hat{\theta}_i^{FH,trans}$ estimates is slightly lower than the national estimate for the male and female model, leading to an upward adjustment from benchmarking. Slud and Maiti (2006) discuss bias-corrected small area estimation formulas for the Fay-Herriot model in the context of a logarithmically transformed data. In case of an arcsine transformation, rather than back-transforming the linear model predictions, one could calculate an additive adjustment for the bias as $E(\hat{\theta}_i^{direct}) - E(\hat{\theta}_i^{FH,trans})$, where $E(\cdot)$ is the unconditional expectation under the model. One additional line of research might be to explore the above mentioned bias correction and to extend the MSE estimation to the adjusted predictors. Another line for further work could be to investigate machine learning approaches like random forest for the prediction of socio-demographic indicators and compare them with small area methods.

We have also presented first discussions regarding the time-intensive cleaning, processing and handling of the mobile phone data and available software. However, this can be only a first step in this direction. From a long-run perspective it is necessary to build platforms with open software/ algorithms for NSIs. The aim of such platforms can be twofold: first, NSIs can use code and software to work with large data sources and, second, NSIs can potentially access passively collected data of private companies in a safe environment.

The use of mobile phone covariates has some drawbacks as well. First, additional uncertainty in the mobile phone data arises from the fact that the coverage of the mobile phone tower differs and is unknown. To the best of our knowledge we are not aware of an established way to handle a potential overlap of tower coverage. Second, landlines and the use of internet-based mobile communication services such as Skype, WhatsApp or Viber may cause distortion in communication patterns. However, for Senegal the distortions may be less strongly because of a stagnating landline penetration rate of 2.8% (GSMA, 2015). In addition, the all-time downloads of messaging applications are extremely low in Senegal compared to other countries (e.g. WhatsApp 124,818 and Viber 95,891 on iOS as of December 18th 2014 - extracted from Priori Data). Nevertheless, some types of users may systematically be excluded. Modelling these users is another avenue for further research.

Supplementary material A

A.1 Additional description: mobile phone covariates

Table A.1 describes the covariates used in the paper. The variables are split by categories to ease the understanding of their calculation and origin. Hourly covariates have been calculated on hourly call detail records, daily covariates on aggregated daily call detail records and so on. The variables in the category *interactions* take every single interaction for the year 2013 into account. The covariates are first calculated on a tower level for the year 2013 and then the median is applied for the higher geographic levels like communes and regions. For instance, the covariate *ic_sms_work_ratio* for a tower is the ratio of incoming SMS during 9am to 5pm over all incoming SMS for the year 2013 based on hourly call detail records.

Additionally to the variables described in Table A.1 we created covariates with the open-source python toolkit bandicoot (http://bandicoot.mit.edu) (Montjoye et al., 2013). A list of these variables can be found at the bandicoot website.

A.2 Design-based simulation for unemployment

The results presented in Table A.2 split by the 191 in-sample, the 210 out-of-sample and the 30 out-of-covariate communes. The table reports summary statistics of the RMSE and Bias of the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) over communes. The performance of the FH Bench and NLRS Bench is very comparable regarding Bias and RMSE for the in-sample and out-of-sample communes and outperforms the NL Bench estimator in this particular simulation study. For the out-of-covariate communes, where the covariates are obtained by geographically weighting as described in Section 2, all benchmarked model-based estimators (FH Bench, NL Bench, and NLRS Bench) reveal on average a small positive bias. In addition, we point out that the results of the benchmarked estimators (FH Bench, NL Bench, and NLRS Bench) are very similar to the non-benchmarked estimators (FH Trans, NL, and NLRS) because the average of the commune level estimates required only a small adjustment to meet the national estimate for the country.

Table A.1: **Mobile phone covariates.**

Name	Covariate	Description
Distance		
dist2d	distance to Dakar	The distance to the centroid of the Dakar region in kilometers.
calls_dist_mean	average calls distance	The average distance between towers that were involved in call interactions during the year in kilometers.
sms_dist_mean	average SMS distance	The average distance between towers that were involved in SMS interactions during the year in kilometers.
Interactions		
calls_entropy	entropy of calls	The entropy of calls based on tower to tower interactions throughout the whole year.
sms_entropy	entropy of SMS	The entropy of SMS based on tower to tower interactions throughout the whole year.
calls_isolation	isolation of calls	Total number of towers that a tower had call interactions with. The lower this number, the more isolated a tower is assumed to be in terms of calls.
sms_isolation	isolation of SMS	Total number of towers that a tower had SMS interactions with. The lower this number, the more isolated a tower is assumed to be in terms of SMS.
Based on yearly aggregates		
calls_ratio	calls ratio	The ratio of outgoing calls over incoming calls.
sms_ratio	SMS ratio	The ratio of outgoing SMS over incoming SMS.
vol_ratio	call volume ratio	The ratio of minutes from outgoing calls over minutes from incoming calls.
sms2calls_ratio	SMS to calls ratio	The ratio of outgoing SMS over outgoing calls.
calls2d_ratio	calls to Dakar ratio	The ratio of call interactions where a tower inside the Dakar region was involved over all call interactions.
sms2d_ratio	SMS to Dakar ratio	The ratio of SMS interactions where a tower inside the Dakar region was involved over all SMS interactions.
Based on monthly data		
calls_ratio_var	variance of calls ratios	The variance of the monthly ratios of outgoing calls over incoming calls.
sms_ratio_var	variance of sms ratios	The variance of the monthly ratios of outgoing sms over incoming sms.
vol_ratio_var	variance of call volume ratios	The variance of the monthly ratios of outgoing call minutes over incoming call minutes.
Based on daily data		
og_calls_week_ratio	outgoing calls week ratio	The percentage of calls being initiated during the weekend.
og_sms_week_ratio	outgoing SMS week ratio	The percentage of SMS being sent during the weekend.
og_vol_week_ratio	outgoing call volume week ratio	The percentage of minutes from outgoing calls during the weekend.
ic_calls_week_ratio	incoming calls week ratio	The percentage of calls being received during the weekend.
ic_sms_week_ratio	incoming SMS week ratio	The percentage of SMS being received during the weekend.
ic_vol_week_ratio	incoming call volume week ratio	The percentage of minutes from incoming calls during the weekend.
Based on hourly data		
og_calls_work_ratio	outgoing calls work ratio	The ratio of outgoing calls during 9 am to 5 pm over all outgoing calls.
og_sms_work_ratio	outgoing SMS work ratio	The ratio of outgoing SMS during 9 am to 5 pm over all outgoing SMS.
og_vol_work_ratio	outgoing call volume work ratio	The ratio of minutes from outgoing calls during 9 am to 5 pm over all outgoing minutes.
ic_calls_work_ratio	incoming calls work ratio	The ratio of incoming calls during 9 am to 5 pm over all incoming calls.
ic_sms_work_ratio	incoming SMS work ratio	The ratio of incoming SMS during 9 am to 5 pm over all incoming SMS.
ic_vol_work_ratio	incoming call volume work ratio	The ratio of minutes from incoming calls during 9 am to 5 pm over all incoming minutes.
og_calls_peak_ratio	outgoing calls peak ratio	The ratio of calls being initiated between 3 to 5 am (early peak) over calls being initiated between 10 am to 12 pm (late peak)
og_sms_peak_ratio	outgoing SMS peak ratio	The ratio of SMS being sent between 3 to 5 am (early peak) over sms being sent between 10 am to 12 pm (late peak)
og_vol_peak_ratio	outgoing call volume peak ratio	The ratio of minutes from outgoing calls between 3 to 5 am (early peak) over minutes of outgoing calls between 10 am to 12 pm (late peak)
ic_calls_peak_ratio	incoming calls peak ratio	The ratio of calls being received between 3 to 5 am (early peak) over calls being received between 10 am to 12 pm (late peak)
ic_sms_peak_ratio	incoming SMS peak ratio	The ratio of SMS being received between 3 to 5 am (early peak) over SMS being received between 10 am to 12 pm (late peak)
ic_vol_peak_ratio	incoming call volume peak ratio	The ratio of minutes from incoming calls between 3 to 5 am (early peak) over minutes of incoming calls between 10 am to 12 pm (late peak)

Table A.2: **Performance of benchmarked predictors over communes in design-based simulations.**

191 In-sample communes							
Indictor	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.017	0.030	0.042	0.053	0.069	0.254
	NL Bench.	0.014	0.040	0.049	0.056	0.060	0.262
	NLRS Bench.	0.015	0.029	0.043	0.053	0.070	0.256
Bias	FH Bench.	-0.196	-0.023	0.006	0.007	0.035	0.253
	NL Bench.	-0.103	-0.009	0.005	0.012	0.028	0.171
	NLRS Bench.	-0.204	-0.023	0.005	0.006	0.035	0.255
210 Out-of-sample communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.009	0.032	0.055	0.074	0.104	0.344
	NL Bench.	0.009	0.035	0.061	0.076	0.104	0.322
	NLRS Bench.	0.008	0.031	0.056	0.073	0.103	0.344
Bias	FH Bench.	-0.343	-0.039	0.017	0.013	0.068	0.252
	NL Bench.	-0.321	-0.040	0.014	0.015	0.067	0.284
	NLRS Bench.	-0.344	-0.038	0.013	0.013	0.064	0.253
30 Out-of-covariate communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Bench.	0.010	0.036	0.064	0.081	0.102	0.282
	NL Bench.	0.010	0.043	0.077	0.086	0.101	0.282
	NLRS Bench.	0.010	0.038	0.064	0.082	0.100	0.284
Bias	FH Bench.	-0.168	0.003	0.049	0.042	0.095	0.282
	NL Bench.	-0.150	-0.001	0.051	0.046	0.096	0.282
	NLRS Bench.	-0.165	0.003	0.044	0.042	0.090	0.284

Chapter 2

Intercensal updating using structure-preserving methods and satellite imagery

Part II

Addressing Methodological Challenges in Survey Augmentation

Chapter 3

Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling

3.1 Introduction

Mobile phone metadata has become a popular data source to complement official statistics. When an individual makes a call, sends a message or uses the mobile internet, meta information about this interaction, such as the time stamp and the location, are stored in a database of the mobile network operator (MNO). Researchers exploit those spatio-temporal references for geo-located analysis. One string of research in this field investigates the question whether a certain characteristic such as poverty, literacy or food insecurity is reflected in mobile phone behaviour. Matching this behaviour accurately to a 'groundtruth' - often statistical data from surveys or censuses provided for statistical areas - however, poses a major challenge as the two data sources lack a common reference. In the case of call detail records (CDRs), the geographic reference is provided by the antenna location, often stored as a point coordinate of the physical location of the corresponding base transmitter station (BTS). Due to its simplicity, some scientific literature treat antennas as point coordinates (Schmid et al., 2017). However, the interactions captured by the antenna do not happen entirely at this exact coordinate, but within the coverage area of the antenna - the cell. While an antenna may be located in one statistical area, most of the cell may lie within the neighboring area. The state-of-the-art attempt to address this is to use spatial weights based on the overlapping area size of statistical areas and cells approximated via voronoi tessellation (Pokhriyal and Jacques, 2017; Blumenstock et al., 2015). This approach has three major drawbacks: First, voronoi tessellation perfectly divides the space around BTS locations depending on the distance to the surrounding BTS. This represents a naïve approximation of the true coverage areas as it does not take overlaps, areas without coverage and additional network complexities (multiple antennas per site/BTS, directionality of antennas, varying frequency bands etc.) into account (Ricciato et al., 2017). For example, roughly 90 million people in Africa in 2019 were still not connected to any mobile network hinting at major holes in the coverage (The Economist Intelligence Unit, 2019). Second, even though the concept of 'home-locating' subscribers to specific BTS offers a network-based alternative to the statistical concept of 'usual place of residence', it is not reflected within cells. As the weights are based on area sizes, the voronoi tessellation implicitly assumes that individuals/households are homogeneously distributed within cells, which in most cases does not hold true. For example, a lake would receive the same importance in the creation of area-level mobile phone metadata aggregates as an equally sized built-up area. Third, as mobile stations (MS, generally defined as a combination of device and SIM card) and antennas communicate via modulated radio signals whose propagation paths depend on a range of factors such as the weather, coverage areas are stochastic by nature. More elaborate approaches to model coverage ranges of mobile networks exist (Ricciato et al., 2017; Phillips et al., 2013), especially in the field of radio propagation modelling native to electrical engineering, however, they often require detailed information on the area's topology, a number of technical details concerning the network infrastructure and additional information from passive monitoring systems, which mobile network operators are generally highly reluctant to share and in the latter case often not capable to collect.

3.1.1 Contributions

Acknowledging this, I divide my methodological contribution in this paper in two parts: First, I propose the use of settlement information extracted from publicly available satellite imagery to account for within-cell heterogeneity within the mobile network when linking statistical data with mobile phone metadata. Building on this, the second part of the methodology takes advantage of scenarios where additional technical specifications are available in order to address the issues for holes, overlaps and non-linearities within the mobile network using propagation-based modelling. My main contributions are as follows:

1. The idea of using settlements retrieved from publicly available satellite imagery as a common reference for statistical units such as households and 'home-located' MS in order to calculate weights for mapping mobile phone metadata and statistical data based on settlement counts in scenarios where MS counts are not available. This way, within-cell heterogeneity is addressed.

2. A propagation-based approach to account for overlaps, holes and non-linearities in coverage service provision - in case additional information on the network infrastructure are available.

3. A large-scale simulation study on a synthetic population grid to systematically compare the accuracy of different mapping approaches and their effects on predictive perfor-

mance.

4. A real-world application that demonstrates the impact of the mapping choice on outcomes in later analysis.

3.1.2 Datasets

In the application, I revisit the simulation study of Schmid et al. (2017) published in 2017 in the *Journal of the Royal Statistical Society Series A* on fine-granular unemployment estimates from mobile phone metadata in Senegal in order to investigate the effects of different mapping schemes on the unemployment outcomes. Therefore, I re-run the original simulation with the difference that I implement multiple mapping schemes to derive area-level covariates from CDRs. Specifically, I use behavioural indicators and SIM card counts extracted from CDRs provided by the major Senegalese MNO *Sonatel* in the context of the D4D 2014 challenge for the whole year of 2013 and aggregated on the level of BTS, for which the exact geo-coordinates are also provided (de Montjoye et al., 2014). The behavioural indicators are generated using the popular open-source Python module *Bandicoot* (de Montjoye et al., 2016). Further, I use population counts from the full 2013 general population and housing census (*RGPHAE 2013*) available for the NUTS 4-level of Senegal - the *communes* - on the website of *ANSD*, the National Statistical Office of Senegal. Commune-level unemployment information are generated from a 10 % sample of RGPHAE 2013. Unemployment information in RGPHAE 2013 are self-reported. Geographic information on the administrative boundaries are available for communes and above. The settlement-based weights I present in this paper use data on human settlement areas in Senegal extracted from the Global Urban Footprint (GUF) project (Esch et al., 2017) of the German Aerospace Center (DLR) at a resolution of 0.4 arc seconds, which is approximately 12m x 12m. The GUF project used 180,000 TerraSAR-X and TanDEM-X images collected during the period of 2011 - 2012 (with some data from 2013/14 to fill gaps) to create black and white abstractions where white pixels represent human settlements with a true positive rate (accuracy to correctly detect human settlements) of 85 % on average, with 68 % at lowest and 98 % at heighest. GUF data for Senegal is provided as a single black and white .tif-file with a resolution of 55568 x 39459 pixels (see Fig 3.1). All datasets used in this study are available for research purposes under the conditions of the respective data use agreements.

3.1.3 Related work

Increasing processing capabilities have propelled the use of satellite imagery in official statistics. The United Nations Department of Economic and Social Affairs (2009) recommends using satellite imagery to prioritize and check geospatial processes such as the delineation of enumeration areas during census preparation. It further supports the construction of population grids as a common spatial reference system as proposed by Stevens et al. (2015) and Freire et al. (2016). Various studies have used remote sensing, sometimes in combination with mobile phone metadata, to estimate key statistical indicators such as economic growth (Henderson et al., 2012; Chen and Nordhaus, 2011; Pinkovskiy and Sala-i Martin, 2016), population density (Leyk et al., 2019; Bonafilia et al., 2019; Harvey, 2002; Steinnocher et al., 2019) or poverty

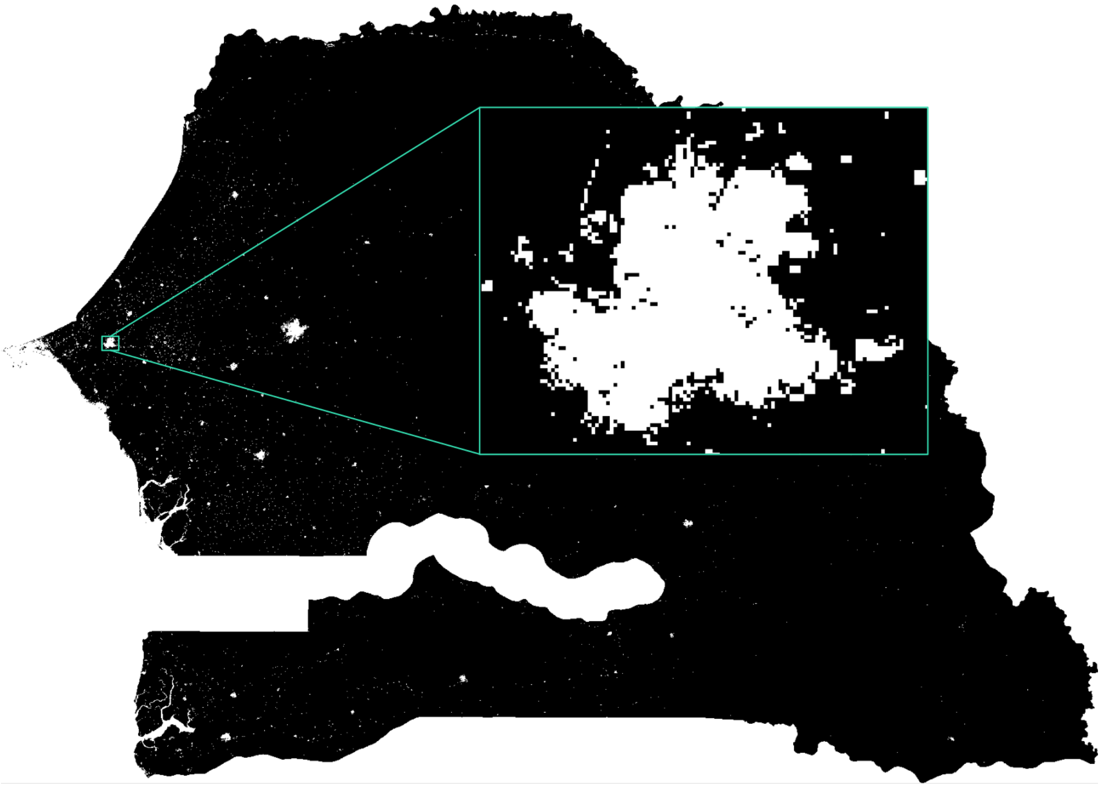


Figure 3.1: **Settlements in Senegal provided as b/w image by the GUF project.** Lower resolution built-settlements extents data reprinted from WorldPop (2018) under a CC BY license, with permission from WorldPop, original copyright 2018, are used in this figure for illustrative purposes.

(Pokhriyal and Jacques, 2017; Jean et al., 2016; Weidmann and Schutte, 2017). Work in that field most closely related to this study uses settlement information extracted from satellite imagery in combination with radio propagation models for application in cost-benefit analysis concerning additional infrastructure investments (Oughton, 2019). While Oughton (2019) also uses population counts from official statistics to estimate the latent demand for mobile services, the author neither investigates the effects of different coverage mapping techniques on the results nor does he use mobile phone metadata for statistical purposes.

In addition, the last decade has seen an impressive amount of research on proposing the use of mobile phone metadata for official statistics foremost in the hope to overcome the limiting relationship of sample size and data collection costs. Blondel et al. (2015) provide an excellent overview on the use of mobile phone metadata that also covers its application for statistical purposes. Use cases to produce more frequent, more granular and/or more timely data on a wide range of statistical topics have been identified. For example, Deville et al. (2014); Khodabandelou et al. (2018); Botta et al. (2015); Douglass et al. (2014) and Ricciato et al. (2017) use mobile phone metadata to investigate population dynamics for more frequent population and tourism statistics. Lu et al. (2012) and Gundogdu et al. (2016) apply the question on the whereabouts of a population to the post disaster setting. Mobility aspects such as commuting and travelling routines have been looked at in more detail by Schneider et al. (2013); Wesolowski et al. (2013); Matamalas et al. (2016); Iovan et al. (2013); Janzen et al. (2018) and Taylor (2016). By exploiting both mobility and (social) network characteristics of mobile phone metadata, Wesolowski et al. (2012); Rubrichi et al. (2018); Tizzoni et al. (2014); Le Menach et al. (2011); Frías-Martínez et al. (2011); Lima et al. (2013) and Park et al. (2018) and Bakker et al. (2019) use mobile phone metadata to model disease spreading and integration, respectively. Mobile usage patterns have been explored to provide fine granular insights on socio-demographic indicators such as multi-dimensional poverty (Pokhriyal and Jacques, 2017; Blumenstock et al., 2015), literacy (Schmid et al., 2017; Sundsøy, 2016) and economic vulnerability (Blumenstock et al., 2018; Bruckschen et al., 2019). While most of these studies have mapped mobile phone metadata and groundtruth data using point-to-polygon allocation or voronoi tessellation, very few studies have applied more elaborate approximation schemes. Ricciato et al. (2017) propose a methodology based on maximum likelihood estimation that uses cell footprints provided by one or multiple MNOs in combination with location data from passive monitoring systems to acquire more accurate measures on the density of MS. The authors run a simulation study on a 100x100m synthetic population grid to compare the proposed methodology against voronoi-based coverage maps. However, the methodology requires very detailed information from the involved MNOs, e.g. on the cell footprints and the signalling data that may prove difficult to acquire in practice (see Section 3.2.1). Further, while the authors rightly assume a multinomial distribution of the MS counts, finding appropriate distributions for the wide range of behavioural covariates appears less trivial. In order to simplify and improve the coverage mapping process, members of the European Statistical System as part of the *ESSnet Big Data* project are currently developing *mobloc* (Tennekes, 2018) - an *R* package that implements the free space path loss propagation model using technical specifications of antennas as input parameters. However, neither Ricciato et al. (2017) nor Tennekes (2018) systematically evaluate different coverage mapping techniques on statistical modelling approaches using real-world data.

3.2 Background

3.2.1 Mobile phone metadata

Mobile networks not only transport data for communication purposes, they also generate data for reasons such as network auditing, billing, maintenance and service provision. Some of this meta information is created in interaction with user equipment such as MS. There are four main caveats of using mobile phone metadata for population statistics in general. All of them have in common that they are active areas of current research. First, the customer base of an MNO constitutes a non-representative population sample with unknown sampling design. The consequences are varying sampling rates, i.e. locally changing market shares and parts of the population being structurally excluded from the sample such as children, elderly and the very poor. Second, the unit of observation - i.e. the MS, device, the SIM card and/or the subscriber - does not perfectly match the unit of interest, which is the individual or household, as phone sharing schemes or multi-SIM uses illustrate. Common approaches to account for these two caveats are calibration and/or reconstructing the sampling design empirically. Third, mobile phone metadata lacks the statistical concept of *usual residence* - a concept frequently used in official statistics to determine the geo-location of an individual/household defined as the place where an individual has lived or intends to live for a period of at least 6 or 12 months (OECD, 2013). Different approaches to approximate the *home location* of an MS exists (e.g. night-time home location defined as the most frequently used cell by an MS between 7pm and 7am during a certain time window), however, the definitions do not map perfectly introducing uncertainty in further analysis (Vanhoof et al., 2020). Fourth, coverage areas cannot be pinpointed as radio propagation is dynamic and stochastic by nature. Propagation models of various complexity exist to provide approximations as coverage ranges can generally vary from couple of hundred meters to over 40km.

