# The ABCDE of Big Data: Assessing Biases in Call-detail records for Development Estimates *

Gabriel Pestre[†]
Data-Pop Alliance

Emmanuel Letouzé[‡]
Data-Pop Alliance

Emilio Zagheni[§]
University of Washington, Seattle

March 15, 2016

Revised Draft prepared for presentation at the
Annual World Bank Conference on Development Economics 2016

**Abstract**

This article contributes to improving our understanding of biases in estimates of demographic indicators, in the developing world, based on Call Detail Records (CDRs). CDRs represent an important and largely untapped source of data for the developing world. However, they are not representative of the underlying population. We combine CDRs and census data for Senegal in 2013 to evaluate biases related to estimates of population density. We show that: (i) there are systematic relationships between cell-phone use and socio-economic and geographic characteristics that can be leveraged to improve estimates of population density; (ii) when no 'ground truth' data is available, a difference-in-difference approach can be used to reduce bias and infer relative changes over time in population size at the subnational level; (iii) indicators of development, including urbanization and internal, circular, and temporary migration, can be monitored by integrating census data and CDRs. The paper is intended to offer a methodological contribution and examples of applications related to combining new and traditional data sources to improve our ability to monitor development indicators over time and space.

*This paper benefitted from the financial support of the *Agence française de développement* (AFD), which is gratefully acknowledged. The Call Detail Records (CDRs) were made available by Orange / Sonatel within the framework of the Data for Development (D4D) Challenge. Permission was obtained for their use in the context of Data-Pop Alliance's work on sample selection bias correction, and authors are grateful to Orange / Sonatel for facilitating access to these CDRs. The Census datasets from Senegal's 2013 *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage* (RGPHAE) were provided by the *Agence Nationale de la Statistique et de la Démographie* (ANSD). The authors wish to thank the ANSD for its help in accessing and using the Census data.

[†]gpestre@datapopalliance.org

[‡]eletouze@datapopalliance.org

[§]Corresponding author: emilioz@uw.edu

# 1 Introduction

This paper addresses a critical yet widely under-researched question for using 'Big Data' to produce socio-economic and demographic estimates (talker referred to as 'development estimates'): most data have not been produced for research purposes and the sample is typically not representative of the underlying population. Specifically, we use what has become one of the most sought-after kinds of Big Data—cell-phone data, in particular call detail records (CDRs) made available as part of the 2014 Orange Data for Development (D4D) Senegal Challenge—to estimate population densities in Senegal and contrast the results with the latest Census to understand the nature and size of the biases.

This paper builds on and feeds into the now large body of research that has leveraged Big Data—understood both as new kinds of passively emitted data about people's actions and interactions and as powerful computing techniques—in general, and CDR analytics in particular, to infer a wide range of social, economic, and demographic indicators. Many papers have come out of the D4D Senegal Challenge—and indeed the winning paper was precisely about using these digital breadcrumbs to infer socio-demographic indicators [4]. However, to our knowledge, this is the fist time that Senegal's 2013 Census has been used for analytical purposes in conjunction with CDRs, to improve our understanding of sample selection bias and how to address it.

The attention paid to Big Data by the development policy and research communities in recent years has sprung from two main sets of factors that can be lumped under the broad categories of supply and demand. On the demand side is the dearth of good—i.e. reliable, timely, disaggregated—data on development processes and outcomes that is believed to constrain development policy and programming. This has been compared to a "statistical tragedy" in the case of Africa—in reference to the continent's "growth tragedy" of the 1990s [9]. The availability of reliable, up-to-date, disaggregated data covering a wider range of human development dimensions has improved over time—notably thanks to the Demographic and Health Surveys (DHS) programs in the mid-1980s and the Millennium Development Goals (MDG) monitoring framework since 2000, but many data gaps remain as the world faces the uphill tasks of achieving the Sustainable Development Goals (SDGs).

The supply side refers to the expectations raised by the "Data Revolution"—and of "Big Data for development" as a general and fast emerging field of practice. In less than a decade since the early years when it was shown that GDP could be inferred "from outer space" using light emissions, Big Data's potential to address some of the world's most acute challenges has spurred a great deal of both excitement and skepticism. The debates surrounding the topic have many facets that go beyond the scope of this paper—including political, ethical, and legal [14, 15]—and with time a growing consensus is found on some of the most contentious issues.

But excessive claims about the real state of knowledge [21] and too little attention to the central issue of statistical representativeness have continued to fuel skepticism amongst many traditional social scientists—chief of which demographers. We hope

this paper will show promising avenues for better assessing and addressing sample selection bias in Big Data sources and help spur the interest of demographers and other social scientists in fulfilling Big Data's potential and producing methods to generate development estimates.

The rest of this paper is organized as follows. Section 2 presents the context and rationale for the paper in more depth; section 3 introduces the data sources that we used: CDRs and Census data for Senegal in 2013; section 4 presents the results that we obtained in terms of improving our understanding of the biases in CDRs for estimating population density in the context of the developing world; section 5 discusses the possibility of applying these results to lower levels of disaggregation; section 6 presents a difference-in-difference approach to reduce biases when estimating relative changes in population size over time. An illustrative example for the region of Dakar is offered. Finally, section 7 provides a short discussion of our work, the implications, and the next steps.

