Freie Universität Berlin

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

Timo Schmid
Fabian Bruckschen
Nicola Salvati
Till Zbiranski

School of Business & Economics
Discussion Paper

Economics

2016/9

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

Timo Schmid[*], Fabian Bruckschen[*], Nicola Salvati[**], and Till Zbiranski[*]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany
[**]Dipartimento di Economia e Management, Università of Pisa, Pisa, Italy

Abstract

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 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.

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, 2003). 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) and Molina and Rao (2010) 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: 38% for women and 62% for men, ANSD (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 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 are 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 2 we describe the DHS survey and the mobile phone data including the cleaning and preparation. In Section 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 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 5. Finally, in Section 6 we conclude the paper with some final remarks and discuss limitations of the proposed approach. Additional results are presented in the supplementary materials.

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.

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 10 sections covering socio-demographic indicators (like age and date of birth, schooling, literacy, 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 ANSD (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 households 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) (ANSD, 2012).

Figure 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* (38%) in Senegal is lower than the proportion of literate men (62%). The results are consistent with the official published results of the ANSD (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. Figure 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 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 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.

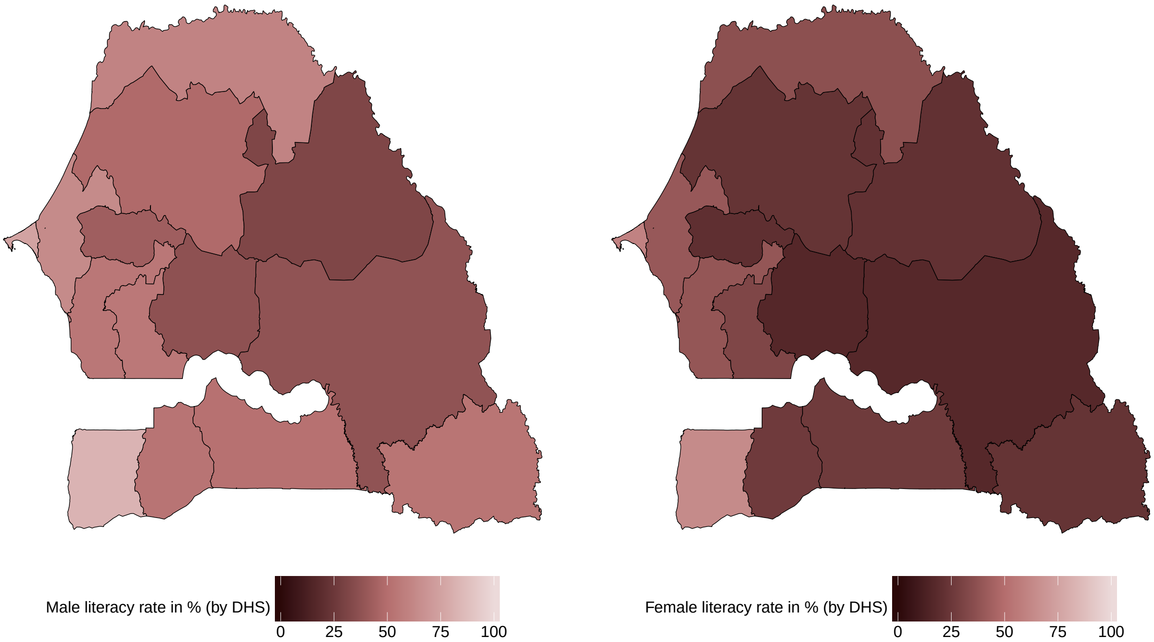

Figure 1: Estimates for the literacy rate by gender on regional level based on DHS survey 2011.

Table 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

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.

2.2.1 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 SMS, which tower received

