HARVARD HUMANITARIAN INITIATIVE

mit media lab

FLOWMINDER.ORG

DATA POP ALLIANCE

The ABCDE of Big Data: Assessing Biases in CDRs for Development Estimates

Gabriel Pestre (Data-Pop Alliance)
Emmanuel Letouzé (Data-Pop Alliance)
Emilio Zagheni (University or Washington, Seattle)

Presented by Gabriel Pestre
Cologne, Germany
17 May 2016

Data-Pop Alliance
is a global coalition on Big Data and development created by the Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute that brings together researchers, experts, practitioners, and activists to promote a people-centered Big Data revolution through collaborative research, capacity building, and community engagement. As of February 2016, Flowminder Foundation has joined Data-Pop Alliance as its fourth Core Member.

BACKGROUND & CONTEXT

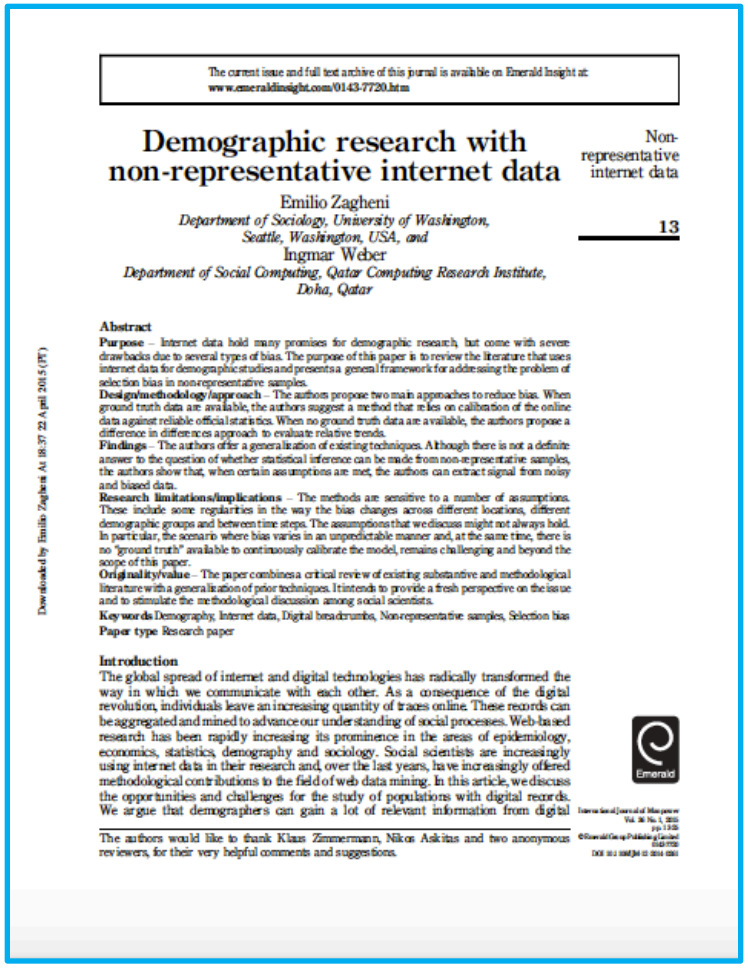
1

Motivations of the work:

- Methodological
- Political

- P. Deville, 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.

- Zagheni & Weber (2015). Demographic research with non-representative internet data.



The current issue and full text archive of this journal is available on Emerald Insight at: www.emeraldinsight.com/0143-7720.htm

Demographic research with non-representative internet data

Non-representative internet data

13

Emilio Zagheni
Department of Sociology, University of Washington, Seattle, Washington, USA, and
Ingmar Weber
Department of Social Computing, Qatar Computing Research Institute, Doha, Qatar

Abstract

Purpose – Internet data hold many promises for demographic research, but come with severe drawbacks due to several types of bias. The purpose of this paper is to review the literature that uses internet data for demographic studies and presents a general framework for addressing the problem of selection bias in non-representative samples.