# 2 Contextualization and motivation: the promise and pitfalls of Big Data-based development estimates

As mentioned in the introduction, the idea of using Big Data to produce development estimates has been traced back to a frequently-cited paper which found that light emissions picked up by satellites could track GDP growth and proposed that they could supplement national accounting in data-poor countries [12]. This finding has been validated in other sources[1], but there is also evidence that this relationship can fade once the penetration of electric lighting approaches saturation [13, 20].

In more recent years, the high rate of growth of cell-phones penetration and use around the world, as well as the richness of information about the individual and collective behavior of users embedded in CDRs, have made them the focus of a large number of scientific articles and debates. In particular, the fact that people move with their cell-phones has given rise to a whole strand of research attempting to infer population movement and distribution, both in crisis and non-crisis contexts [8, 3, 10].

The value or having estimations of population movement, distribution, and potentially structure by age and gender, before, during and after a natural hazard, is very high. Promising applications have been developed. For example, Pastor-Escuredo et al. [17] studied how people moved before and after the major 2009 floods in the Mexican state of Tabasco. These reconstructions of the flood's impact were validated against the assessment of the flood area from Landsat-7 images, as well as official figures on the number of displaced people. Other examples in the cases of post-earthquake mobility

[1]See for example, Chen and Nordhaus [5] and Olivia et al. [16] (who use 'gold standard' data on electrification and economic growth for 5,000 sub-districts in Indonesia between 1992 and 2008).

analysis using CDRs include the case of Rwanda [3], and Flowminder's work in Haiti and in Nepal [10].

However, inferring population-wide estimates from a sample of the population (i.e. the subset of people who own and use a cell-phone at any given point in time) requires an understanding of how well that sample represents the population it is drawn from. This kind of problem with digital data was first exposed in some length for the specific case of relying on crowdsourced data by Patrick Ball, Jeff Klingner, and Kristian Lum in March 2011 [2]. In this context, data from volunteers were used to infer building damage in Haiti after the 2010 earthquake, but the signals actually painted a misleading picture because proximity to damage correlated strongly (negatively) with people's willingness and ability to report it. The bulk of reports, controlling for building location, indeed came from *less* damaged zones [2]. A similar problem was raised after Hurricane Sandy hit the New York and New Jersey areas of the United States in 2012, when most of the Tweets about Sandy came from Manhattan—an area which was hit much less than other parts of New York and New Jersey [6].

The issue is of course not limited to crisis contexts where the event itself affects the sample and its behavior. The possibility of making population-wide inferences from data from digital devices and services in the absence of additional information is limited by the simple fact that not everybody has access to these devices and services. Moreover, and critically, access and usage are not randomly distributed.

Overall, three main sources of biases can be identified:

1. *Selection bias*: people who use a cell-phone or who sign up with a specific carrier are not necessarily representative of the underlying population; some people text instead of calling, so they don't show up in call lists.
2. *Compositional changes*: the characteristics of the people in the sample change over time: some people start using their phone with the provider, some stop using a phone.
3. *Behavioral changes*: people change the way they use cell-phones over time for various reasons, and users may use their cell-phones in different ways during the week or during the weekend, when on vacation or at work, etc.

Valid estimates could be drawn from non-representative samples, without any correction, in some specific cases. For example, if the underlying population thinks in the same way about a specific issue. For instance, asking the micro-subset of billionaires whether they prefer dining with friends or plowing a field will likely yield similar results to those found in the population at large. But in most cases the 'signals' coming from the non-representative fraction of the population will not provide a good picture of the experience or perspectives of its whole because different people think and act differently.

Well-understood bias can nonetheless be accounted for and corrected to some extent by understanding how sample selection bias may skew the representation of each group in

the sample (i.e. the dataset) and 'unskewing' the data by giving more weight to certain observations (i.e. entries in the dataset). This is commonly referred to as sample bias correction. The key is to bring in other variables (such as the proportion of people in each age group who use the technology in question, in the case of most digital data) and use them to account for the under- or over-representation of certain groups within the sample, in order to get a better picture of the whole population.

Early work in this area has been done by Zagheni and Weber using data provided by Yahoo! about email user accounts. In their 2012 paper "You are where you E-mail" [22], the authors propose a method for studying human migration patterns based on geographic information for a large sample of Yahoo! e-mail messages, self-reported demographic information of Yahoo! users, migration rates for 11 European countries gathered by Eurostat from national statistical agencies, and international statistics on Internet penetration rates by age and gender. Based on IP address, they determine the country from which a user sends the most emails, then study how that location changes over time across all users.

In order for these observations to provide a reliable estimate of global migration, Zagheni and Weber apply a sample bias correction method reflecting the following assumptions:

- "when Internet penetration is very high, then the population of Yahoo! users is highly representative of the entire population" [22];
- there is an "over-representation of more educated and mobile people in groups for which the Internet penetration is low" [22].

They subsequently divide the observations according to gender, age group, and country, and apply a different correction factor to each group—the lower Internet penetration, the greater the expected bias, and the more correction is needed. They then calibrate their preliminary emigration estimates for 11 European countries against data on age-specific emigration rates published by Eurostat for those countries in 2009, in order to estimate a shape parameter for each subgroup. From these shape parameters, they then estimate emigration rates, by age and gender, for a large number of countries and discuss their results for two cases, the United States and the Philippines [22].

The reliability of their correction factor depends a lot on the availability of ground truth data to calibrate the shape parameter. As the authors point out, using data for European countries with relatively high Internet penetration rates means that there is a larger uncertainty for developing countries in their model:

> "estimates for countries with high Internet penetration rates are not very sensitive to changes in [the shape parameter] $k$, whereas estimates for countries with low Internet penetration rates are. Since we only have statistics from European countries, the likelihood function with respect to the parameter $k$ tends to be fairly flat, meaning that we have rather high uncertainty" [22].