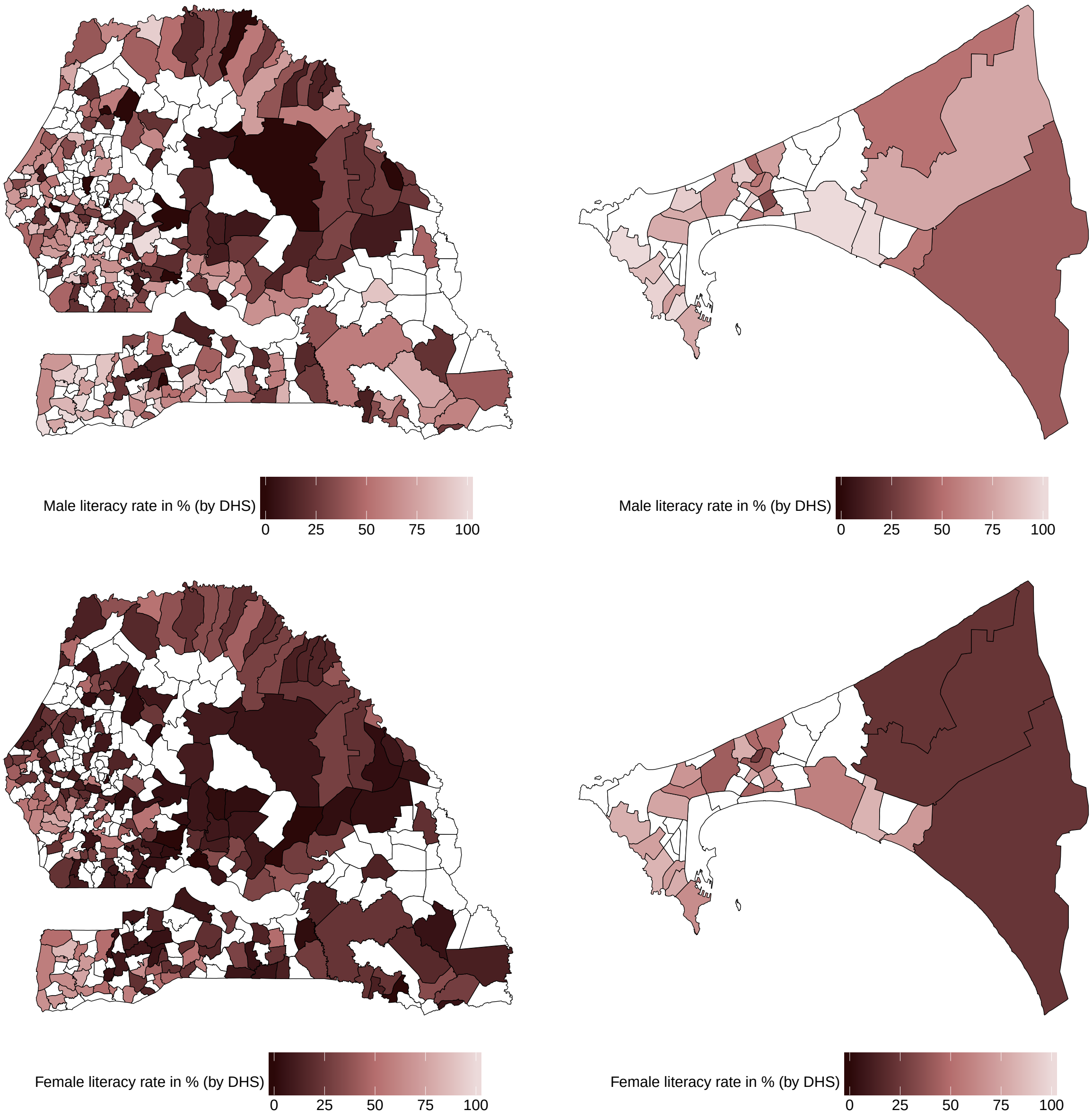

Figure 2: Estimates for the literacy rate by gender on commune level based on DHS survey 2011: Senegal (left panel) and Dakar (right panel).

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 calls and SMS as well as the duration of each call for every tower on an hourly basis for the year 2013. Table 2 shows the head of an preprocessed dataset for the usage of SMS. The first column is

Table 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

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.

2.2.2 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

Table 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

two towers. The covariate *distance-to-dakar* measures the distance from each tower to a centroid of the region Dakar. According to Smith et al. (2013) 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

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

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* (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

$$Entropy(t_i) = -\sum_{\substack{j \neq i \\ j=1}}^{1666} p(t_i, t_j) \cdot log\big[p(t_i, t_j)\big], \tag{2}$$

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 *http://bandicoot.mit.edu/docs/reference/index.html*. 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.

2.2.3 First descriptive statistics

Figure 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 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 3 shows summary statistics of the number of mobile phone towers over the communes. The mean number of towers per

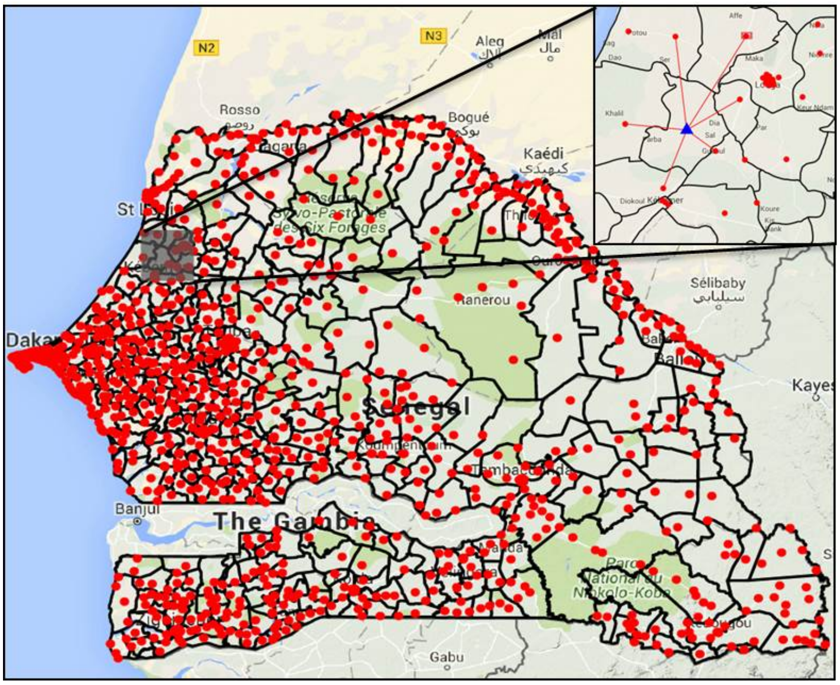

Figure 3: Location of mobile phone towers in Senegal.

commune is 4.1 with a maximum of 60. Although Figure 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 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.

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 corrected 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. MSE estimation is performed by a parametric bootstrap approach.