Most scientific studies in the context of international development and official statistics use CDRs - logs of interactions such as calls, text messages or internet use containing attributes of the MS, the network and the connection - as a basis for further analysis. The advantages of CDRs compared to other mobile phone metadata such as Visitor Location Registers (VLRs) or other signalling data are threefold: First, they provide fine-grained geographical resolution through cell-level identifiers. Second, they provide information both on the mobility and the (social) network of the MS. Third, CDRs are fairly easy to access and to use in analysis as the storage of essential attributes adheres to global standards such as *3gpp 32.295*. However, in addition to the aforementioned general caveats of mobile phone metadata there are important caveats specific to CDRs: Social network information extracted from CDRs are increasingly incomplete due to a shift towards app-based communication (e.g. Whatsapp and Facebook messenger). Mobility patterns are fragmented as locations are logged only during active MS use - again a case of non-random sampling. Some MNOs are able to extract more detailed information on the location of an MS and its app usage e.g. for geo-fencing purposes or app-based pricing schemes through trilateration of signalling data and deep packet inspection, respectively. This, however, requires specific hardware equipment and software capabilities, which not every MNO has. Consequently, these type of information are rarely available to researchers.

3.2.2 Radio propagation modelling

Radio propagation modelling has been subject to research for decades. Coverage mappings in mobile networks are generally used for network planning purposes (Oughton, 2019; Oughton et al., 2019). Looking at Phillips et al. (2013) is highly recommended as they provide an excellent overview on coverage mapping methods. In general, radio propagation modelling techniques in mobile networks largely focus on estimating the path loss L_p a radio signal incurs en route between a transmitter tx and a receiver rx. Together with the output power of the transmitter P_{tx}, the gains through directivity and efficiency of the involved antennas G_{tx} and G_{rx} and their respective technically-incurred losses L_{tx} and L_{rx}, it defines the *link budget* - the received power P_{rx} usually expressed logarithmically in decibel per milliwatt (dBm).

$$P_{rx} = P_{tx} + G_{tx} + G_{rx} - L_{tx} - L_{rx} - L_p \tag{3.1}$$

Since all RHS parameters except L_p are either known in advance due to the choice of the technical equipment (i.e. G_{tx} and L_{tx}) or hardly observable (i.e. G_{rx} and L_{rx}), I assume $G_{tx} + G_{rx} - L_{tx} - L_{rx} = 0$ in the following, leading to a simplified link budget defined as:

$$P_{rx} = P_{tx} - L_p \tag{3.2}$$

Intuitively, Eq. 3.2 thus states that the signal strength observed on a MS solely depends on the output power of the connected antenna and the loss in signal strength that occurs along the way between antenna and MS. Given the abundance of available models, I follow the guidance of the European Conference of Postal and Telecommunications Administrations (CEPT) on radio propagation simulation for mobile services and opt for the widely popular extended HATA model (Green and Wang, 2002), named after Masaharu Hata, the author of the 1980 landmark study on the ”Empirical Formula for Propagation Loss in Land Mobile Radio Services” (Hata, 1980). It is derived from the COST-231 HATA model (Damasso, 1999), which in turn builds on the original HATA (Hata, 1980) and Okumura model (Okumura et al., 1968). They all have in common that they are empirical models to estimate the median path loss between a transmitter and a receiver based on real-world measurements. The HATA model extends the Okumura model by distinguishing between urban, suburban and rural settings, thus accounting for different levels of mean attenuation due to obstacles and changes in terrain. The COST-231 HATA model increases the frequency range of the original HATA model. The extended HATA model is applicable for settings with frequencies f between 30-3000 MHz, distances d between 0-100km, transmitter heights h_{tx} between 30-200m and receiver heights h_{rx} between 1-10m. The general form of the extended HATA model L_p^{EH} consists of a loss function L for the median path loss and a path loss variation term V drawn from a log-normal distribution that accounts for the stochastic nature of radio propagation[1].

$$L_p^{EH}(f, d, h_{tx}, h_{rx}, env) = L(f, d, h_{tx}, h_{rx}, env) + V(\mu, \sigma, d) \tag{3.3}$$

As an example, I provide the path loss function of the extended HATA model L_p^{EH} for distances above 0.1km outdoor in rural areas for frequencies between 150 and 1500 MHz:

$$\begin{aligned}
L_p^{EH} = 69.6 + & \\
& 46.09 * \log_{10} f - \\
& 13.82 * \log_{10} h_{tx} + \\
& (44.9 - 6.55 * \log_{10} h_{tx}) * \log_{10} d - \\
& (1.1 * \log_{10} f - 0.7) * h_{rx} - \\
& 20 * \log_{10}(h_{rx}/10) - \\
& 20 * \log_{10}(h_{tx}/30) - \\
& 4.78 * (\log_{10} f)^2 - \\
& 40.14 + \\
& V(12, 12)
\end{aligned} \tag{3.4}$$

[1] Since model parameters vary depending on the distance, the expected environment env (indoor/outdoor and rural/suburban/urban) and the frequency, the full extended HATA model is not spelled out in this paper, but can be accessed here: https://ecocfl.cept.org/display/SH/A17.3.1+Outdoor-outdoor+propagation

So, for example, an MS 1m above the ground at a line-of-sight distance of 3km in a rural area to an omnidirectional antenna that is 30m above the ground transmitting at the 900 MHz frequency band would experience a path loss of $L_p^{EH} \approx 118dBm$. Assuming a GSM macrocell with an output power $P_{tx} = 43dBm$ using Eq. 3.2 yields a budget for that link, also known as *received signal strength* (RSS), of $P_{rx} = P_{tx} - L_p^{EH} \approx -75dBm$. As a rule of thumb, signals with RSS values above $-80dBm$ are considered excellent, RSS values below $-110dBm$ point to very poor signals.

3.3 Methodology

Usually, statistical data on individuals or households are geo-located to statistical areas via their respective *places of residence*. Further, unit-level data is aggregated to area-level aggregates using some form of weighting factor such as survey weights. For example, the poverty rate of a region can either be calculated as the share of units classified as *poor* among the interviewed residents of the region multiplied by their sampling weight or via sub-regional poverty rates weighted with the respective sub-regional population counts. However, neither the places of residences nor the weights are generally available on the cell-level of a mobile network (as an equivalent to the sub-region). Hence, they need to be estimated.

In mobile phone metadata analysis, the place of residence of an individual/household is usually approximated with the night-time home location of an MS recorded at the cell-level.

To derive survey weight proxies, for example, point-to-polygon allocation assumes equal weights for all cells point-located within a statistical area. Voronoi tessellation uses the area size of the intersection of voronoi tile and statistical area as weighting factor, i.e. 1 km^2 always conveys the same importance in aggregation, no matter whether it is 1 km^2 of sparsely-inhabited desert or 1 km^2 of a densely-populated city.

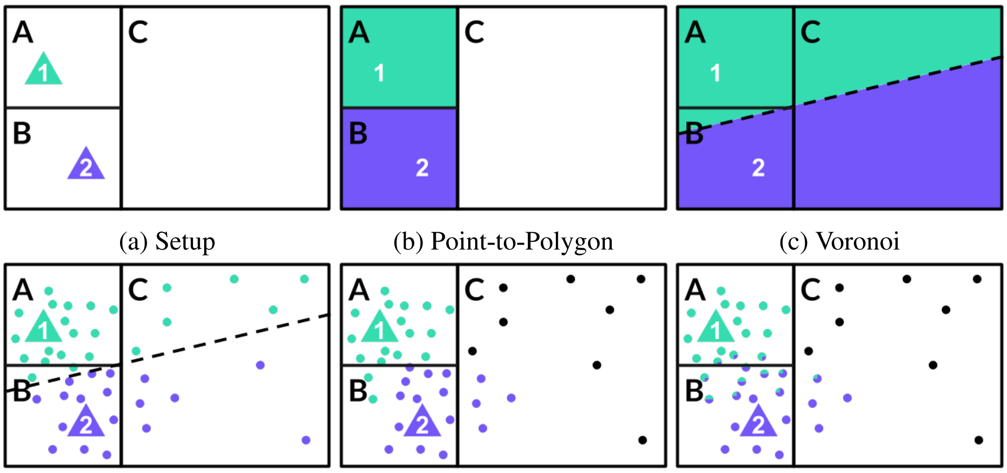
In most cases, the place of residence of an individual/household (thus is approximation alike) is linked to some form of settlement. However, neither the statistical area nor the coverage area of a cell account for that fact. Consequently, the underlying idea behind the proposed methodology is to use human settlement information extracted from publicly available satellite imagery as common geographic reference level for both statistical units such as households and home-located MS. This allows to a) construct weights based on settlement counts and b) refine weights in cases where MS counts, often regarded as highly sensitive information by the MNO, are available. Further, in combination with technical information on the antenna, it allows for an efficient coverage estimation to address the issues of holes and overlaps in a mobile network.

In the following, settlements are denoted as i, BTS as j, statistical areas as t, the number of home-located MS as d, the population count as p, the number of settlements as n and metadata covariates as R. To illustrate the value added of the proposed methodologies, Fig 3.2a and Table 3.1 showcase a typical setup faced when one seeks to augment official statistics with mobile phone metadata: statistical indicators are provided for statistical areas A, B and C. Mobile phone metadata is provided as BTS-level aggregates with the corresponding point locations 1 and 2. To account for that, I treat each cell site that may host multiple antennas as single omnidirectional antenna, calling it *BTS* subsequently. This constitutes a simplification of real mobile networks where usually multiple directional antennas serving on various frequency bands are co-located at the same site that does not necessarily have to be an actual (cell) tower. Although accounting for directionality of antennas as done by e.g. Ricciato et al. (2017) is likely to affect the overall outcome of later analysis by increasing the number of network tiles available for mapping, the challenges for allocating them correctly (holes, non-linearities, overlaps, within-cell heterogeneity) remain. Consequently, it is expected that results from this study also apply to a setup based on directional antennas, thereby justifying the simplifying assumption. Further details on Figs 3.2b - 3.2f are provided in the following subsections.

Table 3.1: **Example of statistical data and mobile phone metadata.**

area_id	poverty_rate
1	0.23
2	0.11

bts_id	# of calls	lon	lat
6453	34050	43.2344	23.2342
8348	1023	50.0988	18.84217



(a) Setup (b) Point-to-Polygon (c) Voronoi

(d) Augmented Voronoi (e) BSA (f) IDW

Figure 3.2: **Popular and proposed mapping schemes.**
Three statistical areas (A-C), two BTS (1-2) and numerous dots representing built-up areas illustrate how different mapping schemes affect the allocation of BTS-level data to statistical data.

3.3.1 Point-to-polygon allocation

For purposes such as model fitting one approach to combine statistical data and mobile phone metadata is to aggregate metadata covariates onto the same geographical level, e.g. statistical areas. To do so, the point-to-polygon approach ($p2p$) treats BTS point locations as such and allocates BTS-level metadata covariates using a binary weighting scheme (see Fig 3.2b and Eq. 3.5).

$$w^{p2p}_{j,t} := \begin{cases} 1 & \text{if } j \subseteq t \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

Consequently, all network traffic handled by a BTS is attributed to one statistical area exclusively, no matter whether it was generated by a home-located MS actually 'residing' in this area or not. In the toy example, but also in the real-world application presented in Section 3.5 this leads to a situation where no metadata covariates are available for certain area, e.g. area C - with negative effects on the final sample size in model fitting.

3.3.2 Voronoi tessellation

In contrast, voronoi tessellation (denoted by superscript v) divides the total space of interest into perfectly disjunct tiles along the equidistant lines between points, in this case the BTS point locations (see Fig 3.2c). The current state-of-the-art procedure is to intersect these tiles - representing approximated coverage areas of BTS - with the statistical areas. The weights to aggregate BTS-level metadata covariates to the respective statistical area are derived from the size of the intersection of tiles a_j and a_t of BTS j and statistical area t, respectively, in relation to the total size of a_t, also expressed as

$$w_{j,t}^{v} := \frac{a_j \cap a_t}{a_t} \tag{3.6}$$

In the toy example of Fig 3.2c, this would reduce to be the intersection of e.g. statistical area A and the voronoi tile of BTS 1 divided by the total area of A. However, as mentioned above, area sizes are used in that approach to approximate the (usually) unknown population counts per intersection by implicitly assuming homogeneous distribution of the population within a given statistical area.

3.3.3 Augmented voronoi tessellation

The proposed settlement-based mapping schemes relax this obviously strong assumption by assuming a homogeneous housing structure instead, i.e. a constant population density per settlement area within a given statistical area. Applied to voronoi tessellation, Figs 3.2c and 3.2d - with settlement areas represented as dots - illustrate the difference. Instead of using the area sizes a_j and a_t to calculate the weights, the "augmented" voronoi tessellation (av) uses the number of settlements per area, denoted as n_j and n_t, respectively.

$$w_{j,t}^{av} := \frac{n_j \cap n_t}{n_t} \tag{3.7}$$

Consequently, statistical area-level covariates can easily be acquired for both approaches using a weighted average (or a weighted median) on BTS-level data.

$$\hat{R}_t = \sum_{j=1}^{J} w_{j,t} R_j \tag{3.8}$$

Going back to the toy example, while BTS 1 covers the smaller part of C in Fig 3.2c, thus receives a smaller weight in the calculation of area-level metadata aggregates, it looks different in Fig 3.2d when comparing the number of settlements, represented by green and purple dots. This way, the proposed methodology accounts for within-cell heterogeneity of the population distribution.

Both voronoi tessellation and augmented voronoi tessellation splits the full space of interest into disjunct tiles. Applied to a mobile network this means ubiquituous coverage and zero redundancies, i.e. all dots are uniquely associated to a specific BTS in the toy example. Again this is a strong assumption that most likely does not hold true in any real-world application. To relax this assumption by introducing holes and overlaps in the network coverage, additional information are necessary that allow for the estimation of coverage measures such as the received signal strength (RSS) at any given point in space. Fig 3.2e exemplifies the consequences: Some settlements are not covered (black dots) and some settlements, even though closer to one BTS, receive a stronger signal from a more distant BTS. Assuming coverages are correctly estimated in Figs 3.2e and 3.2f, it demonstrates that point-to-polygon allocation tends to underestimate the coverage of statistical areas while voronoi tessellation tends to overestimate it.

3.3.4 Propagation-based mapping schemes

Previously presented schemes follow a 'BTS-centric' approach by first determining the respective coverage area of a BTS and then analyzing potential overlaps with other places of interest such as settlements. In contrast, propagation-based schemes follow an 'MS-centric' approach by looking at the connectivity at the place of interest, i.e. the place of usual residence or the home location first and then estimating which (group of) BTS it most likely serves. As outlined in Section 3.2.2, multiple ways exist to estimate the 'connectivity' of an MS, but all require at least information on the distance to the surrounding BTS and additional technical specifications. With that, the serving BTS can be determined at each place of interest, thus allowing for a more nuanced coverage mapping. Here, settlements can provide a common geographic reference for the *place usual residence* and the *home location* alike.

Best server area (BSA)

In mobile networks, an MS usually connects to the antenna that offers the strongest signal. Thus, the settlement-level weight is 1 for the BTS with the strongest signal and 0 otherwise.

$$w_{i,j}^{bsa} := \begin{cases} 1 & \text{if } P_{rx,i,j} = max(P_{rx,i,\cdot}) \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

Links weaker than a certain threshold (e.g. a P_{rx} value below - 110 dBm) can be discarded as they represent 'dead' links. This way the approach accounts for holes in the network coverage. The weights $w_{i,j}$ express the importance of a BTS for a pixel. Similarly to Eq. 3.8, they can be used to determine the statistical area-level covariate estimates $\hat{R}_t$ using a weighted average:

$$\hat{R}_t = \sum_{i=1}^{n_t} \frac{w_{i,j}}{\sum_{i=1}^{n_t} w_{i,j}} R_j \tag{3.10}$$

Due to the binary nature of the weight, $\sum_{i=1}^{n_t} w_{i,j}$ represents the number of settlements with mobile coverage within a given statistical area. In areas with homogeneous network infrastructure and full coverage, the best server approach closely resembles the augmented voronoi tessellation with the difference that path loss increases non-linearly with the distance, i.e. locations very close to the location of a BTS may be served by another, more distant one.

Inverse signal strength

Radio propagation is stochastic by nature. Changing environmental conditions and varying network loads affect the RSS at a given location across time. Consequently, the strongest signal is not always provided by the same BTS. In order to assure quality of service, mobile networks usually exhibit a certain number of overlaps. To account for that, I calculate inverse distance weights (IDW) for each pixel i using the median link budget $P_{rx,i,j}$ as non-linear distance measure (see Eq. 3.11) to the k-nearest antennas. s denotes a tuning parameter, where $s = 0$ reduces $w_{i,j}^{idw}$ to a fixed weight per BTS and a large s can be used to approximate the best server approach.

$$w_{i,j}^{idw} := \frac{v_{i,j}}{\sum_{j=1}^{k_i} v_{i,j}} \qquad \text{with } v_{i,j} := \frac{1}{|P_{rx,i,j}|^s} \qquad \forall j \in k_i \tag{3.11}$$

Here again, $w_{i,j}^{idw}$ can be used to calculate statistical area-level weighted averages of BTS-level mobile phone metadata covariates as presented in Eq. 3.10.

3.3.5 Potential extensions

Depending on data availability, the methodology can further be extended. While MNOs often regard MS counts as highly sensitive information since they reveal a detailed picture of local market shares, they can be used to further refine the weights towards more accurate population counts. Ricciato et al. (2017) presents elaborate approaches to use MS counts and advanced technical network specifications to derive high-resolution population density estimates from signalling data.

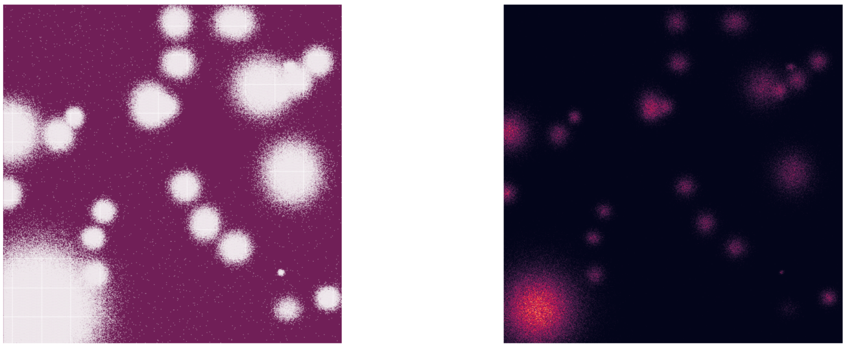
Further, high-resolution population grid estimates such as provided by WorldPop at 100x100m (Stevens et al., 2015) can be used as an alternative to binary settlement data. Here, $\hat{w}_{i,j}$ can be substituted with the estimated population count $\hat{p}_i$ per pixel directly extracted from the image.

3.4 Simulation

In order to evaluate the underlying motivation behind this methodology, i.e. more accurate mapping schemes produce more accurate outcomes, I test the performance of the different mapping approaches in terms of their overlap with the true coverage area and the accuracy of the predictions in a controlled setting with groundtruth information. Therefore, I run a simulation $T = 1000$ times on a synthetic population grid in which I re-distribute individuals, their poverty status, BTS locations and technical BTS specifications randomly. I observe the geographical overlap of the true and the estimated coverage areas, the overlap in home-located settlements and the correlation between the true and the estimated variable of interest (in this case the *poverty rate*). The main challenge in this simulation is to create ”true” coverage areas for each BTS that provide a realistic, but simplified benchmark for this study. Consequently, I opt for the extended HATA model. The choice is motivated by a series of propagation model evaluations using real-world measurements, notably Sharma RK Singh (2010); Abhayawardhana et al. (2005) and Phillips et al. (2011). The stochastic component within the HATA model is disabled in order to isolate the effect of interest.

3.4.1 Setup

I simulate a country including a major city, an uninhabited area such as a large lake or a national park and rural area otherwise using a 1000 x 1000 grid where each quadratic pixel represents an edge length of 100m. The urban area is divided into 16 equally-sized (50 x 50 pixel) small statistical areas, whereas the rural area is divided into 24 larger ones (200 x 200 pixel). I randomly distribute one million individuals across the grid using a multivariate normal distributions with $\mu_x = 10$, $\mu_y = 10$, $\Sigma_x = [50, 0]$ and $\Sigma_y = [0, 50]$ for the urban area (1/2 of the total population) and varying parameter values for the rural centers and a uniform distribution for the remaining rural area. Pixel-level population counts are calculated from individual-level data. Fig 3.3 shows an example of the settlement distribution across space and the corresponding population density.

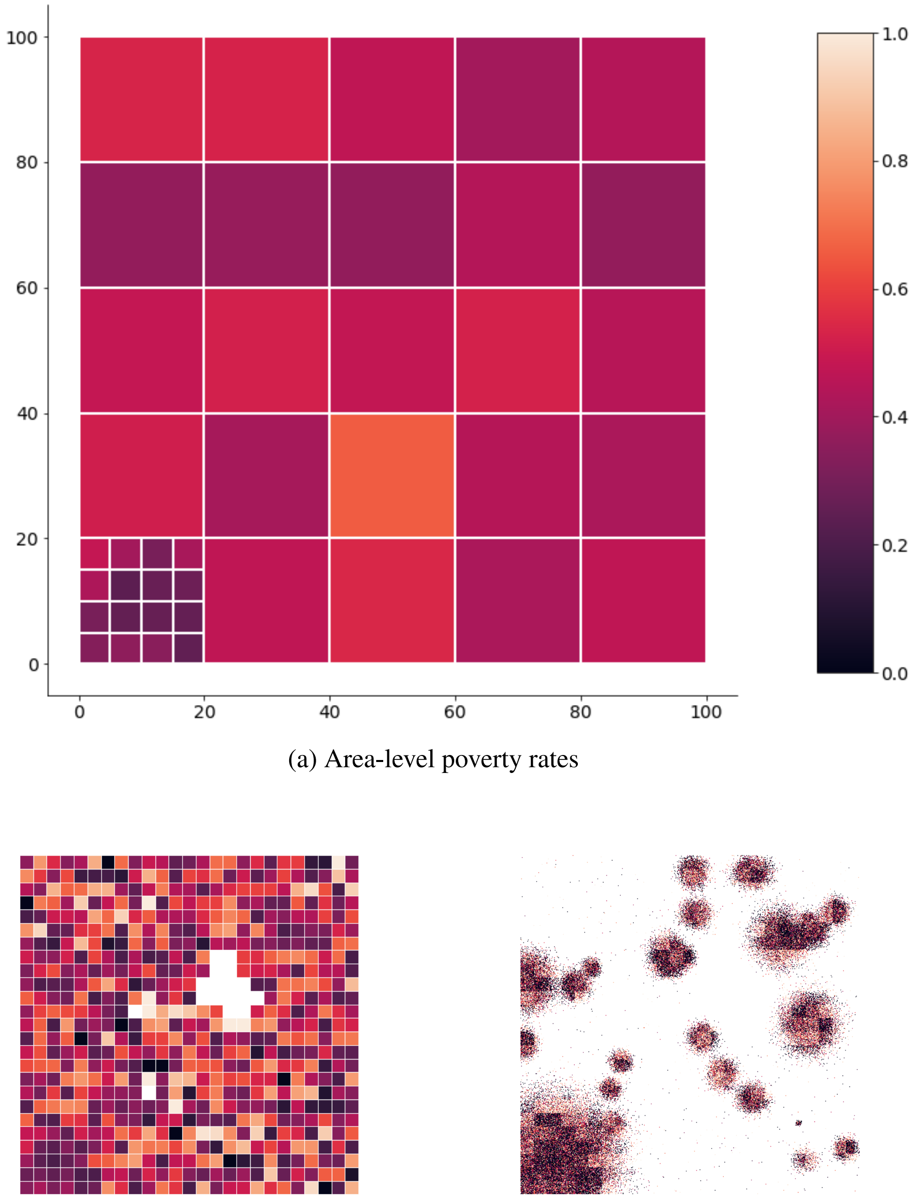


(a) Settlements

(b) Population density

Figure 3.3: **Simulation setup - Settlements.**
(a) shows locations of the built-up areas in a hypothetical country, while (b) shows the corresponding population density in these areas (the brighter the colour, the higher the population density).