A similar observation was made by Deville et al. regarding their estimates of population densities in France and Portugal: "applying the method to low-income countries where penetration rates are increasing rapidly but still exclude an important fraction of the population would require further sensitivity analyses of the impact of phone use inequalities, especially as marginalized populations also are the most vulnerable to disasters, outbreaks, and conflicts" [8].

In other words, correcting for sample selection bias is essential and sample bias correction requires solid ground truth data for calibration. In the next sections, we will introduce the data that we have for Senegal and present the methodology that we propose to address the issue of sample bias in the specific case of population density in Senegal.

# 3 Data

## 3.1 Call Detail Records from Orange

Our analysis uses anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between more than million Orange customers in Senegal between 1 January 2013 to 31 December 2013. These CDRs were released as part of Orange's 2014 Data for Development (D4D) Challenge [7]. The D4D Challenge provided 3 different datasets:

- Dataset 1: One year of site-to-site traffic for 1666 sites on an hourly basis;
- Dataset 2: Fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users;
- Dataset 3: One year of coarse-grained (3rd administrative level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users.

The methodology in this paper employs Dataset 3, which provides the complete call list for the 2013 calendar year for 146,352 users meeting both of the following criteria:

1. Users having interactions on more than 75% of days in the given period.
2. Users having had an average of less than 1000 interactions per week (since users with more than 1000 interactions per week were presumed to be machines or shared phones).

The CDRs provide the point of origin of the call at country's 3rd administrative level (i.e. the *arrondissement*, of which there are 123 in this dataset), which allows us to estimate how many Orange callers were present in each *arrondissement* on a given day (given that the sample in Dataset 3 is representative of Orange subscribers).

## 3.2 Census data from ANSD Sénégal

Our population and demographic data comes from the *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage* (RGPHAE), Senegal's official census, which is carried out by the *Agence Nationale de la Statistique et de la Démographie du Sénégal* (ANSD). The 2013 edition of the RGPHAE was conducted over the 21 day period from November 19 to December 14 of that year.[1]

The ANSD provided us with a one-tenth sample of the Census, which contains datasets on individuals, households, emigration, deaths, agriculture, and livestock. We focused mainly of variables from the individuals and households datasets.

This data is used as a ground truth, to calibrate CDR-based estimates, which is made possible by the fact that the Census and CDR data cover roughly the same period. The goal of this analysis is, on one hand, to develop a methodology for using CDRs to calculate census-like indicators of population and demographic characteristics; and on the other hand to study in depth how the CDRs paint a biased picture of the general population, and how some of those biases can be accounted for to make good use of the CDRs as a complement to the census.

## 3.3 Geolocation of the CDR and Census data

Senegal is administratively divided into 4 levels:

- *région* (region)
- *département* (department)
- *commune / arrondissement / ville* (CAV)
- *commune d'arrondissement / commune rural* (CACR)

It should be noted that the 3rd administrative level, generally referred to as *CAV*, includes includes *communes* and *villes* (generally large towns and cities, respectively), which are administered separately from communes. However, for historical reasons, each *commune* and *ville* generally lies within the geographic boundaries of a single *arrondissement*.

The CDR dataset assumed that the country was neatly partitioned into arrondissements at the 3rd administrative level, and used a scheme with 14 regions, 45 departments, and 123 arrondissements. A number of *communes* and *villes* were therefore groups with their nearest arrondissement in the data we received from Orange. The census data is geolocated to Senegal's 4th administrative level (*commune d'arrondissement / commune rural*), and is divided into 547 such areas.

Using a combination of spatial merges in GIS, consultation of Senegal's laws pertaining to changes in the administrative division of the country [18, 19], and tables


Figure 1: Administrative breakdown of Senegal used by the CDR dataset.

of administrative areas from the GADM database of Global Administrative Areas [11], we were able to map all 547 areas in the census data to exactly one of the 123 areas in the CDR data.

# 4 Estimating Population Density using Call Detail Records

## 4.1 The standard approach for evaluating population density

In the literature about modeling population density using CDRs, the standard approach relies on the following model:

$$\log(P) = \alpha + \beta log(U) + \epsilon \tag{1}$$

Where $P$ is population size for a specific geographic area and time; $U$ is the number of cell-phone users for the respective geographic area and time; $\alpha$ is a scale ratio parameter; $\beta$ is the parameter that describes the superlinear effect in the relationship between users and population size; $\epsilon$ is a random error. The parameters are typically estimated using a regression model (see Deville et al. [8]).

The model described in equation (1) has proven useful in the context of high-income countries, with rather uniform cell-phone penetration rates. The same baseline model performs quite well with our data for Senegal. Figure 2 shows an example of model fit for equation (1) using Census and CDR data for Senegal (2013). With an $R^2$ equal to 0.768, the relationship seems to hold fairly well in the Senegalese context.