3.1 Transformed Fay-Herriot estimator

We assume that the population U, consisting of N units, is divided into m disjunct 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 $\boldsymbol{y}$ 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. A design unbiased estimator for the population mean θ_i of the variable of interest $\boldsymbol{y}$ in area i is given by

$$\hat{\theta}_i^{direct} = \frac{1}{N_i}\sum_{j=1}^{n_i} \omega_{ij} y_{ij}, \tag{3}$$

and $Var(\hat{\theta}_i^{direct})$ denotes the corresponding variance. 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.

$$\text{Sampling model (first stage)}: \quad \hat{\theta}_i^{direct} = \theta_i + \varepsilon_i \tag{4}$$

$$\text{Linking model (second stage)}: \quad \theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i, \tag{5}$$

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 $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon_i})$. Furthermore, ε_i is estimated based on the design of the survey and known, for instance, $\varepsilon_i = Var(\hat{\theta}_i^{direct})$. The random effects u_i are assumed to be independently normally distributed with $u_i \sim N(0, \sigma_u^2)$. 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

$$\hat{\theta}_i^{direct} = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i + \varepsilon_i. \tag{6}$$

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 restricted maximum likelihood (Datta and Lahiri, 2000; Rao, 2003). The EBLUP under the FH model is obtained by

$$\begin{aligned} \hat{\theta}_i^{FH} &= \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \hat{u}_i & (7)\\ &= \gamma_i\hat{\theta}_i^{direct} + (1-\gamma_i)\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, & (8)\end{aligned}$$

where $\gamma_i = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \sigma^2_{\varepsilon_i})^{-1}$ denotes the shrinkage factor for area i. 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):

$$\hat{\theta}_{i,out}^{FH} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}. \tag{9}$$

The MSE of the EBLUP in (7) can be obtained by analytic solutions following Prasad and Rao (1990) and Datta et al. (2005).

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. Following Carter and Rolph (1974) and Raghunathan et al. (2007) we use arcsine transformation in modeling. Let $\boldsymbol{y}$ 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 $\sigma_{\varepsilon_i}^2 = 1/(4\tilde{n}_i)$, where $\tilde{n}_i$ stands for the effective sample size (Carter and Rolph, 1974). 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 (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

$$\hat{\theta}_i^{FH,trans} = f^{-1}(\hat{\theta}_i^{FH}) = \sin^2(\hat{\theta}_i^{FH}) \quad \text{for } i = 1, ..., m, \tag{10}$$

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

For the MSE estimating of $\hat{\theta}_i^{FH,trans}$ we use a parametric bootstrap procedure following Gonzalez-Manteiga et al. (2008). The steps are as follows:

1. for given $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u^2$ 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,

$$\hat{\theta}_i^{*,(b)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + u_i^* + \varepsilon_i^* \tag{11}$$

   and the corresponding bootstrap population

$$\theta_i^{*,(b)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + u_i^*. \tag{12}$$

3. Using the bootstrap sample, we estimate the model parameters in (6). Based on the estimated model parameters from the bootstrap sample, we compute the corresponding FH estimator (7) in area i, $\hat{\theta}_i^{FH,(b)}$.
4. Using the B bootstrap samples, the MSE estimator of $\hat{\theta}_i^{FH,trans}$ is given by

$$\hat{\text{MSE}}(\hat{\theta}_i^{FH,trans}) = \frac{1}{B} \sum_{b=1}^{B} \left(f^{-1}\{\hat{\theta}_i^{FH,(b)}\} - f^{-1}\{\theta_i^{*,(b)}\} \right)^2. \tag{13}$$

The properties of this bootstrap scheme are empirically evaluated in Section 5. Furthermore we use the MSE estimates of $\hat{\theta}_i^{FH,trans}$ for the benchmarked transformed FH estimator introduced in Section 3.2.

3.2 Benchmarked transformed Fay-Herriot estimator

Although the model-based estimator in (10) provides estimates for all communes (small areas) in Senegal, the aggregated estimates on national level can differ substantially 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.

We seek for a benchmarked FH estimator $\hat{\theta}_i^{FH,bench}$ such that

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

where

$$\alpha = \sum_{i=1}^{m} w_i \hat{\theta}_i^{direct}.$$

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

$$\hat{\theta}_i^{FH,trans,bench} = \hat{\theta}_i^{FH,trans} + \frac{\sum_{i=1}^{m} w_i^2}{\phi} \left(\alpha - \sum_{i=1}^{m} w_i \hat{\theta}_i^{FH,trans} \right) \frac{w_i}{\phi_i} \quad \text{for } i = 1, ..., m. \tag{14}$$

There are several way to define the weight ϕ_i (Datta et al., 2010). For instance, $\phi_i = w_i/\hat{\theta}_i^{FH,trans}$ leads to a ratio adjustment of the FH estimator, where small areas with larger estimates will receive a larger adjustment and vice versa. As the $\hat{\theta}_i^{FH,trans}$ is restricted to $[0,1]$, we define the weights by $\phi_i = w_i/\hat{\text{MSE}}(\hat{\theta}_i^{FH,trans})$. That means that small areas with higher variability in terms of MSE will receive a larger adjustment. Note that the benchmarked FH estimator for (7) is defined analogously.

For the MSE estimating of $\hat{\theta}_i^{FH,trans,bench}$ we also apply a parametric bootstrap procedure following Gonzalez-Manteiga et al. (2008). The steps are as follows:

1. for given $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u^2$ 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 $\hat{\theta}_i^{*,(l)}$ and the corresponding bootstrap population $\theta_i^{*,(l)}$ according to equations (11) and (12) respectively.
3. Using the bootstrap sample, estimate the transformed FH estimator (10) and the corresponding MSE estimator (13) in area i. Note that this step involves B bootstrap replications described in (13).
4. Compute the corresponding benchmarked transformed FH estimator (14) in area i, $\hat{\theta}_i^{FH,trans,bench,(l)}$.
5. Using the L bootstrap samples, the MSE estimator of $\hat{\theta}_i^{FH,trans,bench}$ is given by

$$\hat{\text{MSE}}(\hat{\theta}_i^{FH,trans,bench}) = \frac{1}{L} \sum_{l=1}^{L} \left(\hat{\theta}_i^{FH,trans,bench,(l)} - f^{-1}\{\theta_i^{*,(l)}\} \right)^2. \tag{15}$$

The MSE estimation for the proposed benchmarked transformed FH estimator is computationally demanding because it involves $B \cdot L$ bootstrap replications.

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 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 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.

