The World Bank Economic Review, 34(Supplement), 2020, S89–S97
doi: 10.1093/wber/lhz039
Supplementary Article

The ABCDE of Big Data: Assessing Biases in Call-Detail Records for Development Estimates

Gabriel Pestre, Emmanuel Letouzé, and Emilio Zagheni

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.

JEL classification: C82

Keywords: Big Data, national statistics, demographic indicators, call detail records, sample bias correction, digital breadcrumbs

1. Context and Objective

This paper aims to shed light on a critical, yet under-researched, challenge for using Big Data to produce development estimates: the challenge is that most of these data have not been collected for inference purposes, such that the sample of people who self-select into using certain devices, services, and/or providers, and are typically not representative of their underlying population, especially in developing countries. This study uses cell-phone call detail records (CDRs) made available as part of the 2014 Orange Data

Gabriel Pestre is an MS candidate in Data Science at Harvard University; his email address is gpestre@gmail.com. Emmanuel Letouzé is the director and co-founder of Data-Pop Alliance; his email address is eletouze@datapopalliance.org. Emilio Zagheni (corresponding author) is director of the Max Planck Institute for Demographic Research; his email address is zagheni@demogr.mpg.de. This paper benefited from the financial support of the World Bank Group, 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.

for Development (D4D) Senegal Challenge to estimate population size down to the third administrative level in Senegal. The results are then contrasted with the latest census to study the nature and size of the biases.

This study builds on, and adds to, the large body of research that has leveraged the richness and growth of CDRs (with cell-phone penetration reaching 90 percent in developing countries) (International Telecommunication Union 2014) to infer development indicators (Henderson, Storeygard, and Weil 2009; Chen and Nordhaus 2011; Kulkarni et al. 2011; Blumenstock 2012; Deville et al. 2014; Olivia et al. 2014; Smith-Clarke, Mashhadi, and Capra 2014; Flowminder 2015). Many papers came out of the D4D Senegal Challenge (Orange 2012, 2014)—and the winning paper was precisely about inferring socio-demographic indicators (Bruckschen, Schmid, and Zbiranski 2014). However, the present study seems to be the first time Senegal's 2013 Census has been used along with CDRs to improve understanding of sample selection bias.

Bias can be corrected by understanding the representation of subgroups in the sample and "unskewing" the data by giving more weight to certain observations—based on age groups, for instance, especially when dealing with data reflecting technological use. Zagheni and Weber (2012) used data provided by *Yahoo!* about email user accounts to estimate emigration rates in combination with user-reported demographic information and international statistics on Internet penetration rates by age and gender. They found that their estimates depended on the availability of ground truth data to calibrate their "correction factors," with greater reliability for European countries with relatively uniformly high Internet penetration rates and larger uncertainty for developing countries with lower and more skewed penetration rates (Zagheni and Weber 2012).

More recently Deville et al. (2014, 15891), discussing their analysis of population densities in France and Portugal based on CDRs, noted that "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."

This paper aims to help develop better sample bias correction methods in the specific context of developing countries with lower cell-phone penetration rates, using traditional data (Senegal's census) in conjunction with new data sources (CDRs) to estimate population size, taking into account the different usage patterns of various user groups.

2. Data Sources and Preparation

Call Detail Records from Orange

This study used anonymized CDRs of phone calls and short message service exchanges between more than a million Orange customers in Senegal between January 1, 2013, and December 31, 2013, released for Orange's 2014 D4D Challenge (de Montjoye et al. 2014). In particular, the study focused on a dataset of coarse-grained mobility data for 146,352 randomly sampled users. The CDRs provide the point of origin of all calls at country's third administrative level (the *arrondissement*). Users who had interactions on less than 25 percent of days or more than 1,000 interactions per week were removed from the datasets provided, the former because of the risk of reidentification and the latter because they were assumed to be machines or shared phones. These CDRs make it possible to estimate how many Orange callers were present in each *arrondissement* on a given day.