In the next step, I randomly assign a poverty rate to each pixel. First, I generate a 4x4-pixel poverty grid for which I calculate the population density (see Fig 3.4b). In order to account for differences in the poverty rate between urban and rural areas, I randomly draw from a uniform distribution with values between 0 and 1 and multiply it with the inverted normalized population density. This poverty rate serves as the mean μ for randomly assigning poverty rates to settlements within the respective grid area using a normal distribution $N(\mu, \sigma)$ with $\sigma = 0.5$. Values below 0 and above 1 are windsorized. This two-step procedure tries to limit good predictive performances for areas not actually covered due to inference facilitated by the same underlying data generating process. Further, I assume that every inhabitant has one and only one MS and that there exists an indicator derived from mobile phone metadata that perfectly correlates with the true poverty rate of a given set of MS. Consequently, deviations in the correlation between the poverty rate captured via the ”true” coverage area and the poverty rate captured via the estimated coverage area exclusively originate in their coverage mismatch.



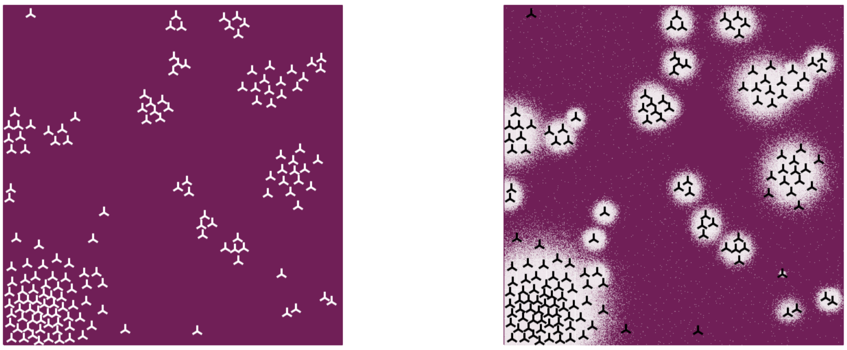
(a) Area-level poverty rates

(b) Grid-level poverty rates

(c) Settlement-level poverty rates

Figure 3.4: **Simulation setup - True poverty rate.**

In order to create a mobile network on top of that structure, I use a clustering algorithm based on the population density (see Fig 3.5b). BTS are distributed across the country at a ratio of roughly 1 BTS per 5,000 inhabitants in urban areas and 1 BTS per 10,000 inhabitants in rural areas. This results in 100 urban and 50 rural BTS in this simulation. BTS are interpreted as omnidirectional antennas and assigned specific heights, frequencies and output powers. The specifications vary more strongly in the urban area in order to reflect the greater complexity of network topology generally found in metropolitan areas. Since the HATA model requires a classification of areas into *urban*, *suburban* and *rural*, I use those 50% of BTS with the smallest number of pixels associated to them by the clustering algorithm used above as *urban* and those 5% of BTS with the largest number of pixels as *rural*, *suburban* otherwise. At the end, BTS heights are between 15 - 60 m with frequencies at 900 MHz and 2100 MHz and output power between 40 and 47 dBm. The MS height is fixed at 1m above ground level.



(a) BTS locations

(b) BTS locations and settlements

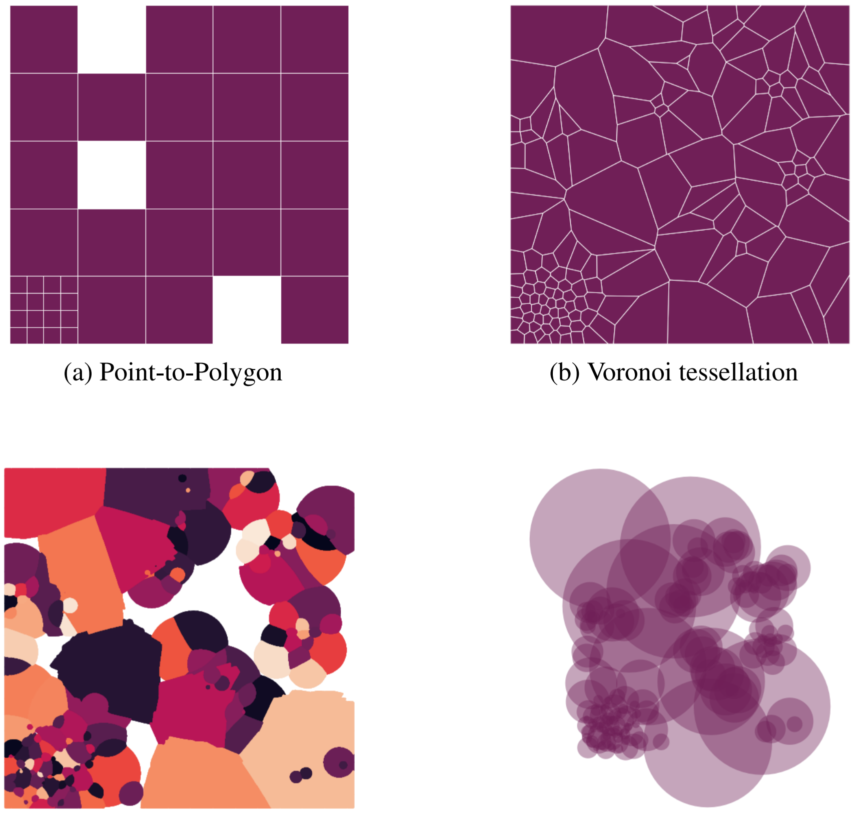
Figure 3.5: **Simulation setup - BTS locations.**

Based on these technical specifications, the true coverage areas and the true home locations of the settlements using the extended HATA model are calculated and used to create benchmark estimates of the true poverty rate. The results are then compared against estimates from point-to-polygon allocation, voronoi tessellation, augmented voronoi tessellation and BSA and IDW approaches of a naïve ('simple') version of the extended HATA model that does not know the exact technical BTS specifications, but makes an educated guess based on publicly available information such as the frequencies used in the country and the location of urban centers. Fig 3.6 exemplifies how the approaches differ in terms of geographical coverage.

The results are compared in three different ways: How much do they overlap geographically? How much do they overlap in terms of home-located settlements? How well do they predict the true poverty rate of a given statistical area?

3.4.2 Results

Table 3.2 shows the best performing approach in each round across round for all five performance indicators. Performance differences between voronoi tessellation versus the augmented voronoi tessellation and the augmented voronoi tessellation versus the HATA (BSA) approach showcase the relative contribution of settlement weighting and radio-propagation modelling, respectively. As expected, the simple HATA model clearly outperforms the other mapping approaches in terms of overlap, both geographically with the true coverage area (see Table 3.3) as well as concerning the home-located settlements (see Table 3.4). As the settlement-based approaches do only affect the calculation of weights and not of the coverage area, the coverage results are identical for voronoi tessellation and augmented voronoi tessellation and for the two HATA approaches, respectively. However, this advantage is not reflected to a similar extent in



(a) Point-to-Polygon

(b) Voronoi tessellation

(c) HATA (BSA)

(d) HATA (IDW)

Figure 3.6: **Coverage areas exemplified.**

the predictive performance.

Table 3.2: **Best performing approach by round across rounds (in %).**

Mapping	Coverage		Prediction		
	Geography	Settlements	R^2	Bias	RMSE
Point	0.0	0.0	27.5	28.2	28.6
Voronoi	0.0	0.07	2.6	9.9	2.1
Aug. Voronoi (GUF)			35.5	33.1	36.8
HATA (GUF, BSA)	100.0	99.3	29.7	13.7	27.7
HATA (GUF, IDW)			4.7	15.1	4.8

Table 3.3: **Geographical overlap with true coverage area (in %).**

Mapping	Total	Rural	Suburban	Urban
Point	25.8	15.3	30.9	22.3
Voronoi	30.7	14.1	25.5	37.0
Simple HATA	55.3	80.1	62.1	46.7

Interestingly, the HATA (IDW) approach performs poorly in prediction in contrast to the

Table 3.4: **Overlap with true home-located settlements (in %).**

Mapping	Total	Rural	Suburban	Urban
Point	16.9	44.3	15.9	14.9
Voronoi	54.2	56.8	60.7	48.0
Simple HATA	59.7	87.6	66.5	50.6

HATA (BSA) approach. This is due to the fact that the poverty rate in the true coverage area is calculated based on a deterministic home location, i.e. it is calculated from a constant set of settlements. This coincides directly with the mode-based HATA (BSA) approach, however, it does not reflect most real-world settings, in which stochastic radio propagation and overlapping coverage areas lead to situations where the captured poverty rate by the BTS is sourced from varying sets of settlements. The HATA (IDW) approach addresses this setup. Consequently, it is expected that the differences between these two approaches at least diminish in the application with real-world data in Section 3.5. Also, deviations of the HATA (BSA) approach from the benchmark exclusively originate in the technical misspecifications as the true coverage area is calculated from a correctly specified HATA model. The network complexity faced in real-world settings is expected to further undermine the accuracy of propagation-based mapping schemes.

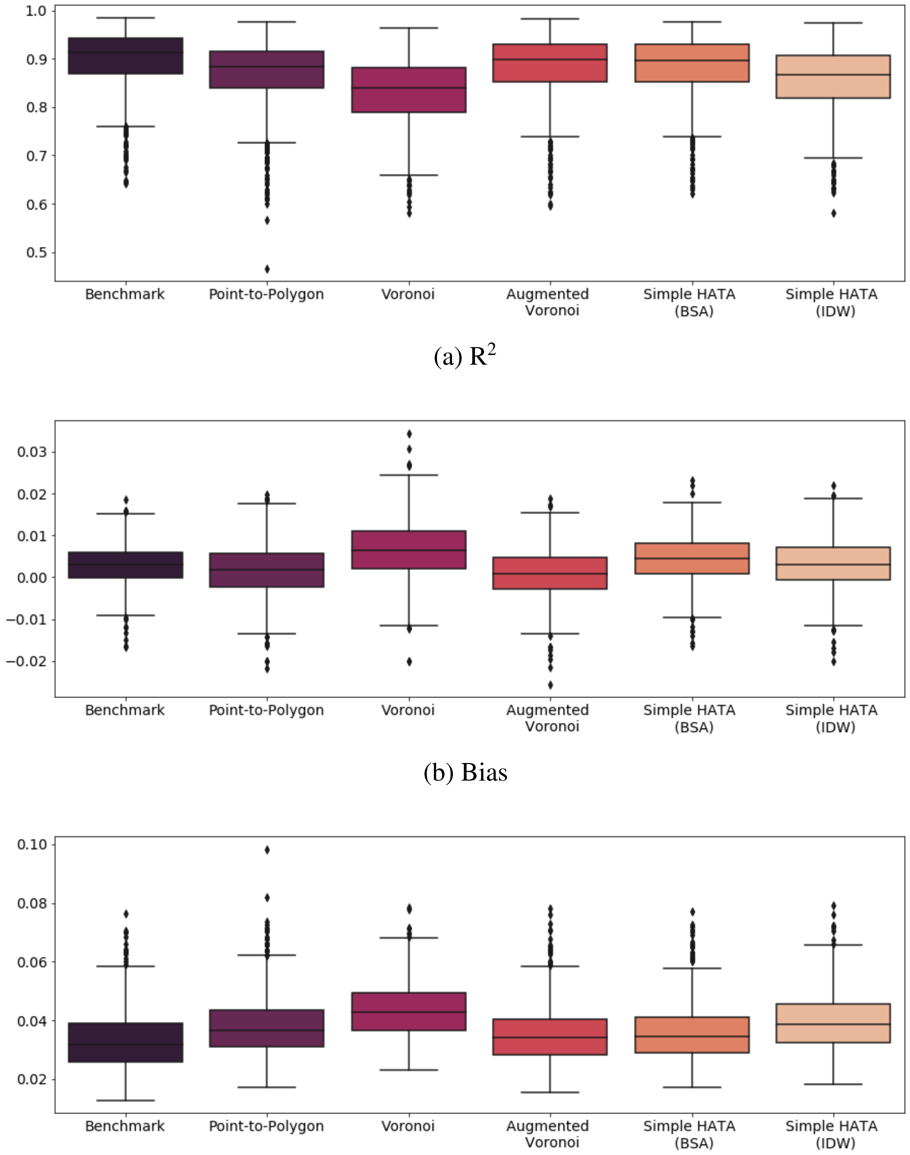
Looking at the performance of the two voronoi approaches in Table 3.2 the value added of using settlement information becomes apparent. Recalling the setup, the simulation assumes error-free human settlement identification. This, again, may not hold true in a real-world application as some buildings may not be detected while some detected buildings may not be inhabited. Consequently, it is expected that the difference between thee two voronoi approaches will be less stark in the application.

Fig 3.7 shows the distribution of the three performance indicators across rounds for those statistical areas for which every mapping scheme can provide estimates. On average, this reduces the underlying set of observations from 40 to 32 (see the sample sizes in Table 3.5). The result for the true coverage area are represented as *benchmark* for the other approaches as it estimates the settlement-level poverty rates actually captured by the respective BTS. Consequently, the benchmark should provide the upper bound for the R^2 and the lower bound for the bias and the RMSE in each round. Deviations thereof may only be due to spurious correlation.

Table 3.5: **Area-level correlation of estimated and true poverty rate & sample size.**

Mapping	ρ	n	ρ_{Rural}	n_{Rural}	ρ_{Urban}	n_{Urban}
Benchmark	0.905	40	0.734	24	0.971	16
Point	0.930	36	0.828	20	0.940	16
Voronoi	0.873	40	0.622	24	0.966	16
Aug. Voronoi	0.896	40	0.715	24	0.966	16
Simple HATA (BSA)	0.897	40	0.717	24	0.957	16
Simple HATA (IDW)	0.885	40	0.670	24	0.962	16

The sample size difference also explains the difference between the performance of the



(a) R^2

(b) Bias

(c) RMSE

Figure 3.7: **Estimating the true poverty rate for statistical areas.**
Distribution of the three performance metrics adjusted R^2, bias and RMSE with the estimated poverty rate using the true coverage area, i.e. built-up areas perfectly allocated to BTS, as 'Benchmark' across 1000 simulation runs.

point-to-polygon approach in terms of correlation in Table 3.5 vis-à-vis the performance metrics, especially in rural areas. Point-to-polygon allocation does not provide poverty estimates for 8 out of 40 statistical areas, on average, as they do not host a BTS (cf. Figs 3.6b). As both poverty rate and BTS allocation is linked to the population density by design, it can be expected that the predictive performance for rural areas not hosting a BTS are poor as they are generated from different underlying distributions.

However, this does not fully explain the performance differences between the approaches. On one hand, statistical areas are quite large, thus most of the BTS experience little overlaps in their true coverage area with other statistical areas. Consequently, the statistical area provides a decent approximation for the coverage. In contrast, simple voronoi tessellation with geographical weights tends to overemphasize the importance of remote areas as a) it assumes to cover areas for which data is actually not captured and b) BTS are usually located in close proximity to populated areas while serving remote areas further away as a side effect of it. This may be especially relevant in situations with large between-variation among statistical areas, strong population clusters and imperfect mobile network coverage. While b) is accounted for in the simulation, only approx. 0.1 % of the settlements are not covered by the network. Although this in line with the mobile network coverage in most countries, it can be expected that propagation-based schemes that account for holes in the mobile network outperform established approaches in setups with poor coverage.

3.5 Application

In their 2017 study on estimating literacy rates in Senegal published in the *Journal of the Royal Statistical Society Series A*, Schmid et al. (2017) use point-to-polygon allocation to map BTS point locations to statistical areas (*communes*). I revisit the design-based simulation of the study and extend it with four alternative mapping schemes, notably voronoi tessellation, satellite-augmented voronoi tessellation and the herein presented propagation-based coverage estimation methods using the best server area approach and the inverse signal strength weights. I compare the outcomes of all five schemes in terms of bias, root mean squared error (RMSE) and adjusted R^2.

3.5.1 Situation in Senegal

The application draws on real-world data from Orange-Sonatel for the year of 2013 (de Montjoye et al., 2014). During that time, the MNO operated mainly on the GSM 900 (2G) band with some UMTS 2100 (3G) deployments in urban centers. A large share of on-net traffic (approx. 91 % of overall traffic vis-à-vis a market share of approx. 57 %) during that year suggests a high prevalence of dual SIM use. It is expected that in this setting a negligible share of SIM cards are used by IoT devices others than MS. Coverage advantages in rural areas suggest dual-SIM use to be a phenomenon of more densely populated areas. The country exhibits little irregularities in the terrain: The highest point of Senegal being approx. 648 m above sea level is located at its southern border. The lowest point constitutes the sea level. Urban built-up areas with multi-storey buildings are predominantly limited to downtown Dakar. Most of the country is dominated by savanna with sparse high-grown vegetation.

3.5.2 Original study

In their design-based simulation, Schmid et al. (2017) implement a stratified two-stage cluster sample design similar to the one used in large-scale household surveys such as the Demographic and Health Survey (DHS) using a 10 % random sample of a pseudo-population as sampling frame, the 431 communes of Senegal as primary sampling units (PSUs) and the 14 regions of Senegal as strata. The authors combine the constructed 'survey' data with covariates extracted from mobile phone metadata on the level of communes in order to evaluate different small area estimation techniques using the *unemployment rate* as target variable of choice. The 72 available covariates are calculated on the subscriber-level using the Python library *Bandicoot* (de Montjoye et al., 2016). The subscriber-level covariates are allocated and aggregated to a BTS using the most frequently used BTS by a subscriber between 7pm and 7am as the *home location*. The BTS-level covariates are then allocated and aggregated using point-in-polygon allocation. Variable selection is performed backwards on large communes using the Bayesian Information Criterion. The covariates are used to generate small area unemployment rate estimates using a transformed Fay-Herriot model. Finally, Schmid et al. evaluate the small area estimates against the 'true' pseudo-population aggregates in 500 simulation runs using bias and RSME for a) communes covered by the survey (*in-sample*) b) communes not covered by the survey (*out-of-sample*) and c) communes without covariates from mobile phone metadata. For additional details on the setup of the original study, I refer to Schmid et al. (2017).

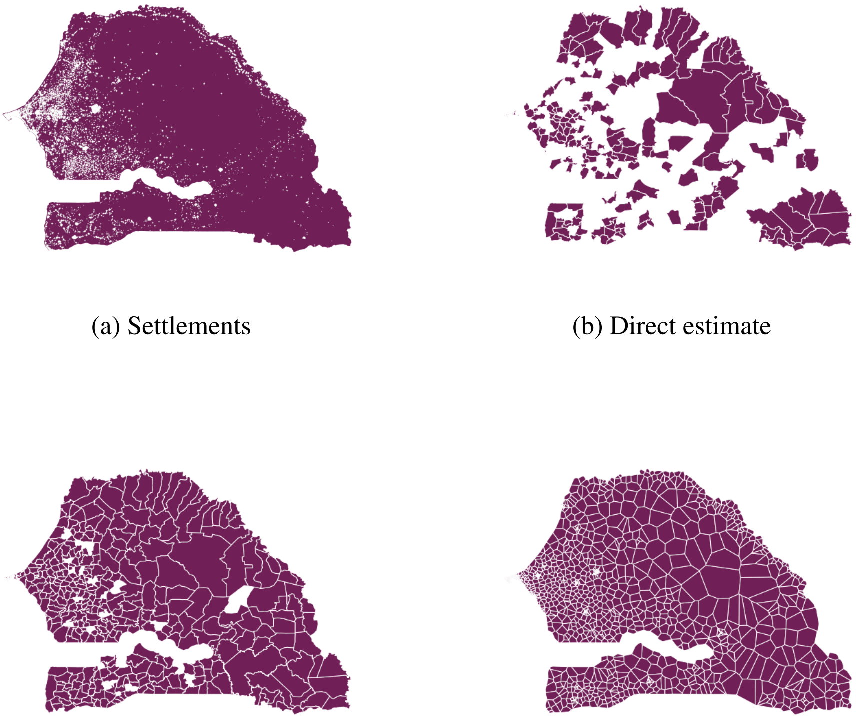
3.5.3 Extensions

I re-run the simulation of the original study five times thereby only varying the commune-level matrix of covariates as inputs. Specifically, I create five distinct sets of commune-level covariates beforehand by applying different mapping schemes during the aggregation process of the BTS-level data of the original study. First, I use the point-to-polygon allocation used in the original study. Second, I apply a standard voronoi tessellation to extract spatial weights proportional to the geographical overlap of tile and statistical area as described in Section 3.3.2 since it is used in most other studies in this field. Third, I augment the voronoi tessellation with settlement information from GUF by taking the number of white pixels (representing (part of) a settlement) within each section as a weight for commune-level aggregates to account for within-cell heterogeneity. Fourth, I implement the extended HATA (BSA) model as presented in 3.3 and GUF data. In densely populated areas, this approach closely resembles voronoi tessellation, however, it allows for holes in the network and for non-linear relationships between signal strength and distance. Fifth, I use inverse signal strength weights - HATA (IDW) - to capture the stochastic nature of a link.

Comparing Figs 3.8c and d to the direct estimator (Fig 3.8b) shows the benefits of augmenting survey data with mobile phone metadata: providing estimates for small areas not originally covered by the survey. Looking at settlements in Fig 3.8a, it is noteworthy that one commune - Thietty in the region Kolda - does not appear to host any settlement identified as such in GUF data. While official population numbers do not support this view, it underlines the fact that information extracted from satellite imagery, e.g. settlement classifications, are subject to some degree of uncertainty.

3.5.4 Assumptions

In contrast to point-to-polygon allocation and voronoi tessellation, the extended HATA model requires additional technical antenna specifications, notably the antenna and receiver height, the frequency and the transmitter power. As additional information are not available in the original study, I make following assumptions: I fix both the antenna height h_{tx} and the receiver height h_{rx} at the lower bound of the extended HATA model, which is 30 m and 1 m,



(a) Settlements

(b) Direct estimate

(c) Point-to-Polygon

(d) Voronoi

Figure 3.8: **Commune-level coverage areas in Senegal.**
Areas for which estimates of indicators of interest are available are coloured in red. Lower resolution built-settlements extents data reprinted from WorldPop (2018) under a CC BY license, with permission from WorldPop, original copyright 2018, are used in (a) for illustrative purposes.

respectively, both located outdoors with line-of-sight and a transceiver installed above the roof. As most of Senegal is flat without high multi-storey buildings except in downtown Dakar and in large parts no high-grown vegetation this assumption appears reasonable. Further, I fix the frequency in rural areas at 900 MHz and in urban centers at 2100 MHz and I interpret BTS as omnidirectional antennas with an output power of 45 dBm. This is clearly a simplification of the actual network topology, especially in urban areas with a mix of directed micro and macro cells. However, in Senegal in 2013, 4G has not yet been introduced and Orange-Sonatel was operating 3G (on the 2100 MHz frequency band) only in urban areas. The remaining country was served with 2G technology on the 900 MHz band. Comparing own estimates with coverage area estimates for 2G in 2017 published by Sonatel (2019) allows for a rough sanity check for the assumptions.

While Senegal offers an official classification of *rural* and *urban* on the commune-level, it is imperfect for the purposes of this study, as it takes a wide variety of non-network-specific factors into account. This leads to a situation where places with a high population density, e.g. Touba Mosque, are classified as *commune rurale*. Instead, I use BTS density per km^2 as a proxy for urbanity with a threshold of 1. Communes with more than one BTS per km^2 are classified as *urban*, those 50% of the communes with the lowest site density are classified as *rural*, the remaining communes are classified as *suburban*. This represents a more network-oriented measure of urbanity and is also in line with the area type classification of the HATA model.

3.5.5 Results

Similar to Table 3.2 in the simulation, Table 3.6 shows which mapping scheme performed best across the 500 evaluation rounds. Confirming initial findings of Section 3.4, there is no clear winner. While point-to-polygon allocation performs best in out-of-sample predictions in terms of RMSE (54.0 % of the rounds), it performs poorest in in-sample predictions. One possible explanation is that the lower average number of predictors used across rounds reduces the effects of overfitting. While HATA (IDW), HATA (BSA) and the augmented voronoi approach perform well across performance metrics, the overall difference between the approaches is limited (see Fig 3.9 and Table 3.7).