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. The model selection in this paper is done by using the classic Akaike information criterion (AIC) based on a linear model. Although we are aware of more complex methods for Fay-Herriot model selection discussed in Marhuenda et al. (2014) we used an simple approach which is implemented in standard statistical software. Based on the wide range of the mobile phone covariates discussed in Section 2 we identified the final set of covariates by a stepwise selection procedure using the AIC. The final model on commune level for the variables *literacy women* and *literacy men* include 26 and 30 mobile phone covariates with an adjusted R^2 of 68% and 52% respectively.

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 (6) by gender. As discussed in Section 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 4.1 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 ANSD (2012) in Senegal and show an high value of the design effect of the direct estimator using DHS survey 2011.

Table 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 4 shows normal probability plots of level 1 and level 2 residuals 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

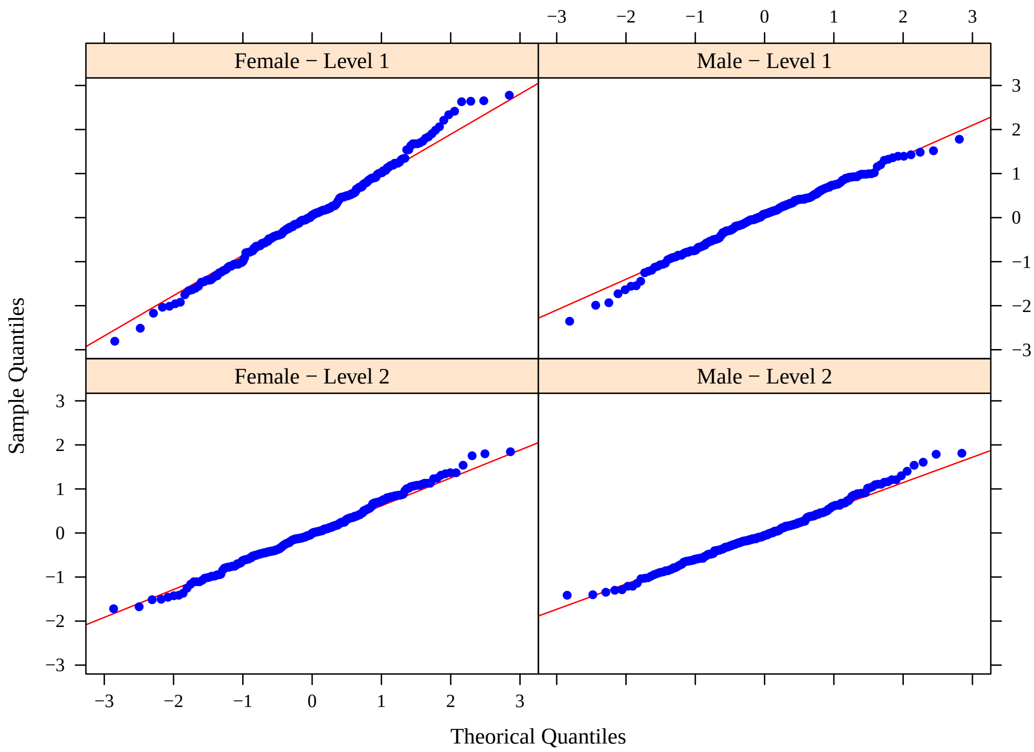

Figure 4: Normal probability plots of level 1 and level 2 standardized residuals (top down) for the female model (left panel) and for the male model (right panel).

standardized residuals (p-values: male model = 0.3099 and female model = 0.3378) and level 2 standardized residuals (p-values: male model = 0.2589 and female model = 0.4803). Using the transformed Fay-Herriot model (10) may be advisable for estimating the literacy of women and men.

4.2 Small area estimates on commune level

Estimates of the literacy rate by gender for each commune are calculated by using the transformed FH estimator (10) (FH Trans) and by the benchmarked transformed FH estimator (14) (FH Bench). MSE estimation for the FH Trans and FH Bench is implemented with the parametric bootstrap approaches discussed in Section 3. We performed $B = 200$ replicates for the bootstrap of the FH Trans (13) and $B = 200$ with $L = 200$ replicates for the bootstrap of the FH Bench (15). 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 5 reports the distribution of estimated literacy rates for women in the communes in Senegal, the corresponding estimated RMSE and the coefficient of variation (CV). 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. Our first observations is that the estimates for the literacy rate are higher for the FH Bench compared to the FH Trans. The reason is that the aggregated FH Trans estimates (36.1%) on national level slightly underestimate the national share of literate women (38%). However, the differences are more pronounced for the out-of-sample communes. In order to investigate the reason for these

Table 5: Distribution of the female literacy rates, estimated RMSE and coefficient of variation over communes in Senegal.

233 In-sample communes							
Indictor	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Point Est.	Direct	0.000	0.105	0.234	0.298	0.474	0.839
	FH Trans.	0.022	0.150	0.252	0.297	0.435	0.770
	FH Bench.	0.024	0.164	0.269	0.313	0.450	0.823
RMSE	FH Trans.	0.021	0.048	0.057	0.059	0.067	0.239
	FH Bench.	0.016	0.041	0.051	0.050	0.057	0.263
CV	FH Trans.	0.049	0.150	0.237	0.284	0.363	1.043
	FH Bench.	0.061	0.140	0.192	0.214	0.262	0.685
168 Out-of-sample communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Point Est.	FH Trans.	0.000	0.124	0.219	0.268	0.367	0.913
	FH Bench.	0.000	0.133	0.241	0.286	0.390	0.931
RMSE	FH Trans.	0.001	0.051	0.063	0.062	0.074	0.122
	FH Bench.	0.004	0.036	0.049	0.045	0.055	0.077
CV	FH Trans.	0.069	0.184	0.317	0.426	0.432	5.710
	FH Bench.	0.040	0.147	0.207	0.263	0.274	2.245
30 Out-of-covariate communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Point Est.	FH Trans.	0.095	0.157	0.198	0.212	0.241	0.553
	FH Bench.	0.103	0.165	0.207	0.221	0.251	0.570
RMSE	FH Trans.	0.035	0.043	0.047	0.047	0.051	0.066
	FH Bench.	0.031	0.039	0.044	0.043	0.047	0.056
CV	FH Trans.	0.110	0.218	0.236	0.242	0.278	0.367
	FH Bench.	0.095	0.184	0.217	0.211	0.240	0.304