Census Data from ANSD Sénégal

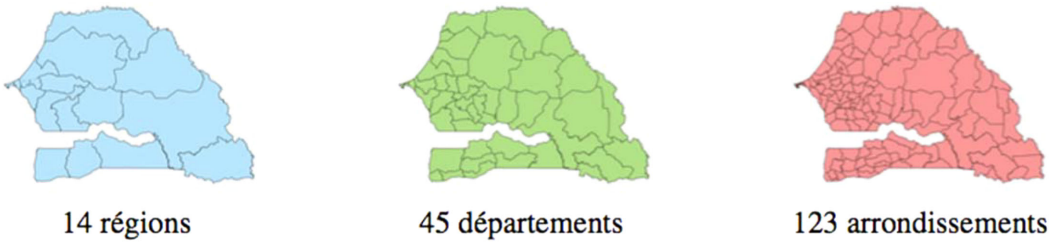
The demographic data come from the 2013 *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage* (RGPHAE), Senegal's official census carried out by the *Agence Nationale de la Statistique et de la Démographie du Sénégal* (ANSD). This RGPHAE was conducted from November 19 to December 14 of that year (Agence Nationale de la Statistique et de la Démographie du Sénégal 2013).

The ANSD provided a one-tenth sample of individual and household responses to the census. These data are used as a ground truth, to calibrate CDR-based estimates, which, importantly, cover roughly the same period.

Geolocation of the CDR and Census Data

Senegal is divided into four administrative levels: région (region); département (department); commune/arrondissement/ville (CAV); commune d'arrondissement/commune rural (CACR). The third administrative level (CAV) includes *communes* and *villes* (generally large towns and cities, respectively) administered separately from communes. However, for historical reasons, most *communes* and *villes* lie within the geographic boundaries of a single *arrondissement*. The CDR dataset used a 3-tier scheme with Senegal's 14 regions, 45 departments, and 123 *arrondissements*, grouping *communes* and *villes* with their nearest *arrondissement*. The census data in contrast is geolocated to Senegal's 4th administrative level, "CACR," made up of 547 small areas. Using a combination of spatial merges in GIS, extensive consultation of Senegal's laws on changes in the administrative division of the country (République du Sénégal 1996, 2013), and tables of administrative areas from the GADM database of Global Administrative Areas (GADM database of Global Administrative Areas 2015), it was possible to map all 547 areas in the census data to exactly one of the 123 areas in the CDR data (see fig. 1).

Figure 1. Administrative Breakdown of Senegal Used by the CDR Dataset



Source: The figure was produced by the authors using boundary data from GADM database of Global Administrative Areas (2015).

3. Methodology

The Standard Approach for Evaluating Population Size from CDRs

Following the standard approach in the literature (Deville et al. 2014), the following model was estimated:

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

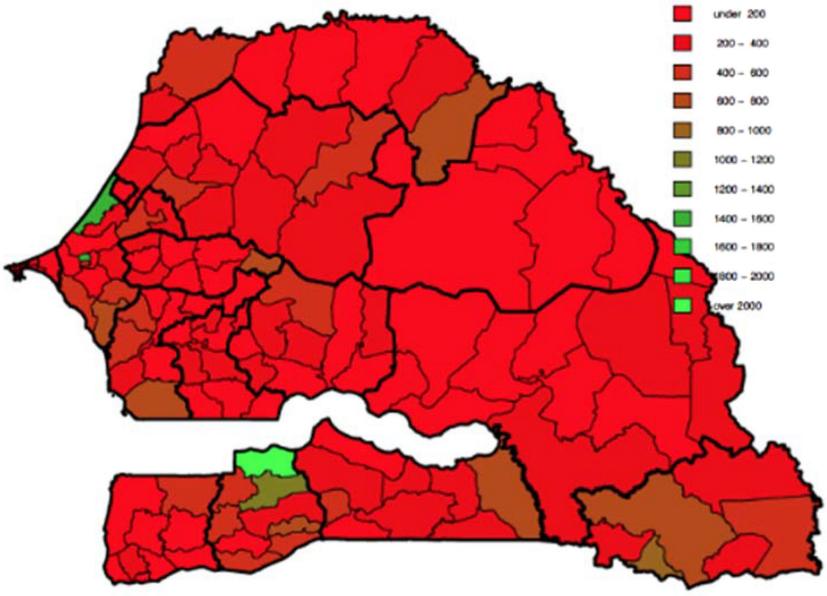
where P is population size for a specific geographic area and time; U is the number of cell-phone users for the respective geographic area and time; α is a scale ratio parameter; β is the parameter that describes the superlinear effect in the relationship between users and population size; and ϵ is a random error. The model described in equation (1) has proven useful in high-income countries with high and rather uniform cell-phone penetration rates. The same baseline model performs quite well with this study's data for Senegal. Figure 2 shows an example of model fit for equation (1) using these data. With an R^2 of 0.768, the relationship seems to hold fairly well in Senegal.