However, when we looked into errors and spatial correlations, we observed systematic pattensrns. For example, figure 3 shows ratios of population size and number of callers at the arrondissement level in Senegal. These values can be interpreted as the inverse of cell-phone penetration rates. An initial exploratory analysis indicates that there are some patterns in the distribution of cell-phone penetration rates, with clusters that

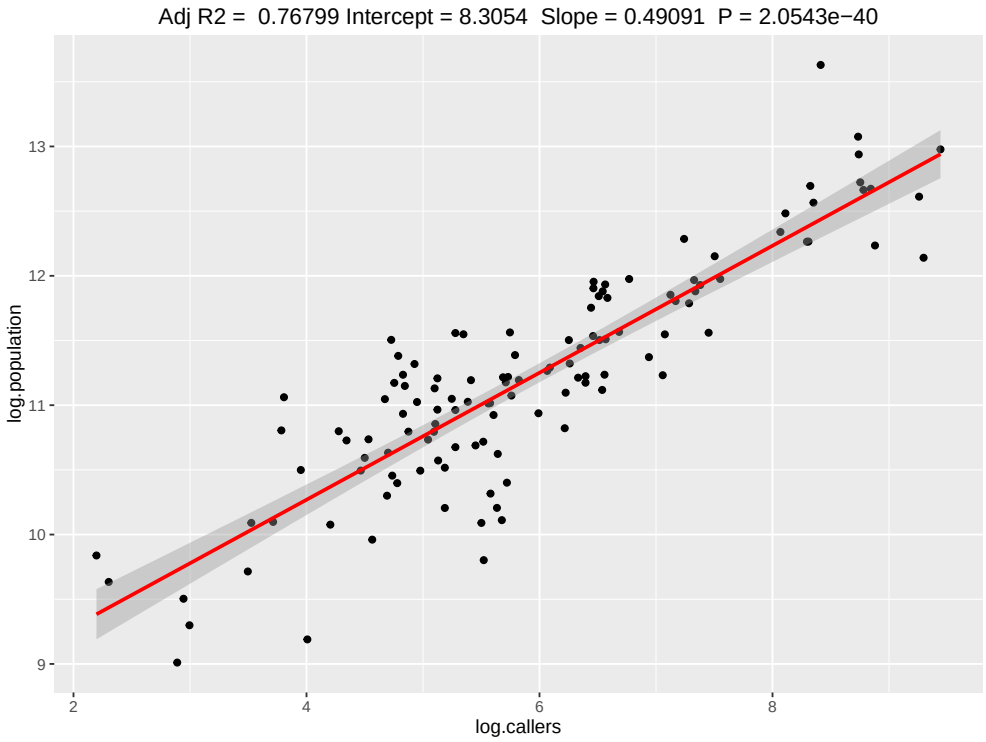


Figure 2: Fit for the regression model of population size on number of callers (log-log scale) for Senegal (2013).

form for areas that are geographically close, or that have the same type of urban vs rural setting, or with similar demographic characteristics. In the next section, we will discuss an illustrative example and some analyses related to this issue.

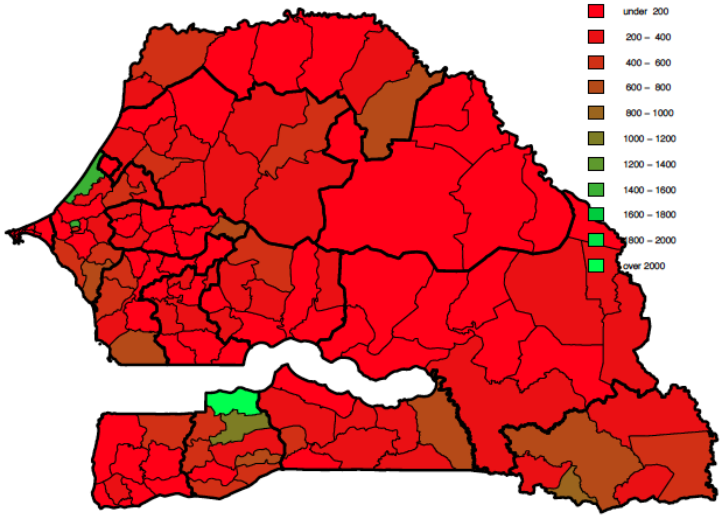


Figure 3: Ratios of population size and number of callers in our CDR sample, for arrondissements in Senegal (2013).

## 4.2 Using Census data to identify patterns of bias

Census data provide valuable information about socio-demographic and economic characteristics of each *arrondissement* in Senegal. This information can be leveraged to understand whether there are systematic biases in the relationship between population size and number of callers.

Figure 4 shows the relationship between the number of callers and the actual resident population at the *arrondissement* level in Senegal. Figure 5 and figure 6 show the corresponding relationships at the *département* and *région* levels, respectively. The data points are color-coded to indicate the average age of the population in each administrative, based on Census information. It is relevant to observe that there are systematic differences across age groups. The red data points, for areas with younger populations, lie mostly above the regression line. Conversely, the green data points, for areas with older populations, lie mostly below the regression line. In other words, using the standard model of equation (1) leads us to underestimate population size/density in regions with younger populations, and overestimate it in regions with older population age structure. This holds true at the *arrondissement*, *département*, and *région* levels.