differences, we had a closer look at the estimated RMSE of the FH Trans in Table 5. As expected, the estimated RMSE are smaller for the in-sample communes compared to the out-of-sample communes. As the weights for the FH Bench are defined by $\phi_i = w_i/\hat{\text{MSE}}(\hat{\theta}_i^{FH,trans})$ we expect that communes with higher variability in terms of RMSE will receive a larger adjustment. This is also confirmed by Figure 5. The plot shows the differences between the FH Bench and the FH Trans (solid line) in relation to the size of the estimated RMSE of the FH Trans (dashed line) on commune level. We can note that (i) the adjustments due to the benchmarking are larger than zero for all communes; (ii) the adjustments are proportional to the size of the estimated RMSE; (iii) the adjustments for the in-sample communes are smaller compared to the out-of-sample communes because of the smaller RMSE indicated by the dashed line. In order to save space, the corresponding table and figures for the estimated literacy rates for men on commune level are reported in the supplementary materials.

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 for women and men. To assess the resulting estimates of the FH Bench for female and male literacy we compare the estimates with the direct unbiased estimates in Figure 6. The figure shows the FH Bench versus the direct estimates for the literacy rate by gender

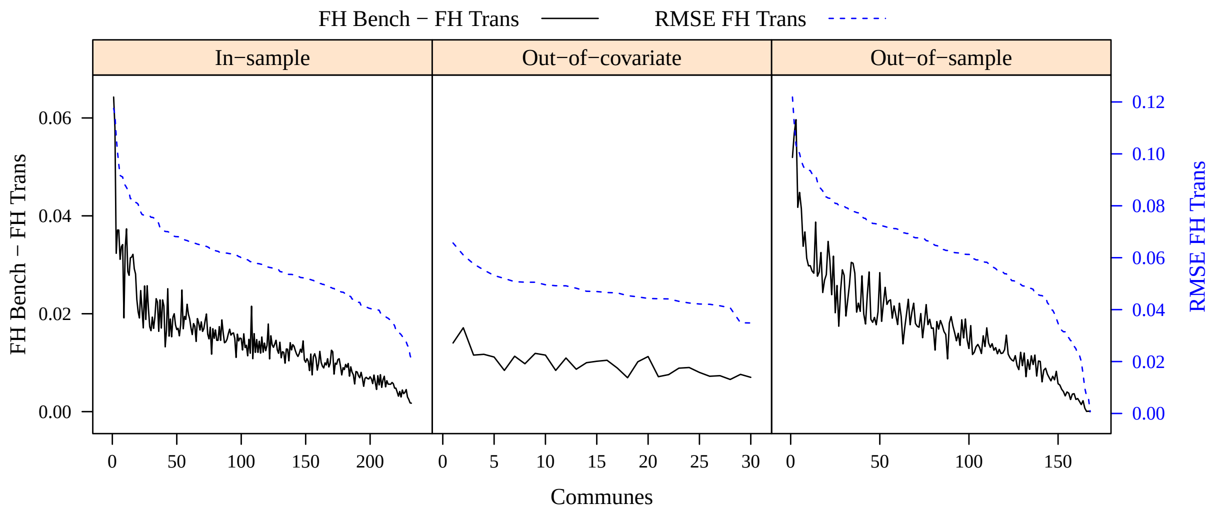

Figure 5: Differences between the FH Bench and the FH Trans on commune level for the female model.

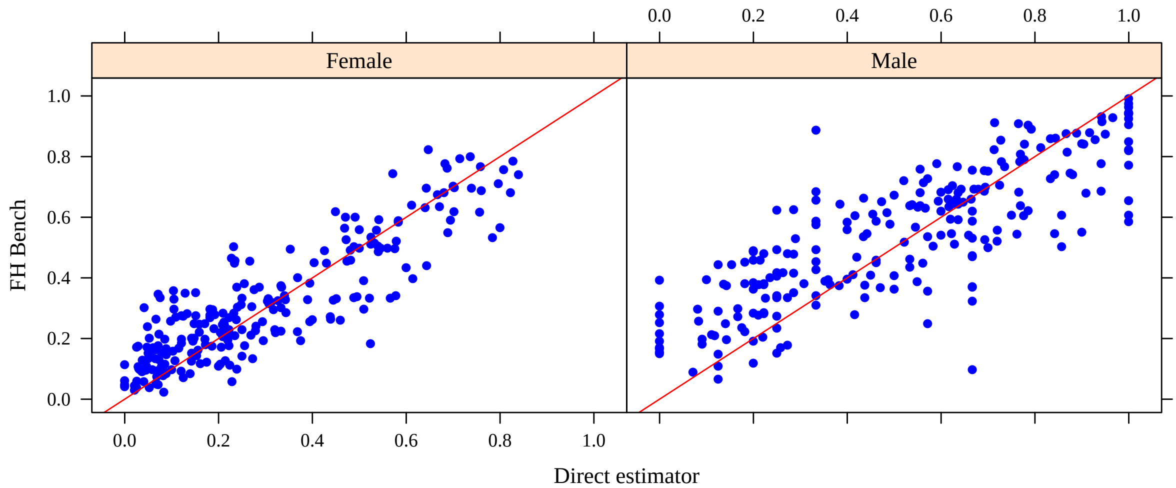

Figure 6: Estimates for literacy for women (left panel) and men (right panel): direct estimates vs. model-based estimates based on the FH Bench.

on commune level. We expect that the estimates of the FH Bench are similar to the direct estimates especially in communes with larger sample size. We observe that the direct estimates and the FH Bench behave similar for the female model. In contrast, the model-based estimates differ more compared to the direct estimates for the male model. This is because the sample size in the women's questionnaire is almost three times as big as the one in the men's questionnaire. We note that the model-based estimators are larger/ smaller compared to the direct estimator for small/ large values respectively.