Design/methodology/approach – The authors propose two main approaches to reduce bias. When ground truth data are available, the authors suggest a method that relies on calibration of the online data against reliable official statistics. When no ground truth data are available, the authors propose a difference in differences approach to evaluate relative trends.

Findings – The authors offer a generalization of existing techniques. Although there is not a definite answer to the question of whether statistical inference can be made from non-representative samples, the authors show that, when certain assumptions are met, the authors can extract signal from noisy and biased data.

Research limitations/implications – The methods are sensitive to a number of assumptions. These include some regularities in the way the bias changes across different locations, different demographic groups and between time steps. The assumptions that we discuss might not always hold. In particular, the scenario where bias varies in an unpredictable manner and, at the same time, there is no "ground truth" available to continuously calibrate the model, remains challenging and beyond the scope of this paper.

Originality/value – The paper combines a critical review of existing substantive and methodological literature with a generalization of prior techniques. It intends to provide a fresh perspective on the issue and to stimulate the methodological discussion among social scientists.

Keywords Demography, Internet data, Digital breadcrumbs, Non-representative samples, Selection bias

Paper type Research paper

Introduction

The global spread of internet and digital technologies has radically transformed the way in which we communicate with each other. As a consequence of the digital revolution, individuals leave an increasing quantity of traces online. These records can be aggregated and mined to advance our understanding of social processes. Web-based research has been rapidly increasing its prominence in the areas of epidemiology, economics, statistics, demography and sociology. Social scientists are increasingly using internet data in their research and, over the last years, have increasingly offered methodological contributions to the field of web data mining. In this article, we discuss the opportunities and challenges for the study of populations with digital records. We argue that demographers can gain a lot of relevant information from digital

The authors would like to thank Klaus Zimmermann, Nikos Askitas and two anonymous reviewers, for their very helpful comments and suggestions.

Orange Data for Development (D4D) Challenge 2014 dataset

- One year of coarse-grained mobility data at individual level for 146,352 randomly sampled users in Senegal in 2013.
- Only users meeting both the following criteria were included:
  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.

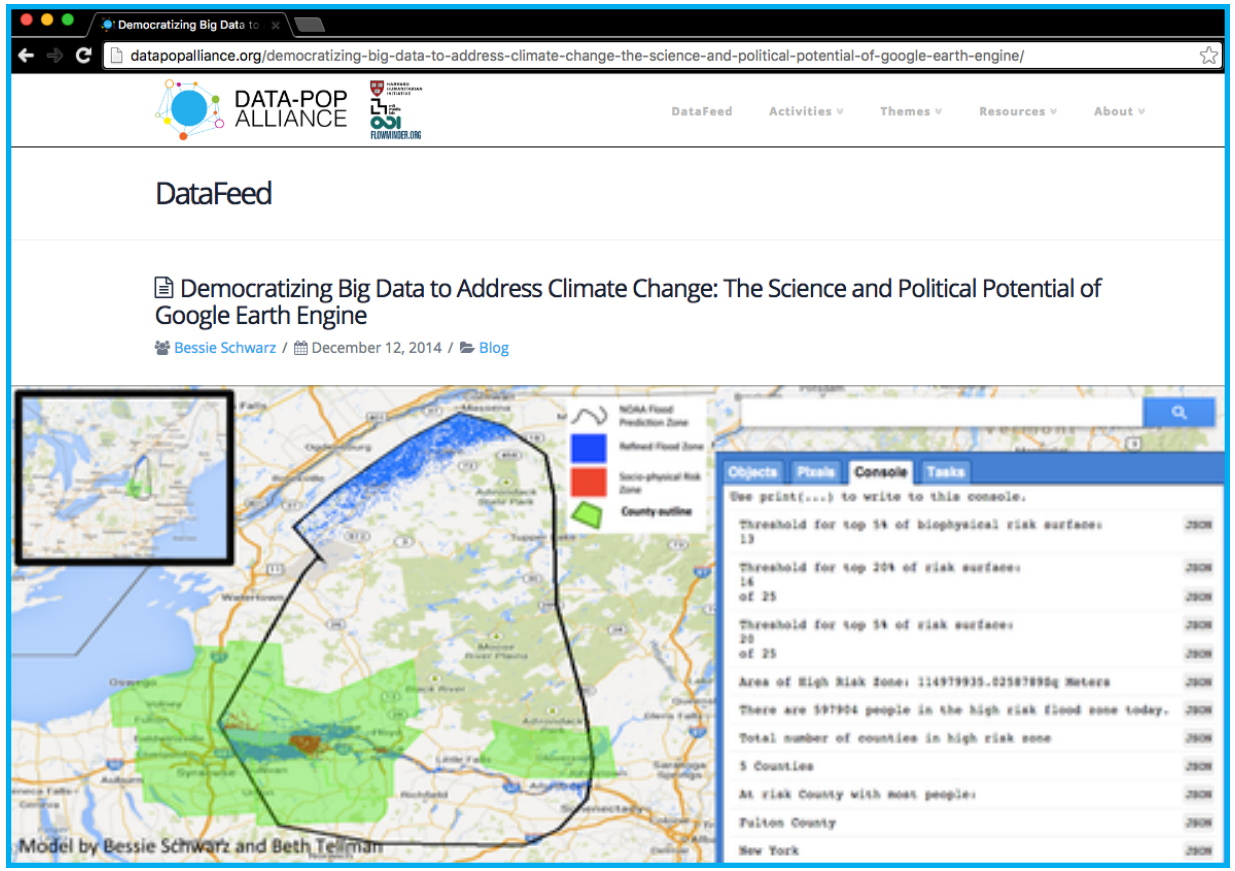
***Agence Nationale de la Statistique et de la Demographie du Sénégal* (ANSD) official census**