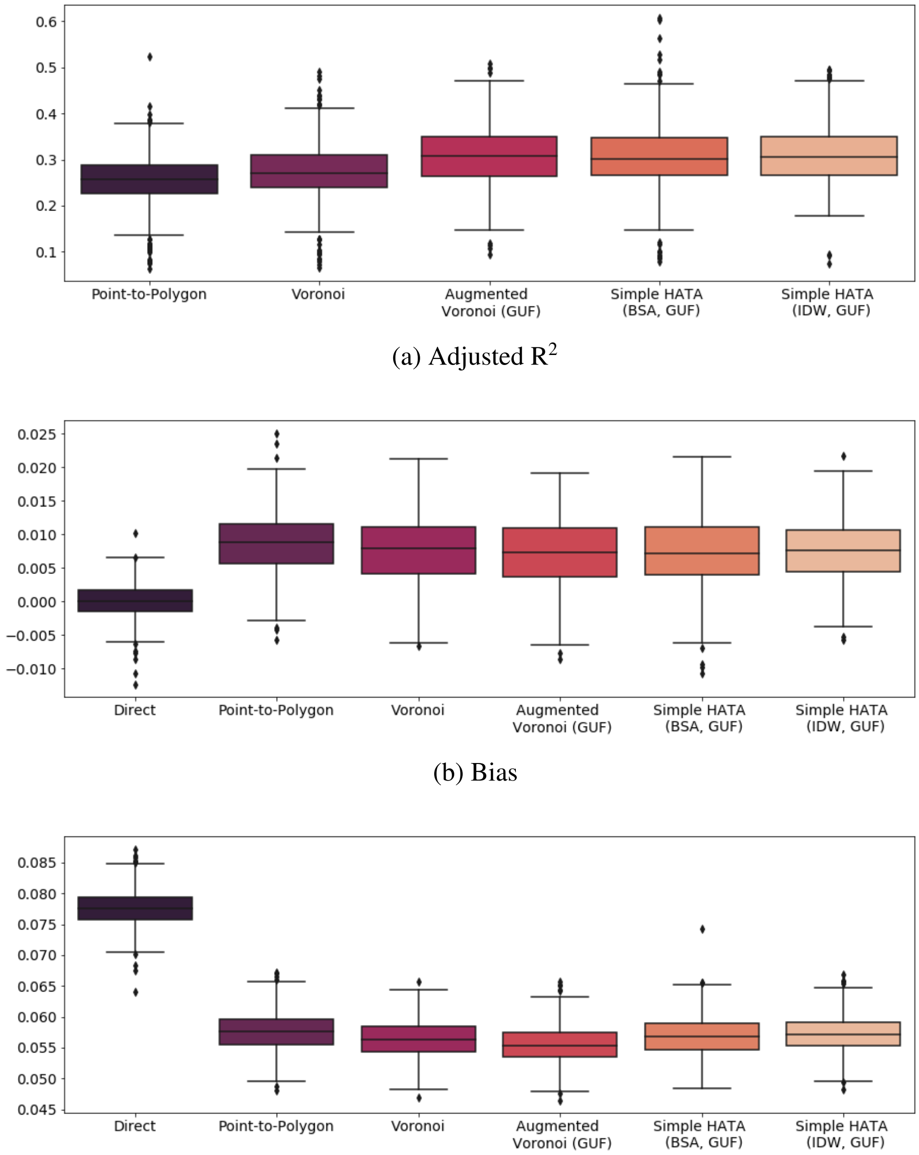
Table 3.6: **Best performing approach by round across rounds (in %).**

Mapping	Adj. R^2	Bias		RMSE		Avg. # of
	in	in	out	in	out	predictors
Point	6.0	16.4	23.2	12.6	54.0	4.2
Voronoi	10.2	21.0	16.6	21.6	22.8	5.0
Aug. Voronoi (GUF)	27.2	22.6	18.0	33.2	5.2	6.5
HATA (GUF, BSA)	27.0	21.8	16.4	18.0	7.8	6.4
HATA (GUF, IDW)	29.6	18.2	25.8	14.6	10.2	6.2

Table 3.7: **Correlation with true unemployment rate and sample size in Senegal.**

Mapping	ρ	n	ρ_{in}	n_{in}	ρ_{out}	n_{out}	ρ_{ooc}	n_{ooc}
Point	0.535	431	0.765	192	0.320	210	0.355	29
Voronoi	0.542	431	0.778	196	0.313	235	-	0
Aug. Voronoi (GUF)	0.519	431	0.780	195	0.280	233	0.586	3
HATA (GUF, BSA)	0.511	431	0.770	194	0.269	232	0.670	5
HATA (GUF, IDW)	0.527	431	0.781	196	0.308	234	-	1

In contrast, urban communes do not perform significantly better than rural ones as suggested by the simulation results. Table 3.8 shows, similar to Table 3.5 for the simulation, the correlation between the actual and predicted commune-level unemployment rates. Fig 3.10 shows an orientation along the diagonal signalling overall good fit. A possible explanation is that the structural relationship of mobile phone metadata covariates and the unemployment rate is captured more robustly for rural areas as they constitute 385 out of 431 communes in Senegal. To test this explanation, Tables B.1 and B.2 in the Supplementary material B.1 show the results for in-sample and out-of-sample predictions by commune status, respectively.



(a) Adjusted R^2

(b) Bias

(c) RMSE

Figure 3.9: **Evaluation of poverty rate estimates for in-sample communes.**
Distribution of the three performance metrics adjusted R^2, bias and RMSE across 500 simulation runs on a comparable set of communes. The typical trade-off between the bias and the variance of a small area estimator vis-à-vis the direct survey estimator becomes apparent.

While urban communes outperform rural ones in in-sample prediction they fare worse for in the out-of-sample setting, thus supporting the aforementioned hypothesis.

While settlement-based mapping schemes exhibit improvements in the model fit compared to point allocation or voronoi tessellation, they do not translate into major efficiency gains in terms of bias and rmse (see Fig 3.9b and c). Possible reasons are threefold: There is a significant classification error in the settlement data. The complete absence of settlements in Thietty, Kolda, support this assumption. As a cross-check, I re-run the analysis with an alternative source of settlement information. Specifically, I use high-resolution population density estimates from WorldPop (Stevens et al., 2015), however, it does not lead to gains in efficiency

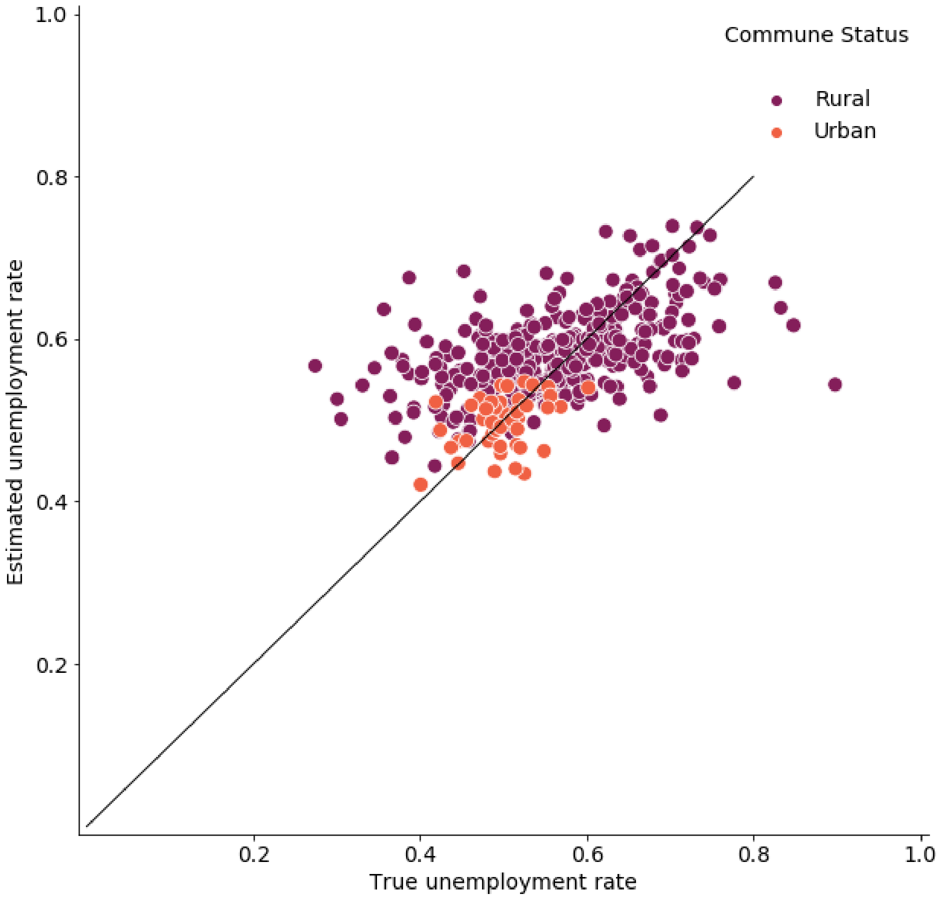


Figure 3.10: **True vs. estimated unemployment rate by commune status for a single simulation run.**

Table 3.8: **Area-level correlation of estimated and true unemployment rate & sample size.**

Mapping	ρ	n	ρ_{Rural}	n_{Rural}	ρ_{Urban}	n_{Urban}
Point	0.535	431	0.507	385	0.527	46
Voronoi	0.542	431	0.519	385	0.469	46
Aug. Voronoi	0.519	431	0.495	385	0.411	46
Simple HATA (BSA)	0.511	431	0.487	385	0.374	46
Simple HATA (IDW)	0.527	431	0.510	385	0.369	46

(cf. Table B.3 in Supplementary material B.1). Second, there is high spatial auto-correlation, thus little structural difference between the densely and sparsely populated areas in terms of the variable of interest - here unemployment - so even though latter are overemphasized in the calculations, it does not affect the outcome predictions. Here, I re-run the application with alternative variables of interest, i.e. the *literacy rate* and the population count (cf. Tables B.4 and B.5 in Supplementary material B.1); again, without significant efficiency gains versa point allocation and voronoi tessellation. Third, there is little within-area variation of the population density so that geographic weights and settlement-based weights are very similar. The correlation coefficient between the weights of the two voronoi approaches confirm that with $\rho = 0.98$. Also, I use the 100 meters x 100 meters population estimates from WorldPop to extract commune-specific variation coefficients. For 76.8 % of the communes, the within-commune variance is below 1, for 4 % it is above 100 with a maximum at 3553.4.

In general, the value added of using propagation-based mapping schemes appears to be negligible in this application, even though official coverage area estimates by Sonatel (2019) hint at the abundant presence of both overlaps and holes in the mobile network. A potential explanation is that the simplified HATA model is misspecified to an extent where the introduced errors cancel out the potential benefits. Looking at the specifications used in the application, this is most likely due to an underestimation of the coverage as the augmented voronoi approach closely resembles the upper bound for an overestimation using the HATA (BSA) within a - by assumption - largely homogeneous network.

3.6 Conclusion

Augmenting official statistics with mobile phone metadata still faces multiple methodological challenges, one of them is finding a common reference unit. As record-linkage on the individual-level presents considerable privacy risks a common procedure is to combine aggregates of these two disparate data sources on a geographical level. However, the stochastic nature of radio propagation makes it difficult to pin down coverage areas of the mobile network. Based on this study the good news is that it does not have to be complicated if supervised learning / prediction is the goal. While propagation-based models can help to refine the accuracy of coverage area estimation, it does not greatly impact the quality of the outcome predictions. One reason is that usually cells are located in a way that they provide a good service to as many MS as possible. As radio signals fade over distance, this means they are in close proximity to areas with high demand, i.e. densely populated places. Mapping schemes, in turn, mainly differ from each other when looking at the limits of a cell. However, most of the traffic which is correlated with statistical data for training/prediction is generated nearby, so the differences between mapping schemes become less relevant. Also, while geographical weights as used in most applications in this field ignore heterogeneity occurring within the cells, the corresponding statistical areas are often significantly larger. Therefore, cross-border cells, which could actually profit from weighting schemes that take within-cell heterogeneity into account, occur less frequent. In addition, cells and administrative (thus often statistical) areas are intimately linked via population clusters as both tend to be centered around them.

However, this study just provided initial evidence to inform future mapping choices and could be extended in multiple ways: First, both in the simulation and the application directional antennas are combined to omnidirectional antennas. While this is motivated by the typical data availability in real-world applications, it is of course a strong simplification of the actual network topology. As the lower bound of spatial heterogeneity captured is given by the number of unique areas resulting from intersecting coverage areas and statistical areas, studies such as Ricciato et al. (2017) have shown that moving from an BTS-oriented to a cell-oriented analysis could greatly affect analysis, especially via potential increases in sample size. However, it needs further investigation how refined mapping schemes can add further value, particularly in the presence of measurement uncertainty, to supervised learning setups in cell-level analysis. Second, the study used comparatively simple empirical propagation models based on real-world measurements largely ignoring actual environments. More advanced propagation models exist, however, they require significantly more computing resources that could limit their applicability as they take the physical surrounding via digital surface models into account. Nevertheless, investigating this constitutes an interesting path for further research.

Supplementary material B

B.1 Cross-checks of application results

B.1.1 Rural-urban performance differences

As stated in the application of the paper, urban communes do not perform significantly better than rural ones as suggested by the simulation results. A possible reason could be that the estimation of urban areas is less robust as there are less urban communes than rural ones in Senegal. Tables B.1 and B.2 show the results for in-sample and out-of-sample predictions by commune status, respectively.

Table B.1: **In-sample area-level correlation of estimated and true unemployment rate & sample size.**

Mapping	ρ	n	ρ_{Rural}	n_{Rural}	ρ_{Urban}	n_{Urban}
Point	0.765	191	0.745	176	0.789	16
Voronoi	0.778	196	0.759	180	0.786	16
Aug. Voronoi	0.780	195	0.762	179	0.777	16
Simple HATA (BSA)	0.770	194	0.750	178	0.783	16
Simple HATA (IDW)	0.771	196	0.751	180	0.784	16

Table B.2: **Out-of-sample area-level correlation of estimated and true unemployment rate & sample size.**

Mapping	ρ	n	ρ_{Rural}	n_{Rural}	ρ_{Urban}	n_{Urban}
Point	0.320	210	0.275	180	0.308	30
Voronoi	0.313	235	0.283	205	0.297	30
Aug. Voronoi	0.280	233	0.246	203	0.211	30
Simple HATA (BSA)	0.269	232	0.238	202	0.141	30
Simple HATA (IDW)	0.308	234	0.285	204	0.225	30

B.1.2 Classification error in the settlement data

Even though GUF data is supposed to have a true positive rate of 85 % on average, with 68 % at lowest and 98 % at heighest, two Senegalese communes host no settlements. This hints at the presence of classification error in the settlement data that may lead to efficiency losses in the estimation process. To investigate it further, I use an alternative source of settlement information. While GUF data is generated from satellite imagery only (180.000 single TerraSAR-X/TanDEM-X image products) at 0.4 arcseconds, WPG data is based on Random Forest-based dasymetric mapping approach using a wide variety of input data including land-cover information from MERIS imagery, night-time lights, distance information to various thematic land-cover classes etc at approx. 3 arcseconds.

Table B.3: **Best performing approach for *unemployment rate* by round across rounds (in %).**

Mapping	Adj. R^2	Bias		RMSE		Avg. # of
	in	in	out	in	out	predictors
Point	4.6	12.4	15.2	9.0	40.8	4.2
Voronoi	6.2	13.0	10.0	12.2	15.4	5.0
Aug. Voronoi (GUF)	18.2	16.4	10.4	17.6	2.2	6.5
HATA (GUF, BSA)	17.2	16.8	10.2	9.4	4.6	6.4
HATA (GUF, IDW)	18.6	9.8	15.2	6.0	6.4	6.2
Aug. Voronoi (WPG)	12.0	11.4	13.2	13.0	7.2	6.0
HATA (WPG, BSA)	12.8	14.0	14.0	17.2	5.4	5.9
HATA (WPG, IDW)	10.4	6.2	11.8	15.6	18.0	5.0

B.1.3 Additional outcomes of interest

The following two tables provide results for two additional outcomes of interest, in this case the literacy rate and the population count.

Table B.4: **Best performing approach for *literacy rate* by round across rounds (in %).**

Mapping	Adj. R^2	Bias		RMSE		Avg. # of
	in	in	out	in	out	predictors
Point	59.2	14.6	93.8	70.4	77.4	5.6
Voronoi	1.0	13.0	0.6	0.8	0.2	5.1
Aug. Voronoi (GUF)	4.2	10.2	1.0	10.6	19.4	5.3
HATA (GUF, BSA)	6.8	7.4	0.8	1.6	0.4	5.5
HATA (GUF, IDW)	1.0	5.4	1.8	1.8	0.8	4.9
Aug. Voronoi (WPG)	14.6	10.4	1.8	3.6	0.8	5.6
HATA (WPG, BSA)	8.2	7.2	0.2	10.8	0.8	5.5
HATA (WPG, IDW)	5.0	41.2	0.0	0.4	0.2	5.5

B.2 Replicating the simulation

The complete code for replicating the simulation can be found in the following GitHub repository: `https://github.com/tilluz/geomatching_open`. No additional files are required.

Table B.5: **Best performing approach for *population count* by round across rounds (in %).**

Mapping	Adj. R^2	Bias		RMSE		Avg. # of predictors
	in	in	out	in	out	
Point	23.4	17.8	5.4	14.6	0.0	6.1
Voronoi	0.0	0.2	85.4	0.2	3.4	3.2
Aug. Voronoi (GUF)	0.0	27.8	0.0	2.0	0.0	3.0
HATA (GUF, BSA)	0.2	10.2	0.0	39.4	0.0	3.8
HATA (GUF, IDW)	73.8	36.8	0.0	26.8	0.0	9.0
Aug. Voronoi (WPG)	1.2	4.0	0.0	15.6	5.2	5.3
HATA (WPG, BSA)	0.4	3.2	0.2	1.4	11.2	4.8
HATA (WPG, IDW)	1.0	0.0	9.0	0.0	80.2	4.2

B.3 Replicating the application

The complete code and necessary data for replicating the application can be found as well as in the following GitHub repository: `https://github.com/tilluz/geomatching_open`. The replication results may slightly differ from the results presented in this study as it cannot be ensured that the 10%-sample of the census data provided by the statistical office of Senegal are identical to the one used in this study. Following data sources not available in the repository are necessary in order to run the code:

- *spss_car_individus_10eme_dr.sav*: The 10% sample of the population part of the RGPHAE 2013. Access can be requested via the microdata portal of the statistical office of Senegal (ANSD): `http://anads.ansd.sn/index.php/catalog/51/`

- *SITE_ARR_LONLAT_EXACT.csv*: This file contains the exact tower locations of SONATEL in 2013. Access can be requested as stated in the data availability statement of this study. To facilitate replication, a file with slightly randomized antenna locations is provided (*SITE_ARR_LONLAT.csv*). Keep in mind this may affect the final outcomes. The exact locations have to be requested as stated in the data availability statement.

- *sen_ppp_2013.tif*: This file contains the population density estimates of Senegal for the year 2013. The data can be downloaded at the WorldPop website: `https://www.worldpop.org/doi/10.5258/SOTON/WP00645`

- *senegal.tif*: This file contains the GUF data for Senegal at 0.4 arcseconds. Access can be requested for scientific, non-commerical purposes via the website of the German Aerospace Center DLR: `https://www.dlr.de/eoc/en/PortalData/60/Resources/dokumente/guf/DLR-GUF_LicenseAgreement-and-OrderForm.pdf`

The application is written in Python and R. The file and folder names indicate the required order of execution. Files 04 - 07 may not be run unless access to individual-level CDRs is available. It may be necessary to align directory paths to make the code run properly. Please ensure that more than 20GB RAM is available for this analysis. If needed, please contact the corresponding author for support.

Chapter 4

Releasing survey microdata with exact cluster locations and additional privacy safeguards

4.1 Introduction

Since almost hundred years, sample surveys are dominating knowledge generation in empirical research. The advantages of survey sampling are obvious: with an appropriate sampling design representative results for a population can be collected by surveying only a fraction of it. With computer assistance, the time from collecting data to publishing results can be sped up significantly (Granello and Wheaton, 2004). Two trends, however, increasingly challenge the way data is collected via surveys. On the one hand, the growing demand for fast and granular information drives up sample size and thus costs. As a response, recent years have seen a large amount of academic research on augmenting surveys with secondary data from non-traditional data sources such as social networks, mobile phones or remote sensing in order to overcome shortcomings in coverage, frequency and granularity with applications in fields as diverse as population dynamics (Stevens et al., 2015; Leasure et al., 2020), socio-demographic analysis (Pokhriyal and Jacques, 2017; Schmid et al., 2017; Subash et al., 2018; Fatehkia et al., 2020; Chi et al., 2022), policy targeting (Blumenstock, 2018; Aiken et al., 2022), environmental mapping (Grace et al., 2019) and health research (Brown et al., 2014; Arambepola et al., 2020). This augmentation is usually done via geographic matching, i.e. combining area-level averages (Koebe, 2020). Since the number of matched areas corresponds to the sample size for subsequent supervised learning tasks, finding the smallest common geographical denominator is essential to avoid running into small sample problems. However, this is not always trivial as sample surveys usually provide data only for a fraction of small geographic areas. On the other hand, digital transformations across various sectors such as health care have led to an explosion of digital personal data. It is the abundance of secondary data that amplifies re-identification risks in published surveys as some of the information could be used to link pseudoanonymized survey responses back to the actual respondents (Armstrong et al., 1999; Kroll and Schnell, 2016; West et al., 2017). Together with new privacy regulations such as the European General

Data Protection Regulation (GDPR) this calls for additional precautionary measures to safeguard the individual's privacy. For aggregated data releases, the introduction of differential privacy has provided a solid mathematical framework to manage re-identification risks independent of a potential attacker's capabilities or prior knowledge (Dwork, 2008). With regard to microdata dissemination strategies, a common de-identification practice today is a combination of deletion and perturbation procedures, which include removing (unique) identifiers such as first and last name and replacing the individual's true location with aggregated (i.e. area-level) and randomized information (see e.g. Andrés et al. (2013); Templ (2017); de Jonge and de Wolf (2019)).

For example, in the Demographic and Health Survey (DHS), a major global household survey program, urban survey clusters are re-located within a 2km-radius and rural clusters within a 5km-, sometimes even 10km-radius (Burgert et al., 2013). This location privacy procedure has two main advantages: it does not affect the quality of the remaining (non-spatial) survey information and it reduces the need for other privacy safeguards, e.g. deleting or perturbing sensitive information. However, it does not provide a similar rigorous measure for privacy protection as already small sets of attributes can quickly increase the chances of re-identification, even in incomplete, pseudonymous datasets (Rocher et al., 2019). In addition, it obviously affects the utility of the published data when it comes to matching with auxiliary data as this type of analysis relies on the congruence of its geographic links (Elkies et al., 2015; Warren et al., 2016; Blankespoor et al., 2021; Hunter et al., 2021).

In that regard, advances in synthetic data generation have introduced new ways to narrow the void between information loss and privacy protection. These methods allow for the generation of synthetic records that resemble the real data by reproducing relationships learned from the latter. While all approaches have in common that they try to capture the joint distribution in the original data, the ways to do so vastly differ. For example, Drechsler et al. (2008) and Heldal and Iancu (2019) use imputation processes to decompose the multidimensional joint distribution into conditional univariate distributions. Alfons et al. (2011) and Templ et al. (2017) use parametric models in combination with conditional re-sampling to synthesize hierarchical relationships. As an alternative to these fully parametric approaches, Reiter (2005) and Wang and Reiter (2012) make use of classification and regression trees (CART), while more recently, Li et al. (2014); Zhang et al. (2017); Rocher et al. (2019); Sun et al. (2019); Torkzadehmahani et al. (2019); Xu et al. (2019) and others have used Bayesian networks, Generative Adversarial Networks or copulas to capture the underlying linear and non-linear relationships between the attributes.