4.3 Literacy rates by gender in Senegal

Having assessed the results of the estimators from a statistical perspective, we now discuss the results of the benchmarked transformed FH in the context of female and male literacy in Senegal. Figure 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). In order to simplify the interpretation of the results, Figure 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 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 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 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. The seven regions and the corresponding literacy rates for women are displayed in Figure 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.

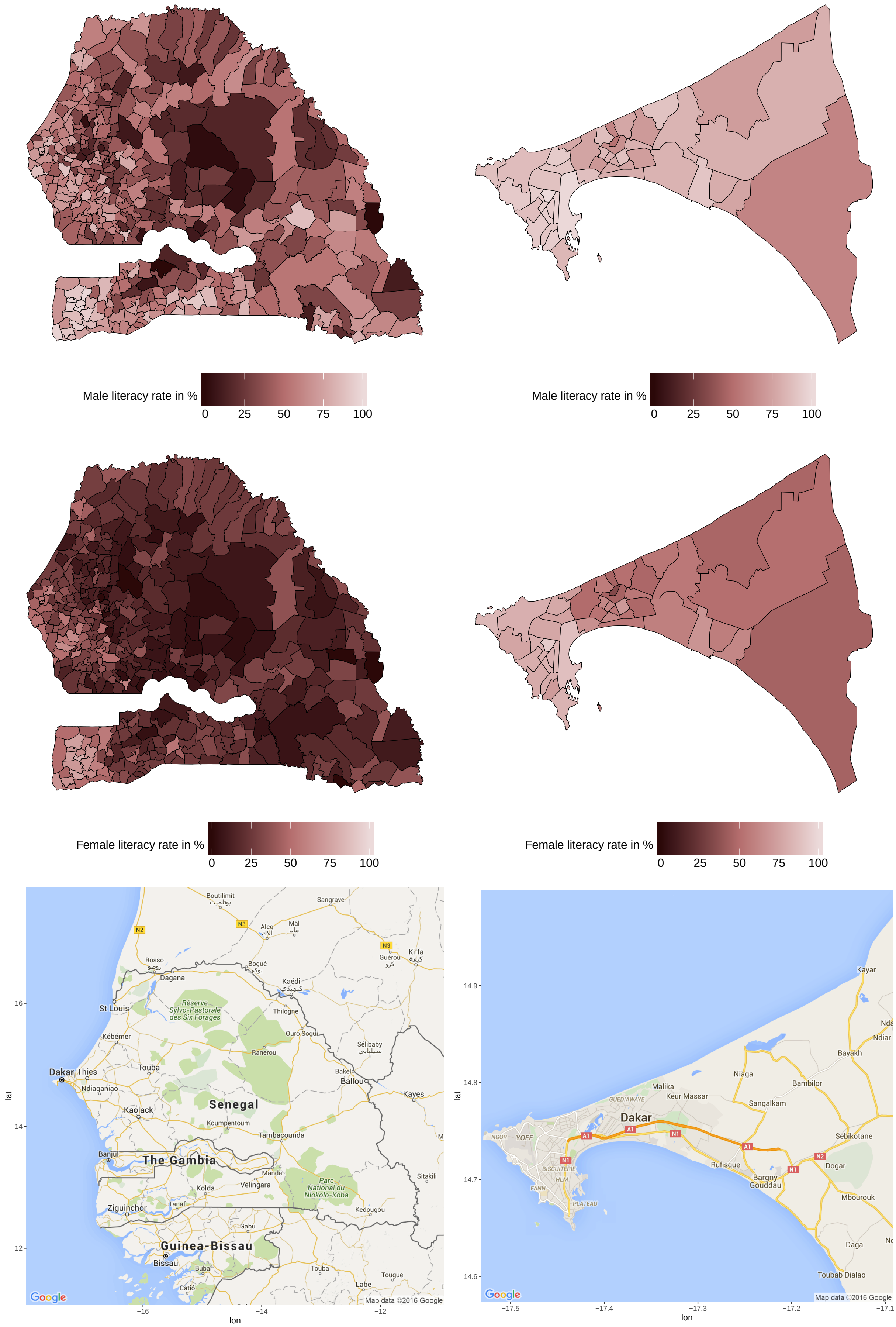

Figure 7: Estimates for the literacy rate by gender on commune level based on a benchmarked FH model: Senegal (left panel) and Dakar (right panel).