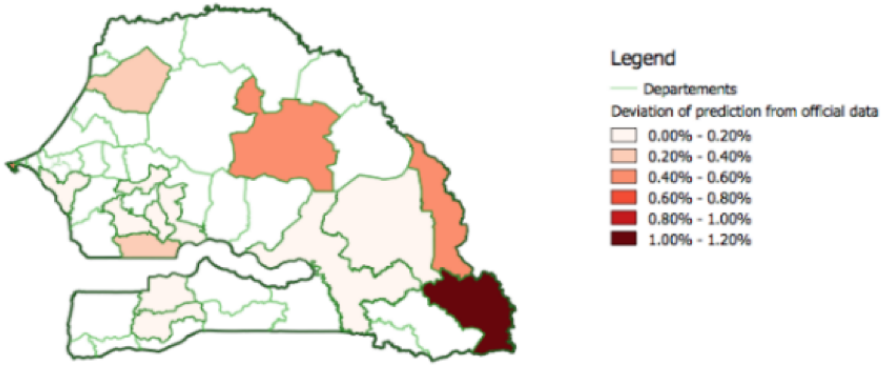
- 1/10th sample of the *Recensement General de la Population et de l'Habitat, de l'Agriculture et de l'Elevage* (RGPHAE).
- The 2013 edition of the RGPHAE was conducted over the 21 day period from November 19 to December 14.

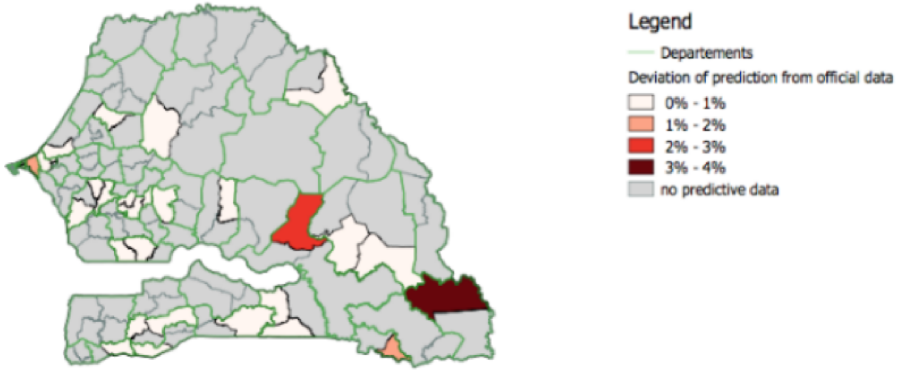
1

Social and physical vulnerability to flooding

- Ongoing work of Data-Pop Alliance research affiliates Bessie Schwarz and Beth Tellman