The challenge for data producers is to define adequate microdata dissemination strategies that allow users to satisfy their needs, i.e. release survey microdata that can be used for statistical analysis and that are compatible with other sources of information allowing to answer new and more detailed research questions and – at the same time – it must be ensured that the identities of the respondents are protected.

In that regard, the Spatial Data Repository of the DHS program (Burgert-Brucker et al., 2018) is a good example for facilitating new types of research by combining survey microdata with geospatial covariates and gridded interpolation surfaces. However, also those products are based on perturbed cluster locations, thus incurring a certain information loss.

In this paper, we propose an alternative microdata dissemination strategy that leverages the utility of the original microdata with additional privacy safeguards through copula-generated synthetic data. Specifically, we propose to adopt a strategy of publishing two sets of micro-level survey data: first, the original microdata stripped of geographic identifiers below the strata-level. Second, synthetic microdata including the true cluster locations. We show in an experiment using Costa Rican census data from 2011 and satellite-derived auxiliary information from WorldPop (WorldPop, 2018) that we can reduce the re-identification risk vis-à-vis common spatial perturbation procedures, while maintaining data utility for non-spatial analysis and improving data utility for spatial analysis.

From the plethora of options, we choose copulas as our synthetic data generation approach. Copulas facilitate fine-tuning as they allow us to model the marginal distributions separately from the joint distribution. Dating back to 1959 (Sklar, 1959) with diverse applications since, their theoretical properties are well understood. In comparison with alternatives like GANs, copula-based synthetic data generation has lower computational cost (Sun et al., 2019) and it is easier to interpret (Kamthe et al., 2021). Furthermore, the procedure is in general less cumbersome, in comparison with the steps followed by Alfons et al. (2011) to generate the synthetic population data AAT-SILC (*Artificial Austrian Statistics on Income and Living Conditions*. Finally, copulas are also attractive for data producers such as National Statistical Offices as only new nationally representative margins are required to update the synthetic microdata file (cf. Koebe et al. (2021)). In addition, well-documented open-source tools such as the Synthetic Data Vault (MIT Data To AI Lab, 2022) are available to users with important features such as data transformation and constraints specification.

4.2 Results

4.2.1 Geomasking to obfuscate true survey locations

We consider a survey D as a random sample with sample size n_D from a given population C. Our unit of observations are individuals i living together in a household h. Each individual is described by a set of attributes denoted as $X_1, X_2, \dots X_d$. While different sampling designs are possible, we assume a commonly used complex design for larger household surveys such as the DHS: a stratified two-stage cluster design. In the first stage, the primary sampling units (PSUs) - usually enumeration areas from the latest census denoted as j - are selected for each stratum s with a probability proportional to (population) size π_j. In the second stage, households within each selected PSU are sampled with a fixed probability $\pi_{h|j}$. Consequently, the sampling weights defined as the inverses of the household-level inclusion probabilities are given for each stratum separately by:

$$w_{hj} = \frac{1}{\pi_{hj}}, \qquad \pi_{hj} = \pi_{h|j} * \pi_j \quad \text{with} \quad \pi_j = \frac{n_s}{N_s}. \tag{4.1}$$

PSUs, called *clusters* in the following, are geo-located as point locations r_j via their geographic centroids. In the following, we describe the original survey attributes together with the original geo-locations of the clusters as our *true* survey. The true survey builds our starting point for further anonymization approaches, notably the geomasking approach and the copula-based synthetic data generation approach. We follow the geomasking methodology outlined in Burgert et al. (2013) by perturbing the centroids of the selected clusters within a given larger administrative area l using a rejection sampling procedure described in Algorithm 1:

Algorithm 1: Geomasked survey: DHS cluster displacement algorithm

```
for $j \in D$ do
    while $r_j^{masked} \notin l_{r_j}$ do
        angle $\leftarrow U_{[0,360]} * \frac{\pi}{180}$ ;        /* Random displacement angle */
        if $j$ is Urban then
            dist $\leftarrow U_{[0,2000]}$ ;    /* Random displacement distance (in
              meters) for urban clusters */
        end
        if $j$ is Rural then
            if $j$ is selected as 1% of rural clusters then
                dist $\leftarrow U_{[0,10000]}$ ; /* Random displacement distance for
                  1% of rural clusters */
            else
                dist $\leftarrow U_{[0,5000]}$
            end
        end
        $r_{x,j}^{\text{masked}} \leftarrow r_{x,j} + \text{dist} * \cos(\text{angle})$ ; /* Displaced x-coordinate */
        $r_{y,j}^{\text{masked}} \leftarrow r_{y,j} + \text{dist} * \sin(\text{angle})$ ; /* Displaced y-coordinate */
    end
end
```

We denote the masked point locations of the selected clusters with the superscript *masked*. As the overall inclusion probability for a household is not affected by geomasking, direct estimates and corresponding variances for area-level aggregates l and above remain the same. However, this does not hold for area-level aggregates smaller than l. In the following, we describe the original survey attributes together with the masked locations of the clusters as our *geomasked* survey.

For our experiment using Costa Rican census data from 2011, point locations for the corresponding enumeration areas are not available. Therefore, we randomly sample them from the smallest available area denoted with k - in our experiment the districts (at the same time the zip codes) in Costa Rica. The zip code therefore corresponds to the smallest geographic identifier available in the survey. Through the displacement procedure, roughly 30% of the clusters are assigned to a new zip code, representing approx. 30% of the sampled individuals in each simulation round.

4.2.2 Copula-based synthetic data generation

As an alternative to geomasking, we use synthetically generated survey attributes for protecting the respondents' privacy while keeping the true point locations of the selected clusters. To do so, we fit a Gaussian copula model on the original survey attributes X_d and sample from the learned joint distribution for each cluster individually with the originally sample size n_j. A copula allows to describe the dependence structure - also called *association structure* - independently from the marginal distributions (also called *allocation structure*). Several copula families are available. We focus on the Gaussian copula that allows us to represent the association structure of random variables irrespective of their true distribution through a multivariate standard normal distribution (Patki et al., 2016). Since we also assume the marginals to be normally distributed, which may certainly constitute a mis-specification for some of the variables, we regard the results rather as a lower bound in terms of goodness-of-fit. Further, a copula is uniquely defined only for continuous variables (Jeong et al., 2016), meaning that in principle, copulas cannot model non-continuous variables. Since socio-economic surveys are largely made up of categorical variables, data transformation, e.g. via one-hot encoding, is needed. In addition, we impose constraints on the marginals to account for censoring (e.g. to avoid negative synthetic age records) or between-variable dependencies (e.g. female and male household members need to add up to the total household size) via rejection sampling.

Thus, the process to generate synthetic data from a survey dataset $\tilde{D}$ with transformed categorical attributes (details of the data transformation are described in Algorithms 3 and 4 in Section 4.4) using a Gaussian copula model is summarized in Algorithm 2:

Algorithm 2: Synthetic survey: Copula-based synthetic data generation algorithm

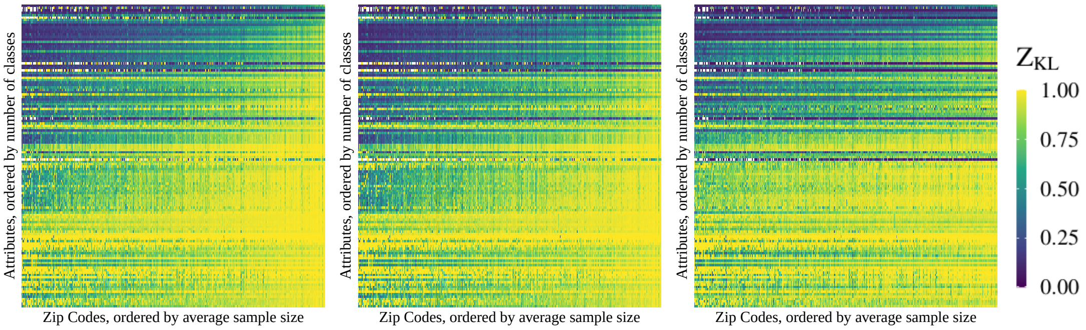
```
Input $\tilde{D} = (\tilde{X}_1, \ldots, \tilde{X}_d)$
Output $\tilde{S} = (\tilde{Y}_1, \ldots, \tilde{Y}_d)$
for $s \in \tilde{D}$ do
    $\Psi \leftarrow$ Estimated marginal distributions of $\tilde{X}$ with $\psi_d \sim \mathcal{N}(\mu, \sigma^2)$
    $\Sigma \leftarrow$ Estimated covariance matrix of $\Psi$
    $U \leftarrow F(\Psi)$ ;            /* Probability integral transforms */
    $C^G_\Sigma(u_1, \ldots, u_d) \leftarrow \phi_\Sigma\big(\phi_1^{-1}(u_1), \ldots, \phi_d^{-1}(u_d)\big)$ ;        /* d-dimensional
     Gaussian copula */
    for $j \in \tilde{D}_s$ do
        for $i \leftarrow 1$ to $n_{D_j}$ do
            while $y^{\{i\}}$ not meets constraints do
                $\mathbf{v} \sim \mathcal{N}(\mu, \Sigma)$
                $\tilde{\mathbf{y}}^{\{i\}} \leftarrow F^{-1}\big(\phi_1(v_1), \ldots, \phi_d(v_d)\big)$ ;        /* Convert back to
                 original space */
            end
        end
    end
end
```

ϕ_Σ is the cumulative distribution function (cdf) of a multivariate normal distribution with $\mathcal{N}(\mu, \Sigma)$ and ϕ_d the cdf of a standard normal distribution. By fitting our model to the true survey, it learns the parameters of both the allocation and association structure, i.e. of the marginal distributions Ψ and the multivariate Gaussian copula $C^G_\Sigma(u_1, \ldots, u_d)$. Based on these learned relationships, new synthetic records $\tilde{\mathbf{y}}^{\{i\}}$ are sampled from the multivariate probability function $c^G_\Sigma(\mathbf{u})$ using the inverse probability integral transform for each component $F_d^{-1}(u_d)$ (cf. Janke et al. (2021)). Since we sample in our experiment for each cluster individually to ensure a synthetic cluster-level sample size of exactly n_j, we use the parameters of a conditional multivariate normal distribution. In case no conditions are applied, the scenario is simplified to drawing from a multivariate standard normal distribution. In the following, we call the synthetic attributes together with the true cluster locations our *synthetic* survey. Further details about the copula-based synthetic data generation procedure can be found in the Section 4.4 and in Nelsen (2007).

Figure 4.1 provides a first impression on the overall goodness-of-fit of the three different survey datasets. Specifically, Figures 4.1a - 4.1c show the normalized KL divergence Z_{KL} of the survey attributes **y** from the true census attributes **x** $\in C$ defined as

$$Z_{KL}(F_{d,k}(X_{d,k})||F_{d,k}(Y_{d,k})) = \frac{1}{1 + D_{KL}(F_{d,k}(X_{d,k})||F_{d,k}(Y_{d,k}))} \tag{4.2}$$

averaged across simulation runs for each attribute d and zip code k, respectively.



(a) True survey

(b) Geomasked survey

(c) Synthetic survey

Figure 4.1: **Normalized Kullback-Leibler divergence (in bits) from the true census distribution for each attribute and zip code, averaged across 100 simulation rounds.** The attributes on the y-axis are ordered by their respective number of classes, the zip codes on the x-axis are ordered by their average sample size across simulation rounds. Values close to one (yellow) represent little divergence from the true census distribution and therefore indicate a high goodness-of-fit. The number of attribute classes range from two to 111. Across attributes and zip codes, the true survey scores best with $Z_{KL} = 0.76$ in total, followed by the synthetic survey with $Z_{KL} = 0.74$ and the geomasked survey at $Z_{KL} = 0.73$.

Clearly visible is a gradient from the top left to the bottom right indicating that the overall goodness-of-fit of the sample distributions improve the larger the underlying sample sizes and the lower the number of classes per categorical attribute. In addition, as expected, attributes with high levels of non-response (visible through the white spots across the horizontal axis) are stronger affected by sampling and anonymization compared to attributes with little or no non-response.

4.2.3 Population uniqueness of survey respondents

To approach the utility-risk trade-off in (pseudo)-anonymized microdata, we define two risk-related measures: the population uniqueness of the survey respondents and the re-identification risk of a sensitive attribute in the original data using the perturbed data. We define population uniqueness $\Xi_{\mathbf{x}}$ as the share of survey respondents being unique in the population C for a given set of attributes $\mathbf{x} = (X_1, \ldots, X_d)$:

$$\Xi_{\mathbf{x}} = \frac{1}{n_D} \sum_{i}^{n_D} \mathbb{1}_{\mathbf{x}^{\{i\}}} \quad \text{with} \quad \mathbb{1}_{\mathbf{x}^{\{i\}}} = \begin{cases} 1, & \text{if } \mathbf{x}^{\{i\}} \text{ unique in } C(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

Figure 4.2 shows how $\Xi_{\mathbf{x}}$ changes with the increasing number of attributes across 100 simulation runs. Naturally, the share constantly increases for the true survey. For the geomasked survey, the population uniqueness increases to a level of roughly 70%. Recalling that the only difference between the geomasked survey and the true survey is the perturbed zip code, the remaining 30% corresponds to the average number of survey respondents assigned to a new zip code due to the spatial anonymization process. Thus, not considering the zip code in the re-identification effort would let the population uniqueness of the geomasked survey also converge towards 1 similar to the true survey, even though requiring further knowledge on additional attributes. For the synthetic survey, the curve remains almost flat. The initial bump can largely be explained by the probability of a random combination of attributes representing an actual population unique in a small (area) sample size setting. Besides this theoretical argument, synthetic data always provides plausible deniability to the survey respondents. Similarly to our definition, Rocher et al. (2019) use a Gaussian copula model to estimate the empirical likelihood of population uniqueness in incomplete datasets such as D by assuming $\Xi_{\mathbf{x}} \sim \text{Binomial}(\mathbb{1}_{\mathbf{x}^{\{i\}}}, n_D)$ with $\mathbf{x}^{\{i\}} i.i.d.$. While this approach is an excellent alternative to measure the re-identification risk in micro-level survey data when no validation data (in our experiment the 2011 Costa Rican census) is available, it assumes that the individual records are independent and identically distributed, which may be contestable in the presence of hierarchical dependencies and complex sampling designs.

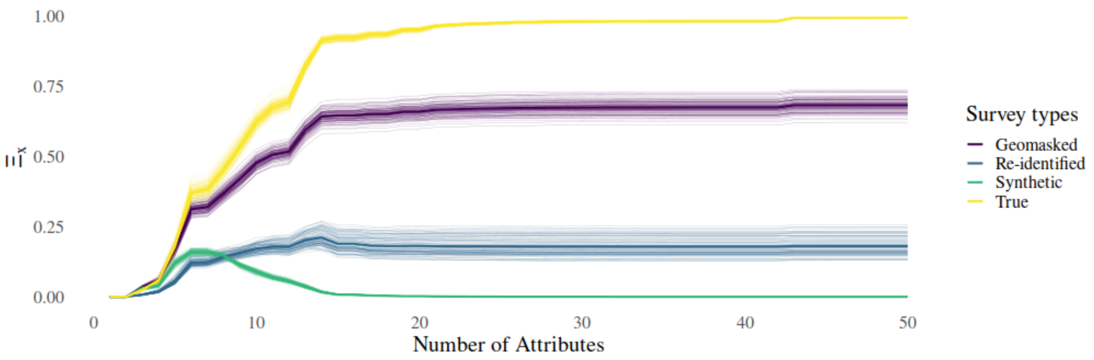


Figure 4.2: **Population uniqueness across survey types.**
Share of population-unique survey respondents for 100 simulation runs. The thick lines represent the average population uniqueness across the 100 simulation runs, the thin lines individual simulation runs. In the *true* survey, no attribute is perturbed. In the *geomasked* survey, the cluster identifier is perturbed. In the *synthetic* survey, all variables but the cluster identifier are perturbed. In the *re-identified* survey, the synthetic survey is used to predict the ”private” attribute – i.e. the cluster – in the original dataset along the lines of the proposed microdata dissemination strategy. The re-identified original survey is then used to calculate population uniqueness. Both the re-identified and the synthetic survey provide significant privacy gains vis-à-vis the other survey types.

Therefore, Figure 4.2 gives a strong indication that geomasking provides little additional safeguards for the respondents' privacy compared to the true survey in the presence of third-party information on a subset of the contained attributes. Hence, we consider an alternative microdata dissemination strategy: instead of publishing original microdata with perturbed cluster locations, we investigate the option of publishing two datasets - original microdata stripped of geographic identifiers below the level of strata and synthetic microdata with the original cluster locations. The choice is motivated by adopting a user-centric perspective: official household survey publications predominantly report on results up to the strata-level as results below are usually considered not representative. Analysis that benefits from below strata-level data often investigates proximity-related questions such as distances to certain locations and surrounding habitat. For the former, cluster locations are of minor importance, for the latter, however, the perturbation procedure introduces significant levels of uncertainty to the analysis (Warren et al., 2016). The alternative microdata dissemination strategy obviously conserves data utility for analysis on the representative level via the first dataset, while the second dataset allows for the accurate capture of proximity-related information. However, two potential shortcomings need to be considered: first, can we use the synthetic dataset to predict the 'private' attribute in the original dataset, i.e. the small area identifier, thus bypassing the privacy protection measures? Second, is the uncertainty we introduce by synthesizing the non-spatial attributes for spatial analysis smaller than the uncertainty from perturbing the cluster locations?

4.2.4 Risk of re-identifying private geocodes

To investigate the first shortcoming, we train a random forest model on the small area identifier - the zip code - in the anonymized surveys for each stratum separately. We use the trained models on the original data to predict the zip code for each record. Finally, we evaluate our predicted label against the original label. In addition, we compare the outcomes to randomly guessing the correct label in order to account for the number of small areas within each stratum. Figure 4.3 shows the median accuracy of the approaches across 100 simulation runs. While we are able to successfully re-construct the original zip code in most cases for the geomasked survey, it does not work much better for the synthetic data than for the random guess.

In our experiment, only one stratum consequently hosts more than ten small areas across all simulation runs, with one stratum hosting only two small areas in some simulation runs, giving the random guess also a good chance to predict correctly. Recalling from Figure 4.2 that roughly 70% of the displaced clusters stay within the same zip code in the geomasked survey, even predicting the sensitive attribute for strata hosting as little as two small areas, average population uniqueness in the synthetic data would not exceed much the 50/50-chance of the random guess%, thus providing better privacy protection in the re-identified original survey than the geomasked alternative.

4.2.5 Utility for survey augmentation

To give an indication about the utility of the different anonymization approaches, we use a setup common in recent academic literature (cf. Pokhriyal and Jacques (2017); Leasure et al. (2020); Schmid et al. (2017)): we augment the surveys with auxiliary information from geospatial (big) data. Specifically, we construct zip code-level aggregates from gridded satellite-derived

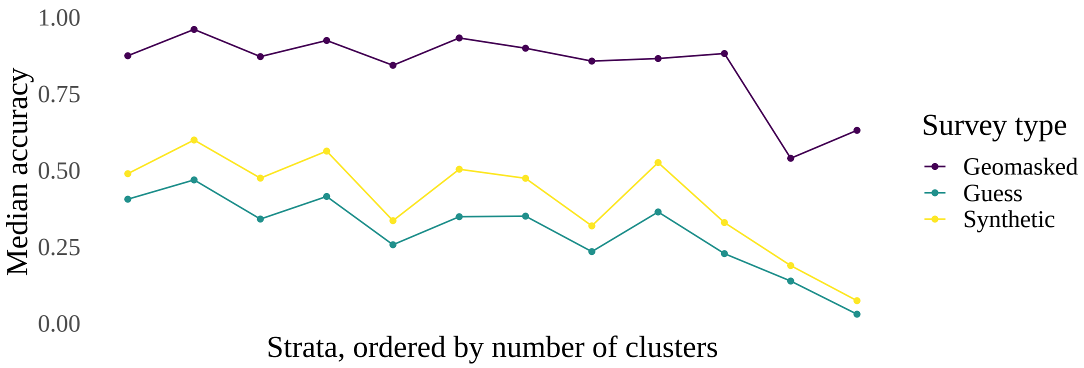


Figure 4.3: **Re-identification of the zip code as private attribute in the true survey for each stratum across 100 simulation runs.**
Accuracy is measured by the share of successfully re-identified zip code labels in the true survey. A random forest model is trained on perturbed data, i.e. the geomasked and the synthetic survey, respectively. We evaluate the results against the true zip code labels in the true survey and compare them against random guesses of the private attribute.

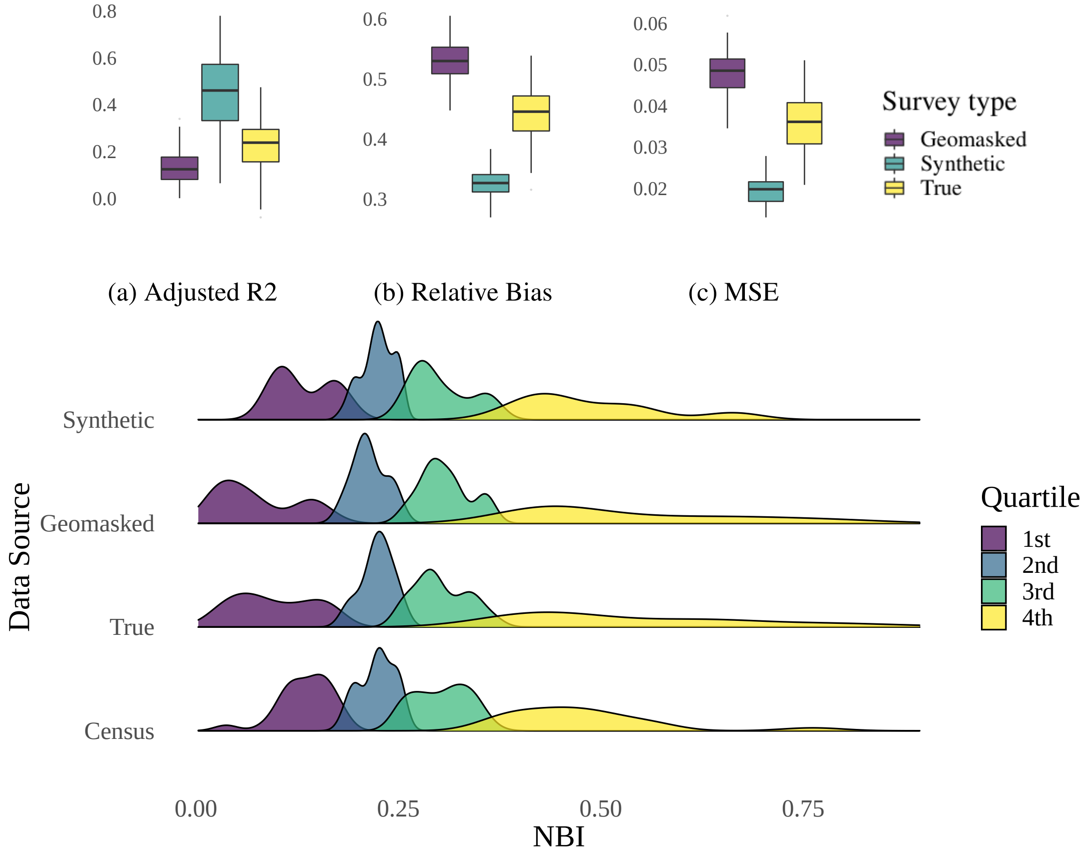
features available from the WorldPop repository (WorldPop, 2018) and combine them with zip code-level survey aggregates to provide predictions, especially for areas not sampled in the survey. As our target variable, we select the Unsatisfied Basic Needs index (*Necesidades Básicas Insatisfechas (NBI)*) - a composite indicator similar to the multidimensional poverty index (MPI) (Méndez and Bravo, 2011; Alkire et al., 2019) used as a key statistical indicator in Costa Rica. Details on the index can be found in the Supplementary Information. We evaluate our predictions against the census in terms of adjusted R^2, bias and the Mean Squared Error (MSE). Figures 4.4a - 4.4c show the performance along these three evaluation criteria across 100 simulation runs.