Initial exploratory analysis confirmed the existence of some systematic patterns in the distribution of cell-phone penetration rates, with clusters forming for areas that are geographically close (see fig. 3), with the same type of urban vs. rural setting, or with similar demographic characteristics.

Using Census Data to Identify Patterns of Bias in CDR-Based Estimates

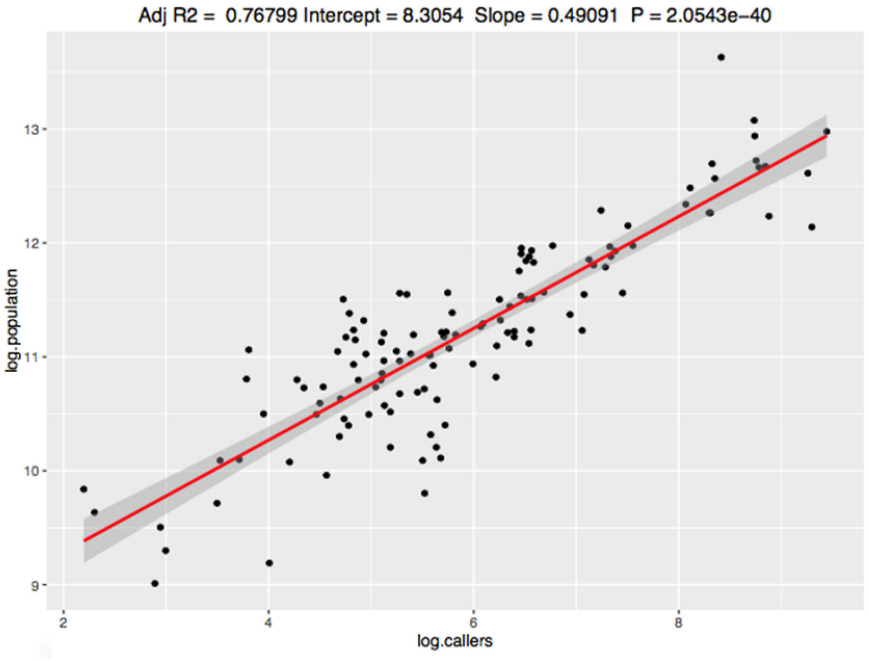
Census data provide socio-demographic information for each *arrondissement* in Senegal. This fact is leveraged to understand whether there are systematic biases in the relationship between population size

Figure 2. Ratios of Population Size in the Census and Number of Callers in the CDR Sample, for *Arrondissements* in Senegal (2013)



Source: Figure created by the authors based on analysis of Orange and ANSD data.
Note: These values can be interpreted as the inverse of the (observed) cell-phone penetration rates for Orange customers in each *arrondissement*.

Figure 3. Fit for the Regression Model of Population Size on Number of Callers (Log-Log Scale) for Senegal (2013), Following Equation (1)

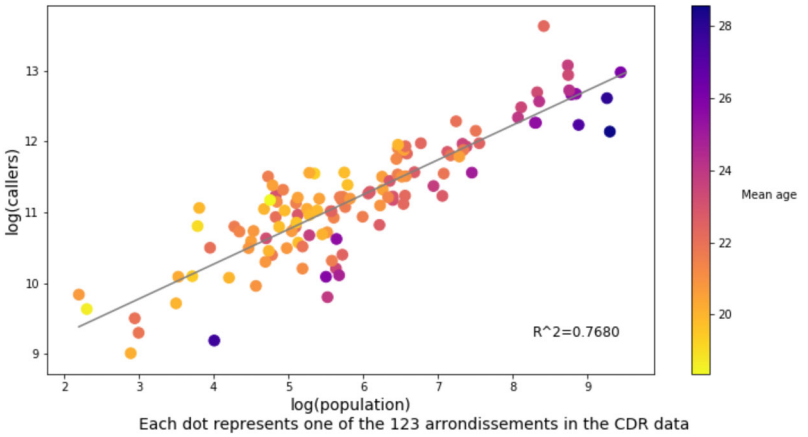


Source: Figure created by the authors based on analysis of Orange and ANSD data.