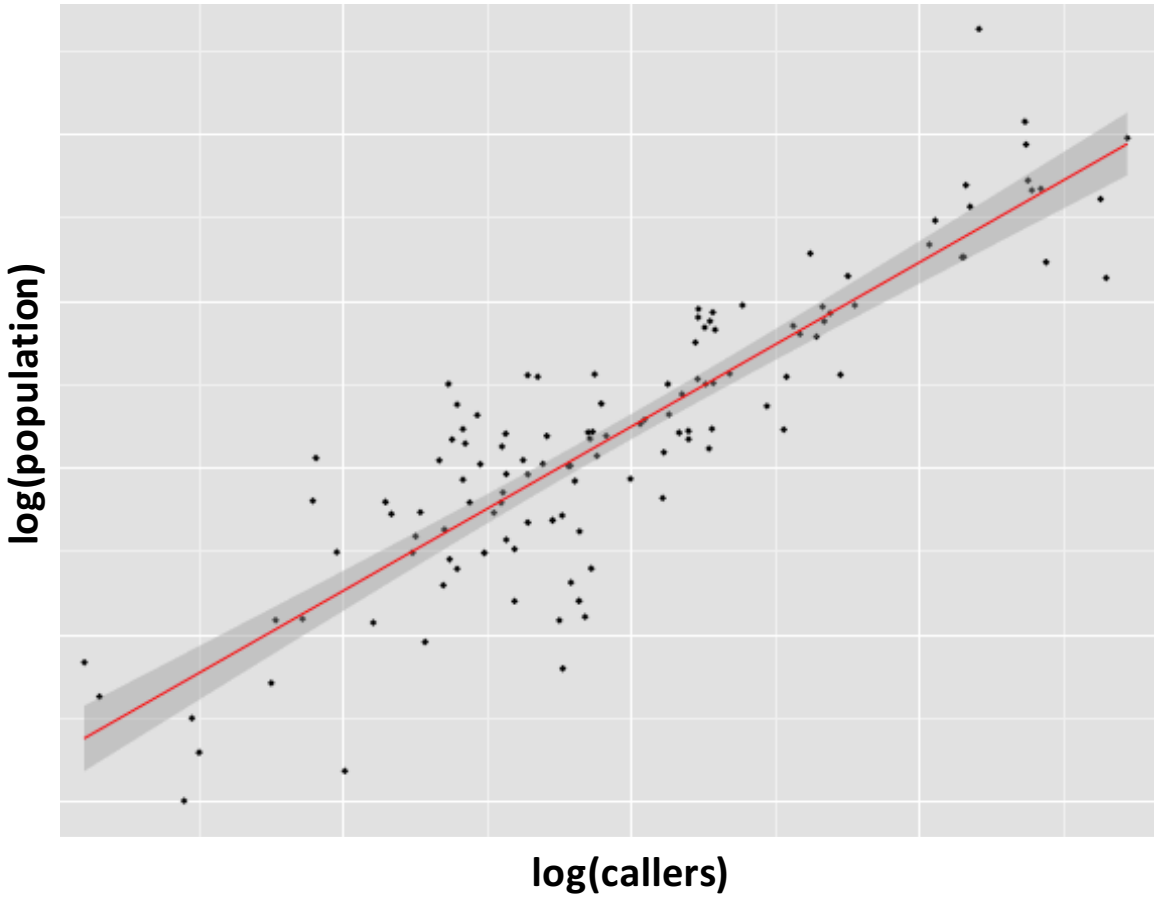




Simple base regression

$$\log(P) = \alpha + \beta \, log(U) + \epsilon$$