Surprisingly, the synthetic approach not only outperforms the geomasked survey, it also provides predictions more in line with the census results than the true survey. A possible explanation could be that the copula approach reduces the impact of outliers on the zip code-specific NBI sample averages. This explanation is supported by Figure 4.4d that shows the distribution of zip code-level NBI averages grouped into quartiles for one simulation run as both the synthetic survey and the census showcase smaller tails in their distributions, respectively.

We run additional experiments to compare the directly synthesized NBI and its underlying indicators with their counterparts computed from synthetic survey variables (see Supplementary Table C.2 and Supplementary Fig. C.4).

4.3 Discussion

In this paper, we proposed and evaluated an alternative data dissemination strategy for micro-level survey data that improves the trade-off between privacy risk and data utility. Specifically, we showed that by publishing two datasets, namely the original survey data with limited geographic identifiers and a synthetically-generated survey dataset with the true cluster locations, re-identification risks can be reduced significantly vis-à-vis popular geomasking approaches without incurring additional losses in terms of data utility for survey augmentation. This could



(a) Adjusted R2 (b) Relative Bias (c) MSE

(d) NBI densities

Figure 4.4: **Performance metrics of survey-based NBI estimates on the zip code-level.** (a) Adjusted R^2 is based on the in-sample zip codes. (b) and (c) are based on the full sample and predictions are evaluated against the census across 100 simulation runs. (d) compares zip code-level NBI averages for a single simulation run.

help mapping initiatives such as WorldPop or GRID3 to improve their products as more accurate spatial data is available. In addition, by separating the marginals from the dependence structure, it provides data producers such as National Statistical Offices also with a useful tool to update the respective synthetic microdata files for the following years by updating the margins with nationally representative new data as sub-nationally representative surveys may only be conducted every few years. In the Supplementary Information, we further investigate the stability of our results by alternating the experiment design.

First, while we chose the strata for the main analysis as they provide 'large-enough' sample sizes at the same time explicitly accounting for at least high-level regional variation, we study in further experiments whether fitting on smaller or larger geographic levels may better capture local variation at the expense of running into the risk of small sample problems or vice versa. Supplementary Fig. C.1 summarizes the results for our copula model being fitted on the whole survey, the twelve strata and the zip code-level, respectively. It shows that by selecting the strata as our fitting level, we strike a balance between the underlying sample size (usually the larger the better) and capturing regional variation (usually the more disaggregated the better) both in terms of utility and risk. In addition, by using subsets of the full microdata for model fitting, the approach becomes computationally tractable also for larger surveys.

Second, since generative models allow us to sample an arbitrary number of synthetic observations, we look at the impact of the synthetic sample size on the outcomes of the survey augmentation experiment, notably the adjusted R^2 and a measure of confidence in the direct survey estimates of the Fay-Herriot model (cf. Section 4.4.2) - the shrinkage factor γ. Supplementary Fig. C.2 shows that with an increasing sample size, γ increases as well, thus shifting more weight to the direct estimate. Even though intuitive as the sampling variance naturally decreases in n_D, at some point it may become misleading with potentially negative effects on the model performance as the synthetic data generating process still relies on the same information conveyed in the true survey with sample size n_D. However, in our experiment the adjusted R^2 does not exhibit a bump, but increases monotonically, thus hinting at little additional explanatory power of our satellite-derived covariates vis-à-vis the area-level direct survey estimate for the in-sample areas.

Third, since our target variable *NBI* is a composite indicator, we compare the different composition levels of the synthetic NBI with the NBI constructed from synthetic data. While the divergence measure shows an overall good fit for the underlying indicators (see Supplementary Table C.2 and Supplementary Fig. C.4), correlations are low, especially for higher-level compositions as the dimensions or the NBI itself.

Lastly, we test alternative encoding schemes for the transformation of categorical data. Also, we relax our assumption of the normally distributed margins by opening up to a wider group of parametric copulas (such as beta, gamma or uniform distributions) selected for each margin individually based on the two-sample Kolmogorov-Smirnov (KS) statistic to study the effect of the specification choice on the normalized KL divergence. Supplementary Fig. C.3 shows that neither the encoding scheme nor the specification of the marginal distributions have large effects on the quality of the synthetically generated data.

Nevertheless, our approach is not without limitations. The copula-based approach towards synthetic data generation largely fails to correctly capture lower-level hierarchical relationships such as *individuals - line numbers - households - houses* from the original data. As said before, since we see our analysis using a naïve Gaussian copula model as providing somewhat a lower bound for improving the utility-risk trade-off by adopting the proposed microdata dissemination strategy vis-à-vis common geomasking approaches, there is much room for improvement. To name a few, latent copula designs can be considered to avoid data transformations, marginal distributions can be modelled non-parametrically, hierarchical structures can be accounted for more rigorously by either modelling the hierarchies separately as suggested by Templ (2017) or by modelling the relationships explicitly.

In addition, synthetic data may - under some circumstances - leak private information, e.g. through the generated value ranges. As a response, differentially-private implementations of existing generative models have been proposed such as PrivBayes (Zhang et al., 2017), PrivSyn (Zhang et al., 2021) and PATE-GAN (Jordon et al., 2019).

4.4 Methods

4.4.1 Fitting Gaussian copulas to survey attributes

As an alternative to geomasking, we use synthetically generated survey attributes for protecting the respondents' privacy while keeping the true point locations of the selected clusters. To do so, we fit a Gaussian copula model on the original survey attributes **x** and sample from the learned joint distribution for each cluster individually with the originally sample size n_j. Therefore, consider our survey D, where X_1 represents a random variable with a continuous marginal cumulative distribution function (cdf) denoted by $F_1(x_1) = P(X_1 \leq x_1)$. For the multivariate case, the joint cdf can be generalized to $F_{1,\ldots,d}(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$.

A copula, firstly introduced in the work of Sklar (1959), is a cumulative density function with uniform marginals between [0,1]. Thus - based on Sklar´s theorem (Sklar, 1959) - when all variables are continuous, the d-dimensional random vector $X_1, \ldots, X_d$ can be defined in a uniform space $[0, 1]^d$, creating a random vector $U_1, \ldots, U_d$ via the probability integral transform $u_d = F_d(x_d)$. In this case, a unique d-dimensional copula $C(u_d)$ exists:

$$C(u_1, \ldots, u_d) = F\big(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\big) \tag{4.4}$$

As motivated in Section 4.2.2, we account for the fact that household surveys largely consist of categorical variables by applying data transformation. Among the plethora of possible encoding schemes, the most common encoding scheme is one-hot encoding, where for each class of a categorical variable a binary dummy variable is created (Benali et al., 2021).

A disadvantage of this option is that it may become computationally challenging and prone to multicollinearity in the presence of variables with a high cardinality, i.e. with a large number of classes, since each possible class creates a new variable (Bourou et al., 2021). Interestingly, there is – to the best of our knowledge – little comprehensive, comparative and conclusive scientific evidence on the properties and performance of different categorical encoding schemes. Therefore, we explore two other well-known alternatives with more favourable computation times: ordinal and frequency encoding. Ordinal encoding uses integers to represent each classes, e.g. from 0 to v, the number of classes in a categorical variable. Assigning an unreal order to nominal variables is the main pitfall of this alternative (Jiang et al., 2020). Frequency encoding – as used in medical imaging (Mansfield and Maudsley, 1977) and similar to the concept of *term frequency* in Natural Language Processing (Aizawa, 2003) – assigns an interval in [0,1] to each class based on and ordered by its proportion of occurrence. Then, it uses the middle point of each interval as float representative of the respective class. Back-transformation is done by assigning a new point to a class via the respective interval it falls into. In this sense, this alternative conveys information of the importance of each class (Sabharwal and Agrawal, 2021). Based on the results of the different encoding schemes shown in Supplementary Fig. C.3, we opt for the frequency encoding scheme in the following. Algorithms 3 and 4 provide details on the chosen scheme.

Algorithm 3: Transform categorical variables

```
Input $D = (X_1, \dots, X_d)$
Output $\tilde{D} = (\tilde{X}_1, \dots, \tilde{X}_d)$

for $X_d \in D$ do
    if $X_d$ is Continuous then
        $\tilde{X}_p \leftarrow X_d$
    end
    if $X_d$ is Non-continuous then
        $\tilde{X}_q \leftarrow \mathrm{T}(X_d)$
    end
end
$\tilde{D} \leftarrow (\tilde{X}_p, \tilde{X}_q) \quad \forall \quad p \in P$ and $q \in Q$
```

Algorithm 4: Back-transform frequency encoded variables

```
Input $\tilde{S} = (\tilde{Y}_1, \dots, \tilde{Y}_d)$
Output $S = (Y_1, \dots, Y_d)$

for $\tilde{Y}_d \in \tilde{S}$ do
    if $\tilde{Y}_d$ is not indexed as variable in Q then
        $Y_p \leftarrow \tilde{Y}_d$
    end
    if $\tilde{Y}_d$ is indexed as variable in $Q$ then
        $Y_q \leftarrow T^{-1}(\tilde{Y}_d)$
    end
end
$S \leftarrow (Y_p, Y_q) \quad \forall \quad p \in P$ and $q \in Q$
```

Thus, the d-dimensional Gaussian copula $C_{\Sigma}^{G}(\mathbf{u})$ is defined as the cdf of a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with $\Sigma \in \mathbb{R}^{dxd}$ represented on the unit cube $[0,1]^d$:

$$C_{\Sigma}^{G}(u_1, \dots, u_d) = \phi_{\Sigma}\big(\phi^{-1}(u_1), \dots, \phi^{-1}(u_d)\big) \tag{4.5}$$

The density of a Gaussian copula is then defined as:

$$c_{\Sigma}^{G}(\mathbf{u}) = \frac{1}{\sqrt{\det \Sigma}} \exp\Big(-\frac{1}{2}\phi^{-1}(\mathbf{u})^T \cdot (\Sigma^{-1} - I) \cdot \phi^{-1}(\mathbf{u})\Big) \tag{4.6}$$

with $\mathbf{u} \in [0,1]^d$, $I \in \mathbb{R}^{dxd}$ being the identity matrix, and ϕ^{-1} being the inverse cumulative distribution function of a standard normal distribution. Σ is a positive semi-definite covariance matrix that we estimate based on Pearson's correlation coefficient ρ (Li et al., 2014).

As noted in Section 4.2.2, we sample for each cluster individually with a sample size of n_j. While rejection sampling could be an option for ensuring only synthetic rows with the respective cluster identifier are selected, it proves computationally inefficient. With copulas being multivariate cdfs, we introduce conditions instead. Hence, we sample from a multivariate normal distribution conditional on cluster j. Thus, our transformed dataset $\tilde{D}$ with one conditional variable becomes $\tilde{D} = (\tilde{\mathbf{X}}_a | \tilde{\mathbf{X}}_b)$ with $\tilde{\mathbf{X}}_a := \tilde{X}_1, \dots, \tilde{X}_{d-1}$ being the transformed attributes to be synthesized and $\tilde{\mathbf{X}}_b := \tilde{X}_d$ being the transformed cluster identifier. The parameters of the respective multivariate normal distributions are thus partitioned into:

$$\tilde{D} = \begin{bmatrix} \tilde{\mathbf{X}}_a \\ \tilde{\mathbf{X}}_b \end{bmatrix}, \ \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \tag{4.7}$$

with $\mu_a \in \mathbb{R}^{d-1}$ and $\mu_b \in \mathbb{R}^1$ and $\Sigma_{aa} \in \mathbb{R}^{(d-1)\times(d-1)}$, $\Sigma_{ab} \in \mathbb{R}^{(d-1)\times 1}$, $\Sigma_{ba} \in \mathbb{R}^{1\times(d-1)}$, and $\Sigma_{bb} \in \mathbb{R}^{1\times 1}$ being the means and positive semi-definite covariance matrices, respectively. Following Algorithm 2, the parameters of our estimated marginal distributions Ψ and of the copula $C_{\Sigma}^{G}(\mathbf{u})$ need to be adapted to mirror the conditionality such that $\Psi_{a|b}(\tilde{\mathbf{X}}_a|\tilde{\mathbf{X}}_b)$ and $C_{\Sigma}^{G}(\mathbf{u}_a|\mathbf{u}_b)$.

Consequently, we sample from $\sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ with:

$$\bar{\mu} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(X_b - \mu_b) \in \mathbb{R}^{d-1} \tag{4.8}$$

and

$$\bar{\Sigma} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{4.9}$$

We iterate the copula-based fitting and sampling procedure for every stratum separately as it allows to better capture sub-national variation using representative sub-samples and as it proves computationally more tractable. For sampling designs with varying household- or individual-level inclusion probabilities (e.g. in the DHS, women - in comparison to men - are usually oversampled), Templ (2017) suggests to sample a synthetic population and re-iterate the sampling procedure to produce valid synthetic sampling weights. As in our design sampling weights are identical across households for a given PSU due to the systematic sampling approach in the second stage, the original sampling weights remain valid. The virtue in our model choice is the relative simplicity, little requirements in terms of ex-ante knowledge about the individual distributions X_d and its computational efficiency. For further experiments on the robustness and sensitivity of our modelling choices, we refer to the Supplementary Information.

4.4.2 Area-level survey augmentation methods

Survey data can be augmented with the use of area-level models, e.g. the Fay-Herriot model (Fay and Herriot, 1979) by linking direct estimators gathered from survey data to relevant auxiliary information. Both, direct estimators, and auxiliary data are aggregated on $k = 1, \ldots, D$ areas. Traditionally, these auxiliary covariates $\mathbf{x}_k$ are obtained from recent censuses, administrative records or other geospatial (big) data sources. In this paper, we make use of satellite imagery features as area-level covariates. The Fay-Herriot is a two-level model, the first part is composed by the sampling model:

$$\hat{\theta}_k^{\text{Dir}} = \theta_k + e_k, \qquad e_k \sim N(0, \sigma_{e_k}^2), \tag{4.10}$$

where the sampling error is represented by e_k and $\hat{\theta}_k^{\text{Dir}}$ is the direct estimator of θ_k (e.g. sample mean). The linking model provides the second part, where relevant area-level covariates are considered:

$$\theta_k = \mathbf{x}_k'\hat{\beta} + u_k. \tag{4.11}$$

Here, the random area effects u_k are assumed to be independent with mean 0 and variance σ_u^2. The empirical best linear unbiased predictor (EBLUP) estimator is given by:

$$\hat{\theta}_k^{\text{FH}} = \gamma_k\hat{\theta}_k^{Dir} + (1 - \gamma_k)\mathbf{x}_k'\hat{\beta} = \mathbf{x}_k'\hat{\beta} + \hat{u}_k, \tag{4.12}$$

with $\gamma_k = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_k}^2}$ denoting the shrinkage factor for each area k. The parameter estimates of this model can be obtained via maximum likelihood (ML) or restricted ML (REML). Note that the shrinkage factor allows to weight in favor of the direct estimator when sampling variances are small; on the contrary the synthetic estimator $\mathbf{x}_k'\hat{\beta}$ receives more weight when the sampling variance is larger. Results on an experiment studying the sensitivity of the shrinkage factor and adjusted R^2 for varying synthetic sample sizes are shown in Supplementary Fig. C.2. Further details on the Fay-Herriot model can be found in Rao and Molina (2015).

Supplementary material C

C.1 Data description

As our reference dataset in this project, we use data from Costa Rica – notably the X$^{\text{th}}$ Population and VI$^{\text{th}}$ Housing Census of Costa Rica, 2011 (Censo Nacional de población y Viviendas de Costa Rica 2011) – to produce three different data file types: First, we draw survey samples from a census population using a stratified two-stage cluster sample design without applying any statistical disclosure control mechanisms. We use these survey samples (called *true* surveys in the study) as starting point for creating file types two and three: By re-assigning clusters to new zip codes based on the displacement algorithm described in Algorithm 1, we perturb the zip code identifier in the true surveys, thereby creating the *geomasked* surveys. Again based on the true surveys, we apply the copula-based synthetic data generation algorithm described in Algorithm 2 to generate synthetic data for each attribute except the zip code, which keeps it original structure. In addition, in order to test the robustness of our specifications, we create additional datasets with alternating data generating process designs. The censuses are carried out every ten years by the national statistic office of Costa Rica (INEC) and collect information of people, households, and dwellings on topics such as access to education, employment, social security, technology necessary for the planning, execution, and evaluation of public policies (Méndez and Bravo, 2011).

Administratively, Costa Rica had in 2011 four disaggregation levels: two zones, six planning regions, 81 cantons and 473 districts (municipalities). The sampling design used for the main National Household Survey (Encuesta Nacional de Hogares, ENAHO) specifies twelve strata - each planning region divided by urban and rural areas. In this case, the strata coincide with the study domains. For our experiment, we use a 10% random sample of the original 2011 census, which can be obtained from the Instituto Nacional de Estadistica y Censos (2022) as a pseudo-population. The smallest geographical information available in this dataset are the 473 districts. In the first stage, we select districts as our PSUs for each stratum separately with a selection probability proportional to population size. In the second stage, we select a minimum of 10 households in each PSU by using simple random sampling without replacement. PSUs with less than 10 households are discarded from this procedure, affecting roughly 4% of all PSUs.

As auxiliary information, we use covariates derived from satellite imagery. Specifically, we use features derived from satellite imagery provided by WorldPop (2018) in our survey augmentation setup. The advantages of using satellite imagery here are five-fold: Data with

N_C	n_D	# of PSUs in C	# of PSUs in D	# of attributes
427830	[7638; 11914]	767	123	106

Table C.1: **Descriptive statistics on the census-derived data across 100 simulation runs**

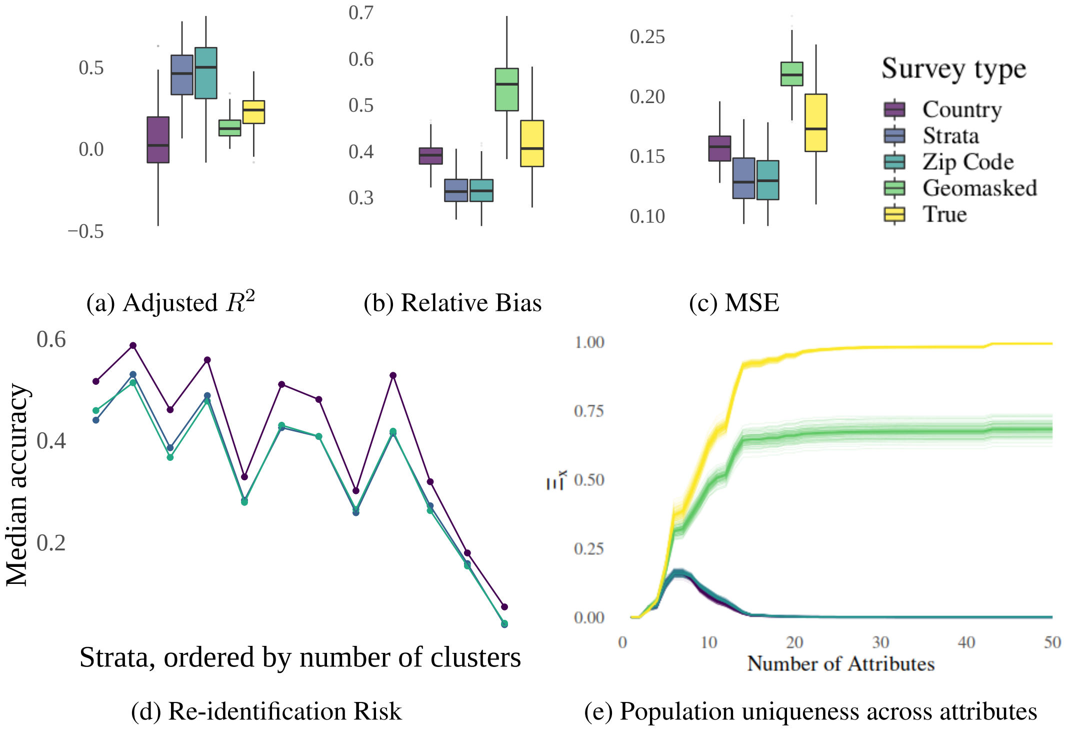
virtually global coverage at high spatial resolutions for frequent time intervals on human-made impact provided in a structured format enables us to extract covariates for all administrative areas in Costa Rica at the time of the census. Therefore, we can use area-level survey augmentation (cf. Section 4.4.2) to provide estimates, especially for areas not covered by the respective survey. WorldPop data are provided in the tagged image file format (TIFF) with a pixel representing roughly a 100m × 100m grid square in an open data repository under CC4.0 licence (WorldPop, 2018). Pixel values are aggregated to the administrative areas of Costa Rica via their centroids. Specifically, we generate area-level averages for the distances to different types of natural areas (e.g. cultivated, woody-tree, and shrub areas, coastlines etc.) and to infrastructure such as roads and waterways, the intensity of night-time lights, topographic information and information on the presence of human settlements.

C.2 Sensitivity of copula vis-à-vis geographic fitting level

In order to study the effect of the geographic level on the copula modelling performed for synthetic data generation, we run Algorithm 2 on the whole survey ('Country), the twelve strata ('Strata') and the roughly 110 zip code areas ('Zip Code'), respectively. Results are provided in Supplementary Fig. C.1. It appears that fitting the copula model on the whole survey limits the ability of the approach to capture regional variations. On the other hand, model fitting on the zip code-level does neither increase the re-identification risk of the zip code identifier as a private attribute and nor affect the overall prediction performance of the outcome variable, hinting at overfitting not being a problem on that level. Striking a balance between underlying sample size and a certain level of disaggregation shows better results. Also, it allows to scale computations to settings with larger samples and more attributes.

C.3 Effects of synthetic sample size on prediction outcomes

Generative models can be used to create synthetic samples of an arbitrary size regardless the amount of underlying data. While the advantages of that are similar to those of other resampling procedures such as bootstrapping (i.e. to estimate the precision of the sample statistics or to perform cross-validation), it can also mislead modelling approaches that 'borrow strength' from auxiliary data by overestimating the strength of the synthetic direct estimates eventually resulting in losses of explanatory power of the model. In our survey augmentation setup, the shrinkage factor γ indicates whether final estimates rather rely on the direct estimates from the synthetic survey or on the satellite-derived covariates for the in-sample predictions depending on the sampling variance. Supplementary Figure C.2 shows that larger sample sizes lead to increasing gamma values (via decreasing sampling variances of the direct estimator), however, not incurring losses in the goodness-of-fit of our estimation model. This hints at the fact that



(a) Adjusted R^2 (b) Relative Bias (c) MSE

(d) Re-identification Risk (e) Population uniqueness across attributes

Figure C.1: **Evaluation metrics for different geographical copula fitting levels.**
(a) - (c) The copula model is fitted on the whole survey (‘Country’), for each of the twelve strata (‘Strata’) and for each of the roughly 110 zip codes (‘Zip Code’) separately. As a reference, the metrics for the geomasked and the true survey are provided as well. (d) The accuracy to successfully re-identify the zip code as a private attribute in the original data using a random forest model trained on synthetic data across fitting levels remains similar. (e) The share of population-unique survey respondents is virtually not affected by the copula fitting level.

the contribution of the auxiliary information to the explanatory power of the model for the in-sample predictions is negligible.