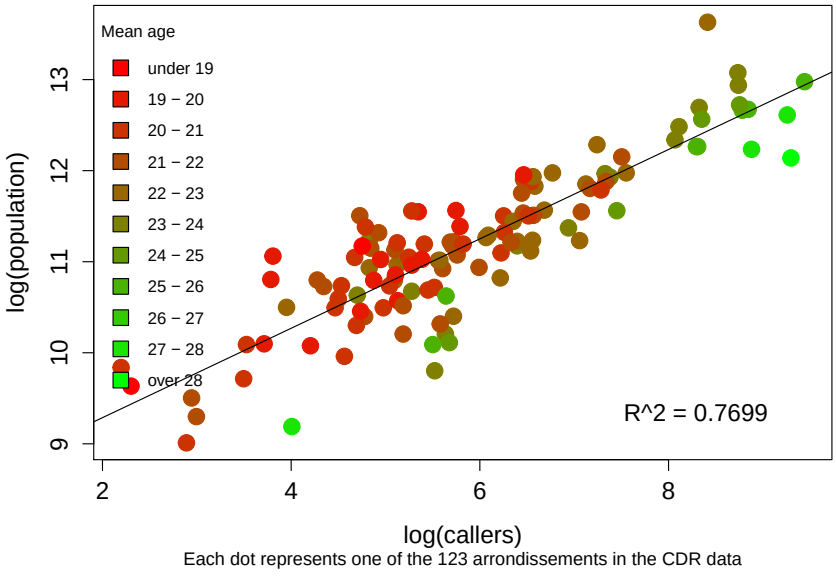


Figure 4: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *arrondissement*.

Including mean population age in the standard model at the arrondissement level, as specified in equation (2), significantly improves the fit ($R^2 = 0.827$):

$$\log(P) = \alpha + \beta\, log(U) + \gamma\, mean.age + \epsilon \qquad (2)$$

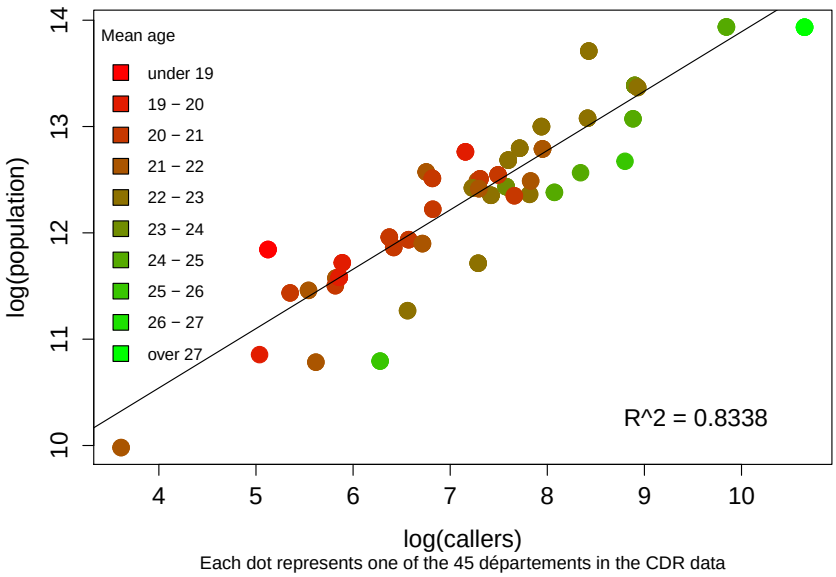


Figure 5: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *département*.

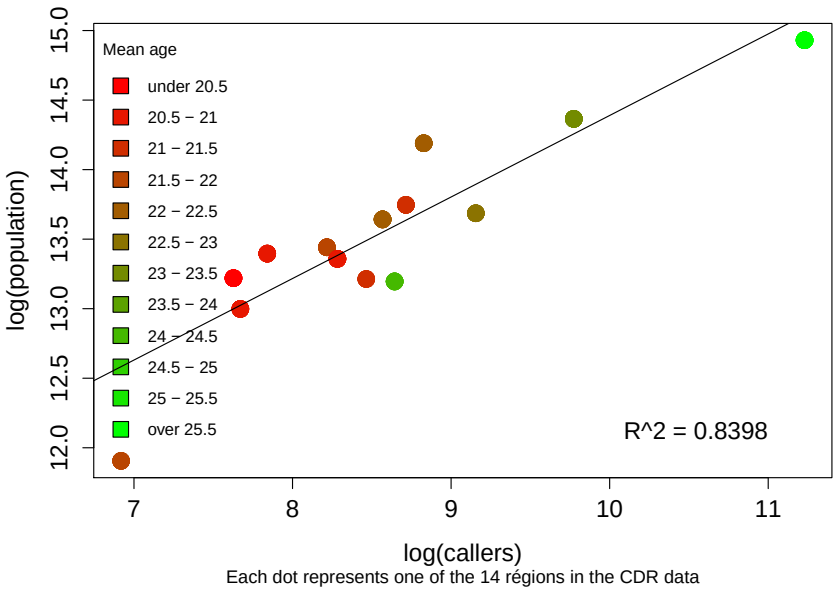


Figure 6: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *région*.

| | |
|---|---|
| Intercept | $10.644^{***}$ |
| | (0.393) |
| log(callers) | $0.597^{***}$ |
| | (0.027) |
| mean population age | $-0.135^{***}$ |
| | (0.021) |

Table 1: Regression coefficients and associated standard errors for the linear model where the dependent variable is the logarithm of population size for each *arrondissement* in Senegal.

As table 1 shows, the coefficient associated to average population age is negative and highly significant, indicating that, all else being equal, estimates of population size based on number of callers for regions with older population structure would be adjusted downwards as expected from the visualization in figure 4.