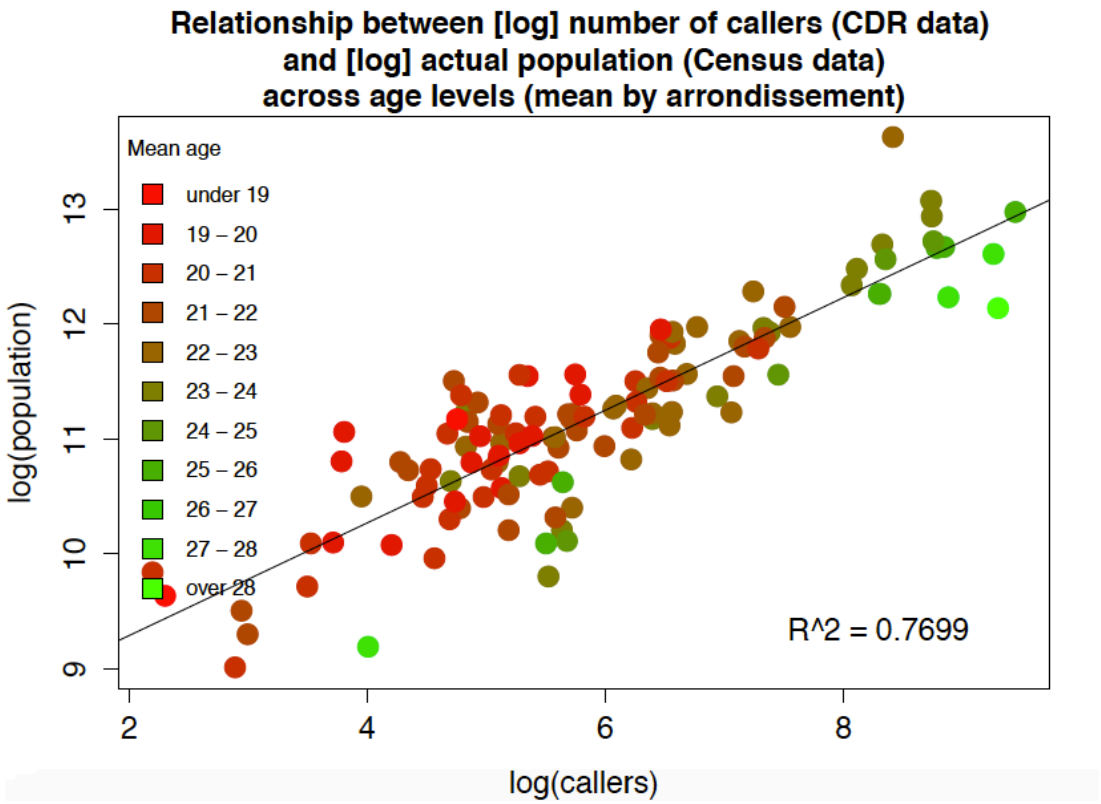
- Coefficients are estimated using a simple linear regression model.
- Adjusted R^2 = 0.768