Table 1. Regression Coefficients (and Associated Standard Errors) for the Regression in Equation (3)

Intercept	10.644^{***}
	(0.393)
Log(callers)	0.597^{***}
	(0.027)
Mean population age	-0.135^{***}
	(0.021)

Figure 4. Relationship between the Number of Callers (from CDR Data) and the Actual Population (from Census Data)



Source: Figure created by the authors based on analysis of Orange and ANSD data.
Note: The data points are color-coded to include the mean population age for each *arrondissement*. Each dot represents one of the 123 *arrondissements* in the CDR data.

and number of callers:

$$\log(P) = \alpha + \beta \log(U) + bias + \epsilon \tag{2}$$

where *bias* is a systematic error in the initial estimate that is accounted for by a given socio-demographic characteristic. The effect of mean age is investigated as follows:

$$\log(P) = \alpha + \beta \log(U) + \gamma\ mean.age + \epsilon \tag{3}$$

Including mean population age at the *arrondissement* level significantly improves the fit ($R^2 = 0.827$). As table 1 shows, the coefficient associated with mean population age is negative and highly significant. The relationship is shown in fig. 4, which illustrates differences across age groups. The points are color-coded to indicate the average resident age in each administrative area, based on census information. Younger areas (lighter) lie mostly above the regression line, and older areas (darker) lie mostly below. In other words, using the standard model of equation (1) would underestimate population size in regions with younger populations, and overestimate it in regions with older populations. This holds true at the *arrondissement*, *département*, and *région* levels.

This suggests that the relationship between number of callers and population size varies systematically according to its demographic characteristics (here, mean age), which can be used for sample bias correction purposes. Beyond age, it is also useful to consider 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).

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	2.80%	-	-
Département level	4.35%	1.90%	-
Région level	7.21%	3.75%	1.51%

Table 3. Mean Absolute Percentage Errors (MAPE) for the Regression in Equation (3)

	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%

Projecting the Regression Coefficients Down to Lower Administrative Levels

To see whether the coefficients calculated at a given administrative level could be projected down to smaller ones, the study calculated regression coefficients and fitted values at the *région* level, then used those coefficients to estimate population at the *département* and *arrondissement* levels. The study used the same procedures at the *département* level and used them to estimate populations at the *arrondissement* level. Finally, the coefficients and fitted values were calculated at the *arrondissement* level. This was done for both the standard model in equation (1) and the model with mean age in equation (2). These population estimates were then compared to the census populations for the corresponding administrative level, and the mean absolute percentage error (MAPE) was calculated in each case (see tables 2 and 3).

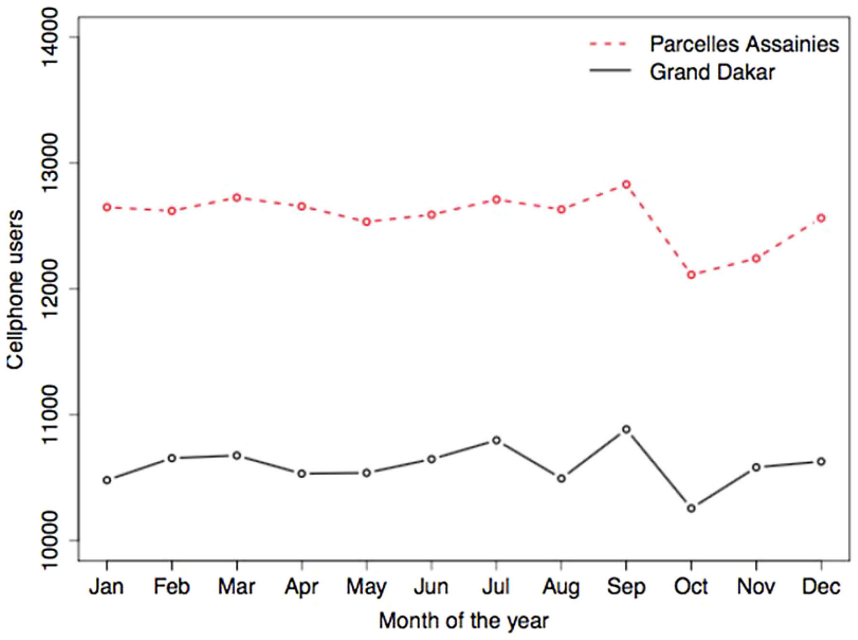
The standard model used in the literature performs poorly here. For instance, when estimating populations at the *arrondissement* level, the MAPE more than doubles when using fits from the *région* level (two levels up) instead of the *arrondissement* level itself—from 2.80 percent to 7.21 percent. Incorporating mean age yields a noticeable improvement: in the same set up, but with equation (3), the MAPE only increases from 2.52 percent to 3.53 percent. This suggests that controlling for characteristics such as age makes extrapolations of population size at smaller geographic levels more robust.