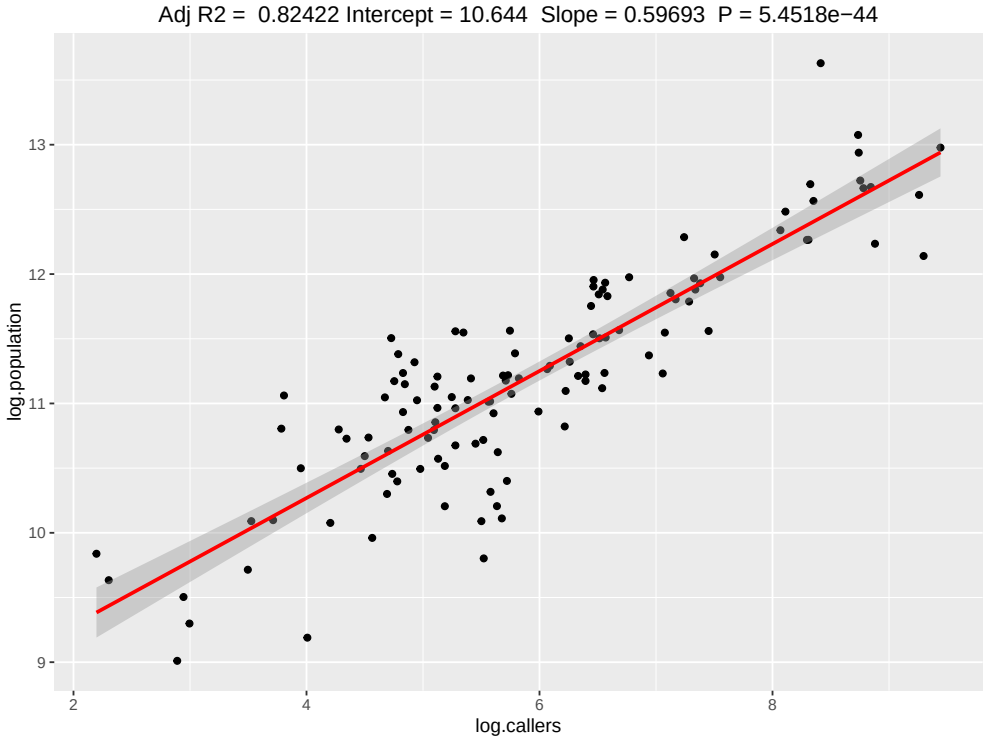


Figure 7: Fit for the regression model of population size on number of callers (log-log scale) for Senegal (2013).

The example that we discussed indicates that there are systematic differences in the relationship between callers and population size. Patterns emerge when rich data sets, like Census data, can be used to complement CDRs. Age is one of the most important demographic characteristics of a population. We showed that variations in age structure distort the relationship between number of callers and population size. Understanding the size and direction of distortions allows us to improve estimates of specific indicators of interest.

So far we have discussed an illustrative example using age structure. Other potential confounding factors that are related to socio-economic characteristics of users (e.g., educational attainment) or behavioral differences in cell-phone use (e.g., differences between weekdays and weekends or between different months of the year) may also be considered. In our case, we do not have access to demographic information about the cell-phone users themselves (as this would have to be provided by the carrier), but controlling for the socio-economic and behavioral characteristics of each administrative area can be leveraged to improve estimates of population density.

# 5 Projecting the regression coefficients down to lower administrative levels

In this section, we explore the possibility of applying the regression results from the previous section across various levels of disaggregation.

In order to see whether the coefficients calculated at a given administrative level could be project down to smaller geographic areas, we calculated regression coefficients and fitted values at the *région* level, then used those coefficients to estimate population at the *département* level and *arrondissement* level. We also calculated the coefficients and fitted values at the *département* level and used them to estimate populations at the *arrondissement* level. Finally, we calculated the coefficients and fitted values at the *arrondissement* level. This was done for both the standard model in equation (1) and the model improved with mean age in equation (2).

We compared these population estimates to the Census populations for the corresponding administrative level, and calculated the mean absolute percentage error (MAPE) in each case. The results are presented in tables 2 and 3.

We notice that with standard model used in the literature, it seems hard to down-project. For instance, when estimating populations at the *arrondissement* level, the MAPE more the doubles, from 2.80% to 7.21%, when using fits from the *région* level (two levels up) instead of the *arrondissement* level itself. However, once we use information about mean age to improve the regression, as in equation (2), the marginal error still increase as we move to smaller geographic areas, but not at the same rate as it does without the additional information. Thus for the second regression, in the same situation, the MAPE only increases from 2.52% to 3.53%.

These results suggest that by adding certain explanatory variables such as age to the regression, we make our ability to extrapolate populations at smaller geographic levels much more robust than in the standard model.

| | Estimates of log(population) at ... | | |
|---|---|---|---|
| | *arrondissement* level | *département* level | *région level* |
| using coefficients fitted at ... | | | |
| *arrondissement* level | 2.80% | – | – |
| *département* level | 4.35% | 1.90% | – |
| *région level* | 7.21% | 3.75% | 1.51% |

Table 2: Mean absolute percentage errors (MAPE) for the regression in equation (1).

| | Estimates of log(population) at ... | | |
|---|---|---|---|
| | *arrondissement* level | *département* level | *région level* |
| using coefficients fitted at ... | | | |
| *arrondissement* level | 2.52% | – | – |
| *département* level | 3.16% | 1.68% | – |
| *région level* | 3.53% | 1.98% | 1.21% |

Table 3: Mean absolute percentage errors (MAPE) for the regression in equation (2).

# 6 Estimating Population change over time using a difference-in-differences approach

In this section we use a difference-in-differences approach to evaluate the extent to which the population density in certain geographic areas changes relative to other areas, and over time.