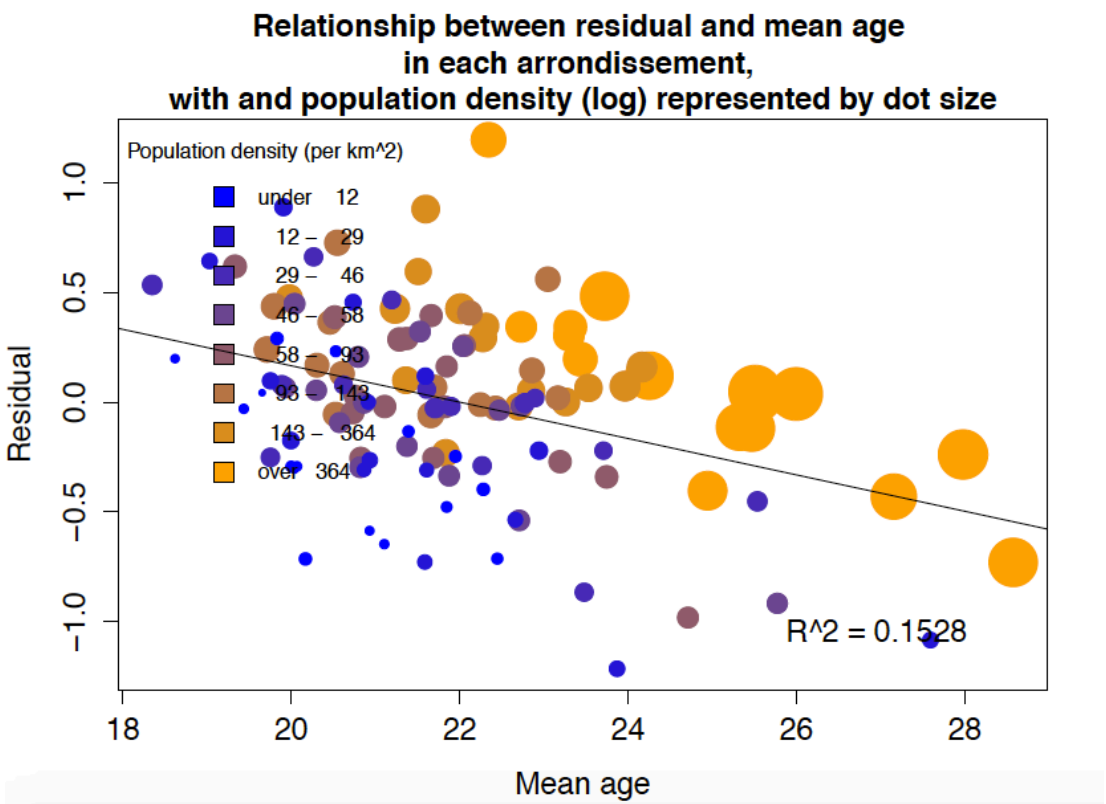


HIGHLIGHT #1 : EFFECT OF AGE

3

What patterns of bias can we identify using census data?





red = younger mean age
green = older mean age

orange = more dense population
blue = less dense population

HIGHLIGHT #1 : EFFECT OF AGE

- Including mean population age in the standard model significantly improves the fit (R^2 = 0.827 ; compared to 0.768 in the base model).

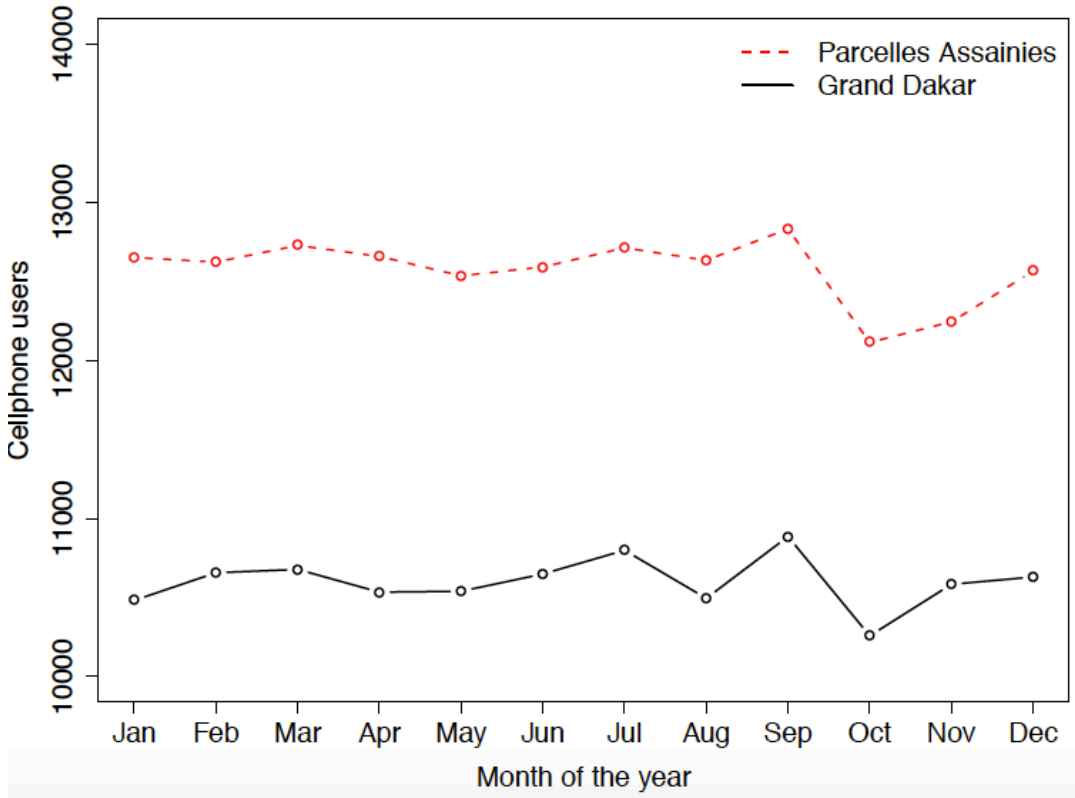
Intercept	10.644***
	(0.393)
log(callers)	0.597***
	(0.027)
mean population age	-0.135***
	(0.021)