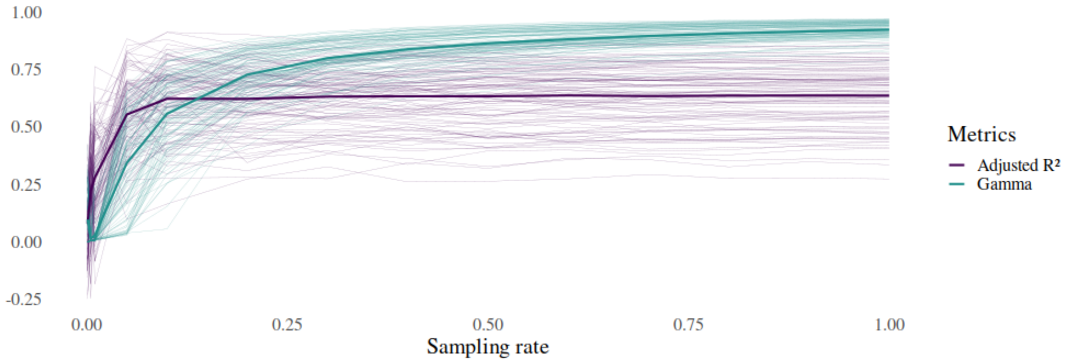


Figure C.2: **Sensitivity of model performance on changes in synthetic sample size.**
Samples are drawn from a synthetic population. The synthetic population is generated using the copula-based approach described in Section 4.2.2. Sample sizes are determined by the sampling rate (shown on the x-axis). Results are evaluated against the true census population. The shrinkage factor γ is averaged across zip codes. The thick lines represent the metric averages across the 100 simulation runs, the thin lines individual simulation runs.

C.4 Choosing marginal distributions & encoding schemes

As already mentioned in Section 4.2.2, assuming normally-distributed margins may represent a misspecification of the true univariate distribution of X_d. In addition, computationally tractable alternatives to one-hot encoding exist. We compare two different ways to model the marginal distributions together with two different encoding schemes. The results are presented in Supplementary Fig. C.3. Measured by the normalized KL divergence averaged across 100 simulation runs, frequency encoding produces slightly better goodness-of-fit of the synthetic data ($Z_{KL} = 0.74$ for frequency encoding versus $Z_{KL} = 0.72$ for ordinal encoding with gaussian marginals). Surprisingly, the naïve assumption of normally distributed marginals outperforms the KS-based parametric marginals with $Z_{KL} = 0.74$ and $Z_{KL} = 0.70$, respectively.

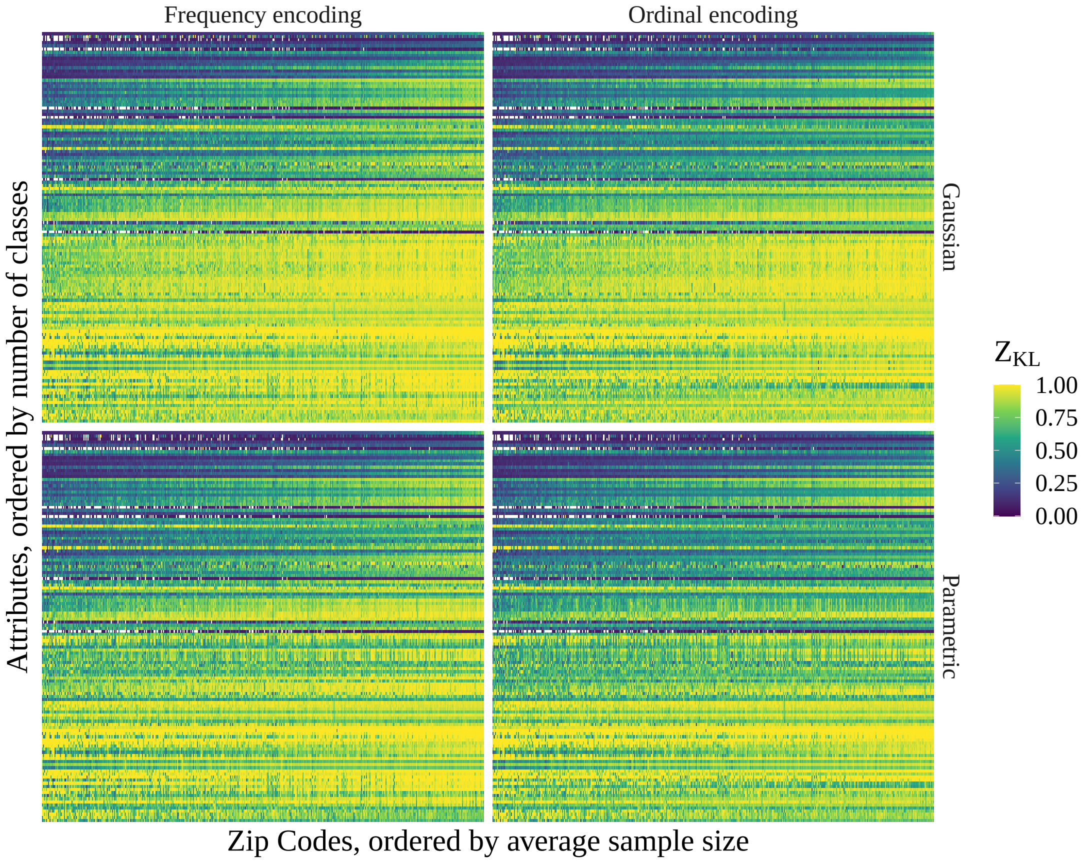


Figure C.3: **Effect of encoding schemes and marginal distribution choice on the overall goodness-of-fit of the synthetic data measured by the normalized KL divergence Z_{KL} (in bits).**

The attributes on the y-axis are ordered by their respective number of classes, the zip codes on the x-axis are ordered by their average sample size across simulation rounds. Values close to one (yellow) represent little divergence from the true census distribution and therefore indicate a high goodness-of-fit.

C.5 Detailed analysis of the NBI as composite indicator

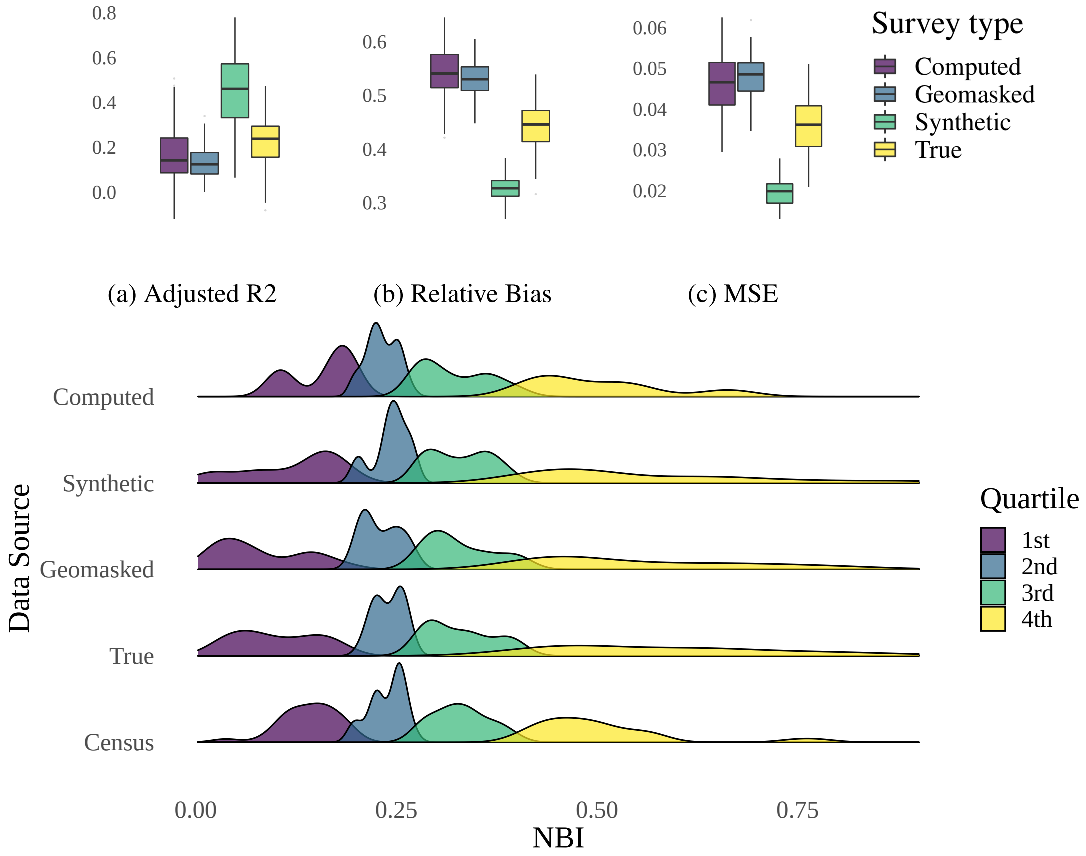
The NBI is a composite indicator computed from approx. 20 underlying survey variables grouped into four dimensions (i.e. access to decent housing (*Acceso albergue digno*), access to a healthy life (*Acceso a vida saludable*), access to knowledge (*Acceso al conocimiento*) and access to other goods and services (*Acceso a otros bienes y servicios*)) using 19 indicators in total. All indicators and dimensions are binary (yes/no). An identified need in one of the indicators leads to a positive needs status in higher dimensions. The sensitivity for false positives is thus assumed to be high for the NBI as a small change (e.g. one year age difference) in one of the 19 underlying variables can turn a NBI-negative to a NBI-positive survey respondent.

Generally, two strategies for computed indicators exist to create synthetic counterparts: a) directly synthesize the computed indicators or b) re-construct the indicator based on synthetic survey variables. While the former is more likely to reflect the original distribution, it may not be consistently decomposable into its underlying indicators; vice-versa holds for the latter. The strength of these effects are largely determined by the complexity and sensitivity of the composite indicator and the overall goodness-of-fit of the synthetic data. Thus, if both approaches produce similar compositions, it can be regarded as a strong indication that the underlying synthetic data also successfully captures relationships across multiple variables in the dataset, not only the composite index. Supplementary Table C.2 shows that this not fully holds for the NBI.

Indicators	# of indicators	Pearson's ρ	Z_{KL}	Incidence
1.x	5	0.42	0.99	100
Dimension 1		0.24	0.98	647
2.x	5	0.22	0.98	85
Dimension 2		0.19	0.98	455
3.x	2	0.02	0.89	507
Dimension 3		0.02	0.84	1845
4.x	7	0.02	0.99	60
Dimension 4		0.03	1.00	622
Composite NBI	19	0.07	0.97	3253

Table C.2: **Relationship between synthetic and computed NBI indicators across 100 simulation runs.**
Indicator-level results (e.g. 1.x) are averaged across indicators. The incidence describes the average number of respondents across 100 simulated surveys with unsatisfied needs in the respective indicator/dimension.

Although the overall number of survey respondents with unsatisfied needs are captured with a high accuracy as measured by the normalized KL divergence Z_{KL} for binary data, the NBI status on the individual level strongly diverges following Pearson's ρ (cf. Supplementary Table C.2). Supplementary Figure C.4 shows that the lack of linear correlation is mainly due to improperly captured relationships in the underlying variables than in the synthetic NBI as the former is outperformed by the latter for survey augmentation expressed in terms of adjusted R^2, bias and MSE. However, it remains on par with the geomasked survey at lower privacy risks.



(a) Adjusted R2

(b) Relative Bias

(c) MSE

(d) NBI densities

Figure C.4: **Performance of the synthetic vs. computed composite NBI.**
(a) - (c) show of the different survey types in our survey augmentation experiment across 100 simulation runs. (d) shows the densities of the composite NBI by quartiles for one simulation run.

Bibliography

Abhayawardhana, V. S., I. J. Wassellt, D. Crosby, M. P. Sellars, and M. G. Brown (2005). Comparison of empirical propagation path loss models for fixed wireless access systems. In IEEE Vehicular Technology Conference, Volume 61, pp. 73–77.

Agence Nationale de la Statistique et de la Démographie (2012). Demographic and Health Survey - Multiple Indicator Cluster Survey (2010-2011).

Agence Nationale de la Statistique et de la Démographie (2013). Rapport Projection de la Population du Senegal 2013-2063. Technical report, Agence Nationale de la Statistique et de la Démographie.

Agence Nationale de la Statistique et de la Démographie (2015). Enquête de Mise à jour du Registre National Unique des Ménages Vulnérables.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). Machine learning and phone data can improve targeting of humanitarian aid. Nature 2022 603, 864–870.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management 39(1), 45–65.

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods & Applications 20(3), 383–407.

Alkire, S., U. Kanagaratnam, and N. Suppa (2019). The Global Multidimensional Poverty Index (MPI) 2019. OPHI MPI Methodological Note 47.

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi (2013). Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 901–914.

Arambepola, R., S. H. Keddie, E. L. Collins, K. A. Twohig, P. Amratia, A. Bertozzi-Villa, E. G. Chestnutt, J. Harris, J. Millar, J. Rozier, et al. (2020). Spatiotemporal mapping of malaria prevalence in madagascar using routine surveillance and health survey data. Scientific Reports 10(1), 1–14.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically masking health data to preserve confidentiality. Statistics in Medicine 18(5), 497–525.

Bakker, M. A., D. A. Piracha, P. J. Lu, K. Bejgo, M. Bahrami, Y. Leng, J. Balsa-Barreiro, J. Ricard, A. J. Morales, V. K. Singh, B. Bozkaya, S. Balcisoy, and A. Pentland (2019). Measuring Fine-Grained Multidimensional Integration Using Mobile Phone Metadata: The Case of Syrian Refugees in Turkey. In Guide to Mobile Data Analytics in Refugee Scenarios, pp. 123–140. Cham: Springer International Publishing.

Bell, W. R. and C. Franco (2015). Borrowing information over time in binomial/logit normal models for small area estimation. Statistics in Transition new series 16(4), 563–584.

Benali, F., D. Bodénès, N. Labroche, and C. de Runz (2021). MTCopula: Synthetic complex data generation using copula. In 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP), pp. 51–60.

Blankespoor, B., T. Croft, T. Dontamsetti, B. Mayala, and S. Murray (2021). Spatial anonymization: Guidance note prepared for the Inter-Secretariat working group on household surveys. Technical report, UN Inter-secretariat Working Group on Household Surveys Task Force on Spatial Anonymization in Public-Use Household Survey Datasets.

Blondel, V. D., A. Decuyper, and G. Krings (2015). A survey of results on mobile phone datasets analysis. EPJ data science 4(1), 10.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076.

Blumenstock, J., M. Callen, and T. Ghani (2018). Why do defaults affect behavior? Experimental evidence from Afghanistan. American Economic Review 108(10), 2868–2901.

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. AEA Papers and Proceedings 108, 72–76.

Bonafilia, D., J. Gill, D. Kirsanov, and J. Sundram (2019). Mapping for humanitarian aid and development with weakly-and semi-supervised learning. Technical report, Facebook.

Boo, G., E. Darin, D. R. Thomson, and A. J. Tatem (2020). A grid-based sample design framework for household surveys. Gates Open Research 4.

Botta, F., H. S. Moat, and T. Preis (2015). Quantifying crowd size with mobile phone and Twitter data. Royal Society Open Science 2(5), 150162.

Bourou, S., A. El Saer, T. H. Velivassaki, A. Voulkidis, and T. Zahariadis (2021, sep). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. Information 12(9), 375–389.

Brown, M. E., K. Grace, G. Shively, K. B. Johnson, and M. Carroll (2014). Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. Population and Environment 36(1), 48–72.

Bruckschen, F., T. Koebe, M. Ludolph, M. F. Marino, and T. Schmid (2019). Refugees in Undeclared Employment - A Case Study in Turkey. In Guide to Mobile Data Analytics in Refugee Scenarios, pp. 329–346. Cham: Springer International Publishing.

Burgert, C. R., J. Colston, T. Roy, and B. Zachary (2013). Geographic Displacement Procedure and Georeferenced Data Release Health Surveys. DHS Spatial Analysis Reports. Technical Report 7, ICF International, USAID, Calverton, Maryland, USA.

Burgert-Brucker, C. R., T. Dontamsetti, and P. W. Gething (2018). The dhs program's modeled surfaces spatial datasets. Studies in Family Planning 49(1), 87–92.

Carter, G. and J. Rolph (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. Journal of the American Statistical Association 69 (348), 880–885.

Casas-Cordero, C., J. Encina, and P. Lahiri (2016). Poverty mapping for the Chilean comunas. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 379–403. Wiley.

Chambers, R., N. Salvati, and N. Tzavidis (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. Journal of the Royal Statistical Society: Series A 179 (2), 453–479.

Chandra, H., N. Salvati, and R. R. Chambers (2015). A Spatially Nonstationary Fay-Herriot Model for small area estimation. Journal of Survey Statistics and Methodology 3(2), 109–135.

Chatterjee, S., P. Lahiri, and H. Li (2008). Parametric Bootstrap Approximation to the Distribution of EBLUP and related Predition Intervals in Linear Mixed Models. The Annals of Statistics 36 (3), 1221–1245.

Chen, X. and W. D. Nordhaus (2011, 5). Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences of the United States of America 108(21), 8589–8594.

Chi, G., H. Fang, S. Chatterjee, and J. E. Blumenstock (2022, jan). Microestimates of wealth for all low- and middle-income countries. Proceedings of the National Academy of Sciences of the United States of America 119(3), e2113658119.

Damasso, L. M. Correia, E. (1999). Digital Mobile Radio Towards Future Generation. Technical Report 11, European Commission, Luxembourg.

Datta, G. S., M. Ghosh, R. Steorts, and J. Maples (2010). Bayesian benchmarking with applications to small area estimation. TEST 20 (3), 574–588.

Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10, 613–627.

Datta, G. S., J. N. K. Rao, and D. D. Smith (2005). On measuring the variability of small area estimators under a basic area level model. Biometrika 92 (1), 183–196.

de Jonge, E. and P.-P. de Wolf (2019). sdcspatial: Statistical disclosure control for spatial data. R package version 0.1.1.

de Montjoye, Y. A., L. Rocher, and A. S. Pentland (2016). Bandicoot: A python toolbox for mobile phone metadata. Journal of Machine Learning Research 17, 1–5.

de Montjoye, Y.-A., Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel (2014). D4d-senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:1407.4885.

Deming, E. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11(4), 427–444.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences 111 (45), 15888–15893.

Douglass, R. W., D. A. Meyer, M. Ram, D. Rideout, and D. Song (2014). High resolution population estimates from telecommunications data. EPJ Data Science 4(1), 1–13.

Dowle, M., T. Short, S. Lianoglou, and A. Srinivasan (2014). data.table: Extension of Data.frame.

Drechsler, J., A. Dundler, S. Bender, S. Rässler, and T. Zwick (2008). A new approach for disclosure control in the iab establishment panel - multiple imputation for a better data access. AStA Advances in Statistical Analysis 92(4), 439–458.

Dwork, C. (2008). Differential privacy: A survey of results. In Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, Volume 4978, pp. 1–19. Springer, Berlin, Heidelberg.

Eagle, N., M. Macy, and R. Claxton (2010). Network Diversity and Economic Development. Science 328, 1029–1031.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.

Elkies, N., G. Fink, and T. Bärnighausen (2015). "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. Population and Environment 37(1), 83–98.

Esch, T., W. Heldens, A. Hirner, M. Keil, M. Marconcini, A. Roth, J. Zeidler, S. Dech, and E. Strano (2017). Breaking new ground in mapping human settlements from space – The Global Urban Footprint. ISPRS Journal of Photogrammetry and Remote Sensing 134, 30–42.

European Union (2020). Copernicus: Europe´s eyes on Earth. `https://www.copernicus.eu/`.

Fatehkia, M., B. Coles, F. Ofli, and I. Weber (2020). The relative value of facebook advertising data for poverty mapping. Proceedings of the International AAAI Conference on Web and Social Media 14, 934–938.

Fay, R. E. and R. A. Herriot (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association 74(366a), 269–277.

Fick, S. E. and R. J. Hijmans (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37(12), 4302–4315.

Ford, L. (2007). 'People treat you as if you are nothing': What good is free education when poverty means children are forced into work? Liz Ford reports from Senegal. The Guardian.

Freire, S., K. MacManus, M. Pesaresi, E. Doxsey-Whitfield, and J. Mills (2016). Development of new open and free multi-temporal global population grids at 250 m resolution. AGILE, 6.

Frías-Martínez, E., G. Williamson, and V. Frías-Martínez (2011). An agent-based model of epidemic spread using human mobility and social network information. In Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011, pp. 57–64. IEEE.

Ghosh, M. and J. N. K. Rao (1994). Small Area Estimation: An Appraisal. Statistical Science 9 (1), 55–93.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati (2012). Semiparametric Fay-Herriot Model Using Penalized Splines. Journal of the Indian Society of Agricultural Statistics 66, 1–14.

Gouvernement du Sénégal (2014). Plan Senegal Emergent. Technical report, Gouvernement du Sénégal.

Gouvernement du Sénégal (2017). Programme National de Bourses de Sécurité Familiale (PNBSF).

Grace, K., N. N. Nagle, C. R. Burgert-Brucker, S. Rutzick, D. C. Van Riper, T. Dontamsetti, and T. Croft (2019). Integrating Environmental Context into DHS Analysis While Protecting Participant Confidentiality: A New Remote Sensing Method. Population and Development Review 45(1), 197–218.

Granello, D. H. and J. E. Wheaton (2004). Online data collection: Strategies for research. Journal of Counseling & Development 82(4), 387–393.

Green, A., S. Haslett, and C. Zingel (1998). Small Area Estimation Given Regular Updates of Census Auxiliary Variables. In Proceedings of the New Techniques and Technologies for Statistics Conference, pp. 206–211.

Green, M. P. and S. S. Wang (2002). Signal propagation model used to predict location accuracy of GSM mobile phones for emergency applications. In Proceedings - RAWCON 2002: 2002 IEEE Radio and Wireless Conference, pp. 119–122. Institute of Electrical and Electronics Engineers Inc.

Groß, M., A. K. Kreutzmann, U. Rendtel, T. Schmid, and N. Tzavidis (2020). Switching between Different Non-Hierachical Administrative Areas via Simulated Geo-Coordinates: A Case Study for Student Residents in Berlin. Journal of Official Statistics 36(2), 297–314.

GSMA (2015). The mobile economy 2015. `http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf`.

Gundogdu, D., O. D. Incel, A. A. Salah, and B. Lepri (2016). Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Science 5(1), 25.

Ha, N. S., P. Lahiri, and V. Parsons (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey. Statistics in Medicine 33 (22), 3932–3945.

Harvey, J. T. (2002). Estimating census district populations from satellite imagery: Some approaches and limitations. International Journal of Remote Sensing 23(10), 2071–2095.

Hata, M. (1980). Empirical Formula for Propagation Loss in Land Mobile Radio Services. IEEE Transactions on Vehicular Technology 29(3), 317–325.

Heil, T. (2014). Are neighbours alike? Practices of conviviality in Catalonia and Casamance. European Journal of Cultural Studies 17 (4), 452–470.

Heldal, J. and D.-C. Iancu (2019). Synthetic data generation for anonymization purposes. Application on the Norwegian Survey on living conditions/EHIS. In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Henderson, C. R. (1950). Estimation of genetic parameters. Annals of Mathematical Statistics 21 (2), 309–310.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring economic growth from outer space. American Economic Review 102(2), 994–1028.

Hunter, L. M., C. Talbot, W. Twine, J. McGlinchy, C. W. Kabudula, and D. Ohene-Kwofie (2021). Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site. Population and Environment 42(4), 445–476.

Instituto Nacional de Estadistica y Censos (2022). X Censo Nacional de Población y VI de Vivienda. Catálogo central de datos. `http://sistemas.inec.cr/pad5/index.php/catalog/113`.

Iovan, C., A. M. Olteanu-Raimond, T. Couronné, and Z. Smoreda (2013). Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In Lecture Notes in Geoinformation and Cartography, Volume 2013-Janua, pp. 247–265. Springer, Cham.

Isidro, M. (2010). Intercensal updating of small area estimates. Ph. D. thesis, Massey University.

Isidro, M., S. Haslett, and G. Jones (2016). Extended structure preserving estimation (ESPREE) for updating small area estimates of poverty. Annals of Applied Statistics 10(1), 451–476.

Janke, T., M. Ghanmi, and F. Steinke (2021). Implicit generative copulas. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Volume 34, pp. 26028–26039. Curran Associates, Inc.

Janzen, M., M. Vanhoof, Z. Smoreda, and K. W. Axhausen (2018). Closer to the total? Long-distance travel of French mobile phone users. Travel Behaviour and Society 11, 31–42.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. Science 353(6301), 790–794.