We chose regions within the Dakar area that have very similar cell-phone penetration rates. Figure 8 shows trends in the average number of cell-phone users for the *arrondissements* Grand Dakar and Parcelles Assainies, over the course of a year. The two regions show trends that are almost perfectly parallel.

Figure 9 shows trends in average number of cell-phone users for the *arrondissements* Grand Dakar and Dakar Plateau. These two *arrondissements* are also part of the greater Dakar area. Dakar Plateau is an important center for commercial activity and tourism.

The trend lines are parallel for the period between August and January. During this period, the trend is very similar to the one observed in figure 8 for Parcelles Assainies. Between February and July, the number of cell-phone users in the Dakar Plateau rapidly increases, suggesting that there might be a seasonal pattern that differentially affects the Dakar Plateau. In order to evaluate the size of the effect, we estimated the following difference-in-differences model:

$$U_i^t = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 G_i T_t + e_{it} \tag{3}$$

where $U_i^t$ is the number of cell-phone users for the regions of Dakar Plateau and Grand Dakar, over time. $G_i$ is an indicator variable that is equal to 1 if the observation is for the region Dakar Plateau, 0 otherwise. $T_t$ is an indicator variable that takes the value

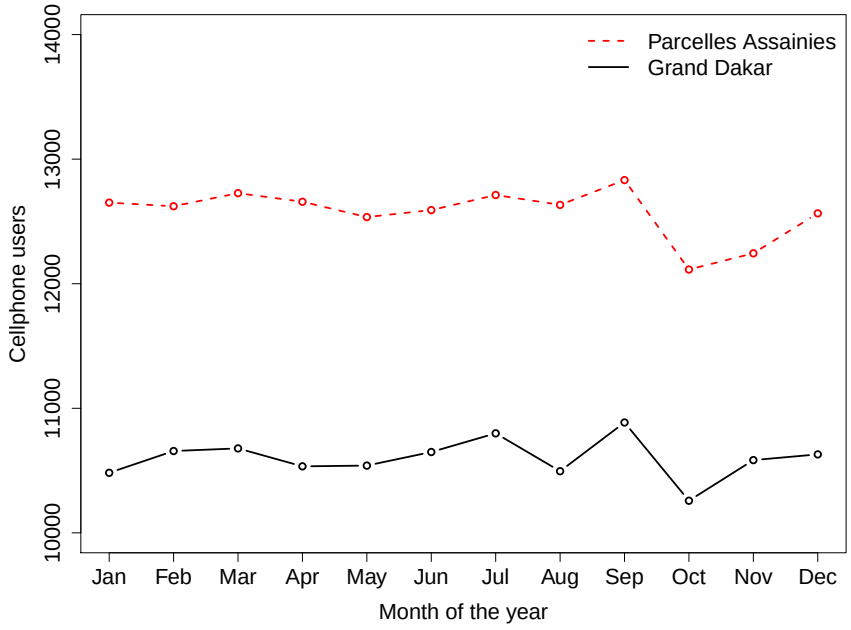


Figure 8: Average number of cell-phone users for the *arrondissements* Grand Dakar and Parcelles Assainies over the course of the year.

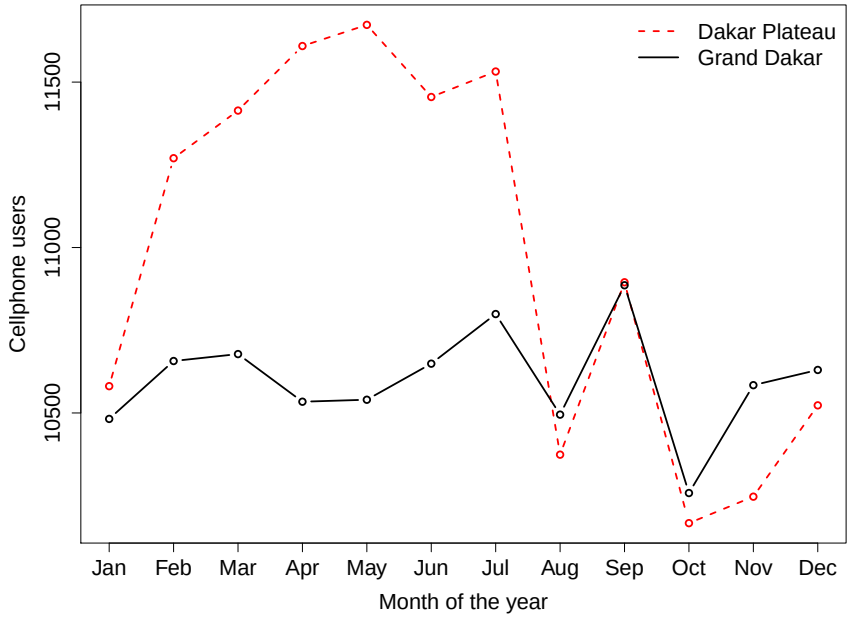


Figure 9: Average number of cell-phone users for the *arrondissements* Grand Dakar and Dakar Plateau over the course of the year.

1 during the period from February to July, 0 otherwise. The difference in difference estimator $\hat{\delta}$ is equal to the estimate for the parameter $\beta_3$. The estimate for $\beta_3$ is 940.67 (s.e. = 154.12) and is highly significant. This is a large change in population size: based on the results form the regression models estimated in the previous section, the change in population size would be in the order of about 100 thousand people.