- The standard model tends to overestimate it in regions with older population age structure, and vice versa.

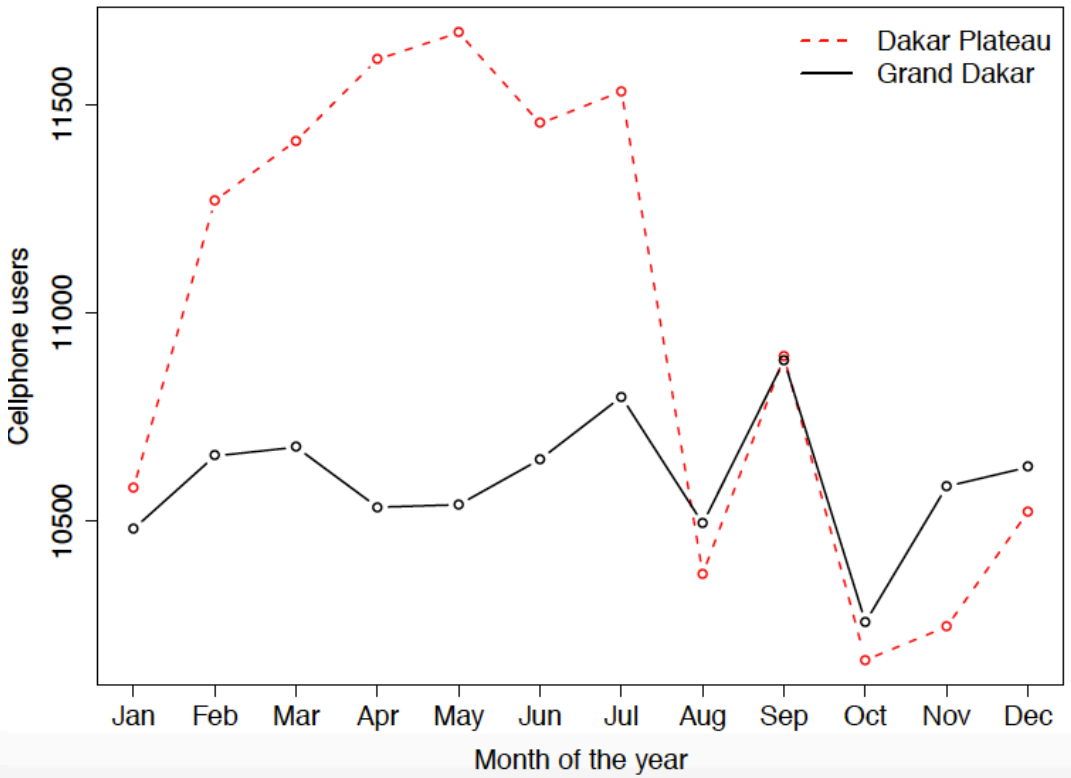
HIGHLIGHT #2 : DIFFERENCE-IN-DIFFERENCE

4

How can we compare trends across time and place even without ground truth?