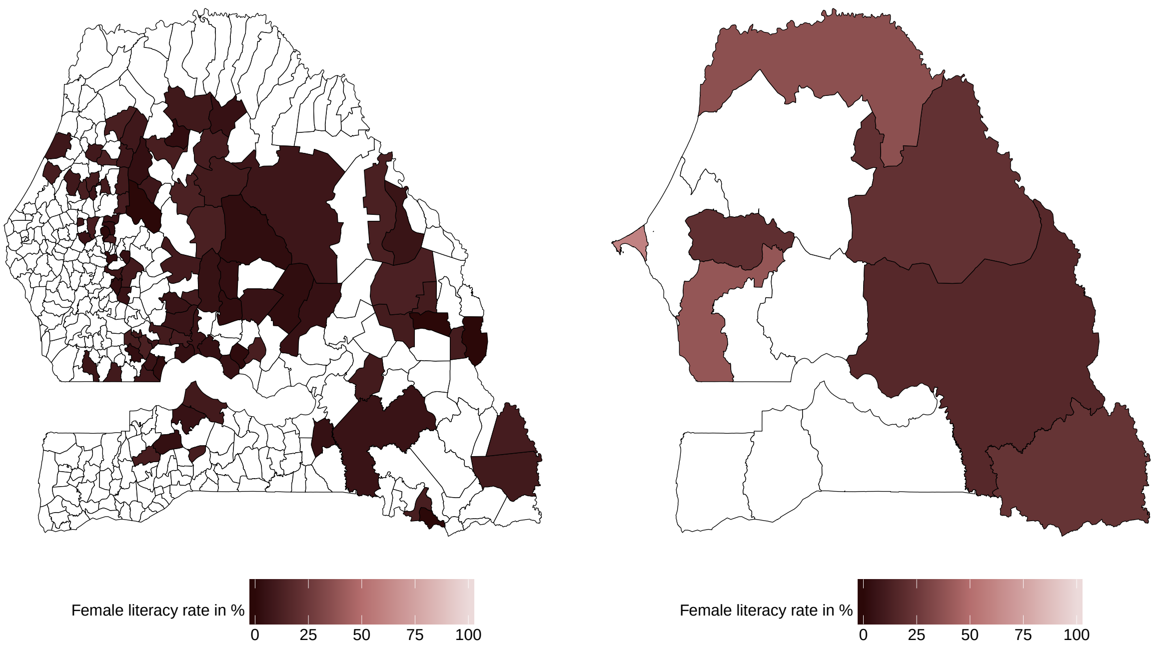

Figure 8: Estimates for the literacy rate for women: 20% of the communes with the lowest literacy rate (left panel) and seven regions in Senegal identified by the ANSD for the PAJEF project (right panel).

Hence, the use of the proposed approach may enable NSIs and governmental organisations to make sound strategic decisions regarding the best places for investigating 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.

5 Design-based simulation for unemployment

The analysis of literacy rates by gender in Section 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 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. We further evaluate the performance of the proposed MSE estimators presented in Section 3.

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 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

Table 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

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). The summary statistics of the sample sizes over communes are also provided in Table 6.

We investigate the estimators presented in Section 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 (6). The covariates were selected by using the Akaike information criterion (AIC) and hold fixed for the simulation study. The R^2 was on average around 30% depending on the selected sample. We evaluate three estimators for the share of unemployment in the communes in the simulation. These are the direct estimator (3), the transformed FH estimator (10) (FH Trans) and the benchmarked transformed FH estimator (14) (FH Bench). The estimators are implemented by computationally efficient algorithms using R. The codes are available from the authors upon request.

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

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

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

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 7 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 estimators 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 for the sampled communes. The performance of the FH Trans and FH Bench is very comparable regarding Bias and RMSE for the in-sample and out-of-sample communes. For the out-of-covariate communes, where the covariates are obtained by geographically weighting as described in Section 2, both estimators reveal on average a small positive bias. Note that the FH Trans and the FH Bench results are not directly comparable as the FH Bench fulfils the benchmarking constraint. However, the results of both approaches are quite 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. The results from the study indicate that combining mobile phone covariates with survey data can lead i) to gains in efficiency compared to the direct estimator and ii) to reasonable results for communes with zero sample sizes.

MSE estimation for the FH Trans and FH Bench is implemented with the parametric bootstrap approaches discussed in Section 3. In particular, we used $B = 500$ replicates for the bootstrap of the FH

Table 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	0.013	0.053	0.068	0.077	0.090	0.273
	FH Trans.	0.013	0.035	0.045	0.053	0.061	0.194
	FH Bench.	0.013	0.035	0.045	0.054	0.062	0.181
Bias	Direct	-0.021	-0.002	-0.001	-0.000	0.001	0.012
	FH Trans.	-0.129	-0.019	0.002	0.006	0.027	0.137
	FH Bench.	-0.124	-0.013	0.005	0.012	0.031	0.146
210 Out-of-sample communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Trans.	0.017	0.042	0.065	0.088	0.112	0.514
	FH Bench.	0.014	0.043	0.065	0.089	0.114	0.495
Bias	FH Trans.	-0.510	-0.052	0.014	0.006	0.059	0.280
	FH Bench.	-0.490	-0.041	0.023	0.017	0.067	0.292
30 Out-of-covariate communes							
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RMSE	FH Trans.	0.025	0.038	0.060	0.083	0.125	0.267
	FH Bench.	0.020	0.039	0.067	0.086	0.128	0.275
Bias	FH Trans.	-0.163	-0.025	0.047	0.038	0.096	0.267
	FH Bench.	-0.155	-0.018	0.054	0.045	0.107	0.274

Trans (13) and $B = 300$ with $L = 300$ replicates (in total 90,000 replicates) for the bootstrap of the FH Bench (15). Table 8 presents the results for the bootstrap MSE estimators and reports the mean and median value of the commune-specific relative Bias (RB) and relative RMSE (RRMSE). We treat the empirical MSE (over Monte-Carlo replications) as the *true* MSE. We observe that on average the proposed

Table 8: Performance of MSE estimators over communes in design-based simulations

		In-sample		Out-of-sample		Out-of-covariate	
Estimator		Median	Mean	Median	Mean	Median	Mean
FH Trans.	RRMSE (%)	19.20	22.79	46.78	58.08	44.63	51.34
	RB (%)	6.27	2.14	9.44	19.21	-0.40	3.46
FH Bench.	RRMSE (%)	21.51	23.80	47.22	59.49	47.30	54.72
	RB (%)	7.07	2.59	10.26	20.93	-11.31	3.69

bootstrap approaches for the FH Trans and FH Bench work quite well and are almost unbiased for the in-sample communes. As expected the variability (in terms of RRMSE) is higher for the out-of-sample communes. There are some out-of-sample communes were the parametric bootstrap estimators shows strong under- or overestimation. However, this is not new in literature and is also reported by Chandra et al. (2015). More detailed results are available from the authors on request.