Although further investigation would be required to determine the reasons behind these changes in population size, we expect two main factors contribute to these differentials. The first one is temporary migration of workers, who spend part of the year in the city, during the dry months, and part of the year in the countryside, during the rain seasons, as they help family members with agricultural production. As second potential explanation is related to flows of tourists. In any case, this comparison of Grand Dakar, Dakar Plateau, and Parcelles Assainies demonstrates that, even in the absence of ground truth, CDRs can be used to infer relative changes over time in population size at the subnational level.

# 7 Conclusions and Discussion

This paper builds on and feeds into a growing body of research on how Big Data can be leveraged to produce socio-economic and demographic estimates. Using CDR data from Orange's 2014 Data for Development Senegal Challenge and Census data from Senegal's 2013 *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage*, we investigated the possibility of combining traditional and new data sources to understand patters of bias in CDRs and to improve estimates of demographic indicators based on CDR data.

Our results demonstrate that many of the potential sources of bias in a CDR dataset can be better understood and accounted for, given sufficient ground truth. Starting from a simple log-log model relating number of callers to population data from the Census, we looked for other variables in the Census that had similar values across areas where the model consistently over- of under-estimated the population size.

We observed, for instance, significant differences in the relationship between number of callers and population density for *arrondissements* with different age structures, and used this to improve the predictive power of our model. We also explored how well the regression fits from a given geographic level can be projected down to a lower geographic level. Although our model is fairly simple—in this case, we take the log-log model traditionally found in the literature, and add only age—it appears to be fairly robust to this sort of extrapolation. We then developed an approach based on a difference-in-differences regression to evaluate relative changes over time at the subnational level, demonstrating that some inference about population size is still possible in the absence of ground truth.

This model is a first step to show that accounting for sample selection bias is possible, given sufficient ground truth data, and it could be extended in multiple directions

depending on context. We hope that this work will show promising avenues for better assessing and addressing sample selection bias in Big Data sources and help spur the interest of demographers and other social scientists in fulfilling Big Data's potential and producing methods to generate development estimates.

# References


[1] Agence Nationale de la Statistique et de la Démographie du Sénégal. Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage, 2013.

[2] P. Ball, J. Klingner, and K. Lum. Beneblog: Technology Meets Society: Crowdsourced data is not a substitute for real statistics, Mar. 2011.

[3] J. E. Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development*, 18(2):107–125, Apr. 2012.

[4] F. Bruckschen, T. Schmid, and T. Zbiranski. Cookbook for a socio-demographic basket: Constructing key performance indicators with digital breadcrumbs. In *Data for Development Challenge Senegal, Book of Abstracts: Scientific Papers*, pages 122–131, MIT Media Lab, Cambridge, MA, 2014.

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

[6] K. Crawford. The Hidden Biases in Big Data. *Harvard Business Review*, Apr. 2013.

[7] Y. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *CoRR*, abs/1407.4885, 2014.

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

[9] W. Easterly and R. Levine. Africa's Growth Tragedy: Policies and Ethnic Divisions. *The Quarterly Journal of Economics*, 112(4):1203–1250, Nov. 1997.

[10] Flowminder. Nepal Population Estimates as of May 1, 2015, May 2015.

[11] GADM database of Global Administrative Areas. Version 2.8, Nov. 2015.

[12] J. V. Henderson, A. Storeygard, and D. Weil. Measuring Economic Growth from Outer Space. Technical Report w15199, National Bureau of Economic Research, Cambridge, MA, July 2009.

[13] R. Kulkarni, K. Haynes, R. Stough, and J. Riggle. Light based growth indicator (LBGI): exploratory analysis of developing a proxy for local economic growth based on night lights. *Regional Science Policy & Practice*, 3(2):101–113, 2011.

[14] E. Letouzé, P. Vinck, and L. Kammourieh. The Law, Politics and Ethics of Cell Phone Data Analytics. *Data-Pop Alliance*, Apr. 2015.

[15] S. McDonald. Ebola: A Big Data Disaster - Privacy, Property, and the Law of Disaster Experimentation. *The Centre for Internet and Society*, Jan. 2016.

[16] S. Olivia, J. Gibson, L. K. Brabyn, and G. Stichbury. Monitoring economic activity in Indonesia using night light detected from space. In *The 12th Indonesian Regional Science Association Conference*, Makassar, Indonesia, 2014.

[17] D. Pastor-Escuredo, A. Morales-Guzmn, Y. Torres-Fernndez, J.-M. Bauer, A. Wadhwa, C. Castro-Correa, L. Romanoff, J. G. Lee, A. Rutherford, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, and M. Luengo-Oroz. Flooding through the lens of mobile phone activity. *arXiv:1411.6574 [cs]*, pages 279–286, 2014.

[18] République du Sénégal. Loi n$^{o}$ 96-06 du 22 mars 1996 portant Code des Collectivités locales, 1996.

[19] République du Sénégal. Loi n$^{o}$ 2013-10 du 28 décembre 2013 portant Code général des Collectivités locales, 2013.

[20] C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 511–520. ACM Press, 2014.

[21] Sustainable Development Goals Blog. UN Adviser underlines importance of partnership with mobile-communications industry to achieve Sustainable Development Goals, Feb. 2016.

[22] E. Zagheni and I. Weber. You Are Where You e-Mail: Using e-Mail Data to Estimate International Migration Rates. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 348–351, New York, NY, USA, 2012. ACM.