Average number of cellphone users for the arrondissements Grand Dakar and Parcelles Assainies over the course of the year



Average number of cellphone users for the arrondissements Grand Dakar and Dakar Plateau over the course of the year

HIGHLIGHT #2 : DIFFERENCE-IN-DIFFERENCE

How can we compare trends across time and place even without ground truth?

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

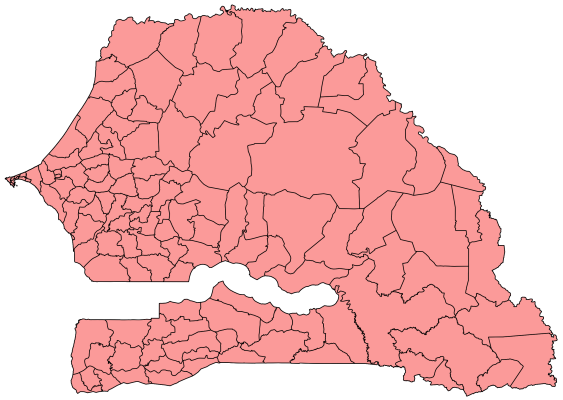
- The difference in difference estimator $\hat{\delta}$ is equal to the estimate for the parameter β. The estimate for β is 940.67 (s.e. = 154.12) and is highly significant.
- This is a large change in population size: based on the results form the regression models estimated in the previous section,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

HIGHLIGHT #3 : PROJECTING DOWN TO LOWER ADMINISTRATIVE LEVELS

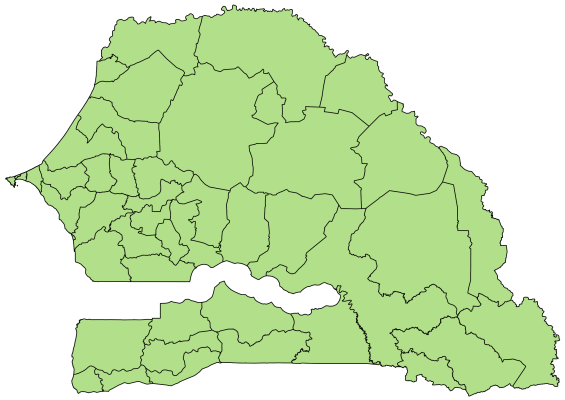
5

Does having ground-truth for larger areas help us estimate for smaller areas?

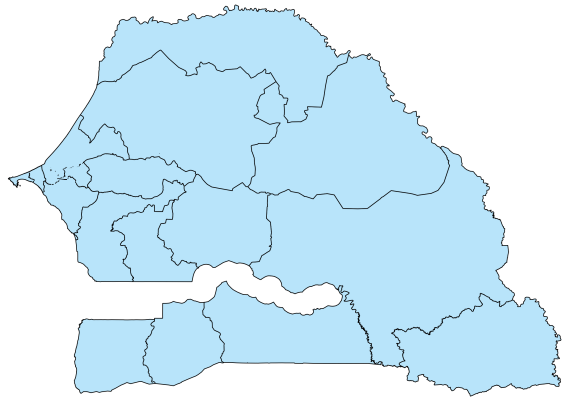
Mean absolute percentage error (MAPE)		Estimates of log(population) at ...		
		... arrondissement level	... département level	... région level
using coefficients fitted at ...	... arrondissement level	2.80%	-	-
	... département level	4.35%	1.90%	-
	... région level	7.21%	3.75%	1.51%



123 arrondissements



45 départements



14 régions

CONCLUSION & NEXT STEPS

- Continue exploring sample bias with additional census variables (education level, rural/urban, gender).
- Additional work on how CDRs can be used to improve indices of social and physical vulnerability to flooding in Senegal.

Thank you