Estimating Population Change over Time Using a Difference-in-Differences approach

In this section a difference-in-differences approach is used to evaluate the extent to which the population size in certain geographic areas changes relative to other areas over time. Three *arrondissements* in the greater Dakar area were chosen that have very similar cell-phone penetration rates on a yearly basis: Grand Dakar (GD), Parcelles Assainies (PA), and Dakar Plateau (DP). Dakar Plateau is an important center for commercial activity and tourism.

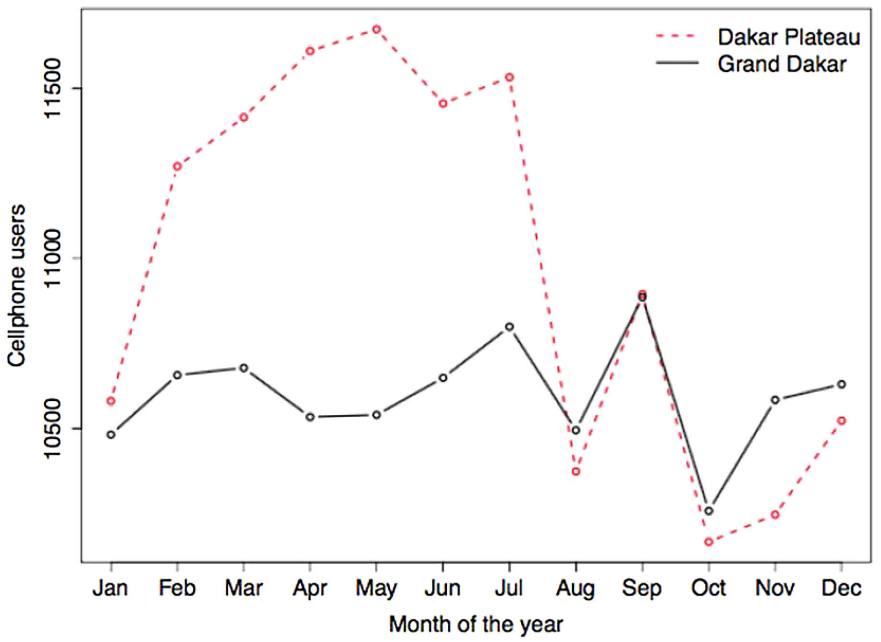
Figure 5 shows trends in the average number of cell-phone users present in GD and PA over the course of a year. The trends are almost perfectly parallel. Figure 6 shows trends in the average number of cell-phone users present in GD and DP. The trend lines are parallel for the period between August and January. During this period, the trend is very similar to the one observed in fig. 5 for PA. However, between February and July, the number of cell-phone users in DP rapidly increases, suggesting a seasonal pattern that would differentially affect DP. To evaluate the size of the effect, the following difference-in-differences model

Figure 5. Average Number of Cell-Phone Users for the *Arrondissements* Grand Dakar and Parcelles Assainies over the Course of the Year



Source: Figure created by the authors based on analysis of Orange and ANSD data.

Figure 6. Average Number of Cell-Phone Users for the *Arrondissements* Grand Dakar and Dakar Plateau over the Course of the Year



Source: Figure created by the authors based on analysis of Orange and ANSD data.

was estimated:

$$U_i^t = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 G_i T_t + e_{it} \quad (4)$$

where U_i^t is the number of cell-phone users for the regions of DP and GD, over time. G_i is an indicator variable that takes the value "1" if the observation is for the DP *arrondissement* and "0" otherwise. T_t is an indicator variable that takes the value "1" during the period from February to July and "0" otherwise. The difference-in-differences estimator $\hat{\delta}$ is equal to the estimate for the parameter β_3. The estimate for β_3 is 940.67 ($s.e. = 154.12$), and is highly significant. Using the regression parameters that were estimated in the previous section), this suggests a change in population size on the order of 100,000 people.