Jeong, B., W. Lee, D.-S. Kim, and H. Shin (2016). Copula-based approach to synthetic population generation. PloS ONE 11(8), e0159496.

Jiang, D., W. Lin, and N. Raghavan (2020). A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques. IEEE Access 8, 197885–197895.

Jiang, J. and P. Lahiri (2001). Empirical best prediction for small area inference with binary data. Annals of Institute of Statistical Mathematics 53, 217–243.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu (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, 2 Massachusetts Avenue, NE, Washington, DC, pp. 75–97.

Jordon, J., J. Yoon, and M. Van Der Schaar (2019). PATE-GaN: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations.

Kamthe, S., S. Assefa, and M. Deisenroth (2021). Copula flows for synthetic data generation. arXiv preprint arXiv:2101.00598.

Khodabandelou, G., V. Gauthier, M. Fiore, and M. A. El-Yacoubi (2018). Estimation of static and dynamic urban populations with mobile network metadata. IEEE Transactions on Mobile Computing 18(9), 2034–2047.

Kish, L. (1965). Survey sampling. New York: John Wiley & Sons.

Koebe, T. (2020). Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. PLoS ONE 15(11 November), e0241981.

Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, and T. Schmid (2021). Intercensal updating using structure-preserving methods and satellite imagery. Journal of the Royal Statistical Society: Series A.

Kroll, M. and R. Schnell (2016). Anonymisation of geographical distance matrices via Lipschitz embedding. International Journal of Health Geographics 15(1), 1–14.

Lahiri, P. and J. Suntornchost (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. Sankhya B 77 (2), 312–320.

Le Menach, A., A. J. Tatem, J. M. Cohen, S. I. Hay, H. Randell, A. P. Patil, and D. L. Smith (2011). Travel risk, malaria importation and malaria transmission in Zanzibar. Scientific Reports 1(1), 93.

Leasure, D. R., W. C. Jochem, E. M. Weber, V. Seaman, and A. J. Tatem (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. Proceedings of the National Academy of Sciences 117(39), 24173–24179.

Leyk, S., A. E. Gaughan, S. B. Adamo, A. de Sherbinin, D. Balk, S. Freire, A. Rose, F. R. Stevens, B. Blankespoor, C. Frye, J. Comenetz, A. Sorichetta, K. MacManus, L. Pistolesi, M. Levy, and A. J. Tatem (2019). Allocating people to pixels: A review of large-scale gridded population data products and their fitness for use. Earth System Science Data Discussions 11(3), 1–30.

Li, H. and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis 101, 882–892.

Li, H., L. Xiong, and X. Jiang (2014). Differentially private synthesization of multi-dimensional data using copula functions. In Advances in database technology: proceedings. International conference on extending database technology, pp. 475–486.

Lima, A., M. D. Domenico, V. Pejovic, and M. Musolesi (2013). Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics. CoRR abs/1306.4.

Liu, B., P. Lahiri, and G. Kalton (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. Survey Methodology 40 (1), 1–13.

Lopez-Vizcaino, E., M. J. Lombardia, and D. Morales (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. Journal of the Royal Statistical Society: Series A 178 (3), 535–565.

Lu, X., L. Bengtsson, and P. Holme (2012). Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences of the United States of America 109(29), 11576–11581.

Luna, A. (2016). Multivariate Structure Preserving Estimation for Population Compositions. Ph. D. thesis, University of Southampton.

Luna, A., L. C. Zhang, A. Whitworth, and K. Piller (2015). Small area estimates of the population distribution by ethnic group in England: A proposal using structure preserving estimators. Statistics in Transition 16(4), 585–602.

MacFeely, S. (2016). The continuing evolution of official statistics: Some challenges and opportunities. Journal of Official Statistics 32(4), 789–810.

Mansfield, P. and A. A. Maudsley (1977). Medical imaging by nmr. British Journal of Radiology 50(591), 188–194.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimatiors using big data sources. Journal of Official Statistics 31 (2), 263–281.

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

Matamalas, J. T., M. De Domenico, and A. Arenas (2016). Assessing reliable human mobility patterns from higher order memory in mobile communications. Journal of the Royal Society Interface 13(121), 20160203.

Méndez, F. and O. Bravo (2011). Costa Rica Mapas de Pobreza 2011. Technical report, INEC Costa Rica, San José, Costa Rica.

MIT Data To AI Lab (2022). The synthetic data vault (SDV). `https://sdv.dev/`.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics 38 (3), 369–385.

Montjoye, Y.-A. d., J. Quoidbach, F. Robic, and A. S. Pentland (2013). Predicting personality using novel mobile phone-based metrics. In International conference on social computing, behavioral-cultural modeling, and prediction, pp. 48–55. Springer.

Nelsen, R. B. (2007). An introduction to copulas. Springer Science & Business Media.

Noble, A., S. Haslett, and G. Arnold (2002). Small area estimation via generalized linear models. Journal of Official Statistics 18(1), 45–60.

OECD (2013). Household definitions in other statistical standards. In OECD Guidelines for Micro Statistics on Household Wealth, pp. 275–277. OECD Publishing.

Okumura, Y., E. Ohmori, T. Kawano, and K. Fukuda (1968). Field Strength and Its Variability in UHF and VHF Land-Mobile Radio Service. Review of the Electrical Communication Laboratory, September-October, 1968 16, 825–873.

Opsomer, J. D., M. Francisco-Fernandez, and X. Li (2012). Model-Based Non-parametric Variance Estimation for Systematic Sampling. Scandinavian Journal of Statistics 39 (3), 528–542.

Oughton, E. (2019). Quantified global broadband strategies for connecting unconnected communities. In TPRC47: The 47th Research Conference on Communication, Information and Internet Policy.

Oughton, E. J., Z. Frias, S. van der Gaast, and R. van der Berg (2019). Assessing the capacity, coverage and cost of 5G infrastructure strategies: Analysis of the Netherlands. Telematics and Informatics 37, 50–69.

Park, P. S., J. E. Blumenstock, and M. W. Macy (2018). The strength of long-range ties in population-scale social networks. Science 362(6421), 1410–1413.

Patki, N., R. Wedge, and K. Veeramachaneni (2016). The synthetic data vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410.

Pestre, G., E. Letouzé, and E. Zagheni (2020). The abcde of big data: assessing biases in call-detail records for development estimates. The World Bank Economic Review 34(Supplement_1), S89–S97.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. Statistical Science 28 (1), 40–68.

Pfeffermann, D. and C. R. Rao (2009). Sample surveys: design, methods and applications. Elsevier.

Phillips, C., D. Sicker, and D. Grunwald (2011). Bounding the error of path loss models. In 2011 IEEE International Symposium on Dynamic Spectrum Access Networks, DySPAN 2011, pp. 71–82.

Phillips, C., D. Sicker, and D. Grunwald (2013). A survey of wireless path loss prediction and coverage mapping methods. IEEE Communications Surveys and Tutorials 15(1), 255–270.

Pinkovskiy, M. and X. Sala-i Martin (2016). Lights, Camera ... Income! Illuminating the National Accounts-Household Surveys Debate. The Quarterly Journal of Economics 131(2), 579–631.

Pokhriyal, N. and D. C. Jacques (2017). Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences of the United States of America 114(46), E9783–E9792.

Porter, A. T., S. H. Holan, C. K. Wikle, and N. Cressie (2014). Spatial Fay-Herriot models for small area estimation with functional covariates. Spatial Statistics 10, 27–42.

Prasad, N. G. N. and J. N. K. Rao (1990). The Estimation of the Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association 85 (409), 163–171.

Pratesi, M. (2016). Analysis of Poverty Data by Small Area Estimation. John Wiley & Sons.

Pratesi, M. and N. Salvati (2009). Small Area Estimation in the Presence of Correlated Random Area Effects. Journal of Official Statistics 25 (1), 37–53.

Purcell, N. J. and L. Kish (1980). Postcensal Estimates for Local Areas (Or Domains). International Statistical Review / Revue Internationale de Statistique 48(1), 3.

Raghunathan, T. E., D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, K. W. Dodd, and E. J. Feuer (2007). Combining Information From Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening. Journal of the American Statistical Association 102 (478), 474–486.

Rao, J. N. K. (1986). Synthetic estimators, SPREE and the best model based predictors. In Proceedings of the Conference on Survey Research Methods in Agriculture, pp. 1–16. US Department of Agriculture Washington, DC.

Rao, J. N. K. (2003). Small Area Estimation. New York: Wiley.

Rao, J. N. K. and I. Molina (2015). Small Area Estimation (2nd Edition ed.). New York: Wiley.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21(3), 441–462.

Ricciato, F., P. Widhalm, F. Pantisano, and M. Craglia (2017). Beyond the 'single-operator, CDR-only' paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing 35, 65–82.

Rocher, L., J. M. Hendrickx, and Y. A. de Montjoye (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 10(1), 1–9.

Rubrichi, S., Z. Smoreda, and M. Musolesi (2018). A comparison of spatial-based targeted disease mitigation strategies using mobile phone data. EPJ Data Science 7(1), 17.

Sabharwal, N. and A. Agrawal (2021). Introduction to word embeddings. In Hands-on Question Answering Systems with BERT, pp. 41–63. Springer.

Schelle, C. (2013). Schulsysteme, Unterricht und Bildung im mehrsprachigen frankophonen Westen und Norden Afrikas. Münster, New York, München and Berlin: Waxmann.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A 180(4), 1163–1190.

Schneider, C. M., V. Belik, T. Couronné, Z. Smoreda, and M. C. González (2013). Unravelling daily human mobility motifs. Journal of the Royal Society Interface 10(84), 20130246.

Searle, S. R. (1971). Linear Models. New York: Wiley.

Sharma RK Singh, P. K. (2010). Comparative Analysis of Propagation Path loss Models with Field Measured Data. International Journal of Engineering Science and Technology 2(6), 2008–2013.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut Statistique de l'Université de Paris 8, 229–231.

Slud, E. V. and T. Maiti (2006). Mean-squared error estimation in transformed Fay-Herriot models. Journal of the Royal Statistical Society: Series B 68 (2), 239–257.

Sonatel (2019). Coverage Map Sonatel 2019.

Steinnocher, K., A. De Bono, B. Chatenoux, D. Tiede, and L. Wendt (2019). Estimating urban population patterns from stereo-satellite imagery. European Journal of Remote Sensing 52(sup2), 12–25.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015). Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. PLoS ONE 10(2), e0107042.

Subash, S. P., R. R. Kumar, and K. S. Aditya (2018). Satellite data and machine learning tools for predicting poverty in rural India. Agricultural Economics Research Review 31(2), 231–240.

Sun, Y., A. Cuesta-Infante, and K. Veeramachaneni (2019). Learning vine copula models for synthetic data generation. Proceedings of the AAAI Conference on Artificial Intelligence 33(01), 5049–5057.

Sundsøy, P. (2016). Can mobile usage predict illiteracy in a developing country? arXiv preprint arXiv:1607.01337.

Taylor, L. (2016). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning D: Society and Space 34(2), 319–336.

Templ, M. (2017). Statistical disclosure control for microdata. Springer.

Templ, M., B. Meindl, A. Kowarik, and O. Dupriez (2017). Simulation of synthetic complex data: The R package simPop. Journal of Statistical Software 79, 1–38.

Tennekes, M. (2018). mobloc: Mobile phone location algorithms and tools. `https://github.com/MobilePhoneESSnetBigData/mobloc_v0.1`.

The Economist Intelligence Unit (2019). The Inclusive Internet Index 2019.

Tizzoni, M., P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza (2014). On the Use of Human Mobility Proxies for Modeling Epidemics. PLoS Computational Biology 10(7), e1003716.

Torkzadehmahani, R., P. Kairouz, and B. Paten (2019). Dp-cgan: Differentially private synthetic data and label generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Tzavidis, N., L. C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A 181(4), 927–979.

UNESCO (2012). Global Partnership for Girls' and Women's Education. `http://www.unesco.org/eri/cp/factsheets{_}ed/SN{_}EDFactSheet.pdf`.

UNESCO (2015). Literacy Project for Girls and Women in Senegal. `http://www.unesco.org/uil/litbase/?menu=4{&}programme=180`.

United Nations Department of Economic and Social Affairs (2009). Handbook on geospatial infrastructure in support of census activities (ST/ESA/STA ed.). New York, USA: United Nations Publication.

United Nations Department of International Economic and Social Affairs (1956). Manual III. Methods for population projections by sex and age. New York, USA: United Nations Publication.

United Nations General Assembly (2015). Res 70/1. Transforming Our World: The 2030 Agenda for Sustainable Development. Technical report, United Nations General Assembly.

Vale, S. (2011). The data deluge: What does it mean for official statistics? In 2nd CSO Administrative Data Seminar, Dublin Castle, Volume 29.

Vanhoof, M., C. Lee, and Z. Smoreda (2020). Performance and sensitivities of home detection on mobile phone data. Big Data Meets Survey Science: A Collection of Innovative Methods, 245–271.

Villalon, L. A. and M. Bodian (2012). Religion, social demand, and educational reforms in Senegal.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. Annals of Applied Statistics 6(1), 229–252.

Warren, J. L., C. Perez-Heydrich, C. R. Burgert, and M. E. Emch (2016). Influence of demographic and health survey point displacements on distance-based analyses. Spatial Demography 4(2), 155–173.

Weidmann, N. B. and S. Schutte (2017). Using night light emissions for the prediction of local wealth. Journal of Peace Research 54(2), 125–140.

Wesolowski, A., N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee (2013). The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society Interface 10(81), 20120986.

Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee (2012). Quantifying the impact of human mobility on malaria. Science 338(6104), 267–270.

West, B. T., A. Kirchner, D. Hochfellner, S. Bender, E. M. Nichols, M. H. Mulry, J. H. Childs, A. Holmberg, C. Bycroft, G. Benson, and F. Hubbard (2017). Establishing Infrastructure for the Use of Big Data to Understand Total Survey Error, Chapter 21, pp. 457–485. John Wiley & Sons, Ltd.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30(4), 377–399.

WorldPop (2018). Global High Resolution Population Denominators Project. `https://dx.doi.org/10.5258/SOTON/WP00644`.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems 32, 1–11.

Yoshimori, M. and P. Lahiri (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. Journal of Multivariate Analysis 124, 281–294.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2017). Privbayes: Private data release via bayesian networks. ACM Transactions on Database Systems (TODS) 42(4), 1–41.

Zhang, L. C. and R. L. Chambers (2004). Small area estimates for cross-classifications. Journal of the Royal Statistical Society: Series B 66(2), 479–496.

Zhang, Z., T. Wang, N. Li, J. Honorio, M. Backes, S. He, J. Chen, and Y. Zhang (2021). PrivSyn: Differentially private data synthesis. In Proceedings of the 30th USENIX Security Symposium, pp. 929–946.

Zhongming, Z., L. Linong, Y. Xiaona, Z. Wangqiang, L. Wei, et al. (2020). Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices. Asian Development Bank.

Summary

Summary in English

Abstract: Constructing socio-demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal

Modern systems of official statistics require the accurate and timely estimation of socio-demographic indicators for disaggregated geographical regions. Traditional data collection methods such as censuses or household surveys impose great financial and organizational burdens for National Statistical Institutes. The rise of new information and communication technologies offers promising sources to mitigate these shortcomings. In this paper we propose a unified approach for National Statistical Institutes in developing countries based on small area estimation that allows for the estimation of socio-demographic indicators by using mobile phone data. In particular, the methodology is applied to mobile phone data from Senegal for deriving sub-national estimates of the share of illiterates disaggregated by gender. The estimates are used to identify hot spots of illiterates with a need for additional infrastructure or policy adjustments. Although the paper focuses on literacy as a particular socio-demographic indicator, the proposed approach is applicable to indicators from national statistics in general.
Keywords: Indicators, Model-based estimation, Official statistics, Small area estimation.

Intercensal updating using structure-preserving methods and satellite imagery

Censuses are fundamental building blocks of most modern-day societies, yet collected every ten years at best. We propose an extension of the widely popular census updating technique *Structure Preserving Estimation* by incorporating auxiliary information in order to take ongoing subnational population shifts into account. We apply our method by incorporating satellite imagery as additional source to derive annual small-area updates of multidimensional poverty indicators from 2013 to 2020 for a population at risk: female-headed households in Senegal. We evaluate the performance of our proposal using data from two different census periods.
Keywords: Multidimensional poverty, Official statistics, Small area estimation, SPREE.

Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling

Mobile sensing data has become a popular data source for geo-spatial analysis, however, mapping it accurately to other sources of information such as statistical data remains a challenge.

Popular mapping approaches such as point allocation or voronoi tessellation provide only crude approximations of the mobile network coverage as they do not consider holes, overlaps and within-cell heterogeneity. More elaborate mapping schemes often require additional proprietary data operators are highly reluctant to share. In this paper, I use human settlement information extracted from publicly available satellite imagery in combination with stochastic radio propagation modelling techniques to account for that. I show in a simulation study and a real-world application on unemployment estimates in Senegal that better coverage approximations do not necessarily lead to better outcome predictions.

Keywords: Mobile networks, Remote sensing, Official statistics, Radio propagation, International development.

Releasing survey microdata with exact cluster locations and additional privacy safeguards

Household survey programs around the world publish fine-granular georeferenced microdata to support research on the interdependence of human livelihoods and their surrounding environment. To safeguard the respondents' privacy, micro-level survey data is usually (pseudo)-anonymized through deletion or perturbation procedures such as obfuscating the true location of data collection. This, however, poses a challenge to emerging approaches that augment survey data with auxiliary information on a local level. Here, we propose an alternative microdata dissemination strategy that leverages the utility of the original microdata with additional privacy safeguards through synthetically generated data using generative models. We back our proposal with experiments using data from the 2011 Costa Rican census and satellite-derived auxiliary information. Our strategy reduces the respondents' re-identification risk for any number of disclosed attributes by 60-80% even under re-identification attempts.

Keywords: Generative models, Statistical disclosure control, Geomasking, Copula, Official statistics, Satellite imagery.

Kurzfassungen in Deutsch

Zusammenfassung: Berechnung sozio-demografischer Indikatoren mittels Mobilfunkdaten für Statistische Ämter: Schätzung von Alphabetisierungsraten in Senegal

Moderne Systeme der amtlichen Statistik erfordern die genaue und zeitnahe Schätzung soziodemografischer Indikatoren für disaggregierte geografische Regionen. Herkömmliche Datenerhebungsmethoden wie Volkszählungen oder Haushaltserhebungen bedeuten für die nationalen statistischen Ämter große finanzielle und organisatorische Belastungen. Die Etablierung neuer Informations- und Kommunikationstechnologien bietet vielversprechende Quellen, um diese Herausforderungen zu überwinden. In diesem Aufsatz schlagen wir einen einheitlichen Ansatz für nationale statistische Ämter in Entwicklungsländern vor, der auf einer Schätzung kleiner Gebiete basiert und die Schätzung soziodemografischer Indikatoren unter Verwendung von Mobiltelefondaten ermöglicht. Die Methodik wird insbesondere auf Mobilfunkdaten aus dem

Senegal angewendet, um subnationale Schätzungen des Anteils von Analphabeten nach Geschlecht aufzuschlüsseln. Die Schätzungen werden verwendet, um lokale Häufungen von Analphabeten zu identifizieren, woraus infrastrukturelle oder politische Anpassungen abgeleitet werden können. Obwohl sich der Aufsatz auf die Alphabetisierungsrate als einen bestimmten soziodemografischen Indikator konzentriert, ist der vorgeschlagene Ansatz auf Indikatoren aus nationalen Statistiken im Allgemeinen anwendbar.

Stichworte: Indikatoren, Modellbasierte Schätzung, Amtliche Statistik, Kleinräumige Schätzung.

Zusammenfassung: Zensusdaten aktualisieren mittels strukturerhaltender Methoden und Satellitenbildern

Volkszählungen sind grundlegende Bausteine der meisten modernen Gesellschaften, werden aber bestenfalls alle zehn Jahre erhoben. Wir schlagen eine Erweiterung der weit verbreiteten Technik zur Aktualisierung von Volkszählungen *Structure Preserving Estimation* vor, indem wir Hilfsinformationen einbeziehen, um laufende subnationale Bevölkerungsverschiebungen zu berücksichtigen. Wir wenden unsere Methode an, indem wir Satellitenbilder als zusätzliche Quelle einbeziehen, um jährliche kleinräumige Aktualisierungen multidimensionaler Armutsindikatoren von 2013 bis 2020 für eine gefährdete Bevölkerungsgruppe abzuleiten: von Frauen geführte Haushalte im Senegal. Wir bewerten den Mehrwert unseres Vorschlags anhand von Daten aus zwei verschiedenen Zählperioden.

Stichworte: Mehrdimensionale Armut, Amtliche Statistik, Kleinräumige Schätzung, SPREE.

Zusammenfassung: Besserer Empfang, bessere Ergebnisse? Mobilfunknetzdaten mit amtlichen Statistiken verbinden anhand von Satellitenbildern und Funkwellenmodellierung

Mobile Sensing-Daten sind zu einer beliebten Datenquelle für Geodatenanalysen geworden, aber die genaue Zuordnung zu anderen Informationsquellen wie statistischen Daten bleibt eine Herausforderung. Beliebte Ansätze für die räumliche Zuordnung wie Punktallokation oder Voronoi-Tessellation liefern nur grobe Annäherungen an die Mobilfunknetzabdeckung, da sie Löcher, Überlappungen und Heterogenität innerhalb der Mobilfunkzellen nicht berücksichtigen. Ausgefeiltere Mapping-Schemata erfordern oft zusätzliche proprietäre Daten, die Netzanbieter selten extern zur Verfügung stellen. In diesem Aufsatz verwende ich deswegen Informationen über menschliche Siedlungsgebiete, die aus öffentlich zugänglichen Satellitenbildern extrahiert wurden, in Kombination mit stochastischen Modellierungstechniken für die Funkwellenausbreitung, um die vorangegangenen Punkte zu berücksichtigen. Ich zeige in einer Simulationsstudie und einer realen Anwendung zu Arbeitslosenschätzungen im Senegal, dass bessere Annäherungen an die Netzabdeckung nicht unbedingt zu besseren Ergebnisvorhersagen führen.

Stichworte: Mobilfunknetzwerke, Fernerkundung, Amtliche Statistik, Funkwellenausbreitung, Internationale Entwicklungszusammenarbeit.

Zusammenfassung: Amtliche Mikrodaten veröffentlichen mit genauen Datenerhebungsstandorten und zuätzlichem Privatsphärenschutz

Haushaltsumfrageprogramme auf der ganzen Welt veröffentlichen detaillierte georeferenzierte Mikrodaten, um die Forschung über die Abhängigkeit der menschlichen Lebensumständen und ihrer Umgebung zu unterstützen. Um die Privatsphäre der Befragten zu schützen, werden Umfragedaten normalerweise (pseudo-)anonymisiert, indem Lösch- oder Störungsverfahren wie die Verschleierung des wahren Ortes der Datenerhebung durchgeführt werden. Dies stellt jedoch neue Ansätze, die Erhebungsdaten mit Hilfsinformationen auf lokaler Ebene ergänzen, vor eine Herausforderung. Hier schlagen wir eine alternative Veröffentlichungsstrategie für Mikrodaten vor, die den Nutzen der ursprünglichen Mikrodaten weitestgehend erhält und mit zusätzlichen Datenschutzvorkehrungen durch synthetisch generierte Daten unter Verwendung generativer Modelle schützt. Wir untermauern unseren Vorschlag mit Experimenten unter Verwendung von Daten aus der Volkszählung von 2011 in Costa Rica und von Satelliten abgeleiteten Hilfsinformationen. Unser Vorschlag reduziert das Reidentifikationsrisiko der Befragten für eine beliebige Anzahl von offengelegten Merkmalen um 60-80%, selbst nach Reidentifikationsversuchen.

Stichworte: Generative Modelle, Statistische Offenlegungskontrolle, Geomasking, Copula, Amtliche Statistik, Satellitenbilder.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Bensheim, July 25th, 2022

Till Koebe
July 25th, 2022