Overall, it seems that the presented bootstrap approaches capture the variability due to the transformation and the benchmarking and have appealing properties regarding stability and bias in this simulation.

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 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*. The maps of these indicators are avaialble from the authors upon request. 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 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 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. One way for further research is to interpret this uncertainty as a measurement error in the covariates and to apply the model proposed by Ybarra and Lohr (2008). 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.

Acknowledgements

First of all, we would like to thank the staff of the Senegalese National Statistics Institute ANSD, especially the Directeur-Général Mr. Aboubacar Sedikh Beye and the Directeur du Management de l'Information Statistique Mr. Mamadou Niang for many fruitful discussions, country insights and data access. Further, we would like to thank Nicolas de Cordes and the Orange Group for the tremendous help with the coordination and logistics of the project as well as Sonatel for providing the call detail records. The project was made possible with the funding of the Bill and Melinda Gates Foundation. The work of Salvati has been developed under the support of the project PRIN-SURWEY http://www.sp.unipg.it/surwey/ (Grant 2012F42NS8, Italy).

References

ANSD (2012). Demographic and Health Survey - Multiple Indicator Cluster Survey (2010-2011). `https://dhsprogram.com/pubs/pdf/FR258/FR258_English.pdf`. Accessed: 2016-02-16.

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

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). *Analysis of Poverty Data by Small Area Estimation*, Chapter Poverty Mapping for the Chilean Comunas, 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*, 109–135.

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.

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.

Dowle, M., T. Short, S. Lianoglou, and A. Srinivasan (2014). *data.table: Extension of Data.frame*. R package version 1.9.6.

Eagle, N., M. Macy, and R. Claxton (2010). Network diversity and economic development. *Science 328*, 1029–1031.

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

Fay, R. E. and R. A. Herriot (1979). Estimation of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74 (366)*, 269–277.

Ford, L. (2007). People treat you as if you are nothing: What good is free education when poverty means children are forced into work? `http://www.theguardian.com/education/2007/aug/21/schoolsworldwide.schools`. Accessed: 2016-04-04.

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.

Gonzalez-Manteiga, W., M. J. Lombardia, I. Molina, D. Morales, and L. Santamaria (2008). Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics & Data Analysis 52 (12)*, 5242–5252.

GSMA (2015). The mobile economy 2015. `http://www.gsmamobileeconomy.com/GSMA\Global\Mobile\Economy\Report\2015.pdf`. Accessed: 2016-01-30.

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

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

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.

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.

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., J. Quoidbach, F. Robic, and A. S. Pentland (2013). *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings*, Chapter Predicting Personality Using Novel Mobile Phone-Based Metrics, pp. 48–55. Berlin, Heidelberg: Springer Berlin Heidelberg.

Montjoye, Y.-A. d., Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. Blondel (2014). D4d-senegal: The second mobile phone data for development challenge. *Working paper arXiv:1407.4885*.

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.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28 (1)*, 40–68.

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. and N. Salvati (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics 25 (1)*, 37–53.

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. (2003). *Small Area Estimation*. New York: Wiley.

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

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

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

Smith, C., A. Mashhadi, and L. Capra (2013). Ubiquitous sensing for mapping poverty in developing countries. *Working paper*.

UNESCO (2012). Global partnership for girls' and women's education. `http://www.unesco.org/eri/cp/factsheets_ed/SN_EDFactSheet.pdf`. Accessed: 2016-03-04.

UNESCO (2015). Literacy project for girls and women in senegal. `http://www.unesco.org/uil/litbase/?menu=4&programme=180`. Accessed: 2016-04-04.

Villalon, L. A. and M. Bodian (2012). Religion, social demand, and educational reforms in senegal. `http://r4d.dfid.gov.uk/PDF/Outputs/APPP/appp-research-report5-avril-2012.pdf`. Accessed: 2016-01-03.

Ybarra, L. M. R. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95 (4)*, 919–931.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2016 erschienen:

2016/1	BARTELS, Charlotte und Maximilian STOCKHAUSEN Children's opportunities in Germany – An application using multidimensional measures *Economics*
2016/2	BÖNKE, Timm; Daniel KEMPTNER und Holger LÜTHEN Effectiveness of early retirement disincentives: individual welfare, distributional and fiscal implications *Economics*
2016/3	NEIDHÖFER, Guido Intergenerational Mobility and the Rise and Fall of Inequality: Lessons from Latin America *Economics*
2016/4	TIEFENSEE, Anita und Christian WESTERMEIER Intergenerational transfers and wealth in the Euro-area: The relevance of inheritances and gifts in absolute and relative terms *Economics*
2016/5	BALDERMANN, Claudia; Nicola SALVATI und Timo SCHMID Robust small area estimation under spatial non-stationarity *Economics*
2016/6	GÖRLITZ, Katja und Marcus TAMM Information, financial aid and training participation: Evidence from a randomized field experiment *Economics*
2016/7	JÄGER, Jannik und Theocharis GRIGORIADIS Soft Budget Constraints, European Central Banking and the Financial Crisis *Economics*
2016/8	SCHREIBER, Sven und Miriam BEBLO Leisure and Housing Consumption after Retirement: New Evidence on the Life-Cycle Hypothesis *Economics*