Although further investigation would be required to determine the reasons behind these changes in population size, this comparison demonstrates that CDRs can be used to estimate relative changes in population sizes over time at the subnational level even in the absence of ground truth.

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

This model is a first step to show that accounting for sample selection bias is possible, and can be extended in multiple directions. This work will hopefully help spur the interest of demographers and other social scientists in fulfilling Big Data's potential and producing methods to generate development estimates.

References

Agence Nationale de la Statistique et de la Démographie du Sénégal. 2013. "Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage." http://anads.ansd.sn/index.php/catalog/51.

Blumenstock, J. E. 2012. "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda." *Information Technology for Development* 18 (2): 107–25.

Bruckschen, F., T. Schmid, and T. Zbiranski. 2014. "Cookbook for a Socio-Demographic Basket: Constructing Key Performance Indicators with Digital Breadcrumbs." In *Data for Development Challenge Senegal, Book of Abstracts: Scientific Papers*, 122–31. Cambridge, MA: MIT Media Lab.

Chen, X., and W. D. Nordhaus. 2011. "Using Luminosity Data as a Proxy for Economic Statistics." *Proceedings of the National Academy of Sciences* 108 (21): 8589–94.

de Montjoye, Y.-A., Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D Blondel. 2014. "D4D-Senegal: The Second Mobile Phone Data for Development Challenge." *CoRR* abs/1407.4885. http://arxiv.org/abs/1407.4885.

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): 15888–93.

Flowminder, N. 2015. "Nepal Population Estimates as of May 1, 2015." https://data.hdx.rwlabs.org/dataset/population-movements-after-the-nepal-earthquake-v-1-up-to-1-may-2015/resource_download/fc242f46-1929-4850-9e79-262e84314d88.

GADM Database of Global Administrative Areas. 2015. "Version 2.8." http://www.gadm.org/.

Henderson, J. V., A. Storeygard, and D. Weil. 2009. "Measuring Economic Growth from Outer Space." NBER Working Paper No. 15199, National Bureau of Economic Research, Cambridge, MA, http://www.nber.org/papers/w15199.pdf.

International Telecommunication Union. 2014. "World Telecommunication Development Conference (WTDC-2014): Final Report." Dubai, United Arab Emirates, http://www.itu.int/en/ITU-D/Conferences/WTDC/WTDC14/Pages/default.aspx.

Kulkarni, R., K. Haynes, R. Stough, and J. Riggle. 2011. "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): 101–13, http://econpapers.repec.org/article/blargscpp/v_3a3_3ay_3a2011_3ai_3a2_3ap_3a101-113.htm.

Olivia, S., J. Gibson, L. K. Brabyn, and G. Stichbury. 2014. "Monitoring Economic Activity in Indonesia Using Night Light Detected from Space." Paper presented at the 12th Indonesian Regional Science Association Conference, Makassar Indonesia, June 2–3.

Orange. 2012. "Data for Development Challenge Côte d'Ivoire." http://www.d4d.orange.com/.

———. 2014. "Data for Development Challenge Senegal." http://www.d4d.orange.com/.

République du Sénégal. 1996. "Loi N° 96-06 Du 22 Mars 1996 Portant Code Des Collectivités Locales." *Journal Officiel*. http://www.gouv.sn/Code-des-Collectivites-locales.html.

———. 2013. "Loi N° 2013-10 Du 28 Décembre 2013 Portant Code Général Des Collectivités Locales." *Journal Officiel*. http://www.gouv.sn/Code-general-des-Collectivites.html.

Smith-Clarke, C., A. Mashhadi, and L. Capra. 2014. "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM Press, 511–20. doi:10.1145/2556288.2557358.

Zagheni, E., and I. Weber. 2012. "You Are Where You E-Mail: Using E-Mail Data to Estimate International Migration Rates." *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12. New York: ACM, 348–51 doi:10.1145/2380718.2380764.